



Systematic Approach -Big Data Analytics in Healthcare

¹J. Priya and ^{2*}R. Anandan

¹Department of Computer Science and Engineering,
Vels Institute of Science,
Technology and Advanced Studies (VISTAS),
Chennai, India.

priyajv14@gmail.com

^{2*}Department of Computer Science & Engineering,
Vels Institute of Science,
Technology and Advanced Studies (VISTAS),
Chennai, India.

anandan.se@velsuniv.ac.in

Abstract

This paper discuss most emerging feature of Big Data tools and challenges related to healthcare community. A collection of massive and complicated datasets, those are tough to process utilizing regular database administration or customary facts- preparing strategies. Big data analytics is the method of groping large as well as varied records units i.e., big facts - - to uncover hidden styles, unknown correlations, market traits consumer options and other beneficial statistics that can help groups make more-knowledgeable enterprise decisions. Big data Analytics is evolving most promising technology to field which generates plenty of data like healthcare related information. It is a sound methodology with great prospective use of the available data for competitive advantages. Big Data have been successful in many areas like healthcare, Banking and finance, telecommunication, risk analysis management, fraud detection management etc. This paper gives a wide outline of methodologies and challenges in big data analytics for the experimenter and practitioners in healthcare.

Key Words:Big data analytics, medical image, HDFS, healthcare, methodology.

1. Introduction

Global digital healthcare in 2012 was anticipated that the data has reached 1 lakh terabyte and is likely to attain 22,204 pebibyte in 2020 [1]. With the promising of current tools and all related strategies, it is forecast that there will be as plenty of information generated as turned into the maximum information that can generate in a planet of earth. The critical and superior aim of big data in healthcare is to provide personalized care to right person at right time. Effectively incorporating and professionally reading numerous forms of healthcare records over a time period can response lots of the imminent healthcare issues [2]. Modern expert systems and intelligent machine-learning systems aren't inherently green or extensible ample to process through huge quantity of information. For instance, for several duration, it's far known that maximum distribution-free method and finite dimensional methods require high estimation cost to place in the worldwide optima [3].

Nowadays most of the hospitals collecting healthcare information in the digitized form like EHR (Electronic Health Record) or EMR (Electronic Medical Record). EHRs or EMRs are real-time patient-centred records that make information available right away and firmly to authorized users. Such large volume of data can be analyzed by using machine learning algorithm such as supervised learning, unsupervised learning, reinforced learning, ensemble learning and deep learning. If we could apply any suitable technique on these data we can get clear data about disease so that prediction about diseases will become easy. Rectification of sicknesses and put in force the accurate treatment can remedy the patients from the illness and get well soon. Big data is regularly numerous information and tends to alternate hastily and transform its nature, making it tough to deal with it. Even now we can't imagine processing of such data without the help of Big Data techniques. Enterprises want tools to assist them in apprehending and analysing healthcare records easily and efficiently.

This significant heap of data often defined as Four-dimensional namely Volume, Velocity and Variety and Veracity (4). The dimension is in fig 1

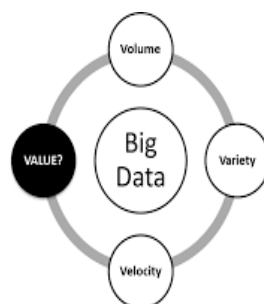


Fig. 1: Dimension of Big Data

Big data can be an aggregate of dependent, unorganized and semi-structured information. Any records that can be saved, accessed and processed in the form of fixed format is called as a structured' records i.e., where the format is well known in advance.

Information stored in a relational database management system is one example of a 'structured' record. Any records with unknown form or the structure is assessed as unstructured information. Usual instance of unstructured information is, a heterogeneous data source containing an aggregate of easy text documents, pictures, motion pictures and so on. Semi-structured information can comprise both the types of structured and unstructured information. Example of semi-structured records is a facts represented in XML record. The formats shown in fig.2

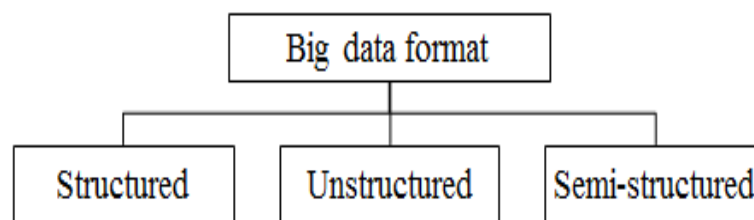


Fig. 2: Big Data Formats

2. Big Data Challenges in Healthcare

E-technological know-how regions are commonly data-extensive therein the standard of their effects upgrades with each volume and quality of information to be had.

However, modern smart machine-learning structures aren't inherently sufficient which finally ends up, in masses of instances, the improving fraction of these dense records undiscovered and underutilized.

It is not a small hassle whilst active techniques not succeed to seize such data enormity. While old-fashioned standards fail to prolong with trade, traditions and early practice turns into insufficient template for deciding what to do next. The significant barriers of big-data analytics in medical field consists of grabbing, warehousing, looking, exchanging and analyzing medical information.

Even the manner of data organisation which collected from different sources and integration of those data also a challenging mission. Especially in the medical field, namely about January 2000, the GenBank repository about nucleic acid series enclosed upon 178 bags of records [6] then the SWISS-PROT database on protein sequences contained in relation to 18 lot records [7,8].

So that in an average these records are become double for each and every 15 months. This is often additionally combined by means of records generated from the innumerable of associated task that examine the organic phenomenon that find out the super molecule structures and the information that interaction amongst these proteins. From that, we will begin to imagine the large quantity of information and variety of records this is often being created every month.

To provide the first-rate of medical treatment, big data analytics are essential in medical field for offering brief aid to sufferers on the way to get relief them from their medical complication at quicker rates and reducing the over dosage of medication.

Present days patients are capable of interact with the people those are residing the globe over, with the assist of recent technology. Those modern technologies are making progressive and maximum availability of the treatment to the pneumoconiotics. With the assist of these services sufferers can able, to get world class treatment with no time and they can also lessen their treatment cost by efficient ways. Examining the hospitals, preserving the cleanliness also reduces the scattering illnesses faster [6].

The information like anatomy and functioning details of organ can get from image called medical image by using these images practitioner can detect the disease states.

For example CAT scan is used for finding diseases in heart, lungs, pelvis, bones and liver. Mainly it is used for find different types of cancers and tumours and so forth. Furthermore, MRI (Magnetic Resonance Imaging) is used to provide a detailed image of tissues and organs. This is used to evaluate blood vessels, bones and joints, spinal injuries and so forth. Similar to this many types of strategies are there. To enhance these strategies moreover machine learning approaches are employed.

Many regions in fitness care including prognosis, diagnosis, and viewing are able to be advanced through utilising computing intelligence [10]. To improve the understandability of pictured contents is the intention of clinical image analytics [11].

Privacy and ownership of data is big issue. There is always conflict between patient and doctor's data.

Big data computing organization represent major classes depends on how records are examined with concern to time restraints [9]. First, instruction execution of huge quantity of stored information, which are in secondary disk, and not using a time constraints for example, MapReduce and GraphLab. Next, execution of in-memory information in actual-space of time or quick stage of time for example, hurricane, SAMOA. All the challenges are listed in Table 1.

Table 1: Challenges Facing Medical Image Analysis

Challenges	Description and possible solutions
Preprocessing	Clinical images suffer from different kinds of noise/artifacts and missing facts. Noise reduction, artifact elimination, lacking records coping with, evaluation adjusting, and so forth ought to enhance the high-quality of pictures and growth the performance of processing methods. Using multimodal information can be beneficial for this motive [22–24].
Compression	Decreasing the quantity of records while retaining essential records including anatomically relevant facts [23, 29, 25].
Parallelization/real-time realization	To finish analysis or processing develop strategies called scalable or parallel [27].
Registration/mapping	Aligning successive images or frames get scanning or any medical images like PET, X-Ray etc. [26, 27].
distribution/Authorization/anonymization	Reliability, isolation, and secrecy of data must be confined [23, 28–30].
Segmentation	Description of physical structure which includes bones and vessels [12, 27, 31].
	Locating dependencies/patterns amongst multimodal records and/or the facts captured at distinct time points if you want to boom the accuracy of diagnosis, prediction, and overall performance of the system [10, 11, 15, 32].
Data integration/mining	
Validation	Reviewing the overall presentation or correctness of the scheme/approach. Affirmation may be biased or unbiased. For the previous, annotated information is commonly needed [33–35].

3. Big Data Tools for Healthcare

Healthcare facts are collectively a set of Genomic records, medical facts and functioning records. Prepared EHR, formless EHR and clinical pictures are novel effects of medical information. Social networking records and the information generated through mobility sensor data are branch of functioning records. All open-source schemes are freely accessible on cloud like Hadoop and MapReduce with the aid of strengthen the software of big data analytics in medical field. Whilst in comparison with conventional data analytics and the user boundary for the gear is throughout unique even as the algorithms used are comparable. Tools are listed below in Table 2.

Table 2: Policy and Tools used For Big data analytics in healthcare

Platform and Tool	Specification
Hadoop Distributed File System (HDFS)	HDFS will divide massive data into small elements and accumulate those in dispensed nodes. Mainly it is developed to provide fault tolerant file system.
MapReduce	Processing part of HDFS is MapReduce. It processes the massive information which is stored in nodes.
PIG and PIG Latin	PIG is a platform for analyzing the massive data which contains high-level language. The language used for PIG platform is PIGLatin. By using PIG platform we can retrieve structured or unstructured information.
HIVE	HIVE is a query processing language like SQL. Even Hive and SQL are more symmetric to each other. It runs on Hadoop architecture.
JAQL	JAQL is also a query language designed to work on massive structured, semi-structured and unstructured information. It is especially designed for JSON (Java Script Object Notation) query processing.
ZOOKEEPER	Zookeeper is a project in Apache architecture which provides distributed services among clusters in Hadoop platform. Usually it distributes the information like configuration details, naming etc. It is an open-source software
Apache HBase	Apache HBase is a database which is used to store huge amount of information like real time healthcare data. It provides random access to the stored data. HBase run on top of the HDFS. HBase is a column-oriented database. It works on no-SQL.
Cassandra	It is also used for storing distributed data like HBase. Both are working in No-SQL information only, even there is no difference in the scheme of storing data. Only the difference is in an accessing method. It is an open-source.
Apache Oozie	Oozie is a server based workflow scheduler to coordinate various jobs in Hadoop. Furthermore, it is an open source scheduler.
Apache Lucene	Apache Lucene is a technology which is functioning as search engine. It is completely developed in Java. It is mainly used for adding searching feature in any website. It is also a freeware.
Apache Avro	Apache Avro is a system which is used for data exchanging with serialization. Avro will be language neutral.
Mahout	Mahout is a sub project of Lucene. It is also useful for searching but it is for advanced search. By using Mahout we can create scalable Machine learning techniques.

4. Hadoop Distributed File System (HDFS)

HDFS are also called as Nutch Distributed File System (NDFS). By using HDFS massive amount of data can store on file system. This file system is entirely developed in Java language. In case of different types of failure, HDFS will detect faults very quick and it will lead the system to recover automatically. Even 90% of GS (Google File System) architecture implemented by using HDFS.

MapReduce: As from the name MapReduce algorithm is going to reduce the complexity of processing in Hadoop. Just it will divide the massive information, which is stored in a cluster, into a number of small nodes and going to process on those. Later on it is going to recombine to get the overall result.

Pig and PigLatin: It is mainly used to analyze the massive data. PigLatin is a language which is used for Pig. Users can perform read, write and process the data by implementing their own functions with the help of the language called PigLatin.

Hive: Hive is designed especially for processing structured information. It placed in the top layer of HDFS. By using Hive can able to retrieve information with the help of queries like SQL. It is not developed for OLTP it is for OLAP. Hive is well-known, scalable and flexible.

Zookeeper: In a distributed environment the difficult task is co-ordination and management of services. This difficult task can achieve easily by Zookeeper. By using Zookeeper we can easily access applications in robust manner.

Jaql: It is used to access the information from file using query language designed to work on massive structured, semi-structured and unstructured information. It is especially designed for JSON (Java Script Object Notation) query processing. It converts high level query into low level queries.

HBase: Apache HBase is a database which is used to store huge amount of information like real time healthcare data. It provides random access to the stored data. HBase run on top of the HDFS. HBase is a column-oriented database. It works on no-SQL.

Cassandra: It is also used for storing distributed data like HBase. When we want high availability and scalability without giving up the performance then Cassandra is suitable. It provides high availability through high fault tolerance.

Oozie: Oozie is a scheduler for distributed environment to schedule the jobs in Hadoop environment. The sub-tasks of single complex task will combine by using Oozie. By Oozie multiple jobs can execute parallel to each other.

Lucene: Apache Lucene is a technology which is functioning as search engine.

It is completely developed in Java. It is mainly used for adding searching feature in any website. It is also a freeware. The content of Lucene file may be either SQL or Non-SQL.

Avro: Avro facilitates exchange of data among various nodes with serialization property. Users can exchange their information by using any languages, i.e., Avro will be language neutral.

Mahout: Mahout is a sub project of Lucene. It is also useful for searching but it is for advanced searching. Mahout is used by Facebook, Twitter, Yahoo and Adobe.

5. Concluding Remarks

In this paper we presented detail information about the big data challenges and tools for the healthcare data management. Big data analytics convert medical data by way of enhancing the consequences by a way of applying effective healthcare analytics. By using big data analytics medical advisors can able to take speed decisions and correct decisions. As in lots of other industries, information accumulating and control is getting larger, and specialists want assist in the matter. Especially in the field of healthcare the doctors have to take an accurate result at a right time compared to other fields. Nowadays, as we said earlier, bulky volume of medical data are generating every day, only with the help of traditional method a practitioner cannot able to provide any solution to save a patient's life. There may be a demand for big data analytics in healthcare centres than ever earlier than. So that with the trending field called big data analytics the practitioner can take perfect decisions at a right time on right patients

References

- [1] Hersh, et al. 2011. Health-care hit or miss? Nature, pp.470-476
- [2] Jensen et al., 2012, Translational Genetics "Mining electronic health records: towards better research applications and clinical care."
- [3] OY. Al-Jarrah et al., September 2015, Big Data Research-volume 2, "Efficient machine learning for big data: A Review" (Elsevier), pp.87-93.
- [4] IBM, (2015, September 25). Big data at the speed of business: Big data and analytics
- [5] NCBI, accessed on 2nd Aug. 2014, Genebank statistics.
- [6] W.Raghupathi et al., 2014, Big data analytics in healthcare: promise and potential, Health Information Science and Systems.
- [7] H.Elshazly, et al., 2013, "Hybrid system for lymphatic diseases diagnosis", Advances in Computing, pp.343-347.

- [8] F. Ritter, et al., 2011, "Medical image analysis," IEEE, pp. 60–70.
- [9] A.Baldominos, et al., 2014, "A scalable machine learning online service for big data real-time analysis", IEEE CIBD, pp.1-8.
- [10] D.Markonis, 2012, "Using MapReduce for large-scale medical image analysis," IEEE Conference on "Healthcare Informatics, Imaging and Systems Biology".
- [11] H.Yang, et al., 2010, Frontiers in Human Neuroscience, vol. 4, "A hybrid machine learning method for fusing FMRI and genetic data: combining both improves classification of schizophrenia".
- [12] OAJ.Toro et al., 2014, "Multi atlas-based segmentation with data driven refinement", The IEEE- BHI '14, pp. 605–608.
- [13] A. Tsymbal, et al., 2014, "Towards cloud-based image-integrated similarity search in big data," in Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI '14), pp. 593–596, IEEE, Valencia, Spain.
- [14] W. Chen, et al., 2010, "Intracranial pressure level prediction in traumatic brain injury by extracting features from multiple sources and using machine learning methods", IEEE BIBM '10, pp. 510–515.
- [15] R. Weissleder, 2006 "Molecular imaging in cancer", Science, vol. 312, no. 5777, pp. 1168–1171.
- [16] T. Zheng, et al., 2014, "Full-range in-plane rotation measurement for image recognition with hybrid digital-optical correlator", Optical Engineering, vol. 53, Article ID 011003.
- [17] L. Ohno-Machado et al., 2012, "iDASH: integrating data for analysis, anonymization, and sharing", Journal of the American Medical Informatics Association, vol. 19, no. 2, pp. 196–201.
- [18] C.-T. Yang et al., 2010, "Implementation of a medical image file accessing system on cloud computing", CSE' 10, pp. 321–326.
- [19] C. O. Rolim, et al., 2010, "A cloud computing solution for patient's data collection in health care institutions", ETELEMED '10, pp. 95–99, IEEE.
- [20] C.-C. Teng, et al., 2010, "A medical image archive solution in the cloud", IEEE ICSESS '10, pp. 431–434.
- [21] JG.Wolff, 2014, "Big data and the SP theory of intelligence", IEEE, vol-2, pp.301–315.
- [22] SW.Jun, et al., 2012, "ZIP-IO: architecture for application-specific compression of Big Data", FPT '12, pp.343–351.

- [23] B. Jalali ET AL., 2014, "The anamorphic stretch transform: putting the squeeze on 'big data'", Optics and Photonics News, vol-25, pp.24–31.
- [24] D. Feldman, et al., 2012, "The single pixel GPS: learning big data signals from tiny corsets", SIGSPATIAL '12, pp. 23–32.
- [25] L. Chiron, et al., 2014, "Efficient denoising algorithms for large experimental datasets and their applications in Fourier transform ion cyclotron resonance mass spectrometry", vol. 111, pp. 1385–1390.
- [26] AC.Gilbert, et al., 2014, "Recent developments in the sparse fourier transform: a compressed fourier transform for big data", IEEE, vol. 31, pp.91–100.
- [27] WY.Hsu, 2015, "Segmentation-based compression: new frontiers of telemedicine in telecommunication", Telematics and Informatics, vol. 32, pp.475–485.
- [28] FPM.Oliveira et al., 2014, "Medical image registration: a review", Computer Methods in Biomechanics and Biomedical Engineering, vol. 17, pp.73–93.
- [29] L. Qu, et al., 2015, "3D registration of biological images and models: registration of microscopic images and its uses in segmentation and annotation", IEEE Signal Processing Magazine, vol. 32, pp.70–77.
- [30] Ulutas, et al., 2011, "Medical image security and EPR hiding using shamir's secret sharing scheme", Journal of Systems and Software, vol. 84, pp.341–353.
- [31] Satoh, et al., 2013, "Teleradiology network system on cloud using the web medical image conference system with a new information security solution", Medical Imaging, vol. 8674.
- [32] CK.Tan, et al., 2011, "Security protection of DICOM medical images using dual-layer reversible watermarking with tamper detection capability", Journal of Digital Imaging, vol. 24, pp.528–540.
- [33] F.Wang, et al., 2011, "Hadoopgis: a high performance query system for analytical medical imaging with map reduce", USA.
- [34] N.Koutsouleris, et al., 2012, "Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the fepsy study", Schizophrenia Bulletin, vol. 38, pp. 1234–1246.

