Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 85 (2016) 16 – 22

Procedia Computer Science

International Conference on Computational Modeling and Security (CMS2016)

Impact of Processing and Analyzing Healthcare Big Data on Cloud Computing Environment by Implementing Hadoop Cluster

Sreekanth Rallapalli[a,*],Gondkar RR[b],Uma Pavan Kumar Ketavarapu[c]

[a]Research Scholar,R&D Center,Bharathiyar University,Coimbatore,India
[b]Professor, AIT, Bangalore, India[c]Associate Professor, AIMS, Bangalore, India

Abstract

The critical challenge that the healthcare organizations are facing is to analyze the large-scale data. With the rapid growth of various healthcare applications, various devices used in healthcare generate varieties of data. The data need to be processed and effectively analyzed for better decision making. Cloud computing is a promising technology which can provide on-demand services for storage, processing and analyzing the data. The traditional data processing systems no longer has an ability to process such huge data. In order to achieve a better performance and to solve the scalability issues we need a better distributed system on cloud environment. Hadoop is a framework which can process large scale data sets on distributed environment. Hadoop can be deployed on cloud environment to process the large scale healthcare data. Healthcare applications are being supplied through internet and cloud services rather than using as traditional software. Healthcare providers need to have real time information to provide quality healthcare.This paper discuss on the impacts of data processing and analyzing large scale healthcare data on cloud computing environment.

Keywords: Big Data;Cloud Computing;Cluster;Hadoop;Healthcare.

1. Introduction

Healthcare providers across the world are migrating to use the software applications as a service which is provided by most of the cloud providers. As Healthcare information is confidential and to provide better quality healthcare, various stakeholders should exchange the patient information, clinical information in a secured way. 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. Today most of the software applications are being deployed in the data centers[1]. In order to perform complex computations cloud computing is a dominant architecture which can efficiently perform large-scale data computation by providing the scalable resources[2]. Leading research organizations like Gartner Inc and market intelligence firms like IDC reports that Big data and Cloud computing are emerging technologies in today’s Business Intelligence market.

*Corresponding author. SreekanthRallapalli, Tel:00-91-9008716563
E-mail:rsreekanth1@yahoo.com

Peer-review under responsibility of the Organizing Committee of CMS 2016
doi:10.1016/j.procs.2016.05.171

As per IBM reports everyday 2.5 quintillion bytes of data is created and 90% of data is created in the past 2 years[3]. As per McKinsey analysis Big data is changing the paradigm of health care and new insights of the data is creating new value for the payers and providers[4]. So far health care data is available in form of Electronic Medical Records (EMR), Electronic Health Records (EHR), and Patient Medical Records (PMR). The collection of digitized medical records is increased over a period of time created large data sets. The large data sets of health care need to be effectively analyzed and try to resolve the various problems the industry is facing. Quality health care, reduction in medical cost, efficient decision making to provide the appropriate healthcare, finding patterns for unnecessary hospital readmissions are some of the issues can be resolved by Big data technology. New value from the large data sets can be analyzed by building effective analytical tools that assist patients, physicians and various stake holders in health care. Cloud computing has resolved most of the health care data related issues such as standardization of exchange of health care records, privacy, and network security. Security is the major issue while sharing the Health care data in the clouds. Access control mechanism which will support sharing of the EHR[5] will address the issue.

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 analyzed on the hadoop clusters in cloud computing environment to solve various issues of the industry.

2. Sharing of Healthcare Data on Clouds

Healthcare providers across the globe are now willing to move the data to the cloud for reducing the operational costs. Adoption of cloud computing is still not popular in many countries due to various challenges like security, authentication and access control mechanisms. Researchers have provided various solutions to overcome the challenges[5, 6].Many of the physicians across the globe do not have proper information while dealing with patient. The patient needs to carry all his past history records and then explain about his past medical data. This sometimes led to inaccurate decision based on patient records. Various organizations have developed integrated solutions to meet the above said challenges. But it requires a new infrastructure to develop in the healthcare organizations [7].

To overcome the issues of security it is proposed to have a segregated network for Electronic Health care Records[8]. Also it is recommended to check for vulnerabilities, providing data loss prevention programs, installation of firewalls, and to provide all national level policies on healthcare.

3. Big Data for Healthcare

Big data in today’s world is identified by 5 V’s, Volume, Velocity, Variety, Veracity and Value[9]. As discussed earlier healthcare is one of the organizations which are generating huge data sets. Analysis of large sets of patient records involve in identifying the patient clusters and correlations. This also involves in developing various models using statistical principles or machine learning techniques[10, 11]. Confidentiality of the healthcare data is one of the major problem the organizations are facing where the healthcare provider need to protect. As healthcare generates complex data base the applications developed in big data should handle such data complex in nature. Big data applications provide solutions for computational bio medicine, genomic data, and other medical data analysis. It is proposed that big data technologies have big potential in the field of health care[12]. The integration of big data analytics and various models in healthcare will demonstrate excellent results.

4. Secured Cloud Framework for Healthcare Big Data

In this section we study the security principles needed for Data management. We need a secured cloud based framework for our healthcare data. Energy efficient and cost saving platforms are needed for large scale data management. Cloud computing provides various advantages such as cost reduction, scalability, flexible in nature, energy efficient and agility[13]. A secured framework can be achieved by instantiating the framework based on identity based encryption, ID-based proxy Re-encryption schemes and Identity based signature encryptions[18]. Fig 1 shows that when the sender encrypts and send the data to cloud server, the data is encrypted based on Identity based encryption

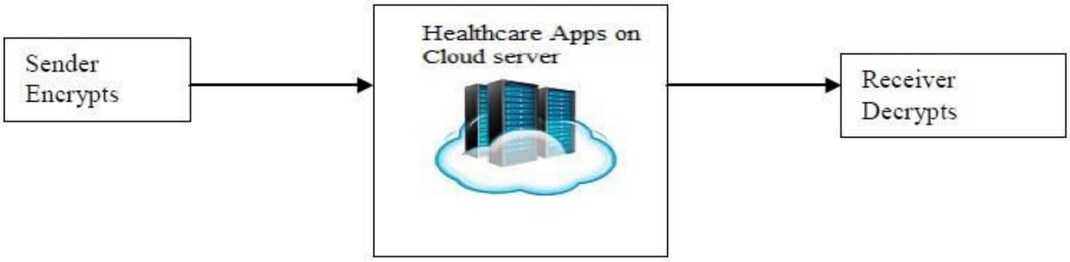

Fig 1: Secure cloud framework for big data

5. Hadoop Clusters

Hadoop clusters are built by the rack servers which are connected to top of rack switch. Uplinks of rack switch are connected to another set of switches which have the equal bandwidth. This forms a cluster in a network. We can setup this cluster on cloud so that the workflow of this cluster will get the required results from the large data sets. In this case we load the EHR data to the cluster and search for a query.

The workflow of the cluster is as follows: Hadoop Distributed File System (HDFS) writes the loaded data into the cluster. The data is analyzed with MapReduce algorithms. HDFS writes the data and saves into the cluster. HDFS reads the results from the cluster. From large data sets of EHR if we need to find how many patients were diagnosed with heart disease, this can be analyzed and processed very quickly using hadoop. Hadoop divides the huge chunks of data into smaller chunks and the process across multiple machines and thus produce the result so quickly. The typical architecture of hadoop cluster is shown in Fig 2.

For faster parallel processing huge data has to be loaded into hadoop cluster for processing. The client is going to break the data into smaller chunks and all the chunks of data are being sent to different machines for processing. To avoid the data loss we make sure that each chunk of data is   running parallel on different machines. In order to prevent data loss and network performance hadoop administrator manually defines the rack number of each slave data node in the cluster.

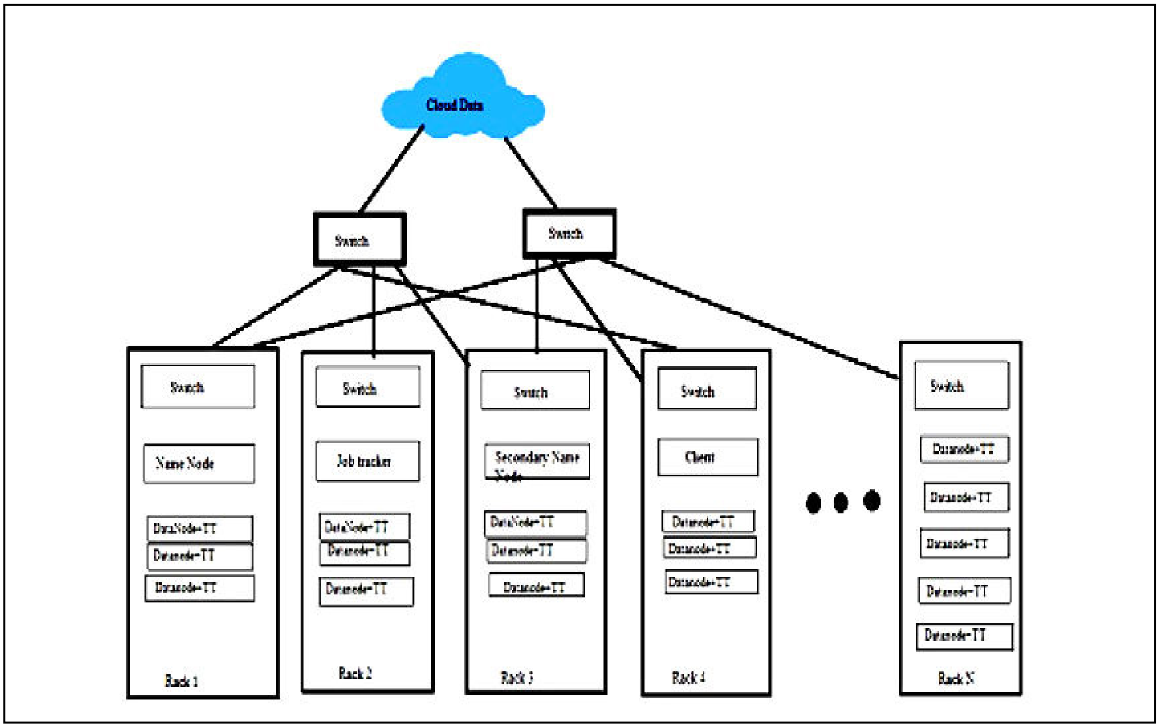

Fig 2: Architecture of Hadoop cluster

EHR data loaded into cluster is divided into data chunks like File A, File B, File C and so on. In this section we present how these data chunks will be loaded into HDFS. Client will consult the name node and then write the block of code to one data node. The data node then replicates the required blocks as decided by the Name node. The same repeats for the next set of blocks till all the chunks of data have been completely processed. Writing of files to HDFS is shown in Fig 3.

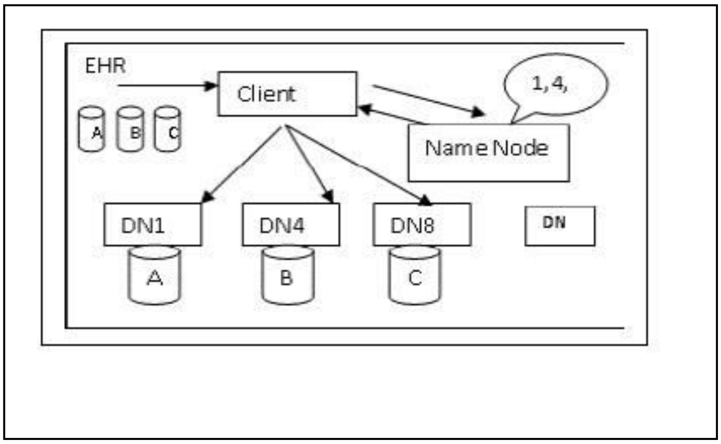

Fig 3: Writing Files to HDFS

6. MapReduce for Big Data Electronic Health Records Processing

An efficient tool for Big Data processing is the MapReduce. Its unique features are simplicity and tasks are reduced to simple level to deliver the output. MapReduce jobs are highly scalable and fault tolerant[14]. Experimentalresults show that MapReduce is slower than two parallel systems by certain degree of factor[15].

There is a need to improve the parameters which affect the performance of the MapReduce. The parameters required for fine tuning the performance of MapReduce is discussed[16]. Various parameters such as I/O interface, Index Structure, Record parsing, grouping algorithms and Block level scheduling can be fine tuned in order to make the MapReduce an efficient tool for Big Data processing.

MapReduce is a flexible data processing tool for large data processing[17]. Health care data being huge can be efficiently processed using the tool. MapReduce can be more flexible by writing efficient algorithms which can group the data and reduce the data to produce efficient results.

7. Model for Data Processing in cloud computing environment based Hadoop Cluster

This section proposes 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. The framework for data processing is shown in fig 4.

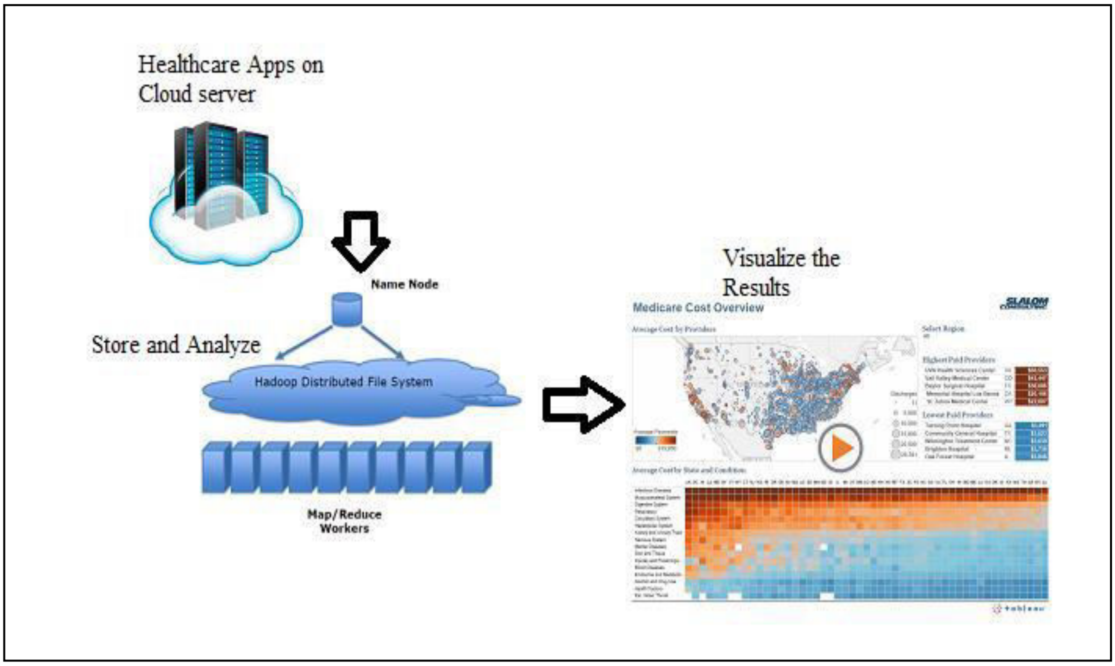

Fig 4: Framework for Healthcare data processing based on hadoop

Big data results 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. Big data analysis on healthcare can reduce the cost of the healthcare and minimize the fraud in healthcare sector. Innovation in healthcare can be improved by new Research and development discovery.

8. Conclusion

Cloud computing is a perfect environment for processing healthcare big data. It is secured for sharing sensitive data like patient health information. Various cryptographic techniques can be implemented for having a better framework for secure sharing of information on cloud. MapReduce can be efficient tool for processing the healthcare data provided the performance can be improved by fine tuning various parameters as discussed. Hadoop clusters can be used for faster parallel processing of the data. Big data in healthcare provides a greater impact on patient by improving the quality of healthcare, giving various options for patient for choosing the right care, right value. Big data can also impact in healthcare by innovations made in biomedicine.

Acknowledgements

I thank my guide who has encouraged me to write this paper. I also thank my institution for providing the infrastructure support.

References

[1] Jing Bi, Zhiliang Zhu, Ruixiong Tian, and Qingbo Wang. 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD), Miami, Florida, PP 370-377,2010.

[2] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li. 2012 IEEE International Symposium on Pervasive Systems, Algorithms and Networks, PP 17-23, 2012.

[3] Improving Decision Making in the world of Big Data. http://www.forbes.com/sites/christopherfrank/2012/03/25/improving-decision making-in-the-world-of-big-data/

[4] Peter Groves, Basel Kayyali,David Knot, Steve Van Kuiken, The Big Data revolution in healthcare, Accelerating value and innovation, McKinsey & Company, January 2013.

[5] Ruoyu Wu, Gail-joon Ahn, Hongxin Hu, Secure sharing of Electronic Health Records in Clouds, 8th International conference on Collaborative Computing, Networking, Applications and worksharing, PA, USA, October 14-17,2012.

[6] M. Li, S.Yu, K.Ren, W.Lou, Securing personal health records in cloud computing. Patient centric and fine grained data access control in multi-owner settings. Security and privacy in communication networks. Pages 89-106, 2010.

[7]IBM and Active Health Management, Active health and IBM pioneer cloud computing approach to help doctors to deliver hight quality cost effective patient care.

[8] Yunyong guo,Mu-hisang Kuo, Tony sahama, Cloud computing for Healthcare Research information sharing, 2012 IEEE 4th International conference oncloud computing technology and science, pp 889-894.

[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.MBishop, Pattern recognition and machine learning Springer New York 2006.

[12]Marco Viceconti, Peter Hunter,Rod Hose, Big Data, big knowledge: big data for personalised healthcare,IEEE journal of Biomedical and Health Informatics.

[13]B. Hayes “Cloud computing ” comun. ACM. Vol 51, no 7, pp, 9-11, 2008.

[14]J. Dean and S. Ghemawat,MapReduce: Simplfiied data processing on large clusters. In OSDI pages 137-150, 2004.

[15]A Pavlo, E. paulson, A.Rasin, D.J Abadi, J.Dewitt, S. Madden, M.Stonebraker, A comparision of approaches to large scale data-analysis. In SIGMOD , Pages 165-178 ACM 2009.

[16]Shwetha Pandey, Dr Vrinda Tokekar, Prominence of MapReduce in Big Data processing, 2014, Fourth International conference on Communication Systems and Network Technologies.

[17]Jeffry Dean and S.Ghemawat, An article on MapReduce, A Flexible Data processing tool. In SIGMOD pages 3(1), 72-75, ACM 2010.

[18]Joonsang Baek, Quang Hieu 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.