International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7488-7492

Apache Spark and Hadoop Based Big Data Processing System for Clinical Research

Sreekanth Rallapalli[1,*], Gondkar R R[2]

[1]Research Scholar, R&D Centre, Bharathiyar University, Coimbatore, Tamilnadu, India.
[2]Professor & Head, CMR University, Bangalore, India.

Abstract

Usage of big data which is related to medical filed is gaining popularity among healthcare services and for clinical research. Medical field is one of the largest areas which is generating enormous amount and varieties of data. Traditional systems are incapable of handling such big data which is characterized by volume, variety, velocity, veracity and values (5 V's). To process this vast amount of data we need a framework which can parallel process the data by utilizing the clusters of commodity hardware. This hardware should be reliable, fault-tolerant. Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. In the Hadoop framework we can develop MapReduce applications which can scale up from single node to thousands of machines. This paper investigates the big data which is used in clinical research to find out the patients with similar patterns and recommend the patients who requires intensive care. Also, the patients can be informed about the future predictions. In this paper we propose a ten-node hadoop cluster to run the distributed mapreduce algorithms. This algorithm shows an efficient data processing with big clinical data. These results can be used to provide efficient and personalized decisions for the patients. The data sets used for the results purpose is taken from MIMIC-III an open source database which is one of the largest repositories of data.

Keywords: Apache Spark, Big data, Hadoop, MapReduce, Machine learning.

INTRODUCTION

Medical data sets are increasing, and the variety of data being generated through hospital systems is becoming complex for processing as lot of data is unstructured. For systems which are currently being used in the hospitals are standalone systems with client server technology. These systems process the data which are structured by using traditional database management system. But when the data sets grow with variety of data being gathered, these systems are not capable to process the data [1,2]. There are lot of developments happened over the years to implement an open source and by using the commodity software applications are built to handle large data sets [3,4].

Apache Hadoop related projects provides framework scalable and fault tolerant analysis for large data sets which utilizes the MapReduce [5] programming model. For clinical research projects we can use hadoop. Hadoop is also used in bioinformatics. Researchers provide details on how hadoop can be used in medical related fields [6,7]. Another important tool that can be used is spark from apache foundation [8]. It is fast and general engine for large scale data processing. Spark and Hadoop are extensively used in medical field and especially have strong impact for cancer research. Hadoop can perform well for analyzing large scale data sets [9]. Researchers and various healthcare organizations now a day are using hadoop for various healthcare services and clinical research projects. Hadoop can also be used in the field of bioinformatics [10]. By using hadoop map reduce a package name called cloudburst has been proposed by Schatz MC.

In the era of Big Data, we have quite good opportunity to utilize the data available in medical field for various decision-making purpose. We can also use the cloud computing technologies for storage of the huge data. Most of healthcare organizations around the world use the Health information systems, such as Electronic Medical Records (EMRs), computerized order entry systems and various administrative systems in daily operations. These systems generate large amount of data. This data can be used as secondary research data in healthcare services. In this paper, we investigate the usage of hadoop and spark based applications to process the large amount of medical data. We discover features of users who are using the Health information systems. Based on structured, semi-structured and unstructured data which is produced by Health information system, an eight-node hadoop cluster is constructed to execute the distributed MapReduce algorithms. Lot of research has been done by using the single node algorithms. But when it is constructed using multi node algorithm, the process of big data for healthcare services is more efficient. Health information systems can be intelligently designed for making personalized recommendations for the patient.

RESEARCH METHODOLOGY

In this section we propose a hadoop and spark based Big data processing system for the healthcare data. The system has four components which will cover applications related to healthcare. The components are spark based Extract-Transform and Load (ETL), multi-node hadoop clusters for the storage of data, processing and managing the data, machine learning based recommendation engine for distributed data and performance coordination service for distributed application using zookeeper. In this rest of the paper we describe each component in detail. The overall system architecture of apache spark and hadoop based data processing system is shown in figure 1.

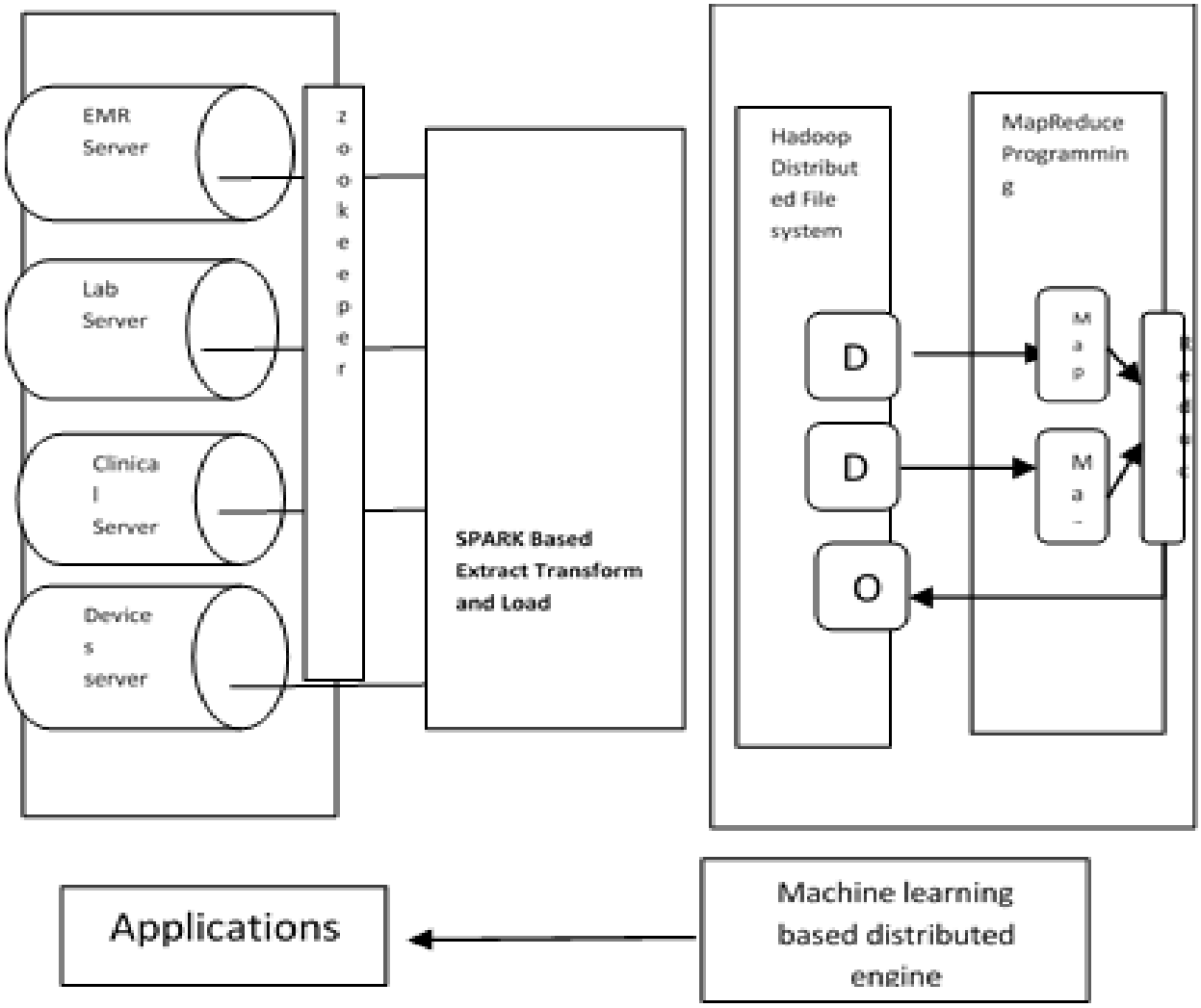

Figure 1: System architecture of Spark based Hadoop system for healthcare data

SPARK BASED EXTRACT, TRANSFORM AND LOAD

In this section we discuss on how Spark based Extract, Transform and Load of the Electronic health records data and various data received from other servers. Apache Spark is built on the concept of distributed datasets, which contain arbitrary Java or Python objects. You create a dataset from external data, and then apply parallel operations to it. The building block of the Spark Application Program Interface (API) is its Resilient Distributed Datasets (RDD). It is based on RDD API. In the RDD API, there are two types of operations: transformations, which define a new dataset based on previous ones, and actions, which kick off a job to execute on a cluster. On top of Spark's RDD API, high level APIs are provided, e.g. Data Frame API and Machine Learning API. These high-level APIs provide a concise way to conduct certain data operations.

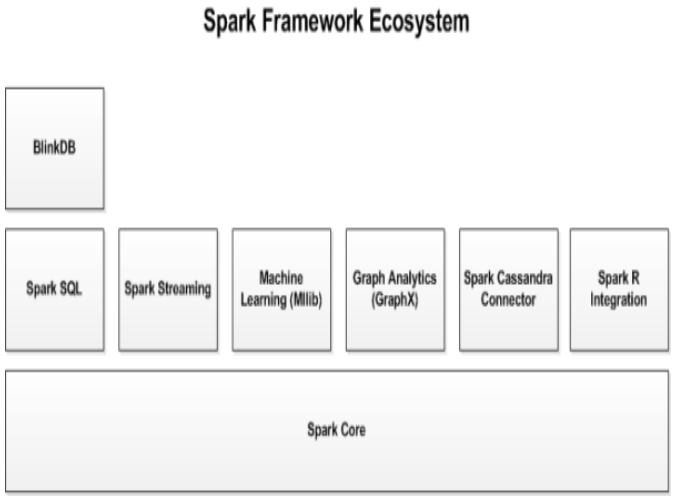

Figure 2: Spark Ecosystem

The spark ecosystem [11] is shown in fig 2. To manage the big data processing requirements with variety of data sets, spark provides a framework and the source of data. Spark will

enable the applications in Hadoop clusters[12] and run up to 100 times faster in memory and 10 times faster when running on a disk. To process the real time streaming data spark streaming library is used. Spark SQL provides the capability to expose the datasets over JDBC API. Common learning algorithms and utilities are included in MLib a machine learning library. For graph and graph-parallel computations GraphX is used. To run interactive SQL queries on large volumes of data BlinkDB a query engine is used. The integration adapters with other products like Cassandra (Spark Cassandra Connector) and R (SparkR). With Cassandra Connector, you can use Spark to access data stored in a Cassandra database and perform data analytics on that data.

The healthcare data are generally stored in relational databases. Relational database which use conventional IT architecture cannot process huge amount of data. To store and process big data we use Hadoop platform. It uses Hadoop Distributed File System (HDFS) for data storage and MapReduce a software framework for developing powerful applications to process big data. Methods proposed by [13] for transferring the data to the system are not efficient. In the proposed system we develop a Spark based ETL module. It is designed to offer efficient bulk data transfer between Hadoop and structured data sets available in Relational data base management system. The structure of spark based ETL is shown in fig 3.

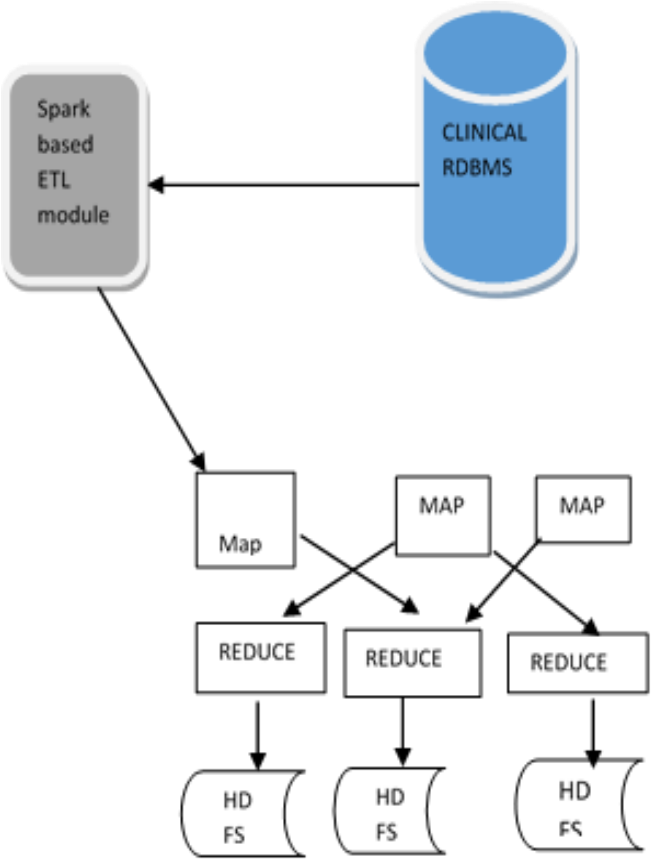

Figure 3: Spark based ETL

MULTI-NODE HADOOP CLUSTERS FOR THE STORAGE OF DATA

In this section we propose the multimode hadoop clusters for storage of large data. Hadoop is an architecture which can store big data. So in the architecture we use multi-node hadoop clusters to storage of the data which can handle any type of data. Hadoop cannot easily access the data stored in relational database. So we need to build an multi-node hadoop clusters for storage of the data. Hadoop Distributed File system (HDFS) based warehouse along with Hive need to be building to query the large datasets. This architecture is shown in fig 4.

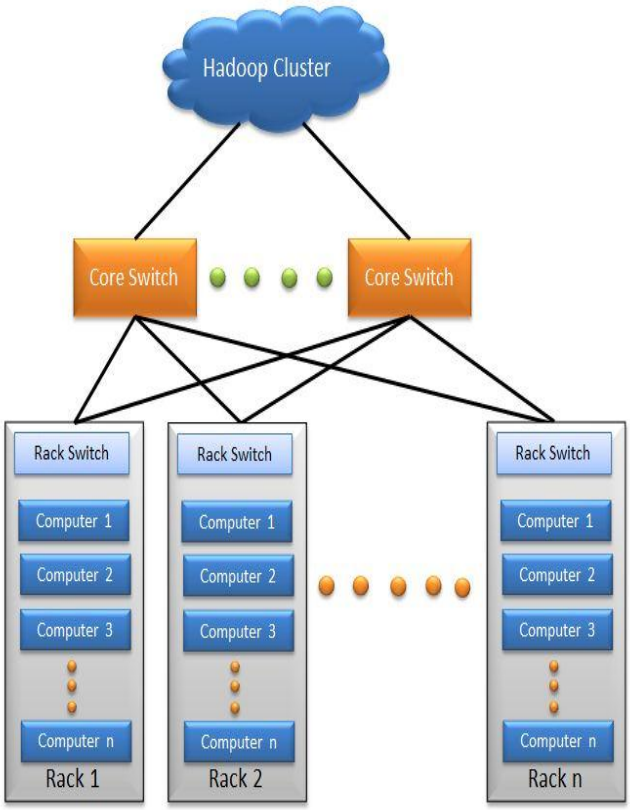

Figure 4: Multi node Hadoop cluster

PROCESSING, ANALYZING AND MANAGING THE DATA BY USING MACHINE LEARNING BASED RECOMMENDATION ENGINE

Big data analysis of clinical data will be carried out by related algorithms on the top of data warehouse. Machine learning library can be provided by Apache Mahout [14]. We can use the various algorithms like collaborative filtering, which includes the classification and clustering. To implement the collaborative filtering algorithms we collect user preferences, find the similar users or items and then calculate recommendations. Clinical data repositories store clinical data. By implementing the multimode hadoop clusters the large clinical data is stored, processed and analyzed. The following code show few settings with the hadoop multimode clusters.

```
<? xml version="1.0" encoding="UTF-8"?>
<? xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://master:9000</value>
</property>
</configuration>
```

Hdfs-site.xml on master machine

```
<? xml version="1.0" encoding="UTF-8"?>
<? xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
<property>
<name>dfs.permissions</name>
<value>false</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>/home/hadoop-2.7.3/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/hadoop-2.7.3/datanode</value>
</property>
</configuration>
```

Hdfs-site.xml on slave machine

The algorithm is assessed by precision and recall metrics for information retrieval and statistical classification to evaluate quality of results.

The measures are defined as

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

Where TP =True Positive, FP = False Positive, FN = False Negative

RESULTS

The datasets for the clinical research are collected from MIMIC – III, the largest repository of medical data sets. We collected the data for 600000 patient's admission data in various hospitals for the last 5 years. Multimode hadoop clusters are used to store the data and it exhibits the highest performance when compared with the single node algorithms.

CONCLUSION

This paper describes the hadoop and spark based design and development of a processing system that can analyze the data for large clinical data sets. This system solves the issues related to collection, storage and analysis of secondary data. Machine learning tools such as Mahout can be used to collect the intelligent data behind the data warehouse. The system designed performs better storage and analysis when compared to single node hadoop cluster.

REFERENCES

[1] Lin, C., Lin, I.-C., and Roan, J., Barriers to physicians' adoption of healthcare information technology: an empirical study on multiple hospitals. J. Med. Syst. 36(3):1965–1977, 2012.

[2] Poon, E. G., Jha, A. K., Christino, M., Honour, M. M., Fernandopulle, R., Middleton, B., Newhouse, J., Leape, L., Bates, D. W., and Blumenthal, D., Assessing the level of healthcare information technology adoption in the United States: a snapshot. BMC Med. Inform. Decis. Mak. 6(1):1, 2006.

[3] Miller, R. H., and Sim, I., Physicians' use of electronic medical records: barriers and solutions. Health Aff. 23(2):116–126, 2004.

[4] Blumenthal, D., Stimulating the adoption of health information technology. N. Engl. J. Med. 360(15):1477–1479, 2009.

[5] Dean, J., and Ghemawat, S., Mapreduce: simplified data processing on large clusters. Commun. ACM 51(1):107–113, 2008. doi:10.1145/1327452.1327492.

[6] Dean, J., and Ghemawat, S., MapReduce: a flexible data processing tool. Commun. ACM 53(1):72–77, 2010. doi:10.1145/1629175/1629198.

[7] Horiguchi, H., Yasunaga, H., Hashimoto, H., and Ohe, K., A userfriendly tool to transform large scale administrative data into wide table format using a mapreduce program with a pig latin based script. BMC Med. Inform. D ecis. Mak. 12:8, 2012. doi:10.1186/1472-6947-12-151.

[8] Liu, B., Madduri, R. K., Sotomayor, B., Chard, K., Lacinski, L., Dave, U. J., Li, J. Q., Liu, C. C., and Foster, I. T., Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. J. Biomed. Inform. 49:119–133, 2014. doi:10. 1016/j.jbi.2014.01.005.

[9] Santana-Quintero, L., Dingerdissen, H., Thierry-Mieg, J.,Mazumder, R., and Simonyan, V., HIVE-Hexagon: high-performance, parallelized sequence alignment for next-generation sequencing data analysis. PLoS One 9(6):11, 2014. doi:10.1371/journal. pone.0099033.

[10] Taylor, R. C., An overview of the Hadoop/MapReduce/HBase framework and its current

applications in bioinformatics. BMCBioinforma. 11:6, 2010. doi:10.1186/1471-2105-11-s12-s1.

[11] Schatz, M. C., CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25(11):1363–1369, 2009. doi:10.1093/ bioinformatics/btp236.

[12] Jun, L., and Peng, Z., Mining explainable user interests from scalable user behavior data. First Int. Conf. Inf. Technol. Quant. Manag. 17: 789–796, 2013. doi:10.1016/j.procs.2013.05.101.

[13] Wang, Z. H., Tu, L., Guo, Z., Yang, L. T., and Huang, B. X., Analysis of user behaviors by mining large network data sets. Futur. Gener. Comput. Syst. 37:429–437, 2014. doi:10.1016/j.future.2014.02.015.

[14] http://mahout.apache.org/