Big Data Analytics for Healthcare Industry

Sujata Ghonmode¹, Hirendra Hajare²
M. Tech Student¹, Assistant Professor²

Department of CSE
Ballarpur Institute of Technology (BIT), Ballarpur, India

Abstract:
In recent years, huge amounts of structured, unstructured, and semi-structured data have been generated by various institutions around the world and, collectively, this heterogeneous data is referred to as big data. The health industry sector has been confronted by the need to manage the big data being produced by various sources, which are well known for producing high volumes of heterogeneous data. Various big-data analytics tools and techniques have been developed for handling these massive amounts of data, in the healthcare sector. In this paper, we discuss the impact of big data in healthcare, and various tools available in the Hadoop ecosystem for handling it. We also explore the conceptual architecture of big data analytics for healthcare which involves the data gathering history of different branches, the genome database, electronic health records, text/imagery, and clinical decisions support system.

Key words: big data; healthcare; Hadoop; MapReduce.

I. INTRODUCTION

Every day, data is generated by a range of different applications, devices, and geographical research activities for the purposes of weather forecasting, weather prediction, disaster evaluation, crime detection, and the health industry, to name a few. In current scenarios, big data is associated with core technologies and various enterprises including Google, Facebook, and IBM, which extract valuable information from the huge volumes of data collected [1–3]. An era of open information in healthcare is now under way. Big data is being generated rapidly in every field including healthcare, with respect to patient care, compliance, and various regulatory requirements. As the global population continues to increase along with the human lifespan, treatment delivery models are evolving quickly, and some of the decisions underlying these fast changes must be based on x data [4]. Healthcare shareholders are promised new knowledge from big data, so called both for its volume as well as its complexity and range. Pharmaceutical-industry experts and shareholders have begun to routinely analyse big data to obtain insight, but these activities are still in the early stages and must be coordinated to address healthcare delivery problems and improve healthcare quality. Early systems for big-data analytics of healthcare informatics have been established across many scenarios, e.g., the investigation of patient characteristics and determination of treatment cost and results to pinpoint the best and most cost-effective treatments [4]. Health informatics is described as the assimilation of healthcare sciences, computing sciences and information sciences in the study of healthcare information. Health informatics involves data acquisition, storage, and retrieval to provide better results by healthcare providers. In the healthcare system, data is characterized by its heterogeneity and variety as a result of the linking of a diverse range of biomedical data sources including, for example, sensor data, imagery, gene arrays, laboratory tests, free text, and demographics [5]. Most data in healthcare system (e.g., doctor’s notes, lab test results, and clinical data) is unstructured and is not stored electronically, i.e., it exists only in hard copies and its volume is increasing very rapidly. Currently, there is a major focus on the digitization of these vast stores of hard copy data. The revolutions of data size are actually creating a problem in order to achieve this goal [6]. The various terminologies and models that have been developed to resolve the problems associated with big data focus on solving four issues known as the four Vs, namely: volume, variety, velocity, and veracity. The various classes of data in healthcare applications include Electronic Health Records (EHR), machine generated/sensor data, health information exchanges, patient registries, portals, genetic databases, and public records. Public records are major sources of big-data in the healthcare industry and require efficient data analytics to resolve their associated healthcare problems. According to a survey conducted in 2012, healthcare data totalled nearly 550pet a bytes and will reach nearly 26 000 pet a bytes in2020[5]. In light of the heterogeneous data formats, huge volume, and related uncertainties in the big-data sources, the task of realizing the transformation of raw data into actionable information is daunting. Being so complex, the identification of health features in medical data and the selection of class attributes for health analytics demands highly sophisticated and architecturally specific techniques and tools.

II. LITERATURE REVIEW

The main difference between traditional health analysis and big-data health analytics is the execution of computer programming. In the traditional system, the healthcare industry depended on other industries for big data analysis. Many healthcare shareholder’s trust information technology because of its meaningful outcomes—their operating systems are functional and they can process the data into standardized forms. Today, the healthcare industry is faced with the challenge of handling rapidly developing big healthcare data. The field of big data analytics is growing and has the potential to provide useful insights for the healthcare system. As noted above, most of the massive amounts of data generated by this system is saved in hard copies, which must then be digitized [7]. Big data can improve healthcare delivery and reduce its cost, while supporting advanced patient care, improving patient outcomes, and avoiding unnecessary costs [8]. Big data
analytics is currently used to predict the outcomes of decisions made by physicians, the out come of a heart operation for a condition based on patient’s age, current condition, and health status. Essentially, we can say that the role of big data in the health sector is to manage data sets related to healthcare, which are complex and difficult to manage using current hardware, software, and management tools. In addition to the burgeoning volume of healthcare data, reimbursement methods are also changing [9]. Therefore, purposeful use and pay based on performance have emerged as important factors in the healthcare sector. In 2011, organizations working in the field of health care had produced more than 150 Exabyte’s of data [10], all of which must be efficiently analysed to be at all useful to the healthcare system [11]. The storage of healthcare related data in EHRs occurs in a variety of forms. A sudden increase in data related to healthcare informatics has also been observed in the field of bioinformatics, where many terabytes of data are generated by genomic sequencing [11]. There are a variety of analytical techniques available for interpreting medical, which can then be used for patient care [12]. The diverse origins and forms of big data are challenging the healthcare informatics community to develop methods for data processing. There is a big demand for technique that combines dissimilar data sources [13].

A number of conceptual approaches can be employed to recognize irregularities in vast amounts of data from different datasets. The frameworks available for the analysis of healthcare data are as follows:

**Predictive Analytics in Healthcare:**
For the past two years, predictive analysis has been recognized as one of the major business intelligence approaches, but its real-world applications extend far beyond the business context. Big data analytics includes various methods, including text analytics and multimedia analytics [14]. However, one of the most crucial categories is predictive analytics which includes statistical methods like data mining and machine learning that examine current and historical facts to predict the future. Predictive methods which are being used today in the hospital context to determine if patient may be at risk for readmission [15]. This data can help doctors to make important patient care decisions. Predictive analysis requires an understanding and use of machine learning, which is widely applied in this approach.

**Machine Learning in Healthcare:**
The concept of machine learning is very similar to that of data mining [4], both of which scan data to identify patterns. Rather than extracting data based on human understanding, as in data mining applications, machine learning uses that data to improve the program’s understanding. Machine learning identifies data patterns and then alters the program function accordingly [16].

**Electronic Health Records:**
EHR represents the most widespread health application of big data in healthcare. Each patient has his/her own medical records, with details that include their medical history, allergies diagnosis, symptoms, and lab test results. Patient records are shared in both public and private Sectors with healthcare providers via a secure information system. These files are modifiable, in that doctors can make changes over time and add3. New medical test results, without the need for paper work or duplication of data.

**III. BIG DATA ANALYTICS ARCHITECTURE FOR HEALTH INFORMATICS:**
Currently, the main focus in big-data analytics is to gain an in-depth insight and understanding of big data rather than to collect it [20]. Data analytics involves the development and application of algorithms for analysing various complex data sets to extract meaningful knowledge, patterns, and information. In recent years, researchers have begun to consider the appropriate architectural framework for health care systems that utilize big-data analytics, one of which uses a four-layer architecture that comprises a transformation layer, data-source layer, big data platform layer, and analytical layer [14]. In this layered system, data originates from different sources and has various formats and storage systems. Each layer has a specific data-processing functionality for performing specific tasks on the HDFS, using the Map Reduce processing model. The other layers perform other task s.i.e., report generation, query passing, data mining processing, and online analytical processing. The main requirement in big-data analytical processing is to bundle the data at high speed to minimize the bundling time. The next priority in big-data analytical processing is to efficiently update and transform queries at a constant time [21].

The third requirement in the big-data analytical processing is to utilize and efficiently manage the storage area space. The last specification of big-data analytics is to efficiently become familiar with the rapidly progressing workload notations. Big-data analytics frame works differ from traditional healthcare processing systems with respect to how they process big data [22]. In the current health care system, data is processed using traditional tools installed in a single stand-alone system like a desktop computer. In contrast, big data is processed by clustering and scans multiple nodes of clusters in the network [23]. This processing is based on the concept of parallelism to handle large medical datasets [24]. Freely available frameworks, such as Hadoop, MapReduce, Pig, Sqoop, Hive, and HBase Avro, all have ability to process the health-related data sets for healthcare systems. Big-data technologies broadly refer to scientific innovations that mimic those used for large datasets [25]. In the first component is the requirement for big data sources for processing. In the second component clusters with a centralized big-data processing infrastructure are at the peak of high performance [24]. It has been observed that the tool mainly available for big-data analytics processing provide data security, scalability, and manageability with the help of the Map Reduce paradigm. In the third component, big data analytics applications have a storage domain to integrate accessed databases that use different applications [26]. In the fourth component, are the most popular big-data analytics applications in healthcare systems, which include reports, Online Analytical Processing (OLAP), queries, & data mining.

![Figure.1. Hadoop System Architecture](http://ijesc.org/)
Figure 1 shows the physical layout architecture of Hadoop which consists of MapReduce, HBase, and HDFS.

**HDFS:** The HDFS was designed for processing big data [21]. Although it can support many users simultaneously, HDFS is not designed as a true parallel file system. Rather, the design assumes a large file write-once/ read-many model that enables other optimizations and relays many of the concurrency and coherency overhead requirements of a true parallel file system [31]. HDFS is designed for data streaming by which large amounts of data are read from disk in bulk [25]. The HDFS block size is 64MB or 128 MB. There are two types of nodes: a name node and multiple data node(s). A single name node manages all the metadata needed to store and retrieve the actual data from the data nodes [13]. No data is actually stored on the name node. Files are stored as block sin proper sequence and these blocks are equal in size [24]. The features of HDFS are its distributed nature and reliability. Storage of metadata and file data is separated. Metadata is stored in name node and application data is stored in data node.

**MapReduce:** Apache Hadoop is often associated with Map Reduce computing. The Map Reduce computation model is a very powerful tool used in many health applications and is more common than most users realize. Its underlying concept is very simple [25]. In MapReduce, there are two stages: amapping stage and a reducing stage. In the mapping stage, a mapping procedure is applied to input data. The reducing phase is implemented when counting also has two stages: a mapping stage that accepts input in key value pairs and generates output in key value pairs and a second reducing stage, in which each phase consists of key-value pairs as input and output [12]. There is a fixed size data segment division step in Hadoop which is called input splits [20]. The Map function generates the value pairs and the key, which are stored in the mapper. Any keys that are the same are merged.

**IV. REFERENCES**


