International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 2, No. 2, June 2017 – August 2017| Pages 156-165

Big Data for Healthcare

Kalaiarasi P, Assistant Professor, Department of CSE, Agni College of Technology, Chennai, India

Abstract:

Lately, tremendous measures of organized, unstructured, and semi-organized information have been produced by different foundations around the globe and, on the whole, this heterogeneous information is alluded to as large information. The wellbeing business area has been stood up to by the need to deal with the enormous information being created by different sources, which are notable for delivering high volumes of heterogeneous information. Different huge information investigation instruments and methods have been produced for dealing with these monstrous measures of information, in the medical services area. In this paper, we examine the effect of enormous information in medical services, and different devices accessible in the Hadoop biological system for taking care of it. We additionally investigate the reasonable design of enormous information examination for medical care which includes the information gathering history of various branches, the genome data set, electronic wellbeing records, text/symbolism, and clinical choices emotionally supportive network.

Introduction

At present medical care frameworks utilize various divergent and persistent observing gadgets that use particular physiological waveform information or discretized indispensable data to give ready systems if there should arise an occurrence of obvious occasions. In any case, such uncompounded methodologies towards improvement and execution of caution frameworks will in general be problematic and their sheer numbers could cause "alert weariness" for both guardians and patients [10–12]. In this setting, the capacity to find new clinical information is obliged by earlier information that has normally missed the mark concerning maximally using high-dimensional time arrangement information. The explanation that these caution components will in general fizzle is essentially on the grounds that these frameworks will in general depend on single wellsprings of data while lacking setting of the patients' actual physiological conditions from a more extensive and more far reaching perspective. Accordingly, there is a need to create improved and more complete methodologies towards considering connections and relationships among multimodal clinical time arrangement information. This is significant in light of the fact that reviews keep on demonstrating that people are poor in thinking about changes influencing multiple signs

Consistently, information is produced by a scope of various applications, gadgets, and topographical exploration exercises for the reasons for climate determining, climate expectation, fiasco assessment, wrongdoing discovery, and the heath business, to give some examples. In current situations, large information is related with center advancements and different undertakings including Google, Facebook, and IBM, which separate important data from the enormous volumes of information collected[1–3]. A time of open data in medical care is currently under way. Enormous information is being produced quickly in each field including medical care, as for understanding consideration, consistence, and different administrative prerequisites. Medical care investors are guaranteed new information from large information, purported both for its volume just as its multifaceted nature and reach. Drug industry specialists and investors have started to regularly investigate large information to get knowledge, however these exercises are as yet in the beginning phases and should be facilitated to address medical services conveyance issues and improve medical care quality. Early frameworks for large information examination of medical care informatics have been set up across numerous situations, e.g., the examination of patient qualities and assurance of therapy cost and results to pinpoint the best and most financially savvy treatments[4]. Wellbeing informatics is depicted as the osmosis of medical services sciences, registering sciences and data sciences in the investigation of medical care data. Wellbeing informatics includes information obtaining, stockpiling, and recovery to give better outcomes by medical care suppliers.

Clinical Image Processing from Big Data Point of View

Clinical imaging gives significant data on life structures and organ work notwithstanding identifying sicknesses states. Also, it is used for organ outline, recognizing tumors in lungs, spinal disfigurement analysis, supply route stenosis identification, aneurysm recognition, etc. In these applications, picture handling procedures, for example, upgrade, division, and denoising notwithstanding AI strategies are utilized. As the size and dimensionality of information increment, understanding the conditions among the information and planning productive, precise, and computationally viable strategies request new PC supported methods and stages. The fast development in the quantity of medical services associations just as the quantity of patients has brought about the more prominent utilization of PC helped clinical diagnostics and choice emotionally supportive networks in clinical settings. Numerous regions in medical care, for example, conclusion, guess, and screening can be improved by using computational insight [28]. The mix of PC investigation with suitable consideration can possibly assist clinicians with improving analytic exactness [29]. The incorporation of clinical pictures with different sorts of electronic wellbeing record (EHR) information and genomic information can likewise improve the precision and diminish the time taken for a conclusion.

- **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 add new medical test results, without the need for paper work or duplication of data.

Classification of neural networks

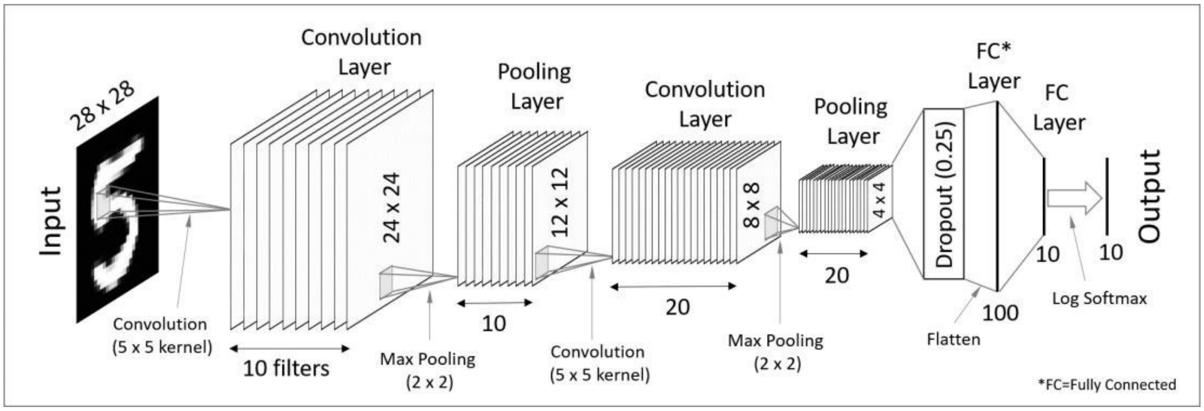
Shallow neural network: The Shallow neural network has only one hidden layer between the input and output.

Deep neural network: Deep neural networks have more than one layer. For instance, GoogLeNet model for image recognition counts 22 layers.

Nowadays, deep learning is used in many ways like a driverless car, mobile phone, Google Search Engine, Fraud detection, TV, and so on.

Convolutional neural networks (CNN)

CNN is a multi-layered neural network with a unique architecture designed to extract increasingly complex features of the data at each layer to determine the output. CNN's are well suited for perceptual tasks.



CNN is mostly used when there is an unstructured data set (e.g., images) and the practitioners need to extract information.

For instance, if the task is to predict an image caption:

- The CNN receives an image of let's say a cat, this image, in computer term, is a collection of the pixel. Generally, one layer for the greyscale picture and three layers for a color picture.
- During the feature learning (i.e., hidden layers), the network will identify unique features, for instance, the tail of the cat, the ear, etc.
- When the network thoroughly learned how to recognize a picture, it can provide a probability for each image it knows. The label with the highest probability will become the prediction of the network.

Hadoop-Based Applications for Health Industry

In light of the fact that healthcare data exists primarily in printed form, there is a need for the active digitization of print form data. The majority of this data is also unstructured, so it is a major challenge for this industry to extract meaningful information regarding patient care, clinical operations, and research. The collection of software utilities known as the Hadoop ecosystem can help the healthcare sector to manage this vast amount of data. The various applications of the Hadoop ecosystem in the healthcare sector are as follows:

- **Treatment of Cancer and Genomics:** We know that human DNA contains three billion base pairs. To fight cancer, it is vital that large amounts of data are efficiently organized. The patterns of cancer mutations and their reactions vary based on individual genetics, which explains the non-curability of some cancer. Oncologists have determined that in recognizing the patterns of cancer, it is important to provide specific treatment for specific cancers, based on the patient's genetic makeup. The Hapdoop technology MapReduce facilitates the mapping of three billion DNA base pairs to determine the appropriate cancer treatment for each particular patient. Arizona State University is working on project to develop a healthcare model that takes individual genomic data and selects a treatment based on identification of the patient's cancer gene. This model provides basis for treatment through big data analysis to improve the chances of saving patients lives.

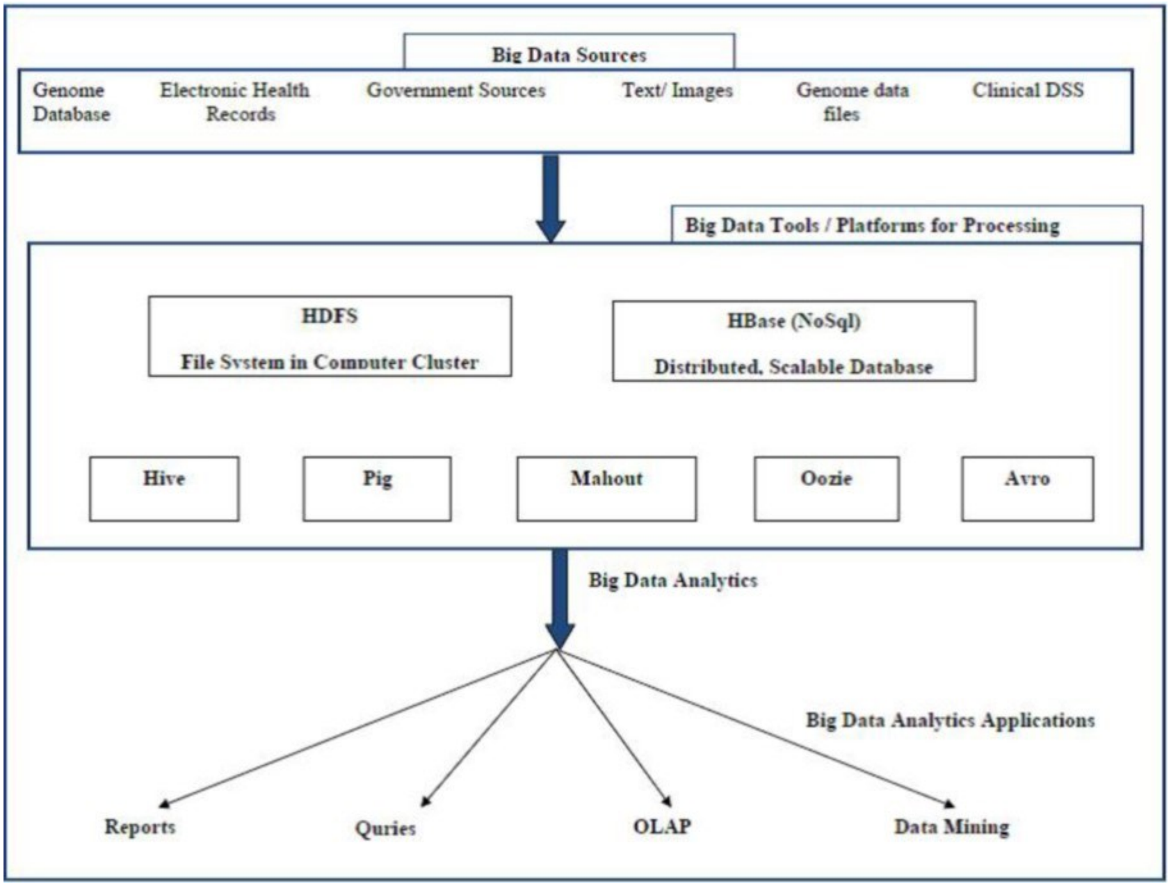
- **Healthcare Intelligence:** Hadoop technology also supports the healthcare intelligence applications used by hospitals and insurance companies. Hadoop ecosystem's Pig, Hive, and MapReduce technologies process large datasets related to medicines, diseases, symptoms, opinions, geographic regions, and other factors to extract meaningful information (e.g., desired age) for insurance companies.
- **Prevention and Detection of Frauds:** In the early faces of big data analytics, health-based insurance groups utilize multiple paths to identify fraud activity and establish methods to prevent medical fraud. With Hadoop, companies use applications based on a prediction model to identify those committing fraud via data regarding their previous health claims, voice recordings, wages, and demographics. Hadoop's NoSQL database is also helpful in preventing fraud related to medical claims at an early stage by the use of real-time Hadoop based health applications, authentic medical claim bills, weather forecasting data, voice data recordings, and other data

sources.

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 MapReduce, 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[22].



- **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 relaxes many of the concurrency and coherency overhead requirements of a true parallel file system.

MapReduce: Apache Hadoop is often associated with MapReduce computing. The MapReduce 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: a mapping stage and a reducing stage. In the mapping stage, a mapping procedure is applied to input data.

The MapReduce programming phase 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. A simplified view of MapReduce is shown in Fig. 5.

- **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 petabytes 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[33].
- **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 MapReduce 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[35], 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[36]. HBase provides parallel data storage via the underlying distributed file systems across commodity servers.

Conclusion

In this paper, we have given an inside and out portrayal and a short review of large information when all is said in done and in medical care framework, which assumes a huge job in medical care informatics and enormously impacts the medical services framework and the huge information four Vs in medical care. We additionally proposed the utilization of a calculated design for tackling medical care issues in huge information utilizing Hadoop-based wordings, which includes the use of the enormous information, created by various degrees of clinical information and the improvement of strategies for investigating this information and to get answers to clinical inquiries. investigation can prompt medicines that are powerful for explicit patients by giving the capacity to endorse suitable prescriptions for every person, instead of those that work for the vast majority. As we probably am aware, huge information investigation is in the beginning phase of improvement and current instruments and strategies can't take care of the issues related with large information. Huge information might be seen as large frameworks, which present colossal difficulties. Consequently, a lot of exploration in this field will be needed to address the issues looked by the medical services framework.

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] J. A. Patel and P. Sharma, Big data for better health planning, in *Proc. International Conference on Advances in Engineering and Technology Research*, 2014, pp. 1–5.

[12] A. E. Youssef, A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments, *International Journal of Ambient Systems and Applications*, vol. 2, no. 2, pp. 1–11, 2014.