Healthcare Big Data Analysis using Hadoop MapReduce

A. **S. Thanuja Nishadi**

Faculty of Graduate Studies, University of Colombo, Sri Lanka

DOI: 10.29322/IJSRP.9.03.2019.p87104
http://dx.doi.org/10.29322/IJSRP.9.03.2019.p87104

Abstract- The large volumes of healthcare big data are rapidly generating all over the world in numerous ways. Therefore, significant amount of money has been allocating for healthcare industry for treatments, diagnosis and other research and development areas in handling healthcare big data. Further, patients are unnecessarily spending time, effort and money, due to lack of telemedicine support. However, the rapid growth of unstructured healthcare data does not support for existing big data analyzing technologies. Therefore, the study suggests Hadoop MapReduce for store and process medical data to avoid the modern issues in healthcare big data analysis.

Index Terms- Big Data, Hadoop, HDFS, Healthcare Big Data, Map Reduce

I. INTRODUCTION

Healthcare big data refers to the vast quantities of data that is available to healthcare providers. As a result of the rapid digitalization of healthcare information and the rise of value-based care, the healthcare sector is searching opportunities for handle data in order to implement strategic business decisions. In order to face the challenges of healthcare big data including volume, velocity, variety, veracity, variability and value, health care systems need to adopt technology capable of handling acquisition, storage, management, analysis and visualization.

Hadoop is an Apache open source framework, primarily used to store and process the large data sets across cluster computing using reliable and scalable methods. In order to facilitate storage and process, Hadoop has two components i.e. Hadoop Distributed File System(HDFS) for storing data and MapReduce Framework retrieval and process data. Moreover, modern healthcare systems need to handle large volumes of batch data which successfully manage by Hadoop.

The aim of the study is to analysis of healthcare big data domain including definitions, life cycle of big data, stream and batch processed data and the current issues of healthcare big data. In addition to that, this will suggest Hadoop as a solution by investigating the features provided by the Hadoop. Furthermore, the current study proposes a model for healthcare big data handling with Hadoop. The section II of the study indicates big data in healthcare, section III of the study express the process of healthcare big data, section IV differentiates the batch and stream processes, section V of the study suggests the Hadoop for healthcare big data handling and finally, in section VI will propose Hadoop big data handling framework.

II. BIG DATA IN HEALTHCARE

Healthcare big data refers to the large volumes and complex of electronic data sets which are difficult to manage by using traditional software and hardware systems (Frost & Sullivan, 2014). However, the healthcare big data is overwhelmingly not only the volumes of it but also due to other factors such as the complexity and diversity of it. Ristevski and Chen (2018) defines the characteristics of big data in 6V model as Volume (large volumes of healthcare data generated), Variety (diversification of data types and sources), Velocity (speed of healthcare data transactions), Veracity (guarantee of the data quality), Variability (consistency of the data over the period) and Value (real added value for the healthcare sector). Therefore, the term big data is not only associated with volumes of it, but also it consists with other complex features.

The global healthcare data is rapidly increasing due to the large volumes of processing and storing of medical transactions by using diversified medical sources of data. The healthcare big data generates in structured, semi-structured and unstructured formats which acquired from primary sources. The recent study conducted by Oracle Company indicates that around 90% of the global data is held in unstructured formats (Farhangmehr, 2014). Further, the sources of big data include Electronic Health Records (eg. Physician notes, lab reports, ECG scans, X-Ray, health sensor devices, medical prescriptions etc.), image processing (Computed Tomography(CT), photo acoustic imaging, ultrasound, molecular imaging, Magnetic Resonance Imaging (MRI), mammography fluoroscopy, positron emission tomography-computed tomography (PET-CT), and X-ray), social media, smart phones and web data (Nambiar et al., 2013). Thus, the modern healthcare big data massively generates all the complex formats.

III. PROCESS OF HEALTHCARE BIG DATA

There are many issues related to healthcare big data from its acquisition to report generation. The big data follows five major processes including data acquisition, data storage, data

management, data analysis and data visualization (Senthilkumar, 2018).

Data Acquisition: The first stage of the biomedical big data life cycle is gathering or collection of data. This is mainly related to acquiring or collecting data from heterogeneous biomedical sources such as structured, semi-structured, and unstructured formats (medical images, medical devices, biomarkers, genomic, general health and clinical data).

However, one of the major challenges in big data is data cleaning which indicates what is useful to capture and what is useless to discard (Zhang et. al., 2015). Further, data provenance is an another real challenge in healthcare big data which used to describe the integrity and history of digital objects, where they came from and how they came to be in their present state or the state of the dada (Curcin, 2017). In health datasets, provenance is used to deliver auditability and transparency, and to achieve trust in a software system (Shen Xu et. al., 2016). Therefore, it will help to indicate the next processing steps due to recordings of the origin and movement of data in the processing pipeline. However, if there is an error happened during one stage, it will affect to all the subsequent analysis stages (Galvic,2015). Furthermore, the third challenge in health care big data is automatic generation of metadata. Metadata provides the information about the meaning of data, terminology, concepts, relationships of data and information about the source of the data provenance (Bilalli et, al., 2015). Therefore, the efficient analytical algorithms are required to understand the provenance of data and process of the vast streaming of data before storing (Zhang et, al., 2015).

Data Storage: Healthcare is demanding more storage space for big data analytics which support for the large volumes of unstructured datasets (Herland et. al., 2014). Thus, it is essential to reveal patterns that is used for diagnosis of diseases and reveal behaviors which may directly influence to patient's health. Further, it is highly required to find a solution for the growth of digital tools which are being brought from both patients and clinicians including wearable devices, applications, EHR etc. Human genome is such a massive area which generates hundreds of gigabytes of data and also sequence of data is doubling every seven to nine months (Kaitoua et. al, 2017). Therefore, traditional data analysis is unfit to manage those systems; hence, it is required to adapt more scalable and reliable methods to ensure high storage capabilities.

Data Management: Managing heterogeneous data is complex due to its diversification. The process of data management ensures the accuracy of data which available in real time and streaming for use and visualization (Wu, et. al., 2014). At this stage, it further ensures the validations of the current data. Not only for that, all the segments of data handling need to manage properly. There are policies related to the data management in some organizations. Proposed Hadoop provides solutions for the identified issues of healthcare big data.

Data Analysis: Methods of analyzing big data is required to supports for heterogeneous data which are inherited from diversified sources. In here, it runs the code or algorithm that makes the calculations that will lead to the actual results. In general, the major types of analysis include descriptive, diagnosis, predictive and prescriptive analysis.

Descriptive Analytics: This is used to describe the current situations and reporting of data that using descriptive tools such as histograms and charts.

Diagnostic Analysis: The aim of this stage is to explain why certain events are occurred and to identify the factors of triggering of those events.

Predictive Analysis: The main activities of this stage are predicting future events, identifying trends and determining the probabilities of uncertain events.

Prescriptive Analytics: Suggestions of suitable actions which leads to optimal decision-making is considered in this stage.

Data Visualization: It is expected to produce multiple outputs in this stage such as visualization of patient health monitoring reports and decision reports. Therefore, visualization needs to support both batch processing and real-time analytics in order to optimize the decision-making and avoid emergencies in medical systems.

IV. BATCH PROCESSING AND STREAM PROCESSING

The processing of big data is classified as two base forms including batch and stream processing (Shahrivari, 2014). Batch processing is based on analyzing data over a period of time but no constraints based on response time (Dean, 2008). In addition to that, batch processing allows for large volumes of data which are collected and stored. Real-Time Processing is used for applications which requires real-time feedback. Thus, modern healthcare systems need to process the transactions of both batch and real-time data.

Batch Processing: In batch processing, it processes blocks of data that already stored over a period of time such as weekly or end of the day. Therefore, the process takes large amount of time for process the data. Further, the batch processing is used the situations which does not suite real time analytics.

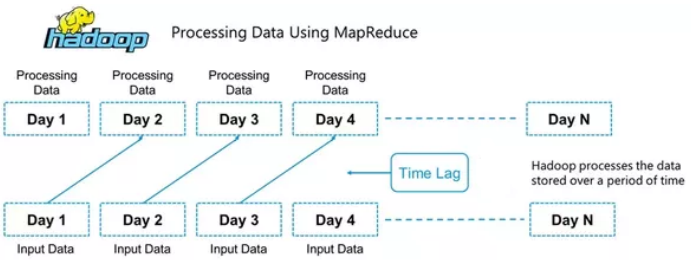

Figure 1: Hadoop Map Reduce for Batch Processing

Hadoop MapReduce is an ideal solution for batch processing, which used to store over a period of time such as hourly, daily, weekly and monthly etc.

V. PROPOSED HADOOP SOLUTION FOR HEALTHCARE BIG DATA

The current healthcare big data generates from numerous sources such as imaging, genomics, clinical data, bio informatics and general health. The proposed framework illustrates how it follows the life cycle of big data, issues generates of it and the proposed batch processing Hadoop MapReduce.

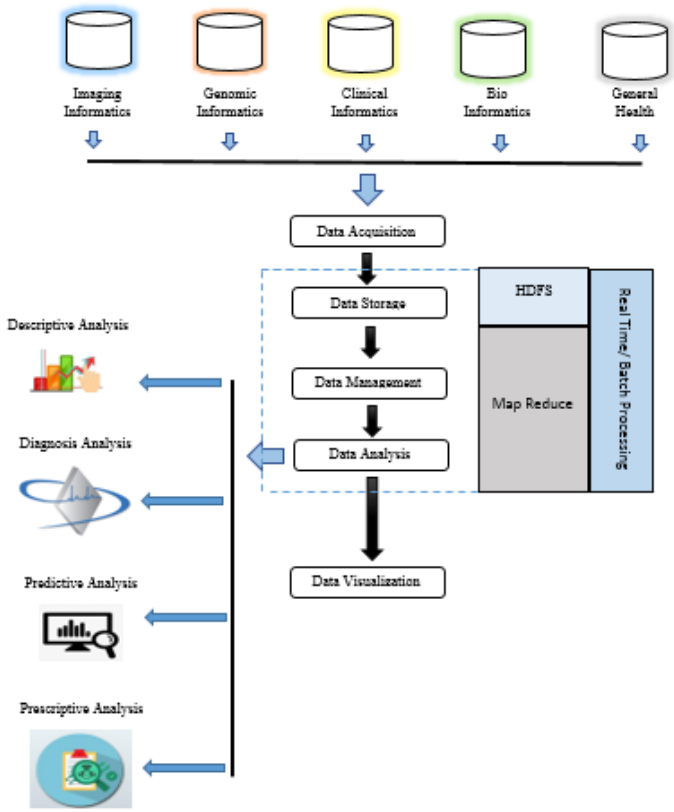

Figure 2: Conceptual Framework of healthcare big data

VI. HADOOP FOR HEALTHCARE BIG DATA HANDLING

Hadoop: Hadoop is an Apache open source framework written in Java which used to process the large data sets across cluster computing. This is basically designed to scale up from single sever to thousands of machines. Hadoop consists with two components including Hadoop Distributed File System(HDFS) for storing data and MapReduce Framework retrieval and processing data.

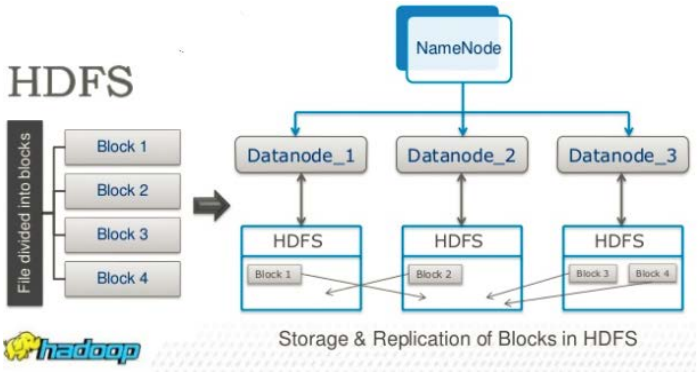

Figure 3: HDFS (Hadoop Distributed File System)

A. HDFS (Hadoop Distributed File System) – HDFS is a file system which designed for storing very large files with streaming data access patterns and running on clusters of commodity hardware. This has a master/slave file architecture, a single name node or a master server which manages the file system namespace and regulates the access to files by clients. Further, the storage on the cluster splits the files into 'blocks' and then stores each block redundantly across the pool of the server. HDFS architecture stores three copies of each file which is called name node or master node, data node and HDFS clients or edge nodes.

Name Node: This is the centrally placed node which indicates the information about Hadoop file system. The name node of HDFS act as a master node because it records the information about the systems such as metadata and attributes and specific location of files. Further, this provides information about newly added recorded, modified recorded and removed nodes.

Data Node: Data nodes are acting as slave nodes. Mainly, data nodes store and retrieve the blocks when requested by name nodes. Data nodes are periodically report back the lists of blocks that stored to the name node. Hadoop support for the feature of fault tolerance by using back up files (persistent state of the file system metadata) and running a secondary name node.

B. MapReduce Architecture- MapReduce is the processing of Hadoop framework which support for two separate

distinct tasks of map job and reduce job. MapReduce is widely adopted solution for handling batch processing data. The data is spitted in small pieces that are distributed among multiple nodes to retrieve intermediate results and once the data processing is over the outcomes will integrated for generate final results.

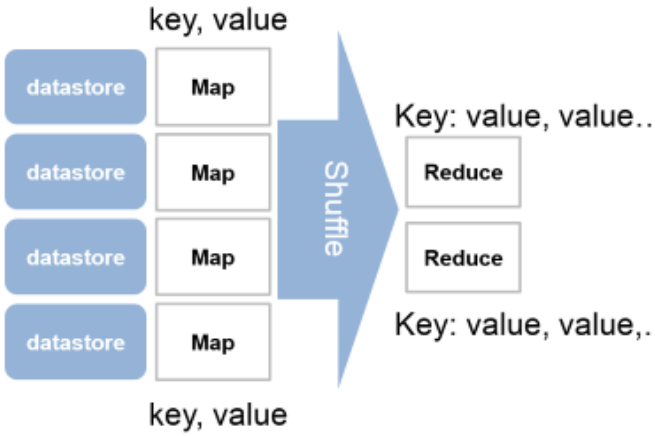

Figure 4: Map Reduce Logic

Map Job is the first stage of MapReduce where a block of data is read and processed to produce key-value pairs as intermediate outputs. In reduce job, the output from the map (key-value pairs) is input to the Reducer. However, the reducer receives the key-value pair from multiple maps. Finally, the reducer aggregates those intermediate key-value pairs into final output.

Logical View:
Map takes one pair of data with a type in one data domain and returns a list of pairs in a different domain.

$$\text{Map } (k1, v1) \rightarrow \text{List } (k2, v2)$$

The Map function applies to every pair (keyed by k1) in the output data in parallel and produces a list of pairs (keyed by k2) for each message. Then, MapReduce framework is able to collect all pairs with same key (k2) from all lists and create groups for each key value.

Then, the Reduce function is applied to each group, which produces a collection of values in the same domain simultaneously.

$$\text{Reduce } (k2, \text{list } (v2)) \rightarrow \text{List } (v3)$$

Each call of Reduce generates either one value v3 or an empty outcome. However, one call allows to return more than one value. But, all these returns are collected as desired result list.

Therefore, Hadoop MapReduce has capability of parallelism and get the output based on key-value pair analysis. Further, this is suitable for large volumes of growing data handling in simplified manner.

VII CONCLUSION

The healthcare big data is rapidly generating due to vast amount of imaging, genomics, clinical and other wearable devices. Therefore, existing big data handling systems unable to cater for current trends of healthcare systems. Significant amount of money has been allocating in the process of treatments, diagnosis and other research and development activities. In addition to that, patients are unnecessarily spending time, effort and money, due to lack of telemedicine support. Healthcare big data need to monitor the its process including creation to visualization. However, the rapid growth of big data, need to manage through the scalable technologies such as Hadoop. Therefore, the research has proposed Hadoop HDFS solution for store and process medical data to avoid the modern issues in healthcare big data analysis.

REFERENCES

1) Bilalli, B., Abello, A., Banet, T.D, Wrembel, R (2015), Towards Intelligent Data Analysis: The Metadata Challenge, Retrieved from http://www.essi.upc.edu/~aabello/publications/16.IoTBD.Besim.pdfon 1st May 2018

2) Curcin V. Embedding data provenance into the Learning Health System to facilitate reproducible research. Learn Heal Syst [Internet] 2017 Apr 1;1(2):e10019–n/a. Available viewed: http://dx.doi.org/10.1002/lrh2.10019

3) Dean J., Ghemawat S. MapReduce: Simplified Data Processing On Large Clusters. *Communications of the ACM*. 2008;51(1):107–113. doi: 10.1145/1327452.1327492.

4) Farhangmehr, F. Statistical Approaches for Big Data Analytics and Machine Learning: Data-Driven Network Reconstruction and Predictive Modelling of Time Series Biological Systems, scholarship, University of California, (2014).

5) Frost & Sullivan, *Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations*, Viewed: 10th Jan 2019, http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technology-complexities-ar.pdf

6) Galvic, B. (2015), Big Data Provenance: Challenges and Implications for Benchmarking, viewed from https://pdfs.semanticscholar.org/9ad4/3adc16c975deb5dfe2d5d1fec815906d3fc9.pdf on 1st May 2018

7) Herland, M., Khoshgoftaar, T. M., & Wald, R, A. review of data mining using big data in health informatics, Journal of Big Data, vol. 1, no. 1, p. 2, 2014

8) Kaitoua, A, Pinoli, P, et al Framework for supporting genomic operations. IEEE Transactions on Computers, 66(3):443–457, March 2017

9) Nambiar. R, Bhardwaj. R, Sethi, A, Vargheese, R, A look at challenges and opportunities of Big Data analytics in healthcare, Proc. 2013 IEEE Int. Conf. Big Data, (2013) 17–22. doi:10.1109/BigData.2013.6691753.

10) Ristevski, B., & Chen, M. (2018). Big Data Analytics in Medicine and Healthcare. Journal of Integrative Bioinformatics, doi: 10.1515/jib-2017-0030

11) Senthilkumar SA, Bharatendara K Rai, Amruta A Meshram, Angappa Gunasekaran, Chandrakumarmangalam S. Big Data in Healthcare Management: A Review of Literature. American Journal of Theoretical and Applied Business. Vol. 4, No. 2, 2018, pp. 57-69. doi: 10.11648/j.ajtab.20180402.14

12) Shahrivari S. Beyond batch processing: towards real-time and streaming big data. *The Computer Journal.* 2014;3(4):117–129. doi: 10.3390/computers3040117

13) Shen Xu, Kecheng Liu, Llewellyn C.M. Tang, Weizi Li. A Framework for Integrating Syntax, Semantics and Pragmatics for Computer-aided Professional Practice: With Application of Costing in Construction Industry. Comput Ind. 2016;83c:28–45.

14) 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.

15) zZhang, X., Hu, Y., Xie, K., Zhang, W., Su, L., & Liu, M. (2015). An evolutionary trend reversion model for Knowledge-Based Systems, doi: 10.1016/j.knosys.2014.08.010

AUTHORS

First Author – A. S. Thanuja Nishadi, MSc. In Information Systems Management (University of Colombo), BSc. (Hons).in BIT (University of Greenwich, UK), thanuja.nishadi@gmail.com