BIG DATA IN HEALTH CARE REVOLUTION – A SURVEY

V. Mageshwari, Dr. I. Laurence Aroquiaraj, T. Dharani

[1]Ph.D Research Scholar, Department of Computer Science, Periyar University, Salem, Tamilnadu, India
[2]Assistant Professor, Department of Computer Science, Periyar University, Salem, Tamilnadu, India
[3]Ph.D Research Scholar, Department of Computer Science, Periyar University, Salem, Tamilnadu, India

ABSTRACT- A foremost research challenge in today's world is data management and decision making. A massive amount of both unstructured and structured data is created every day so managing data has become an essential task. Health Care is one of the major areas where data has been generated tremendously and decision making based on the data has become very crucial. The need for data management and decision making in Health Care has directed the extension of 'Data Mining' to 'Big Data'. Big Data is an advancement of Data Mining, which can be termed as tremendously huge sets of data that may be examined computationally to reveal patterns, trends and associations. Big Data mainly deals with storage and processing of large sets of data. In this paper a survey on Big Data in Health Care has been made which gives us an overview on tools, techniques and algorithms in Big Data used for data management and decision making in Health Care.

Keywords: *Big Data, Data Management, Decision Making, Hadoop, HDFS, MapReduce.*

1. INTRODUCTION

Big Data is an advancement of Data Mining, which can be termed as tremendously huge sets of data that may be examined computationally to reveal patterns, trends and associations [19]. Big data is same as data, but it is bigger in size. The data which is beyond to storage capacity and processing power is termed as 'Big Data'. It mainly deals with storage and processing. By using big data concept the drawbacks regarding structuring a data can be rectified. Having bigger data requires different approaches, techniques, tools and architecture. Hadoop is one of the big data tool which is used in most of the organizations due to its HDFS (Hadoop Distributed File Systems) and MapReduce. The purpose of collecting and processing big data promotes making meaningful decisions. Big data analytics refers to the process of collecting, organizing, analyzing, inspecting, cleaning, transformation and modeling large sets of data to discover patterns and other useful information.

Decision making can be defined as deciding what to do [25]. Clinical decision making comprised of a balance of awareness, experience, knowledge and information gathering by using suitable assessment tools [32]. Clinical decision making is an important practice which usually follows a process of moving from gathering the essential patient's information through to the final decision and outcome. Since the amount of clinical data produced is enormously large, decision making has to be carried out more efficiently in order to get a better outcome. The evidence-based algorithms can be used to improve the clinical decision making.

2. LITERATURE SURVEY

Emad A Mohammed et.al., [1] summarized the modern efforts in clinical big data analytics and highlighted the need of enhancing the outcome of clinical big data analytics tools. They described the possible usage of the Hadoop platform and MapReduce programming framework to process massive amount of clinical data in fields related to medical health informatics.

Osden Jokonya et.al., [2] proposed a new big data framework to support with prevention, progression and control of HIV/AIDS, TB and silicosis (HATS) in the mining industry. They pointed a big data integrated framework which supports the prevention, progression and control of HATS in mining industry. They described a clear big data framework which helps in understanding the connection between HATS in the mining industry. The proposed framework has the potential of addressing the requirements of predictive epidemiology which is significant in forecasting and disease control in mining industry. Thus they laid a base for the use of feasible systems model and big data to tackle the challenges of HATS in mining industry.

Kavitha et al., [3] proposed a Fuzzy C-means Clustering Algorithm which produces a centroid-based clustering commencing a fixed set of examples. The proposed method helps in maintaining all type of health data with low costs, and presents correct intervention to the exact patient at the right time. The proposed method is beneficial for all the components of a HealthCare system like patient, payer, provider and management.

A.Fahad et al., [4] described the concepts and algorithms related to clustering in big data. They summarized a survey of existing clustering algorithms and provided a comparison, both from an empirical and theoretical perspective. As of a theoretical perspective he developed a sorting framework which is based on the major properties pointed out in earlier studies. Empirically, they conducted wide experiments and compared the most representative algorithm from every category by using a large number of valid big data sets. The efficiency of the candidate clustering algorithm is calculated through a lot of validity metrics such as stability, scalability and runtime tests. In addition they highlighted the bunch of best clustering algorithms which are best performing for big data.

3. Big Data

Big data is same to data, but it is bigger in size [19]. The data which is beyond to storage capacity and processing power is termed as 'Big Data' [22]. Since there is a massive amount of both structured and structured data produced every day, it is difficult to process using traditional database and software techniques. In most scenarios the data will be too big or it moves too fast and exceeds current processing capacity. To overcome these drawbacks big data has emerged. Having bigger data requires different approaches, techniques, tools and architecture [22]. Big data is generated by Health Care, Bio-Technology, Social Network, Social Media, Weather Forecasting, Education Data, Banking, Insurance, Finance, Retail & Real Estate and Agriculture. The purpose of collecting and processing all of the Big data assist us in making meaningful decisions.

4. Big Data Analytics

Big data analytics refers to the process of collecting, organizing, analyzing, inspecting, cleaning, transformation and modeling large sets of data to discover patterns and other useful information [19]. The analytical findings can lead to improved operational efficiency, more effective marketing, best customer service and viable advantages over competitor organizations and additional business benefits. Big data analytics can assist organizations to better understand the information contained within the data and it will also help discover the data that is more important to the future business decisions. The types of Big Data Analytics are descriptive, predictive and prescriptive [1].

- Descriptive Analytics – it is to summarize what has happened
- Predictive Analytics – It forecasts what might happen in the future. Probabilistic in nature.
- Prescriptive Analytics – it needs to prescribe an action, so the decision maker can take this information and act.

5. BIG DATA SOLUTIONS IN HEALTH CARE

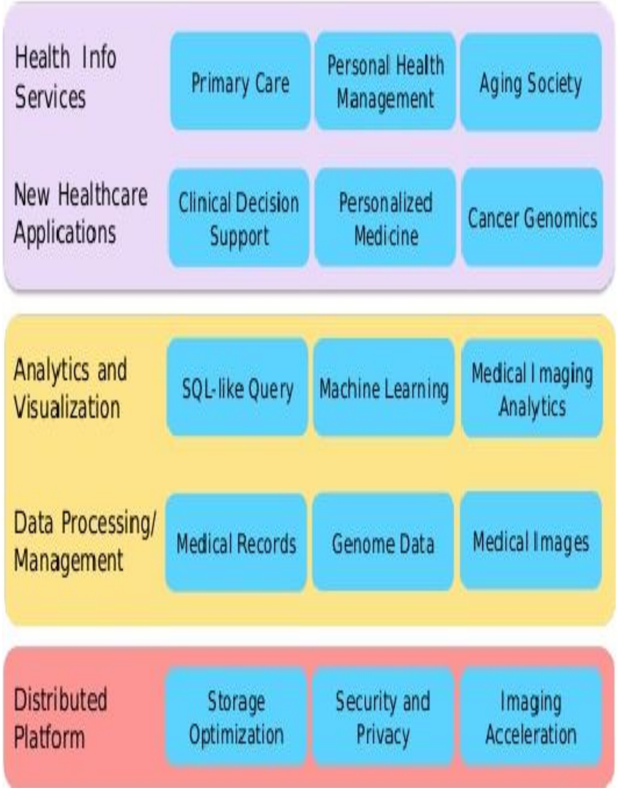


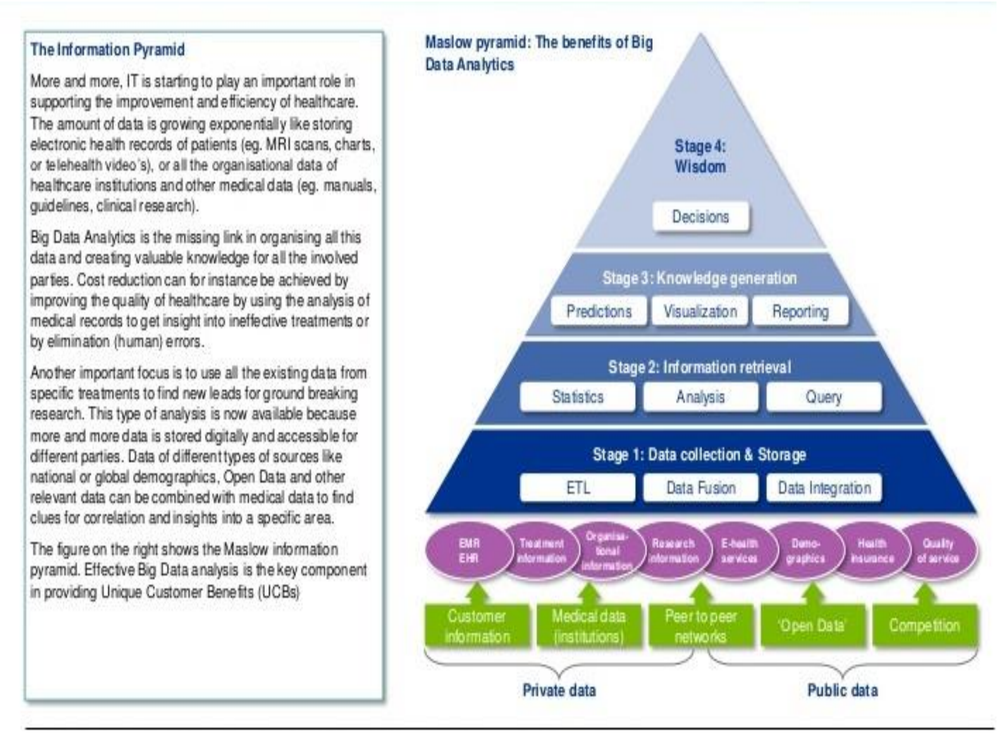
Fig-1: Big Data Solutions for Health Care

Since the number of health records are increasing more than millions and billions the computational technology and infrastructure should be able to provide a cost efficient implementation of

- Parallel Data Processing
- Provide storage for millions and billions of unstructured data sets.
- High availability of systems along with fraud tolerance.

6. Challenges of Big Data in Health Care

During emerging period “Big Data” had only four characteristics such as volume, velocity variety and value [1]. Then it has extended to inclusion of veracity and finally now it has extended to variability and visualization [18].



The Information Pyramid

More and more, IT is starting to play an important role in supporting the improvement and efficiency of healthcare. The amount of data is growing exponentially like storing electronic health records of patients (eg. MRI scans, charts, or telehealth video's), or all the organisational data of healthcare institutions and other medical data (eg. manuals, guidelines, clinical research).

Big Data Analytics is the missing link in organising all this data and creating valuable knowledge for all the involved parties. Cost reduction can for instance be achieved by improving the quality of healthcare by using the analysis of medical records to get insight into ineffective treatments or by elimination (human) errors.

Another important focus is to use all the existing data from specific treatments to find new leads for ground breaking research. This type of analysis is now available because more and more data is stored digitally and accessible for different parties. Data of different types of sources like national or global demographics, Open Data and other relevant data can be combined with medical data to find clues for correlation and insights into a specific area.

The figure on the right shows the Maslow information pyramid. Effective Big Data analysis is the key component in providing Unique Customer Benefits (UCBs)

Fig-2: Big Data in the Healthcare Sector

Volume – Every day a large volume of data is created in Health Care. Now there is more data when compared to before.

Velocity – Data is created in a faster way from various Health Care clinic's that no one has imagined.

Value – With increased volume of data being collected from various medicals center, the important thing is that the data must be relevant to the patient's medical history so that it can be carried out for research purpose.

Variety – Maximum amount of data created is unstructured, that is it comes in a variety of forms such as the data indicating patient record, or dataset that is used for research activities etc.

Veracity – It deals with quality of data. If data is unreliable for expert's to make decisions about patient then nothing can be done with it which leads to quality results.

Variability – The meaning of a data is changing constantly. In health care, the data which is true for one patient may not necessarily mean that it's true for another.

Visualization – Now a days there is a lot of patients and a lot of data is produced which is little difficult to visualize for survey. So data can be made understandable through images such as graph, chart or picture.

7. TOOLS USED FOR BIG DATA

There are a lot of tools available for Big Data. A list of best Big Data tools [34] for developers are listed below

- *Splice Machine* – This tool is SQL 99 compliant with standard ANSI SQL which has a capacity to scale from gigabytes to petabytes.

- *MarkLogic* – This tool is constructed to deal with intense data loads and allow users to access it through real-time alerts and updates.
- *Google Charts* – This tool is a free tool which comes with diverse capabilities for visualizing data from a website such as simple charts or hierarchical tree maps. This tool is simply implemented by embedding JavaScript code on a website.
- *SAP inMemory* – This tool provides an ability to integrate and analyze large workloads of data which is to be analyzed in real-time.
- *Cambridge semantics* – This tool helps to merge data from several sources and customized dashboards to make analysis easy.
- *MongoDB* – This tool is an open-source documental database which is ideal for developers who need to have a specific control over the final outcome.
- *Pentaho* – This tool combines the data integration with business analytics for visualizing, analyzing and combining Big Data.
- *Talend* – This tool is mainly designed for data management through developing, testing and deploying data.
- *Tableau* – The important feature of this tool is its in-memory analytics database & advanced query language.
- *Splunk* – This tool specializes usually in binding machine data created from a number of different sources, such as applications, websites and sensors.

6. CLUSTERING ALGORITHMS IN BIG DATA

There are various algorithms which are used in data mining. As such there are several algorithms which are used in big data. They are listed below

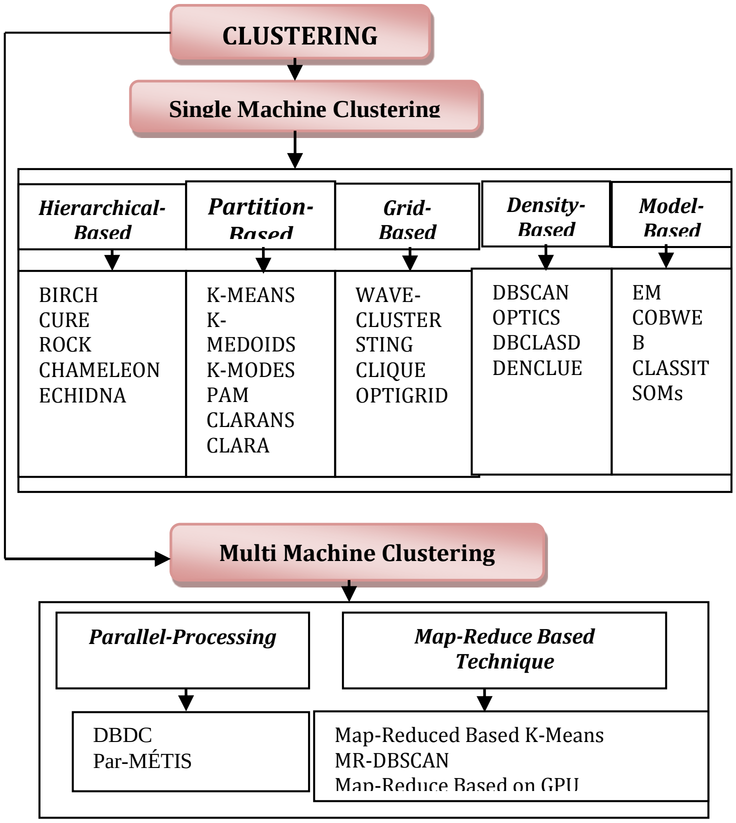


Fig-3: Set of Clustering Algorithms in Big Data

The above are the few clustering algorithms which are used for clustering a group based on some conditions. There are few algorithms which are used only for single machine clustering. For multi-machine clustering parallel processing and Map-Reduced based technique are used.

9. CONCLUSION

Health Care is one of the major areas where data has been generated enormously. So the need for data management and decision making in Health Care has become very important. In this paper a survey on Big Data in Health Care has been made which gives us an overview on tools, techniques and algorithms in Big Data used for data management and decision making in Health Care.

REFERENCE

[1] Mohammed et al., "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends", BioData Mining 2014, 7:22, doi:10.1186/1756-0381-7-22.
[2] Osden Jokonya, "Towards a Big Data Framework for the prevention and control of HIV/AIDS, TB and Silicosis in the mining industry", in CENTERIS 2014 – Conference on ENTERprise Information Systems/ ProjMAN 2014 – International Conference on Project MANagement/ HCIST 2014 – International Conference on Health and Social Care Information Systems and Technologies, Procedia Technology 16(2014) 1533 – 1541, doi:10.1016/jj.protcy.2014.10.175.
[3] V.Kavitha and S. Kannudurai, "HEALTH CARE ANALYTICS WITH HADOOP BIG DATA PROCESSING", in International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST), Vol.2, Special Issue 15, March 2016, ISSN 2395-695X.
[4] A. Fahad, N. Alshatri & Z. Tari, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", in IEEE Transactions on Emerging Topics in Computing, 2014, doi:10.1109/TETC.2014 2330519.
[5] https://www.dezyre.com/article/5-healthcare-applications-of-hadoop-and-big-data/85
[6] https://www.dezyre.com/article/hadoop-components-and-architecture-big-data-and-hadoop-training/114
[7] A. Vararuk I. Petrounias V.Kodogiannis, (2007), 'Data mining techniques for HIV/AIDS data management in Thailand", Journal of Enterprise Information Management, Vol. 21 Iss 1 pp.52-70.
[8] Wei Dai and Wei Ji, "A MapReduce Implementation of C4.5 Decision Tree Algorithm", in International Journal of Database Theory and Application, Vol.7, No.1 (2014) pp.49-60.
[9] Hailu, T.G. (2015), "Comparing Data Mining Techniques in HIV Testing Prediction", Intelligent Information Management, 7, 153-180.http://dx.doi.org/10.4236/iim.2015.73014.
[10] Miss. Harshada S. Deshmukh, Prof.P.L.Ramteke, "COMPARING THE TECHNOLOGIES OF CLUSTER ANALYSIS FOR BIG DATA", in International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 12, December 2015.
[11] Wullianallur Raghupathi and Viju Raghupathi, "Big data analytics in healthcare: promise and potential", in Health Information Science and Systems 20142:3.
[12] Thulasi Bikku, N.Sambasiva Rao and Anand Rao Akepogu, "Hadoop based Feature Selection and Decision Making Models on Big Data" in Indian Journal of Science and Technology, Vol 9(10), DOI: 10.17485/ijst/2016/v9i10/88905, March 2016, ISSN: 0974-5645
[13] http://dx.doi.org/10.11555/2015/370194
[14] https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/hiv-in-your-body/stages-of-hiv/
[15] http://www.aidsmap.com/Viral-load/page/1327496/
[16] http://www.aidsinfonet.org/fact_sheets/view/403
[17]http://www.unaids.org/sites/default/files/media_asset/GARPR_2014_guidelines_en_0.pdf
[18]https://vitalrecord.tamhsc.edu/big-data-health-care-revolution-7-vs-big-data/
[19]http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics
[20]http://www.snia.org/sites/default/education/tutorials/2013/fall/BigData/SergeBazhievsky_Introduction_to_Hadoop_MapReduce_v2.pdf
[21] Boris Lublinsky, Kevin T. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", John Wiley & Sons,Inc., ISBN: 978-1-118-61193-7, ISBN: 978-118-61254-5 (ebk), 2013.

[22] Judith Hurwitz, Alan Nugent, Dr. Fern Halper and Marcia Kaufman, "Big Data For DUMMIES", John Wiley & Sons,Inc., ISBN:979-1-118-50422-2(pbk); ISBN: 978-1-118-64417-1(ebk), 2013.
[23] Dlrk deRoos, Paul C.Zlkopoulos, Roman B. Melnyk Ph.D, Bruce Brown and Rafael Coss, "Hadoop for Dummies", John Wiley & Sons,Inc., ISBN: 978-1-118-60755(pbk); ISBN: 978-1-118-65220-6(ebk), 2014.
[24]GGT Dantanarayana, Tony Sahama and GN Wikramanayake, "Quality of Information for Quality of Life: Healthcare Big Data Analytics", 2015 International Conference on Advances in ICT for Emerging Regions(ICTer) ,24th and 25th August 2015.
[25] http://hrcak.srce.hr/76791
[26]https://hissjournal.biomedcentral.com/articles/10.1186/2047-2501-2-3.
[27] https://www.siam.org/meetings/sdm13/sun.pdf
[28] ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7338161
[29]http://ebiquity.umbc.edu/_file_directory_/papers/808.pdf
[30]http://eecs.wsu.edu/~yinghui/mat/courses/fall%202015/resources/Big%20data%20for%20dummies.pdf
[31]http://faculty.washington.edu/blabob/bob/eBooks/Hadoop%20for%20Dummies%20(eBook).pdf
[32] Vithyaa T, Manivannan K.A survey on healthcare Bigdata clustering and mining. Indian Journal of Engineering, 2016, 13(31), 34-50.
[33]https://www.healthit.gov/policy-researchers-implementers/clinical-decision-support-cds
[34]http://www.cbronline.com/news/big-data/analytics/10-of-the-most-popular-big-data-tools-for-developers-4570483

BIOGRAPHIES

V.Mageshwari received her M.Phil in Computer Science from Bharathiar University, Coimbatore, Tamil Nadu, India in 2015. She is pursuing Ph.D Degree in Department of Computer Science, Periyar University, Salem, Tamil Nadu, India. She has published six articles in International journals, International Conference and National Conference. Her area of interest includes Big Data, Image processing and Data Mining.

Dr. I. Laurence Aroquiaraj has received the M.Sc., Computer Science Degree from Pondicherry University, Pondicherry, India in 2002. He has received the M.Phil., Degree from Manonmaniam Sundaranar University, India in 2003. He received his M.Tech., Computer Science Engineering Degree from Kalinga University, India in 2004, and Master of Computer Applications Degree from Periyar University, India in 2008. He obtained his Ph. D. Degree from the Department of Computer Science, Periyar University, Salem, Tamil Nadu, India in 2015. He is working as Assistant Professor in Computer Science Department, Periyar University, Salem, Tamil Nadu, India. His research interest includes Medical Image Processing, Biometrics, Big Data, Pattern Recognition and Networking.

T.Dharani received her M.Phil in Computer Science from Periyar University, Salem, Tamil Nadu, India in 2013. She is pursuing Ph.D Degree in Department of Computer Science, Periyar University, Salem, Tamil Nadu, India. She has published six articles in International journals and International Conferences. Her area of interest includes Image processing and Data Mining and Big Data.