Elysium Journal
Engineering Research & Management

P-ISSN: 2347-4408
E-ISSN: 2347-4734

ROLE OF BIG DATA ANALYTIC IN HEALTHCARE USING DATA MINING

[1]K.Sharmila [2]R.Bhuvana
Asst. Prof. & Research Scholar
Dept. of BCA & IT, Vels University, Chennai, India
[1]sharmilasenthil@ymail.com
[2]bhuvinavr1981@yahoo.co.in

Abstract— The paper describes the promising field of big data analytics in healthcare using data mining techniques. Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding for healthcare researchers and practitioners. In healthcare, data mining is becoming progressively more popular, if not increasingly essential. Healthcare industry today generates large amount of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices, etc. The large amount of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining provides a set of tools and techniques that can be applied to this processed data to discover hidden patterns and also provides healthcare professionals an additional source of knowledge for making decisions. In the past few eras, data collection related to medical field saw a massive increase, referred to as big data. These massive datasets bring challenges in storage, processing, and analysis. In health care industry, big data is expected to play an important role in prediction of patient symptoms, and hazards of disease occurrence or reoccurrence, and in improving primary-care eminence.

Index Terms— Big data, Analytics, Hadoop, Healthcare,

I. INTRODUCTION

Big data refers to very large datasets with complex structures that are difficult to process using traditional methods and tools. The term process includes, capture, storage, formatting, extraction, curation, integration, analysis, and visualization. A popular definition of big data is the "3V" model proposed by Gartner, which characteristics three fundamental features to big data: high volume of data mass, high velocity of data flow, and high variety of data types. Big data in healthcare refers to electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware; nor can they be easily managed with traditional or common data management tools and methods. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed. The totality of data related to patient healthcare and well-being make up "big data" in the healthcare industry. It includes clinical data from CPOE and clinical decision support systems (physician's written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative data); patient data in electronic patient records (EPRs).

The Table shows the growth of global big data volume and computer science papers on big data since 2009. This table exemplifies that stored data will be in the tens of zettabytes range by 2020, and research on how to deal with big data will grow exponentially as well.

TABLE 1: GLOBAL GROWTH OF BIG DATA AND COMPUTER SCIENCE PAPERS ON BIG DATA

Year	*Data Volume,ZBa,C*	*Conference Paper,CSb,C*	*Journal Papers,CSc*
2009	1.5	12	7
2010	2	26	7
2011	2.3	32	23
2012	3	78	47
2015	8	??	??
2020	44	????	????

a-Data from oracle-Data from Research Trends,cCS, computer science; ZB, zettabytes (1 zettabyte = 1000 terabytes = 106 petabytes = 1018 gigabytes, GB).. Please follow them and if you have any questions, direct them to the production editor in charge of your proceedings (see author-kit message for contact info).

This paper provides an outline of big data analytics in healthcare as it is evolving as a discipline and discuss the various advantages and characteristics of big data analytics in healthcare. Then we define the architectural framework of big data analytics in healthcare and the big data analytics application development methodology. Lastly, it provides examples of big data analytics in healthcare reported in the literature and the challenges that are identified.

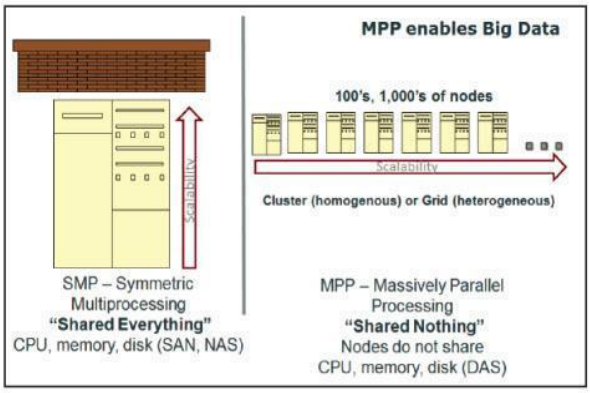
BIG DATA ANALYTICS IN HEALTHCARE:

Health data volume is expected to grow dramatically in the years ahead. It is vitally important for healthcare organizations that profit is not and should not be a primary motivator so it is necessary to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits. The chief application of Big Data in healthcare lies in two distinct areas. First, the

P-ISSN: 2347-4408
E-ISSN: 2347-4734

filtering of vast amounts of data to discover trends and patterns within them that help direct the course of treatments, generate new research, and focus on causes that were thus far unclear. Secondly, the complete volume of data that can be processed using Big Data techniques is an enabler for fields such as drug discovery and molecular medicine.

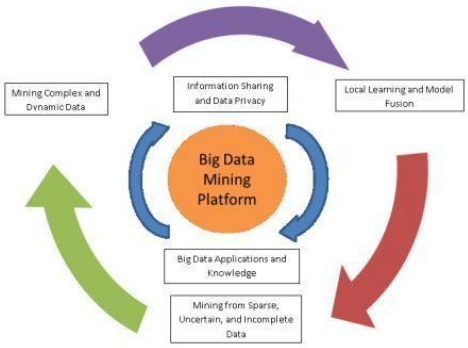
Big data can enable new types of applications, which in the past might not have been feasible due to scalability or cost constraints. In the past, scalability in many cases was limited due to symmetric multiprocessing (SMP) environments,On the other hand, MPP(Massively Parallel Processing) enables nearly limitless scalability. Many NoSQL Big Data platforms such as Hadoop and Cassandra are open source software, which can run on commodity hardware, thus driving down hardware and software costs.



BIG DATA PROCESSING FRAMEWORK:

A Big Data processing framework form a three tier structure and center around the "Big Data mining platform"

Tier I, which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high level semantics, application domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms.



The theoretical framework for a big data analytics project in healthcare is similar to that of a traditional health informatics or analytics project. The key difference lies in how processing is executed. In a regular health analytics project, the analysis can be performed with a business intelligence tool installed on a stand-alone system, such as a desktop or laptop. Because big data is by definition large, processing is broken down and executed across multiple nodes. Furthermore, open source platforms such as Hadoop/MapReduce, available on the cloud, have encouraged the application of big data analytics in healthcare.

PROTAGONIST OF BIG DATA IN HEALTHCARE:

Healthcare and life sciences are the fastest growing and biggest impact industries today when it comes to big data.Disease research is also being supported by big data to help tackle conditions such as diabetes and cancer. The ability to create and capture data is exploding and offers huge potentia to save both lives and scarce resources. Many NoSQL Big Data platforms such as Hadoop and Cassandra are open source software, which can run on commodity hardware, thus driving down hardware and software costs.

DATA MINING CHALLENGES WITH BIG DATA:

The limitations with big data include "adequacy, accuracy, completeness, nature of the reporting sources, and other measures of the quality of the data", and some of the Data Mining Challenges with Big Data in healthcare are inferring knowledge from complex heterogeneous patient sources, leveraging the patient/data correlations in longitudinal records, understanding unstructured clinical notes in the right context, efficiently handling large volumes of medical imaging data and extracting potentially useful information and biomarkers, analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layers of complexity and capturing the patient's behavioral data through several sensors; their various social interactions and communications.

TOOLS USED IN BIG DATA ANALYTICS HEALTHCARE:

The key obstacle in the healthcare market is data liquidity and some are using Apache Hadoop to overcome this challenge, as part of modern data architecture. Hadoop can comfort the soreness caused by poor data liquidity.

For loading the data the tool Sqoop efficiently transfers bulk data between Apache Hadoop and structured datastores such as relational databases. It import data from external structured datastores into HDFS or related systems like Hive and HBase. Sqoop can also be used to extract data from Hadoop and export it to external structured datastores such as relational databases and enterprise data warehouses. Sqoop works with relational databases such as: Teradata, Netezza, Oracle, MySQL, Postgres, and HSQLDB.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

To process the health data Depending on the use case, healthcare organizations process data in batch using Apache Hadoop MapReduce and Apache Pig; interactively with Apache Hive; online with Apache HBase or streaming with Apache Storm.

To analyze the data, the data once stored and processed in Hadoop can either be analyzed in the cluster or exported to relational data stores for analysis there.

CONCLUSION:

The big data has recently helped a major healthcare provider determine its strategy, use cases, and roadmap for utilizing it as part of their strategic plan through 2020. Perficient is currently assisting a client in using Big Data technologies for leveraging medical device data in real time. It is progressing into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Though it is progressing there are some challenges faced by big data analytics are, widespread implementation and guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention. The data is often contained within non-integrated systems, and hospitals and health systems lack the software applications needed to transform this data into actionable clinical information and business intelligence. In future these challenges are to be considered and so we can have health organizations can bring to the forefront better patient care and better business value.

REFERENCES

[1] Data Mining with Big Data, Xindong Wu1, Xingquan Zhu, Gong-Qing Wu, Wei Ding.

[2] Big Data and Clinicians: A Review on the State of the Science, Weiqi Wang, PhD; Eswar Krishnan, JMIR Med Inform 2014 | vol. 2 | iss. 1 | e1 | p.1

[3] Big Data Analytics for Healthcare, Jimeng Sun, Chandan K. Reddy.

[4] Data mining concepts,Ho Veit Lam- Nguyen Thi My Dung May- 14,2007

[4] Data Mining Over Large Datasets Using Hadoop In Cloud Environment,

[5] A Survey on Data Mining Algorithms on Apache Hadoop Platform, DR. A. N.Nandakumar,Nandita ambem2ISSN 2250-2459,ISO9001:2008Certified Journal, Volume 4, Issue 1, January2014)

[6] An Interview with Pete Stiglich and Hari Rajagopal on big data.

[7] Application of Data Mining Techniques to Healthcare Data,Mary K. Obenshain, MAT,,Infection Control and Hospital Epidemiology, August2004.

[8] Kuperman GJ, Gardner RM, Pryor TA, "HELP: A dynamic hospital information system". Springer-Verlag, 1991.

[9] A survey on Data Mining approaches for Healthcare,Divya Tomar and Sonali Agarwal, International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266.

D. S. Kumar, G. Sathyadevi and S. Sivanesh, “Decision Support System for Medical Diagnosis Using Data Mining”, (2011).