An Analytics in Health Record for Medical Analysis using MapReduce

S. Balu[1],
Assistant Professor CSE,
K.S.Rangasamy College of Technology,
Tiruchengode-637215.

M. Tharunkumar[2],
IV B.E CSE,
K.S.Rangasamy College of Technology,
Tiruchengode-637215.

***Abstract*— Like Oxygen, the world is surrounded by data today. The quantity of data that we harvest and eat up is thriving aggressively in the digitized world. Increasing use of new innovations and social media generate vast amount of data that can earn splendid information if properly analyzed. This large dataset generally known as big data, do not fit in traditional databases because of its' rich size. Hospitals need to manage and analyze the health record in big data for better decision making and outcomes. So, big data analytics is receiving a great deal of attention today. In healthcare, big data analytics has the possibility of advanced patient care and clinical decision support.This paper addresses the problem of data quality in electronic patient records using a computerized patient records report system with abstraction of Map reduce of big data technology. We got the data to be processed from traditional system to Hadoop via ETL's. The data what you are going to analyze is a semi-structured data. After uploading their data to cluster anyone can access them again provided they got to be in the cluster or can also use virtual machines that contain the right software to analyze them without any need for conversion. Beyond improving profits and cutting down on wasted overhead, Big Data in healthcare is being used to predict epidemics, cure disease, improve quality of life and avoid preventable deaths. With the world's population increasing and everyone living longer, models of treatment delivery are rapidly changing, and many of the decisions behind those changes are being driven by data. The drive now is to understand as much about a patient as possible, as early in their life as possible – hopefully picking up warning signs of serious illness at an early enough stage that treatment is far more simple (and less expensive) than if it had not been spotted until later..**

***Keywords*— Hadoop; Map Reduc; Healthcare; Clustering; Predictive Analysis**

I. INTRODUCTION

The exponential production of data in recent years has introduced a new area in the field of information technology known as "Big Data". In a clinical setting such datasets are emerging from large-scale laboratory information system (LIS) data, test utilization data, electronic medical record (EMR), biomedical data, biometrics data, gene expression data, and in other areas. Massive datasets are extremely difficult to analyze and query using traditional mechanisms, especially when the queries themselves are quite complicated. In effect, a MapReduce algorithm maps both the query and the dataset into constituent parts. The mapped components of the query can be processed simultaneously or reduced to rapidly return results

Big datasets of clinical, biomedical, and biometric data have been processed successfully using the MapReduce framework on top of the Hadoop distributed file system. An overview of the Hadoop platform, MapReduce framework and its current applications has been reported for the field of bioinformatics. The promise of big data analytics in bioinformatics and health care in general has previously been described. However our review enlarges the scope to the application of the MapReduce framework and its open source implementation Hadoop to a wide range of clinical big data including:

- Publicly available clinical datasets: online published datasets and reports from the United States Food and Drug Administration (FDA).
- Biometrics datasets: containing measurable features related to human characteristics. Biometrics data is used as a form of identification and access control.
- Bioinformatics datasets: biological data of a patient (e.g. protein structure, DNA sequence, etc.).
- Biomedical signal datasets: data resulting from the recording of vital signs of a patient (e.g. electrocardiography (ECG), electroencephalography (EEG), etc.).
- Biomedical image datasets: data resulting from the scanning of medical.

A Map Reduce-based algorithm has been proposed for common adverse drug event (ADE) detection and has been tested in mining spontaneous ADE reports from the United States FDA. The purpose of the algorithm was to investigate the possibility of using the Map Reduce framework to speed up biomedical data mining tasks using this pharmacovigilance case as one specific example. The results demonstrated that the Map Reduce programming framework could improve the performance of common signal detection algorithms for pharmacovigilance in a distributed computation environment at approximately linear speedup rates. The Map Reduce distributed architecture and high dimensionality compression via Markov boundary feature selection have been used to identify unproven cancer treatments on the World Wide Web. The study showed that unproven treatments used distinct language to market their claims and this language was learnable, and through distributed parallelization and state of the art feature selection, it is possible to build and apply models with large scalability.

A novel system known as Group Filter Format has been developed to handle the definition of field content based on a

Pig Latin script. Dummy discharge summary data for 2.3 million inpatients and medical activity log data for 950 million events were processed. The response time was significantly reduced and a linear relationship was observed between the quantity of data and processing time in both a small and a very large dataset. The results show that doubling the number of nodes resulted in a 47% decrease in processing time.

II. RELATED WORK

To enhance the processing of conventional healthcare system, we have a proposed a series of Big Data health Care System by using Hadoop. There are many techniques proposed in order to efficiently process large volume of medical record which has explained below: Aditi Bansal and Priyanka Ghare proposed "Healthcare Data Analysis using Dynamic Slot Allocation in Hadoop". In this paper HealthCare System is analysis using Hadoop using Dynamic Hadoop Slot Allocation (DHSA) method. This paper proposed a framewook which focus on improving the performance of MapReduce workloads and maintain the system. DHSA will focuses on the maximum utilization of slots by allocating map (or reduce) slots to map and reduce tasks dynamically. Wullianallur Raghupathi and Viju Raghupathi has proposed "Big data analytics in healthcare: promise and Potential"

This author proposed the potential and promise of big data analytics in healthcare. The paper provides a broad overview of big data analytics for healthcare researchers and practitioners. 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. Its potential is great; however there remain challenges to overcome.

The use of IT in diagnostic and treatment processes will add to the development of networks of clinical, hospital and health care processes (Smith and Gert van der Pijl). Healthcare management is a growing profession with increasing opportunities in both direct and non-direct care settings.

As defined by Buchbinder and Thompson, direct care settings are those organizations that provide care directly to a patient, resident or client who seeks services from the organization. Non-direct care settings are not directly involved in providing care to persons needing health services, but rather support the care of individuals through products and services made available to direct care settings. The construction of medical information is important to improve the hospital medical care capability, the management decision-making level of health and the hospital operational efficiency.

Nowadays, comprehensive hospital information services and management platform have been established, centering on electronic medical records and clinical pathway. The establishment and use of these information systems played an important role in improving the degree of patient satisfaction, enhancing hospital efficiency and healthcare quality, protecting the safety of healthcare, and reducing healthcare costs.

.

III. DATA STORAGE

HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

A. Features of HDFS

- It is suitable for the distributed storage and processing.
- Hadoop provides a command interface to interact with HDFS.
- The built-in servers of namenode and datanode help users to easily check the status of cluster.
- Streaming access to file system data.
- HDFS provides file permissions and authentication.

B. Namenode

The namenode is the commodity hardware that contains the GNU/Linux operating system and the namenode software. It is a software that can be run on commodity hardware. The system having the namenode acts as the master server and it does the following tasks:

- Manages the file system namespace.
- Regulates client's access to files.
- It also executes file system operations such as renaming, closing, and opening files and directories.

C. Datanode

The datanode is a commodity hardware having the GNU/Linux operating system and datanode software. For every node (Commodity hardware/System) in a cluster, there will be a datanode. These nodes manage the data storage of their system. Datanodes perform read-write operations on the file systems, as per client request.

They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.

D. Block

Generally the user data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and/or stored in individual data nodes. These file segments are called as blocks. In other words, the minimum amount of data that HDFS can read or write is called a Block. The default block size is 64MB, but it can be increased as per the need to change in HDFS configuration.

E. Goals of HDFS

- Fault detection and recovery: Since HDFS includes a large number of commodity hardware, failure of components is frequent. Therefore HDFS should have mechanisms for quick and automatic fault detection and recovery.

- Huge datasets: HDFS should have hundreds of nodes per cluster to manage the applications having huge datasets.
- Hardware at data: A requested task can be done efficiently, when the computation takes place near the data. Especially where huge datasets are involved, it reduces the network traffic and increases the throughput.

IV. ANALYTICS FOR MEDICAL DATA

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job. The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once user write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model. The Algorithm

- Generally MapReduce paradigm is based on sending the computer to where the data resides!
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

A. Map stage

The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

B. Reduce stage

This stage is the combination of the Shufflestage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

- During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
- The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
- Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
- After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

The MapReduce framework operates on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types. The key and the value classes should be in serialized manner by the framework and hence, need to implement the Writable interface.

The fuzzy c-means (FCM) algorithm is a clustering algorithm developed by Dunn, and later on improved by Bezdek. It is useful when the required number of clusters are pre-determined; thus, the algorithm tries to put each of the data points to one of the clusters. What makes FCM different is that it does not decide the absolute membership of a data point to a given cluster; instead, it calculates the likelihood (the degree of membership) that a data point will belong to that cluster. Hence, depending on the accuracy of the clustering that is required in practice, appropriate tolerance mesaures can be put in place. Since the absolute membership is not calculated, FCM can be extremely fast because the number of iterations required to achieve a specific clustering exercise corresponds to the required accuracy.

V. IMAGE PROCESSING

Medical image processing requires a comprehensive environment for data access, analysis, processing, visualization, and algorithm development. MATLAB and Image Processing Toolbox to solve problems using CT, MRI and fluorescein angiogram images. Demonstrations will include the following highlights:

- Volume visualization of a brain MRI image stack
- Measurement of vessel tortuosity
- Video analysis and neural network-based classification of a fluorescein angiogram

VI. CONCLUSION

An integrated solution eliminates the need to move data into and out of the storage system while parallelizing the computation, a problem that is becoming more important due to increasing numbers of sensors and resulting data. And, thus, efficient processing of clinical data is a vital step towards multivariate analysis of the data in order to develop a better understanding of a patient clinical status (i.e. descriptive and predictive analysis). This highly demonstrates the significance of using the Map Reduce programming model on top of the Hadoop distributed processing platform to process the large volume of clinical data. .

The Hadoop platform and the Map Reduce programming framework already have a substantial base in the bioinformatics community, especially in the field of next-generation sequencing analysis, and such use is increasing.

The capability of big data will transform the way today's healthcare providers operate the sophisticated technologies to get knowledge from clinical records and make good decisions. In the nearby future I will see implementation of big data analytics in health care industry.

Big data provides security and privacy. This paper proposes a framework which is aiming that it will improve the performance of MapReduce workloads and at the same time will maintain the fairness.

REFERENCES

[1] Bo Wu and Haiying Shen, (2017) “Exploiting Efficient Densest Subgraph Discovering Methods for Big Data," IEEE Transactions on Big Data.
[2] Dongfang Zhao, Kan Qiao and Zhou Zhou, et al, (2017) “Toward Efficient and Flexible Metadata Indexing of Big Data Systems," IEEE Transactions on Big Data
[3] Gemson Andrew Ebenezer J and Durga S, (2015) “Big data analytics in healthcare: a survey," ARJN Journals on Engineering and Applied Sciences.
[4] Mehraj Ali and John Kumar, (2014) “Implementation of Image Processing System using Handover Technique with Map Reduce Based on Big Data in the Cloud Environment," IAJIT Publications on Big Data.
[5] Mukesh Borana , Manish Giri and Sarang kamble et al, (2015) “ Healthcare Data Analysis using Hadoop," (IRJET)International Research Journal of Engineering and Technology.
[6] Prasan kumar sahoo, suvendu kumar mohapatra and shih-lin wu, (2017) “Analyzing Healthcare Big Data With Prediction for Future Health Condition," IEEE Access on Big Data.
[7] Ravindra Ch, G Rajesh, Annapurna G and Ch Swetha et al, (2015) “Automated Health Care Management System Using Big Data Technology,"(JNCET) Journals of Network Communications and Emerging Technology.
[8] Lalit Malik and Sunita Sangwan, (2015) “MapReduce Framework Implementation on the Prescriptive Analytics of Health Industry,"(IJCSMC) International Journal of Computer Science and Mobile Computing.
[9] Le Dong, Zhiyu Lin, Yan Liang, Ling He and Ning Zhang et al, (2016) " A Hierarchical Distributed Processing Framework for Big Image Data," IEEE Transactions on Big Data.
[10] Shlomi Dolev, Patricia Florissi and Ehud Gudes et al, (2017) “A Survey on Geographically Distributed Big-Data Processing using MapReduce," IEEE Transactions on Big Data.