Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

MapReduce Program to Efficiently Analyse Big Data Electronic Health Records Database using Hadoop Cluster on Amazon Elastic Compute Cloud

Sreekanth R*
Research Scholor, R&D Centre,
Bharathiyar University, Coimbatore, India

Dr. Gondkar RR
Professor, Dept of MCA
AIT, Bangalore, India

Abstract— Data generated by organizations including Health care providers is exceeding from Gigabytes to Petabytes and continue to grow over the period of time. Analyzing such huge data for the organizations is a challenging task. Cloud computing is a promising technology where data can be moved and processed at low cost. Cloud computing provides on-demand and scalable resources for the organizations. Most of the health care organizations today use cloud computing for their data processing needs. Apache hadoop is open source software used to process huge data sets in the distributed computing environment using clusters and commodity hardware. MapReduce is a programming model for processing such huge data sets. Amazon Elastic Compute Cloud (EC2) provides virtual computing environments on a secured network. In this paper we first study how the large data sets can be moved efficiently to cloud for processing. We also study how Amazon EC2 provides an environment to handle the huge data sets. Further we propose a MapReduce Program to efficiently analyse Electronic Health Records (EHR) database.

Keywords— Cloud Computing, Data Sets, Hadoop, MapReduce, EHR.

I. INTRODUCTION

The challenging problem in Big data is to analyse the datasets arriving from multiple sources. Variety of data sources is yet another challenge. MapReduce is the framework used to handle such huge data. For batch processing of large amount of data this is an efficient tool. MapReduce is in useful in number of areas where huge data analysis is required [1]. Many applications which handle the huge data sets have been developed. Apache hadoop [2] is yet open source software which is used to process huge data sets in distributed computing environment. Cloud services are being used as a best option for big data applications. Everyday lot of data is being created (2.5 Exabytes) and the data centers which are processing these big data use cloud servers. Hadoop Distributed File System (HDFS) is built in a unique way such that the files are divided into chunks of data (64 MB/128MB/256MB) and then distributed in data nodes in the hadoop cluster. As we are moving huge data onto the cloud we need fast bandwidth to transfer the data [3].

Unstructured data generating from the devices which monitor patient health care is increasing [4]. These data specific to patients has to be analysed in the cloud computing environment which will help to improve the quality of health care by the health care providers. There are few challenges involved for efficient processing of the huge data sets related to medical data processing as it contains lot of unstructured data, medical image data and other types. The computational power of cloud infrastructure should be efficient to process and extract the information in a secured way [5]. The rest of the paper is organized as follows: Section II focus on Electronic Health Record systems, Section III focus on Big Data and Healthcare, Section IV focus on MapReduce, Section V focus on our proposed MapReduce algorithms to efficiently handle the Big data. Section VI discuss the results on Amazon EC2. Section VII concludes the paper.

II. ELECTRONIC HEALTH RECORD SYSTEMS (EHR)

Electronic Health Record systems (EHR) store the entire patient's medical history information from the time of admission, medical tests performed on the patient, drug prescriptions, readmission information and any other relevant information of the patient. These data has to be accessible and managed by all health care providers [6]. With the increase in number of admissions each year the health care data is voluminously increasing. Health care providers now choose the cloud services for moving the data and processing the information. As patient data is sensitive, confidential secure networks to be used for transmitting and accessing the patient information. As health care data is generated in variety of devices, with high velocity and huge volume the big data solutions are required to solve the problems of storage and processing. There are many big data technologies available to solve these issues. But as health care data need to be handled in a different way we many need to customize according to the specific purpose. Big data analysis in health care data can reduce the costs [7] and improve the quality of health care by providing a personalized health care. There are various standard data sources in healthcare such as Health Level 7 (HL7), Health Insurance Portability and Accountability Act (HIPAA), Digital Imaging and Communication in Medicine (DICOM), National Health Information Network (NHIN) which governs the standards to be adopted while distributing the health care data over the networks. There are various stakeholders in generating the Electronic health records of the patients. Fig 1 shows the list of healthcare partners [8].

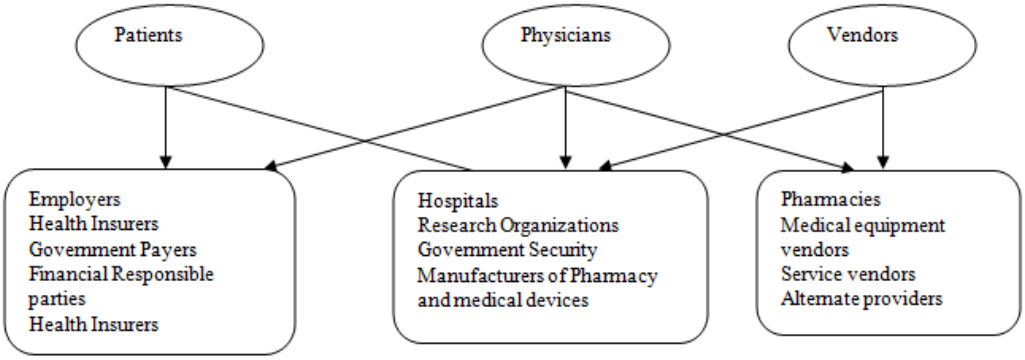

Fig 1: Healthcare Stakeholders

III. BIG DATA AND HEALTH CARE

Big data is a buzz word today in every organization. The global data volume is estimated to be around 40,000 Exabytes the data being doubled once in two years as per IDC report [9]. Demand for Big data analysis is increasing at a high pace. Applications developed in business [10], sciences, health care shows the insight value that big data brings into organizations. Big data encompass five V's: (1) Volume (2) Variety (3) Velocity (4) Veracity (5) Value. The first three V's represent data engineering which consist of collection, storage and moving the data. The fourth V represents the authenticity of the data while the last V focuses on the statistical methods, knowledge extraction and decision making. Healthcare is generating huge data sets as discussed in Section II. The challenge is how to create better applications to make more value out the data sets. The applications should provide the users to compare the quality of health care services available in his/her area. The doctor needs to make a patient referral to find the specialists efficiently and communicate with them easily and securely. The patient information should be readily available to the nurse so that she can help the patient instantly in case of chronological illness. The patients should be able to find the clinical trials relevant to him very easily [11].

There are various analysis methods for the big data. The most appropriate methods with respect to health care are (1) Recommendation system (2) Deep Learning and (3) Network analysis [12].

IV. MAPREDUCE

MapReduce [13,14] is a programming model for processing large data sets. The map function is specified by the users. This function is used to process a (key, value) pair which generates a set of intermediate (key, value) pair. The reduce function merges all associated intermediate values with the key. Programs are automatically parallelized even if those are written in a functional programming and then executed on a large cluster of commodity machines. Partitioning of data is taken care by the run time system. Handling the failures and execution of program across large set of machines, inter-machine communication is taken care by the run time system alone. Utilizing the resources of the distributed system requires the programmers to have extensive experience on parallel and distributed systems. But MapReduce model allows programmers to utilize the resources of distributed system effectively.

The MapReduce execution [14] overview is shown in Fig 2. For huge data processing the main advantages of the MapReduce framework are based on its simplicity of usage, flexibility, fault tolerance and high scalability [13,14,15]. In this paper we propose to have a huge database consists of EHR and use MapReduce framework to efficiently process on cloud environment.

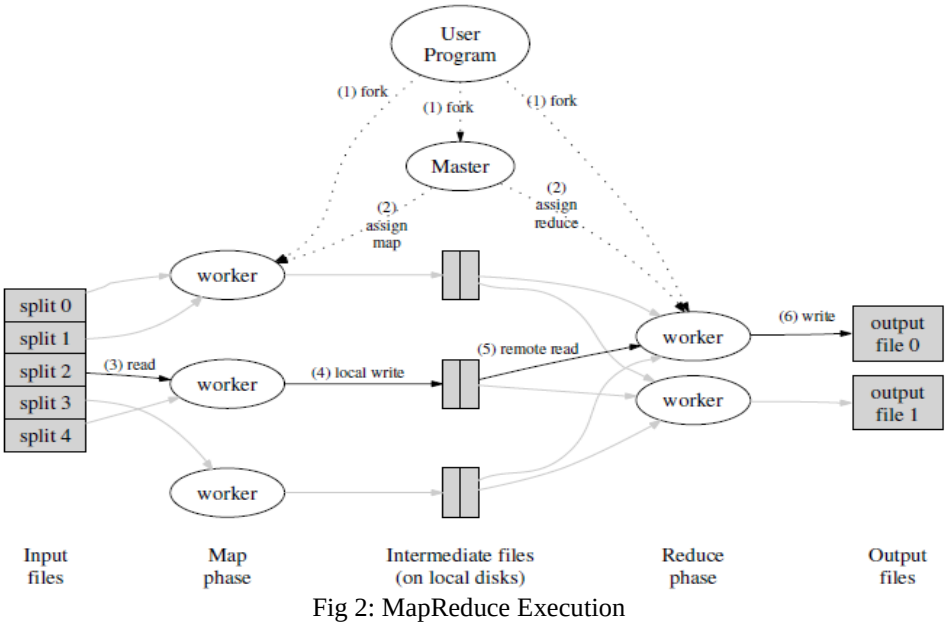

Fig 2: MapReduce Execution

The common notation used in the MapReduce program is the mapper and reducer with the following representation

Map: $(k1, v1) \rightarrow [(k2, v2)]$

Reduce: $(k2, [v2]) \rightarrow [(k3, v3)]$

In the programming for the MapReduce we use two additional elements called partitioners and combiners. Partitioners will specify the task to which an intermediate key-value pair must be copied. Combiners are an optimization in MapReduce which allow for local aggregation that occur before the shuffle and sort phase.

V. PROPOSED MAPREDUCE ALGORITHM

In this section we propose an efficient MapReduce algorithm which can efficiently handle and process Big data Electronic Health records database. This algorithm can be tested on Cloud environment like Amazon EC2 using Hadoop Clusters. We consider here the networks for moving the huge data sets onto cloud with high bandwidth. The main goal of this algorithm will be fast processing of EHR data using Hadoop parallel operation.

Let us consider the following notation for the EHR Big data. For any Big data size ∞, and a set of parameters for the EHR data

$$\theta(0), \theta(1), \theta(2), ...$$

We propose an iterative algorithm that finds the successive series of data for the patient's record to find similarities in these records. The expectation maximization is an iterative algorithm that guarantees

$$\prod_{k=1}^{x} \sum_{y \in Y} \Pr(X = xj, Y = y; \theta(i+1) \geq \prod_{k=1}^{x} \sum_{y \in Y} \Pr(X = xj, Y = y; \theta(i)$$

The algorithm updates with two steps the Expectation Step and Maximization step. The process repeats with new set of parameters and terminates when the parameter remain unchanged. We then write the mapper and combiner class for the proposed MapReduce algorithm. The Input for the algorithm is Data block of the patient Records. The Map function and Reduce function are together combined to produce the results which are written to file.

1: Class MAPPER
2: method INITIALIZE (integer iteration)
3: <S,O> ←READMODEL
4: $\theta, <-< A, B, \pi >< -Readmodelparams(iteration)$
5: method MAP(record id, recordseq x, recordatt y)
6: $\propto \leftarrow FORWARD\ (X, \theta)$
7: $\beta \leftarrow BACKWARD(X, \theta)$
8: $I \leftarrow new\ ASSOCIATIVE\ ARRAY$
9: for each data block D_i
10: $T_i \leftarrow \emptyset$
11: get occurance count $count(\emptyset, Di)$
12: $freq(\emptyset, Di) \leftarrow count(\emptyset, Di)/mod(Di)$
13: if $(Di) > (Dj)$ process the data block
14: Map recordatt ;

The Reduce function for the datablock is as follows

1: Class Reducer
2: set of n-records $Mi(1 \leq i \leq kblock)$
3: $R_{es} \leftarrow \emptyset$
4: for $i \leftarrow 1\ to\ kblock\ do$
5: combine M_i
6: for each n-records<id,x,y>
7:$\emptyset \frac{R}{id,x,y}$
8: return $reducer(\emptyset)$;

VI. RESULTS

Amazon EC2 provides scalable computing resources so that we can deploy various applications for faster processing. As per our needs we can deploy the clusters and then run instances of the applications. In our paper we setup Amazon Webservice Cloud using Elastic Cloud Computing. The Operating system used is 64 bit Ubuntu Server. The other hardware configurations are Intel® Xeon ® CPU E5-2650 @ 4.00 GHZ Processor with 8GB RAM, 1 TB Harddisk. Hadoop clusters are deployed on the servers. We have used the opensource Electronic Health Record Data Sets with 500000 patient records and 50000 observation reports. The database has been loaded into servers on cloud which has Hadoop environment set with the proposed MapReduce algorithm. The results are as follows

No of Nodes in Cluster	Execution time to finish the EHR database reading	Time to Reduce the data and output to file
1	95 Sec	45 Sec
3	73 Sec	35 Sec
5	61 Sec	30 Sec
7	52 Sec	25 ec

VII. CONCLUSION

Today healthcare providers, payers, physicians generate huge data which requires analysis to provide better health care to patients. Patients can have personalized health care and in order to reduce the cost analytics play a major role. In this paper we have seen the importance of Big Data analytics in health care. MapReduce model is an efficient programming paradigm for processing such huge data. Efficient in MapReduce model can be increased by improving the efficiency of algorithm. In this paper we propose an algorithm where the EHR database can be handled efficiently with minimum time taken for reading the database. The time taken to reduce the records and write the results back to the file is also minimized. The MapReduce algorithm run on Hadoop clusters under Amazon EC2.

REFERENCES

[1] Han Hu, Yonggang Wen, Tat-Seng Chua, Xuelong Li "Towards Scalable systems for Big data analytics", IEEE.
[2] Tom White, Hadoop: The Definitive Guide,O'Reily Press.
[3] http:// asperasoft.com/technology/transport/fasp/
[4] Charu C Agarwal, "Managing and Mining Uncertain Data",ed. Springer 2009
[5] Sun Fuquan, Zhang Dawei, Cheng Xu, Liu Chao, Construction of enterprise private cloud storage platform based on Hadoop, Journal of Liaoning Technical University, 2011(3), 112-113.
[6] Haluk Demirkan, A Smart Health care system Framework, IEEE, Sept/Oct 2013.
[7] Uma Srinivasan, Bavani Arunasalam, "Leveraging Big data analytics to reduce health care costs", IEEE, Nov/Dec 2013.
[8] W Liu, E.K Park, Big Data as an e-health service,in 2014 International conference on computing, networking and communication, IEEE 2014.
[9] J Gantz and D Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east", in Proc , IDC, iview IDC, Anal,Future, 2012.
[10] Ryen W White, et al Report on the SIGIR 2013, Workshop on health search discovery, ACM SIGIR, Forum 47 (2) (2013).
[11] AJ Burns, M. Eric Johnson, "Securing Health Information:", IEEE computer society, Jan/Feb 2015.
[12] Tao Huang, Liang Lan, Xuexian Feng, Peng An, Junxia Min, Fudi wang, "Promises and chanllenges of Big data in Health Sciences", Elsevier Big Data Research Vol 2, 2015, pp 2-11.
[13] J.Dean, S.Ghemawat, "Mapreduce:Simplified data processing on large clusters", Communication ACM, Vol 51, No 1, PP 107-113, Jan 2008.
[14] J.Dean, S.Ghemawat, "Mapreduce:Simplified data processing on large clusters", in OSDI'04: Proceedings of the 6th Conference in Symposium on Operating System Design and implementation, USENIX Association, Berkeley, CA, USA, 2004.
[15] J Lin, C. Dyer, "Data-intensive Text processing with MapReduce", Morgan and Claypool, 2010

