BIG DATA ANALYSIS FOR FRAMEWORK COLLECTIONS

1. S. Hendry Leo Kanickam, Assistant Professor,
2.Mukilan, Student,
Information Technology,
St. Joseph's College(autonomous), Trichy, India

Abstract : In recent years, huge amounts stored in particulate server based on 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. They providing Server sector has been confronted by the need to manage the big data being produced by various sources and companies , which are well known for producing high volumes of heterogeneous data and homogeneous server data. Various big-data analytics tools and techniques have been developed for handling these massive amounts of data used to different kind frame models , in the Data Center . In this survey, we discuss the impact of big data to stored in server how will be worked to the framework, and various tools available in the Hadoop ecosystem for handling it. We also explore the conceptual architecture of big data analytics for Data Sciences which involves the data gathering history of different branches, the genome database, electronic health records, text/imagery, and clinical decisions support system.

Index Terms **- Big Data ,models 3Vs,Hadoop System, Data nodes**

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 heath 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. 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 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 analyze 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 totaled nearly 550 peta bytes and will reach nearly 26 000 peta bytes in 2020. 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. BIG DATA HEALTH MODELS

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 shareholders 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. Big data can improve healthcare delivery and reduce its cost, while supporting advanced patient care, improving patient outcomes, and avoiding unnecessary costs. Big data analytics is currently used to predict the outcomes of decisions made by physicians, the outcome 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 healthcare had produced more than 150 exabytes of data[10], all of which must be efficiently analyzed 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. 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. 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 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 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 add new medical test results, without the need for paper work or duplication of data.

III. THE BIG DATA TYPES

Volume: Big data is a term to referring to huge volumes of collected data. There is no fixed threshold for the volume of this data. Typically, the term is used with respect to massive-scale data which must be managed, stored, and analyzed using traditional databases and data processing architecture. The volume of data generated by modern IT and the healthcare system has been growing 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, efficiencies, and services to consumers.

Velocity: Velocity, which represents primaryreason for the exponential growth of data, refers to how fast data is collected[14]. Healthcare systems are generating data at increasingly higher speeds. In the volume and variety of the structured or unstructured data collected, the velocity of the generation of this data after processing requires a decision based on its output.

Variety : Variety refers to the form of the data, i.e., unstructured or structured, text, medical imagery, audio, video, and sensor data. Structured data information includes clinical data (patient record data), which must simply be collected, stored, and processed by a particular device. Structured data comprises just 5% to 10% of healthcare data. Unstructured or semi-structured data includes e-mails, photos, videos, audios, and other health related data such as hospital medical reports, physician's notes, paper prescriptions, and radiograph films **Veracity** : The veracity of data is the degree of assurance that the meaning of data is

consistent. Different data sources vary in their levels of data credibility and reliability.

Fig – 1 Model of Big Data

IV. HADOOP'S TOOLS AND TECHNIQUES FOR BIG DATA

To manage unstructured big data that does not fit into any database, special tolls are needed. To examine this type of big dataset, the IT sector uses the Hadoop platform for a wide variety of methods that have been developed to record, organize, and analyze this type of data[27, 28]. More efficient tools are needed to extract meaningful output from big data. Most of the tools are implemented in the Apache Hadoop architecture including Map Reduce, Mahout, Hive, and others[29]. Below, we discuss the various tools used in processing healthcare big datasets.

Apache Hadoop : The name Hadoop has evolved to mean many different things[23]. In 2002, it was established as a single software project to support a web search engine. Since that time, it has grown into
an ecosystem of tools and applications that are used to analyze large amounts and types of data[30]. Hadoop can no longer be considered to be a monolithic single project, but rather an approach to data processing that radically differs from the traditional relational database model[23]. A more practical definition of the Hadoop ecosystem and framework is the following: open source tools, libraries, and methodologies for "big data" analysis in which a number of data sets are collected from different sources, i.e., Internet images, audios, videos, and sensor records as both structured and unstructured data to be processed The HDFS block size is 64 MB 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. No data is actually stored on the name node. Files are stored as blocks in proper sequence and these blocks are equal in size. 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.

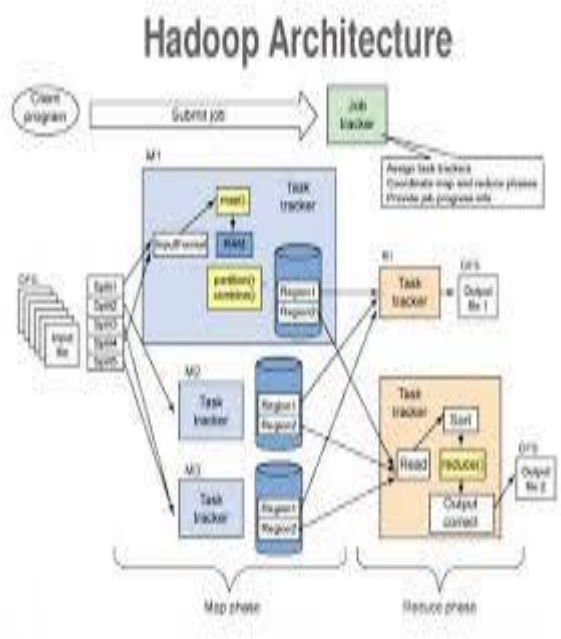

Fig-2 Hadoop Architecture
shows the physical layout architecture of Hadoop which consists of Map Reduce, HBase, and HDFS. HDFS: The HDFS was designed for processing big data. Although it can support many users simultaneously, HDFS is not designed as a true parallel file system. Rather, the design assumes a large file write-once/ read-many model that enables other optimizations and relaxes many of the concurrency and coherency overhead requirements of a true parallel file system. HDFS is designed for data streaming by which large amounts of data are read from disk in bulk.

Map Reduce : 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 Map Reduce, there are two stages: a mapping stage and a reducing stage. In the mapping stage, a mapping procedure is applied to input data. The reducing phase is implemented when counting.

Apache Hive : Hive is a data warehousing layer at the top of Hadoop, in which analyses and queries can be performed using SQL-like procedural language[32]. Apache Hive can be used to perform ad-hoc queries, summarization, and data analysis. Hive is considered to be a de facto standard for SQL based queries over peta bytes of data using Hadoop and offers the features easy data extraction, transformation, and access to the HDFS comprising data files or other HBase storage system.

Apache Pig : Apache Pig is one of the available open-source platforms being used to better analyze big data. Pig is an alternative to the Map Reduce programming tool[34]. First developed by the Yahoo web service provider as a research project, Pig allows users to develop their own user-define functions and supports many traditional data operations such as join, sort, filter, etc.

Apache HBase : HBase is a column-oriented NoSQL database used in Hadoop, in which user can store large numbers of rows and columns. HBase has the functionality of random read/write operations. It also supports record level updates, which is not possible using HDFS. HBase provides parallel data storage via the underlying distributed file systems across commodity servers. The file system of choice is typically HDFS, due to the tight integration of HBase and HDFS. If there is need for a structured low-latency view of the high-scale data stored via Hadoop, then HBase is the correct choice. Its open-source code scales linearly to handle petabytes of data on thousands of nodes.

Fig-3 Pig –Hbase Analysis

Apache Oozie : To run a complex system or tight system design or if there are a number of interconnected stations with data dependencies between them, there is a need for sophisticated technique called Apache Oozie. Apache Oozie can handle and run multiple jobs related to Hadoop. Oozie has two portions: workflow engines that store and execute workflow collections of Hadoop-based jobs and a coordinator engine that processes workflow jobs based on how they are designed in the process schedule. Oozie is designed to construct and manage Hadoop jobs as workflow in which the output of one job serves as the input for a subsequent job. Oozie is not a substitute for the Yarn scheduler. Oozie workflow jobs are represented as Directed Acyclic Graphs (DAGs) of actions[28]. Oozie plays the role of a service in the cluster and clients submit their jobs for proactive or reactive execution.

Apache Avro : Avro is a serialization format that makes it possible for data to be exchanged between programs written in any language[38]. It is often used to connect Flume data flows. The Avro system is schema-based, where the role of a scheme is to perform the read and write operations with the language being independent. Avro serializes the data that have a built-in schema. It is a framework for the serialization of persistent data and remote procedure calls between Hadoop nodes and between client programs and Hadoop services.

Apache Zookeeper : Zookeeper is a centralized system used by applications to maintain a healthcare system and provide organizing and other elements

CONCLUSION

In this paper, we have provided an in-depth description and a brief overview of big data in general and in Data system model, which plays a significant role in hadoop informatics and greatly influences the 3vs system and the big data four Vs in frameworks. We also proposed the use of a conceptual architecture for solving dynamic allocations for problems in big data using Hadoop-based terminologies, which involves the utilization of the big data, generated by different levels of medical data etc and the development of methods for analyzing this data and to obtain answers to medical questions. The combination of big data and analysis system ported.

REFERENCES

[1] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods and analytics, International Journal of Information Management, vol. 35, no. 2, pp. 137–144, 2015.
[2] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "Big Data", Hadoop and cloud computing in genomics, Journal of Biomedical Informatics, vol. 46, no. 5, pp. 774–781, 2013.
[3] C. L. P. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, vol. 275, pp. 314–347, 2014.
[4] M. Herland, T. M. Khoshgoftaar, and R.Wald, A review of data mining using big data in health informatics, Journal of Big Data, vol. 1, no. 1, p. 2, 2014.
[5] D. H. Shin and M. J. Choi, Ecological views of big data: Perspective and issues, Telematics and Informatics, vol. 32, no. 2, pp. 311–320, 2015.
[6] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. Basha, and P. Dhavachelvan, Big data and Hadoop-A study in security perspective, Procedia Computer Science, vol. 50, pp. 596– 601, 2015.
[7] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, Data mining with big data, IEEE transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2014.
[8] S. Sharma and V. Mangat, Technology and trends to handle big data: Survey, in Proc. 5th International Conference on Advanced Computing & Communication Technologies, 2015, pp. 266–271.
[9] R. Mehmood and G. Graham, Big data logistics: A health-care transport capacity sharing model, Procedia Computer Science, vol. 64, pp. 1107–1114, 2015.
[10] D. P. Augustine, Leveraging big data analytics and Hadoop in developing India healthcare services, International Journal of Computer Applications, vol. 89, no. 16, pp. 44– 50, 2014.
[11] MAPR, Healthcare and life science use cases, https:// mapr.com/solutions/industry/healthcare-and-lifescience-use-cases/, 2018.
[12] W. Raghupathi and V. Raghupathi, Big data analytics in healthcare: Promise and potential, Health Information Science and Systems, vol. 2, no. 1, p. 3, 2014.
[13] J. Sun and C. K. Reddy, Big data analytics for healthcare, in Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1525– 1525.
[14] C. Mike, W. Hoover, T. Strome, and S. Kanwal. Transforming health care through big data strategies for leveraging big data in the health care industry, http://ihealthtran.com/iHT2 BigData 2013.pdf, 2013.
[15] J. Anuradha, A brief introduction on big data 5Vs characteristics and Hadoop technology, Procedia Computer Science, vol. 48, pp. 319–324, 2015.
[16] M. Viceconti, P. J. Hunter, and R. D. Hose, Big data, big knowledge: Big data for personalized healthcare, IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1209–1215, 2015.
[17] Y. Sun, H. Song, A. J. Jara, and R. Bie, Internet of things and big data analytics for smart and connected communities, IEEE Access, vol. 4, pp. 766–773, 2016.
[18] A. Jain and V. Bhatnagar, Crime data analysis using Pig with Hadoop, Procedia Computer Science, vol. 78, pp. 571–578, 2016.
[19] T. Jach, E. Magiera, and W. Froelich, Application of Hadoop to store and process big data gathered from an urban water distribution system, Procedia Engineering, vol. 119, pp. 1375–1380, 2015.
[20] C. Uzunkaya, T. Ensari, and Y. Kavurucu, Hadoop ecosystem and its analysis on tweets, Procedia-Social and Behavioral Sciences, vol. 195, pp. 1890–1897, 2015.
[21] S. G. Manikandan and S. Ravi, Big data analysis using Apache Hadoop, in Proc. International Conference on IT Convergence and Security, 2014, pp. 1–4.
[22] V. Ubarhande, A. M. Popescu, and H. Gonzalez-Velez, Novel data-distribution technique for Hadoop in heterogeneous cloud environment, in Proc. 9th International Conference on Complex, Intelligent, and Software Intensive Systems, 2015, pp. 217–224.

Author's Profile:

Mr. S. Hendry Leo Kanickam working as a Assistant Professor in Department of Information Technology ,St. Joseph's College (autonomous) Trichy, India. He received his M.Phil Degree in Bharathidasan University in 2008 and also He is pursuing Ph.D (Computer Science) in Bharathidasan University.

Mr. M. Mukilan is studying II M.Sc Computer Science in the Department of Information Technology ,St. Joseph's College (autonomous) Trichy, India.