IJCS International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
SKRGC Since 2012

http://www.ijcsjournal.com Volume 5, Issue 1, No 25, 2017 ISSN: 2348-6600
Reference ID: IJCS-277 PAGE NO: 1794-1799

Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

A Brief Introduction to Process and Analyze Healthcare Big Data on Cloud Environment

M.Priya M.E.,
Teaching Assistant,
Alagappa Institute of Skill Development.
Alagappa University.
Karaikudi, India.
priyamathymail@gmail.com.

Dr. C.Balakrishnan,
Assistant Professor,
Alagappa Institute of Skill Development.
Alagappa University.
Karaikudi, India.
balasjc@gmail.com

Abstract- With rapid growth of development in healthcare applications,large amount of heterogeneous data are generated by healthcare organizations. It is very difficult to process the huge volume of data by using traditional data processing systems. Healthcare applications are being supplied through internet and cloud services rather than using a traditional software. Without applying proper data analytics methods, these data became useless. The data need to be processed and analyzed effectively for better decision making. The critical challenge that the healthcare organizations are facing is, how to analyze the large-scale data. In this paper we discuss on introduction to the big data characteristics, to process and analyze large scale healthcare data using Hadoop on cloud computing environment.

Keywords: Big data, Cloud, Hadoop, Big data tools, Healthcare.

1. INTRODUCTION

The healthcare organization has generated large amount of data generated from record keeping, compliance and patient related data.In today's digital world, it is mandatory that these data should be digitized.The collection of digitized medical records is increased over a period of time created large data sets. The large datasets of healthcare need to be effectively analyzed and it should be resolvedthe various problems faced by the organizations.Most of the software applications are being deployed in the data centres. As healthcare data is being available in large data sets and in various formats, cloud environment is the efficient way to store and process the data.Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Open source platforms such as Hadoop and MapReduce have encouraged the application of big data analysis in healthcare. Hadoop framework solves most of the problems related to huge data processing. Hadoop applications run on clusters which are built using commodity hardware. One of the important features of Hadoop is fault tolerant. MapReduce is the computing paradigm used in Hadoop as it provides Hadoop Distributed File System (HDFS) which stores data on the nodes. Healthcare data sets need to be analysed on the Hadoop clusters in cloud computing environment. The analyzedresults can bring the appropriate information for the patient where the patient can be informed about the various disease preventions. The patient can have timely and appropriate treatment. He can be informed about the precautions to be taken to prevent various diseases. The information on right provider for the patient can be chosen. The physician can perform a task without doctor.

2. BACKGROUND

2.1 BIG DATA

Big Data is not a new concept but verychallenging.The term “big data” applies to datasets that grow so large that they become difficult to work with using traditional database management systems. The size of big data has expanded beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time. Three main features characterize big data: volume, variety, and velocity, or the three V’s[2]. The volume of the data is its size, while velocity refers to the rate with which data is changing, or how often it is created. Variety regards the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data. Additionally IBM added 4th V which is veracity [3] and some researchers considered value of data as 5th V [1]. Therefore big data in today’s world is identified by 5 V’s, such as Volume, variety, velocity, veracity and value [9].

2.2 BIG DATA IN HEALTHCARE

As mentioned earlier healthcare is one of the organizations which are generating huge data sets. Healthcare is a prime example of how the three Vs of data, velocity (speed of generation of data), variety, and volume [4], are natural aspect of the data it produces.The HealthCare System is overwhelming not only because of massive volumes but also diversity of data types and the speed at which it managed. Big data in healthcare can come from electronic health records, clinical decision support systems, government sources, laboratories, pharmacies, insurance companies often in multiple formats (flat files,.csv, relational tables, ASCII/text, etc.) For the purpose of big data analysis, this data has to be processed or transformed. This data is spread among multiple healthcare systems, health insurers, researchers, government entities, and so forth.Historical approaches to medical research have generally focused on the investigation of disease states based on the changes in physiology in the form of a confined view of certain singular modality of data [12]. Although this approach to understanding diseases is essential, research at this level mutes the variation and interconnectedness that define the true underlying medical mechanisms [13]. After decades of technological improvement, the field of medicine has begun to adapt to today's digital data age. New technologies make it possible to capture vast amounts of information about each individual patient over a large timescale. However, despite the advent of medical electronics, the data captured and gathered from these patients has remained vastly underutilized and thus wasted.

2.3 STORING AND PROCESSING OF HEALTHCARE BIG DATA ON CLOUD

As discussed in section 2.2, large volumes of data are collected in healthcare from various sources. Before processing these huge amounts of data, it needs to be stored in massive storage space.Energy efficient and cost saving platforms are needed for large scale data management. The cloud is increasingly being used to store and process the big data. Cloud computing provides various advantages such as cost reduction, scalability, flexible in nature, energy efficient and agility [4]. Hadoop plays a vital place for storing and processing big data.

3. THE FRAMEWORK DEVELOPMENT

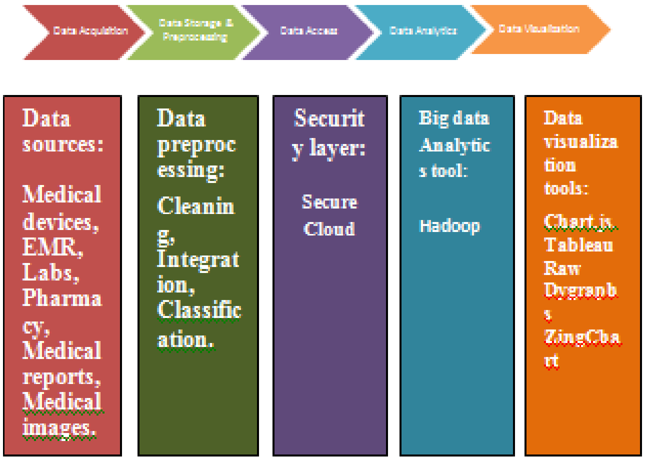
Data Acquisition
Data Storage & Preprocessing
Data Access
Data Analytics
Data Visualization
Data sources:
Medical devices, EMR, Labs, Pharmacy, Medical reports, Medical images.
Data preprocessing:
Cleaning, Integration, Classification.
Security layer:
Secure Cloud
Big data Analytics tool:
Hadoop
Data visualization tools:
Chart.js, Tableau Raw Dygraphs ZingChart

Fig 3.1 framework for processing and analyzing big data in cloud environment using Hadoop

Fig 3.1 shows framework for processing and analyzing big data in cloud environment using Hadoop. Framework defines all activities into five stages. Such as stage 1: Data acquisition, stage 2: Data storage and preprocessing, stage 3: Data access, stage 4: Data analytics and stage 5: Data visualization.

3.1 BIG DATA PROCESSING STAGES (BDPS)

Stage:1 Data Acquisition: It involves the collection of data from various sources and storing it in cloud. Data can be anything such as case history, medical images, social logs, sensor data etc.

Stage:2 Data storage and preprocessing: The data collected from various sources are stored in HDFS. Then data preprocessing activities are carried out. It involves the process of verifying whether there is any junk data or any data that has missed values. Such data needs to be removed.Medicalimagessufferfromdifferenttypesof noise/artifactsandmissingdata.Noisereduction, artifactremoval,missingdatahandling,contrastadjusting,andsoforthcouldenhancethequality ofimagesandincreasetheperformanceofprocessingmethods.Classification involves the filtering of data based on their structure. For example Medical Big data consists of mostly unstructured data such as hand written physician notes. Structured, semi-structured and unstructured data should be classified in order to perform meaningful analysis.

Stage:3 Data Access: The patient needs to carry all his past history records and then explain about his pastmedical data.Many of the physicians across the globe do not have proper information while dealing with patient. This sometimes led to inaccurate decision based on patient records. It requires a new infrastructure todevelop in the healthcare organizations to provide an integrated solution.

Stage:4Data Analytics: It involves performing analysis on the classified data. Large numbers of tools are available to process big data.Big data analysis tools are complex, programming intensive. They have emerged as open-source development tools and platforms. The most established software platform for big data analytics is Apache Hadoop. Hadoop is an open-source software framework used for storing huge volume of data and running applications on clusters of commodity hardware. Hadoop has ability to store and process big data in a distributed environment across clusters of computers using simple programming models. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.Hadoop's distributed computing model processes healthcare big data fast. Hadoop is fault tolerance. If a node fails, jobs are automatically redirected to other nodes in the distributed environment. Multiple copies of all data are stored automatically. Hadoop is flexible and low cost. Hadoop has scalability, so that we can add more number of data.Hadoop consists of basically two factors,

1) Map Reduce
2) HDFS (Hadoop distributed file system)

MapReduce, a popular computing paradigm for large-scale data processing in cloud computing. This framework consist of two function namely map () and reduce (), each having different parameters. Map function contains two parameters i.e. key and value. By default this framework assigns value 1 to all keys. Hadoop uses a specialized scheduling mechanism for allocating task to every node. Scheduling is an important aspect of Hadoop which ensures fair task allocation and load balancing. In heterogeneous clusters the performance of every node differs from all other nodes. To maximize the performance of such clusters and for better resource utilization, the task scheduling should be adaptive.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
SKRGC Since 2012

http://www.ijcsjournal.com Volume 5, Issue 1, No 25, 2017 ISSN: 2348-6600
Reference ID: IJCS-277 PAGE NO: 1794-1799

Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

3.2 MODEL FOR DATA PROCESSING IN CLOUD COMPUTING ENVIRONMENT BASED ON HADOOP CLUSTER

In this section, we discuss current techniques for analyzing big data with emphasis on emerging tool namely MapReduce. Most of the available tools concentrate on batch processing, stream processing and interactive analysis. Most batch processing tools are based on the Apache Hadoop infrastructure. Wepropose a framework for data processing of healthcare data in cloud environments which runs Hadoop clusters. Applications are running on servers with scalability that can fit large number of users.Healthcare data generated can be sent to cloud applications and various servers on cloud can integrate the data together using Hadoop MapReduce and quickly manage to get the results.

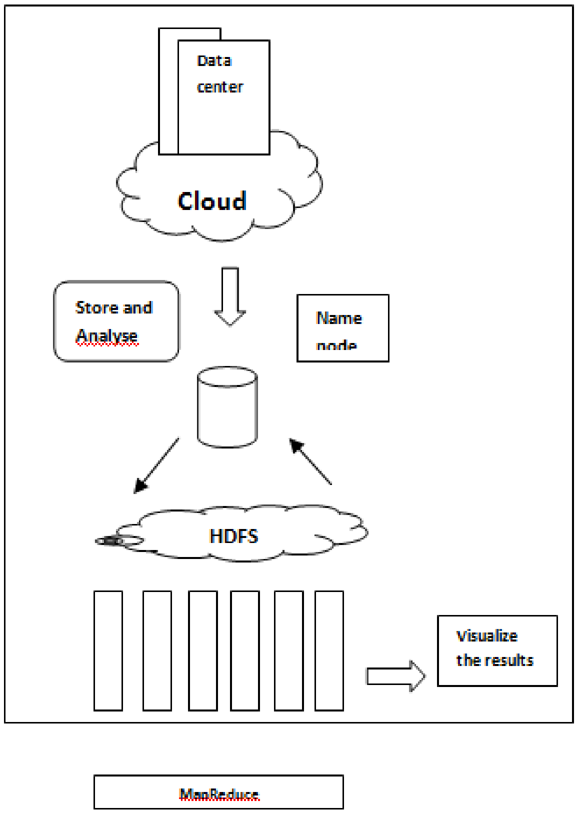

Fig 3.2 Model for data processing in cloud using Hadoop.

Stage:5 Data Visualization: It involves the generation of report based on the data modeling done. Data visualization is a modern branch of descriptive statistics. Various tools are available. Chart.js, Tableau, Raw, Dygraphs, ZingChart.

At the all the stages of BDPS it requires data storage, data integrity and data access control.

4. CONCLUSION

Cloud computing is a perfect environment for processing healthcare big data. Real-time big data analytics is a key requirement in healthcare. A big data analytics platform in healthcare must support the key functions necessary for processing the data. The lag between data collection and processing is addressed. We discussed various stages of data processing in cloud using Hadoop.Big data in healthcare provides a greater effect to improve the quality of healthcare, giving various options for patient for choosing the right care, right value.

REFERENCES

[1] Chang, R.M. Kauffman, R.J., Kwon, Y.(2014), "Understanding the paradigm shift to computational social science in the presence of big data", Decision Support Systems, Vol. 63, pp 67-80.

[2] Fan, S. Lau, R.Zhao. J.L (2015), "Demystifying big data Analytics for business Intelligence through the less of marketing Mix", Big data Research, Vol 2. pp 28-32.

[3] Jagadish, H.V (2015), " Big data and Science: Myths and reality" Big data research, Vol 2, pp 49-52.

[4] B. Hayes "Cloud computing", comun. ACM.Vol 51, no 7, pp, 9-11, 2008.

[5] Victor, N., Lopez, D., &Abawajy, J. H.. Privacy models for big data: a survey. International Journal of Big Data Intelligence, 2016 3(1), 61-75.

[6] Sagiroglu, Sinanc, "Big Data: A Review", 978-1-4673-6404-1/13 IEEE.

[7] JoonsangBaek, QuangHieu Vu, Joseph K. Liu, Xinyi Huang, Yang Xiang, "A secure cloud computing based framework for Big data information management of smart Grid", IEEE transactions on cloud computing, Vol 2, No 2, June 2015.

[8] Singh, S.; Singh, N., "Big Data analytics," Communication, Information & Computing Technology (ICCICT), 2012 International Conference on , vol., no., pp.1,4, 19-20 Oct. 2012

[9] O. Terzo, P. Ruiu, E.Bucci and F.xhafa, "Data as a service", (DaaS) for sharing and processing of large data collections in the cloud. Pp 475-480.

[10] W.B. Nelson, Accelerated testing, statistical models, test plans and data analysis, John Wiley & Sons, 2009.

[11] C.M Bishop, Pattern recognition and machine learning Springer New York 2006.

[12] J. J. Borckardt, M. R. Nash, M. D. Murphy, M. Moore, D. Shaw, and P. O'Neil, "Clinical practice as natural laboratoryfor psychotherapy research: a guide to case-based time-series analysis," The American Psychologist, vol. 63, no. 2, pp. 77–95,2008.

[13] L. A. Celi, R.G.Mark, D. J. Stone, and R.A.Montgomery, "'Big data' in the intensive care unit: closing the data loop," American Journal of Respiratory and Critical Care Medicine, vol. 187, no. 11, pp. 1157–1160, 2013.