International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Healthcare Data Management and Analytics using Big Data Tools

Subham Singh Chauhan, Iraa Sharma, Ishan Kanungo, Gurpartap Singh

Abstract— Data have been expanding enormously in latest years, enormous amounts of structured, unstructured and semi-structured information have been produced in various areas around the globe, collectively known as big data. The health sector has produced enormous amounts of heterogeneous information that must be handled and analyzed. In this paper, we discuss about the characteristics of data generated by healthcare and how to manage this data using big data tools. We also explore tools to analyse this data and discuss the implementations of this data. A conceptual architecture of big data analytics is also given, which includes data cleaning, data injection, data management, data mining, data visualization and data analysis.

Keywords—healthcare, big data, data management, data analytics

I. INTRODUCTION

Now a days we are seeing huge amount of data all around us in the form of text files, images, records, csv files, etc. Data is being generated at extremely high rate. This heterogeneous data is known as big data. This data is of high value.it has wide range of applications. Modern day industries are using this data for making future strategies and polices. Data analytics is back bone of many business firms. Healthcare sector is also generating huge amount of data on daily basis. This data consists of patient records, medicine subscriptions, test reports, etc. There is need of gathering this data and storing on large data bases for data analysis. There are many different tools and frameworks that can be used for management and analysis of such data. In this paper we have discussed about different tools and frameworks that can be used for healthcare analytics. There are many real time implementations of data analytics in different fields which are also discussed in this paper.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Subham Singh Chauhan*, Computer Science and engineering SRM Institute of Science and Technology, Ramapuram Chennai,India
Iraa Sharma, Computer Science and engineering SRM Institute of Science and Technology, Ramapuram Chennai,India
Ishan Kanungo, Computer Science and engineering SRM Institute of Science and Technology, Ramapuram Chennai,India
Gurpartap Singh, Computer Science and engineering SRM Institute of Science and Technology, Ramapuram Chennai,India

II. DATA

The Four primary attributes (shown in Fig. 1) that are associated with big data are volume, velocity, variety, and veracity.

- **Volume:** Big data is a term to referring to huge volumes of collected data. The quantity of this information does not have a set limit. The word is typically used for massive information that needs to be managed, stored and analyzed using traditional databases and architecture for information processing[1]. The volume of data generated by modern IT and the healthcare system has grown and is driven by the reduced costs of data storage and processing architectures and the need to extract valuable insights from data to improve business processes, efficiency and consumer services[2]
- .**Velocity:** Velocity, it is the main reason for exponential information development, referring to the speed of information collection[3]. At increasingly greater speeds, healthcare systems generate information. The velocity of generating this information after processing needs a choice based on its performance in the quantity and range of the structured or unstructured data gathered.

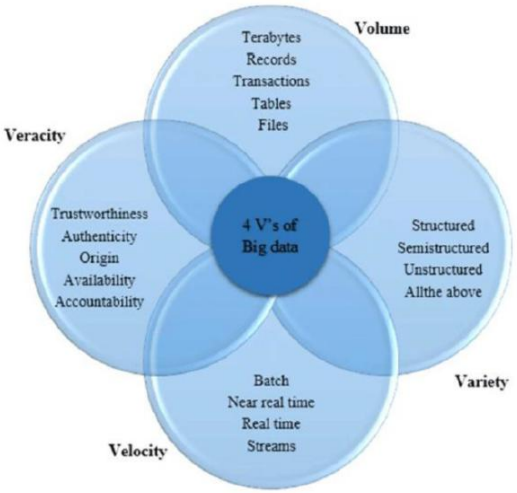

Fig 1 Attributes of data

- **Variety:** Variety Refers to the form of data, unstructured or structured, text, medical imaging, audio, video, and sensor data. Structured data includes clinical data (patient record data) that must simply be collected, stored and processed by a specific device. Structured information includes only 5% to 10% of information on healthcare.

Unstructured or semi-structured data involves emails, pictures, videos, audios, and other health-related information such as hospital medical records, physician notes, paper prescriptions, and radiograph films[4].

- **Veracity:** The veracity of data Different information sources differ in their level of information credibility and reliability[5]. Healthcare analytics are tasked with extracting helpful ideas from this information for the treatment of patients and making the best possible choices.

SYSTEM ARCHITECTURE

Big data analytics is all about gathering in-depth insights of data and understanding trends and patterns in big data[7]. This whole process of analysis begins with collecting data from different sources. After cleaning this raw data, it is injected into distributed databases that run on different frameworks. Valuable data is extracted from these databases and processed for data analytics, known as data mining[8].
Four layered architecture frameworks is appropriate for analysis of healthcare data. Different tools and techniques are used at different layers to perform desire task.

Data gathered from different source can be injected through different modes on the basis on data type. There are different tools like Sqoop, Fumes, Storm that are used for injecting data into database. Many frameworks are used for storing data. Data can be stored in batch file or it can be stored in tables. Different storage frameworks are used to store different type of data.

Data can be extracted from these databases according to our needs. There are many tools like RapidMiner, that are used for mining data. Data analysis is done using many data manipulation tools like hive, pig, MapReduce, spark. Figure 2 is a conceptual representation of architecture of healthcare analysis system.

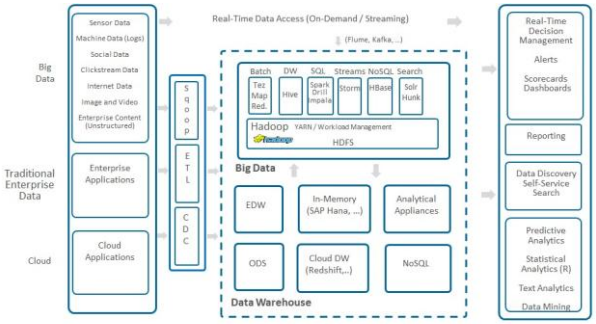

Fig 2 System Architecture

III. DATA CLEANING AND DATA INJECTION

Data from different sources like relational databases and hospital records is gathered for analytics. But, before processing data into database it needs to be clean. For cleaning data there are many big data tools that can be used. And for data injection tools like Sqoop, flume are used.

Excel: Microsoft Excel is a spreadsheet for Windows, macOS, Android and iOS developed by Microsoft. It includes calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. The spreadsheet for these platforms has been widely used, especially since version 5 in 1993, and has replaced Lotus 1-2-3 as the industry standard for spreadsheets.

Open refine: OpenRefine (formerly Google Refine) is a strong instrument to use chaotic information: clean it; transform it from one format to another; and expand it with internet services and external information. OpenRefine is accessible in the following languages: English, Chinese, Spanish, French, Russian, Portuguese, German, Japanese, Italian, Hungarian, Hebrew, Philippine, Cebuano, Tagalog

Sqoop: Apache Sqoop is a strong instrument that conducts Relational Database Management System (RDMS) information extraction functionality and input into Hadoop architecture for query processing. This method utilizes the MapReduce paradigm or other normal level instruments, such as Hive[10], for this purpose. Once placed in HDFS, the data can be used by Hadoop applications.

Flume: Apache Flume is a extremely reliable service for accurate information collection and transition from autonomous computers to HDFS[11]. A number of flume agents that can cross a sequence of computers and places are often involved in information transportation. Flume is frequently used for log files, social media produced information and email messages.

Apache storm: Apache Storm is a real-time computing scheme distributed free and open source. Storm makes it simple to process unbounded information streams reliably, doing what Hadoop did for batch processing for real-time processing. Storm is easy, can be used with any language of programming, and is a lot of fun to use. Storm has many applications: real-time analytics, internet machine learning, ongoing computing, RPC, ETL and more dispersed. Storm is quick: it was clocked by a benchmark at more than a million tuples per second per node. It is scalable, tolerant to errors, ensures the processing of your information and is simple to set up and function. Storm integrates with the technology you already use in the queueing and database.

IV. STORING AND MANAGING DATA

It is very important to store and manage data in secure environment. Data can be stored in different forms like files, images, and tables. So, there are different types of frame works that can be used for storing different types. below mention tools can be used to store and mange big data.

MongoDB: In flexible and JSON-like materials, MongoDB stores information, meaning that fields can differ from the document to the document. The document model maps objects in your application code, making information simple to work with ad hoc queries, indexing, and aggregation in real time provide strong methods of accessing and analyzing your information.

Cassandra: If you need scalability and high availability without compromising efficiency, the Apache Cassandra database is the correct option. The linear scalability of the commodity hardware or cloud infrastructure and the well established defect tolerances make this the ideal platform for critical task information.

Cassandra supports various datacenter replication, provides your customers with reduced latency and a peace of mind to know that you can survive regional outages.

Apache HBase: The NoSQL database used in Hadoop is a column oriented database[12] in which big numbers of columns and rows can be saved by users. HBase has random read / write operations features. It supports updates to record levels that can not be made with HDFS[13]. HBase offers parallel data storage across commodity servers through the underlying distributed file systems. Due to the narrow integration of HBase and HDFS, the file system of choice is usually HDFS[14]. If the high-scale information stored via Hadoop require a structured low latency visual perspective, then HBase is the right option. It scales its open source code linearly to manage thousands of nodes with information petabytes.

Apache Hadoop: Hadoop's name has developed to mean many things[15]. In 2002 it was set up to support a web search engine as a single software project. It is a tool and application ecosystem used since then to evaluate big quantities and data types[26]. The approach to data processing that differs radically from the traditional relational data base model[17] can not be regarded any more as a single monolithic project. The Hadoop ecosystem and framework is defined in more practical terms in the form of open source instruments, Libraries, and Big Data methods for analyzing a number of data sets from distinct sources, i.e. pictures from internet, audio, video, and sensor recordings as organized and unstructured data[18].

Apache Zookeeper: Zookeeper is a centralized system used in health care apps to provide organization and other components between nodes[19] and between them. In big cluster settings, it retains the prevalent items required including configuration data and hierarchic naming space. The distributed processing of Hadoop clusters can be co-ordinatedby various apps. Zookeeper is also responsible for ensuring reliability of applications[20]. If an application master dies, a fresh application master is generated by the zoo maintainer to resume duties.

V. DATA MINING AND DATA VISUALIZATION

Valuable data can be extracted from database using tools like Teradata and RapidMiner. this technique of data extraction is known as data mining. We can also visualize data using different visualization tools. These tools can be used for data mining and data visualization.

RapidMiner: RapidMiner facilitates predictive modeling readiness for information. Exploring information interactively, to assess its health, integrity and quality. Fix prevalent problems such as missing values and outsourcing quickly. Close together various data sets with a single expression editor to generate fresh columns. When lastly prepared, build RapidMiner Studio and Auto Model predictive models or export them to famous enterprise apps like Excel.

Plotly: Plotly's team retains R, Python, and JavaScript's most expanded open source visualization libraries. These libraries interface seamlessly with our company-ready deployment servers to facilitate cooperation, code-free editing and the deployment of production-ready dashboards and applications.

VI. DATA ANALYSIS

The primary distinction between the traditional health assessment and large-scale health analysis is that computer programs are executed. The health industry was dependent in the traditional scheme on other sectors to analyze big data. Because of its significant results, many health-care shareholders trust IT–their systems are functional and can process data in standardized forms. The health sector today faces the challenge of managing large health information quickly. The field of Big Data Analytics is on the rise and can be of use to the healthcare system. As noted above, most large amounts of this system's information are saved on paper and then digitized[21]. Big data may enhance and reduce health care costs while promoting advanced treatment, improving patient outcomes and avoiding unnecessary costs[22]. Big data analytics are now being used to predict the results of doctors ' choices, the result of cardiac surgery for a disease based on era, present situation and health. Basically, we can say that the function of big data is in managing healthcare-related information sets that are complicated and hard to handle using modern software, equipment and instruments. Besides the growing quantity of health information, the techniques for reimbursement change as well[23]. Thus, purposeful utilization and performance-based pay have proven to be significant variables in the health industry. In 2011, healthcare organisations produced more than 150 exabytes of data, all of which need to be analyzed efficiently to be of benefit to the healthcare system. Data are saved in a range of forms linked to healthcare in EHRs. A sudden rise in healthcare computer information was also noted in bioinformatics, where genomic sequence generates many terabytes of information. A number of analytical methods are accessible for medical interpretation and can then be used for patient care. Big data's various origins and forms challenge the healthcare information technology community to develop data handling techniques. There is a large demand for technology combining different sources of information. A number of conceptual methods can be used to detect irregularities in large numbers of information from various information sets.

Hive: Hive is an Hadoop data storage layer where tests and queries can be conducted using the procedure-like SQL language[24]. Ad-hoc queries, summarizations and information analysis can be carried out with Apache Hive. Hive is de facto regarded a standard for SQL-based information queries using Hadoop, providing simple information extraction, conversion and access characteristics including information files or other HBase storage systems[25]. Hive provides a de facto standard to HDFS.

Pig: Apache Pig is one of the open-source systems that are accessible for better big data analysis. The MapReduce programming tool Pig is an alternative[26]. First created as a study project by the Yahoo web service provider, Pig enables users to create their own user-defined functions and supports many traditional information transactions such as joining, sorting, filtering etc.

Hadoop MapReduce: The MapReduce computer often relates to Apache Hadoop.

The Calculation MapReduce model is a very strong toole that is more prevalent than many users know and used in many health apps. Its idea is very straightforward[27]. There are two phases in MapReduce: a mapping phase and a reduction phase. A mapping process is added to the input information during the mapping process. At the end of the count, the reduction stage is implemented[28]. There are also two phases to the MapReduce programming phase: a mapping stage that accept key value pairs and produces output in key value pairs and a second phase reduction, where each phase includes key value pairs as the input and output[29].

VII. IMPLEMENTATIONS

The potential of large-scale data is for it to revolutionize the results of the most appropriate or precise diagnosis for patients and the precision of the health information system[31].

As such, the study of large quantities of data will influence in five ways or "pathways" the medical service structure (see Figure 2). The healthcare system will concentrate on increasing patient results in these areas, as outlined below, and have an direct effect on the patient.

Right Living: Good living means a better, healthier life for the patient[32]. By living well, patients could take the best decisions themselves based on better choices and the improvement of their well-being through the use of data mining. Patients can be involved in the achievement of a good life by selecting the correct route to their daily health, in terms of diet, preventive treatment, exercise and other daily activities[33].

Right Care: This means that patients are treated as appropriately as possible and that all suppliers obtain the same information and are aimed at avoiding redundancy of effort and planning. In the age of large information, this element became more feasible.

Right Provider: Through the combination of information from different sources, such as medical facilities, government health statistics and socioeconomic information, healthcare providers can gain a general perspective of their clients. The availability of this data allows suppliers of human services to perform targeted inquiries and create the capabilities and capacities to identify and provide patients with better treatment choices.

Right Innovation: This way acknowledges the continuing evolution of new diseases, new treatments and new medical products. Progress in patient services revision, for instance, upgrade of drugs, and effective attempts in research and development will also provide fresh methods of promoting health and well-being through the domestic social insurance scheme. For stakeholders, the accessibility of early-test information is essential. These information can be used to define high-potential objectives and techniques for enhancing traditional clinical therapy methods.

Right Value: Providers must be cautious and constantly concerned with their patients to enhance the quality and value of health-related services. In their social insurance scheme, the patients must receive the most useful outcomes. Measures to guarantee smart information use include the identification and destruction of misrepresentation of information, manipulation and waste, and improvement of funds.

VIII. CONCLUSION

In this paper we presented a comprehensive description and brief overview of large-scale data and the health system, which play an important role in the computer sciences and greatly affects the healthcare system and the four-V big data in healthcare. We have also suggested using the Hadoop-based terminologies the use of a conceptual architecture to solve health issues with Hadoop-based big data produced by distinct concentrations of medical information and the creation of techniques to analyze these information and to provide responses to medical issues. The combination of large information and health analytics may lead to medicines that are efficient for individual patients and not for the majority of individuals, by offering the capacity to prescribe medicines that are suitable. Big data analytics, as we know, is in the early stages of development and existing instruments and methods can not solve the Big Data problems. Big data can be seen as large systems that present enormous difficulties. A lot of studies in this area will therefore be needed to address the problems facing the healthcare system.

REFERENCE

1. S. Ghemawat, H. Gobioff, and S. T. Leung, The Google file system, ACM SIGOPS Ope. Syst. Rev., vol. 37, no. 5, 2003
2. Welcome to Apache Hadoop. http://hadoop.apache.org/, 2017.
3. Apache HBase–Apache HBase Home, http://hbase.apache.org/, 2017.
4. J. Dean and S. Ghemawat, MapReduce: Simplified data processing on large clusters, Commun. ACM, vol. 51, no. 1, 2008.
5. M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, park: Cluster computing with working sets, in Proc. 2nd USENIX Conf. Hot Topics in Cloud Computing, Boston, MA, USA, 2010, p. 10.
6. Apache Hadoop, http://hadoop.apache.org/, 2018.
7. A. Katal, M. Wazid, R. H. Goudar, and T. Noel, Big data:Issues, challenges, tools and good practices, in Proc. 6th International Conference on Contemporary Computing, 2013, pp. 404–409.
8. Apache Hive, https://hive.apache.org/, 2018.
9. K. K. Y. Lee, W. C. Tang, and K. S. Choi, Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage, Computer Methods and Programs in Biomedicine, vol. 110, no. 1, pp. 99–109, 2013.
10. Apache Pig, https://pig.apache.org/, 2018.
11. E. Dede, B. Sendir, P. Kuzlu, J.Weachock, M. Govindaraju,and L. Ramakrishnan, Processing Cassandra datasets with Hadoop-streaming based approaches, IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 46–58, 2016.
12. Apache HBase, http://hbase.apache.org/, 2018.
13. Apache Oozie, https://oozie.apache.org/, 2018.
14. Apache Avro, https://avro.apache.org/, 2018.
15. Apache Zookeeper, https://zookeeper.apache.org/, 2018.
16. Apache Zookeeper, https://www.ibm.com/analytics/hadoop/zookeeper, 2018.
17. Apache Yarn, https://yarn.apache.org/, 2018.
18. Apache Sqoop, https://sqoop.apache.org/, 2018.
19. Apache Flume, https://flume.apache.org/, 2018.