Modern Framework for Distributed Healthcare Data Analytics Based on Hadoop

P. Vignesh Raja and E. Sivasankar

Dept. of Computer Science and Engineering, National Institute of Technology, Trichy, India
vigneshrp@cdac.in, sivasankar@nitt.edu

Abstract. Evolution in the field of IT, electronics and networking resulted in enhancements in connectivity and in computation capabilities. Proliferation of miniaturized devices paved way for Body Area Network. Healthcare systems have been going through cycles of modernization as the advent of IT system in the field of medical sciences. Body Area Network is a network of lightweight wearable sensor nodes that sense human body functions. Modern healthcare informatics systems produce lots of data that emerge from sensors. Even though many healthcare informatics systems exist and produce volume of data, such solutions exist in silos. Existing Healthcare IT systems intend to gather multivariate medical data about the patients by the means of electronic format. They capture multi variant types of data, process and store them in a RDBMS. Inferring knowledge from such systems is tedious. This paper aims at proposing a framework for modernizing the healthcare informatics systems. Proposed framework is based on Apache Hadoop platform which is open source and its implementation is distributed in nature as it is deployable at various healthcare centers in different geographic locations.

Keywords: E-Health, Medical Informatics, Hadoop, Map Reduce, Hive, Massive Multivariate Data sets, BigData Analytics.

1 Preliminaries

1.1 Health Informatics Systems

Health informatics is a multidisciplinary field that deals with computer science, information science, biology, medicine and analytics. Evolution in the electronics, IT field fostered various health informatics solutions. Almost all of the healthcare centers are equipped with health informatics solutions [1] [13]. Such solutions generally consists of modules like hospital information management system, telemedicine solutions and others. These systems basically collect data about the patients and medication in text format, also they store x-rays, scan reports in image format. Such implementations eventually result in “data silo”. Data captured by each of the system is local to that environment and they became less reusable. The more the number of health informatics solutions the more is the data they produce. Appropriate mechanism is needed in order to store, process and extract knowledge from the massive amount of

Linawati et al. (Eds.): ICT-EurAsia 2014, LNCS 8407, pp. 348–355, 2014.

data. Exchange of vital medical practices and medication assistances through the present healthcare informatics systems have the following concerns

- Ability to exchange the data externally
- Issues related to data integration standards
- Veracity of the healthcare data.

As a health informatics solution, system should provide mechanism for better understanding of medication compliance. Huge amount of money is spent wastefully in healthcare across worldwide is attributed due to excess hospitalization which is triggered by non-compliance of medication procedures. Such incidents are also caused by situations where it lacks of clinically proven medication solutions.

1.2 Body Sensor Networks

As the miniaturized sensors emerge as the promising method of self-care medication, establishment of Body Area Network or Body Sensor Network is feasible [14][2]. The purpose of "Body Sensor Networks for Healthcare" is to bring the expertise of doctors in the urban areas closer to the patients in the rural areas. It would bridge geographical distances and provide healthcare to those who do not have access to quality healthcare. More than 70 % of the land area is covered with GSM/GPRS for Internet access [2]. Integration of existing healthcare IT system with wearable health monitoring system is niche in the healthcare IT solutions [3]. One of the major limitations for wider acceptance of such system was non-existent support for massive data collection and knowledge discovery.

Implementation of BSN for healthcare is carried out through infrastructure like medical Kiosk that can enable any person to walk into the kiosk and check the vital parameters like ECG, Blood Pressure, and Pulse Rate etc. The data measured at these kiosks is stored in the distributed datastore and is made available to the doctors / Paramedics. The data could range from text (health records, reports) to image (x-ray, ECG images). Data gathered at each of the healthcare centers are stored in the relational database management system. With such system we have variety and volume of data.

RDBMS lacks of fault tolerance, linear scalability, processing of unstructured data and it cannot effectively handle high concurrent reading and writing of database [4]. Also true decentralized approach is required whereby valuable medication information can be shared. Present health informatics systems do not allow exchange of data or when there is need for "healthcare related knowledge" within from the existing data. Along with the massive multivariate data, a mechanism for inferring knowledge from such data is highly required.

2 Proposed Framework

Implementation of new architecture is needed because of the following reasons, the storage required to store healthcare records for large number of patients may results in

terabytes (TB) to petabytes (PB) in a typical distributed environment. Also the new architecture is expected to provide storage of massive structured and unstructured data, parallelism in data processing and high availability with fault tolerance. Performing analytics on the historical data to infer knowledge is also a feature of this framework. Proposed model envisages to infer knowledge from such distributed, multivariate data. This framework enables provisions for "Knowledge on Demand". This framework also have a mechanism to seamlessly interface with the legacy system where from the data is imported to perform analytics. As this framework is built using open source tools, extension of functionalities to suit custom requirements is made possible.

Variety of choices including RDBMS (Postgresql 9.x) with high availability, Pandas the open source Python language based analytics library were considered for performing analytics on the data. Even though the recent versions of Postgresql offer ability to handle large volume of data (up to TBs) [15], a hardware infrastructure of higher configuration is required to support such a high volume of data. Further such RDBMS are prone to mid query faults, whereas the proposed Hadoop based framework offers mid query fault tolerance. Pandas – Python based data analytics library [16] offers many features such as high performance merging and joining, integrated indexing on the given dataset. But most of the data sets using which the analytics is performed, is distributed in nature. Proposed framework provides mechanism to interface with the data sources and it performs analytics using the commodity hardware. So setting up of such a framework does not require the hardware with higher configuration.

Multi variant types of data generated by the existing BSN framework is stored in RDBMS, which is distributed at many locations. This RDBMS based framework cannot be relied on, when the nature of application is with volume, variety of health care data. It is also very cumbersome to manage such data and to perform analytics on such data to infer knowledge [4]. By using Hadoop, Hive existing framework can still be exploited to perform analytics on the data it generates. Hadoop based tool sets are introduced in order to alleviate such limitations.

Following are the few usecases that can be executed using this modern framework. Finding out the age group and demography where a particular disease is spreading. Such information will help the Government in planning and implementing suitable health schemes. When the doctors are dealing with critical medical cases, where immediate medical assistance is required, knowledge inferred using the proposed framework can be used in a near real time (Near real-time does not mean the on-demand data by monitoring the patients, but the provision to assist the users by performing analytics on the existing voluminous multivariate datasets). Another usecase is the analytics of structured and unstructured data to provide knowledge when doctors search for specific issues. Any kind of data (medical records, CT scans, emails) can processed through this framework and doctors can then extract the information that they need based on specific symptoms. Below is the components of the framework.

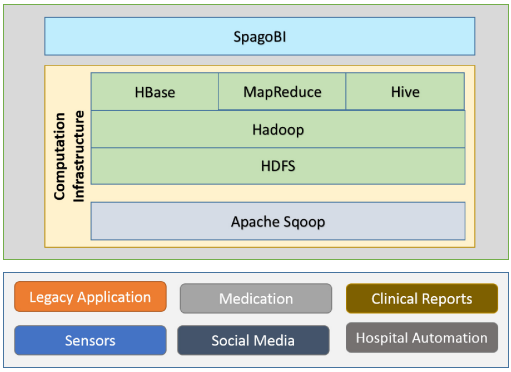

Fig. 1. Proposed Framework

2.1 Apache Hadoop

Apache Hadoop is a software framework and it processes large data sets across clusters. Such clusters could be on commodity hardware. With MapReduce (the default programming model), the application is divided into many small fragments of work (Jobs), each of which can execute or re-execute on any node in the cluster [5]. Jobs submitted to the Hadoop cluster is managed by Job Tracker and Task Tracker. Hadoop is designed to scale up from single servers to thousands of machines, each offering local computation and storage. It also provides a distributed file system, HDFS [6]. Overall functioning of inferring knowledge (analytics) is shown below,

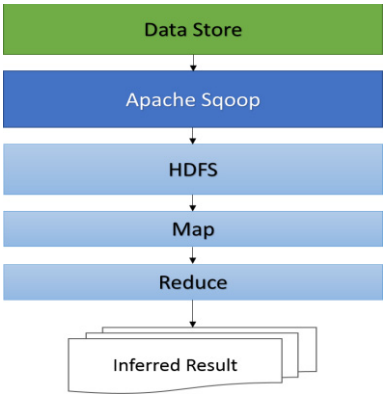

Fig. 2. Map Reduce Process

2.2 Apache Hive and HBase

Hive is a data warehouse infrastructure built on top of Hadoop and it facilitates querying and managing large datasets residing in distributed storage [7]. Proposed framework supports analysis of data from legacy systems, where the data is available in the

RDBMS. Hive supports SQL (HiveQL) like interface, which is familiar to the data professionals. HiveQL basically compiled into map reduce jobs. These jobs are then submitted to the Hadoop for further execution. Extension of existing queries in order to be equipped with tailored map reduce scripts is made feasible by HiveQL. Query plan is generated from the HQL scripts through the execution engine that is available in the Hive. HBase is an open source distributed database. HBase is highly suited for storing non-relational data.

2.3 Apache Sqoop and SpagoBI

Apache Sqoop, tool for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases is also used [8] [9] [12]. Sqoop transfers data from data stores like MySQL, Postgresql and Oracle to Hive and HBase. Proposed framework will also contain SpagoBI to offer BI solutions. SpagoBI is an open source business intelligence suite. Apart from performing query and analytics on large volume of heterogeneous, structured and unstructured data, SpagoBI provides mechanism to extract useful information. SpagoBI can extract information from various platforms such as Hadoop, Hive, HBase and Cassandra. SpagoBI analytical engine can produce charts, reports, thematic maps and cockpits from the information retrieved from such platforms [10] [11].

3 Related Work

3.1 Web Based Health Informatics Application

Existing traditional web based health informatics application is used in order to evaluate this framework. Existing application is based on Java Technologies (J2EE, JSP, and Hibernate). This application is capable interfacing with various sensors that are meant to sense various health factors such as heart beat, blood pressure etc. In its aboriginal form this application is in the state of “Data Silo”. All the data captured by this application is internal to this application. For the purpose of interfacing with external application this system is enabled with Web Services. But as the size of total data increases there happens performance lagging since the data is maintained in the traditional database (Postgresql). This web based application also captures medication for various symptoms, prescribed by the Physicians.

Currently this system lacks the ability to perform analytics (RDBMS systems analytics performance depends on the factors such as types of queries, number of joins and hardware capacity) on the massive datasets. Proposed framework is applied for this scenario with various options and the results are recorded. Performance evaluation on healthcare data analysis using MapReduce and Hive is carried out against the traditional RDBMS system. Experiment results proves the limitations of the RDBMS when it comes to voluminous data to be dealt and eventually RDBMS based solutions fail to perform data analytics. But MapReduce and Hive continue to work for voluminous data whereas Hive has an edge over MapReduce in terms of performance.

As the SpagoBI is highly configurable Business Intelligence suite is made available as the main interface to the underlying computation infrastructure, various health

related information (medication, lab reports etc.) can be easily obtained. SpagoBI is highly customizable to display the rate of change in the health information inferred from the data store via dashboards.

3.2 Test Environment

The following environment is setup to evaluate this framework, whereas the data for experiment is taken from web based health informatics application. The data is obtained using Sqoop and stored in the HBase datastore. Data from the application is about heart beat rate, blood pressure level. Total size of the data is approximately 200 GB. Even though 200 GB is not a really the large one, sample dataset was generated based on this dataset of size 200 GB to result in more GBs. The existing database based system's performance has come down and the mid query faults brought havoc to the execution of analytics jobs.

We have used four machines with Intel i7 processors (up to 3.70GHz) and each machine is equipped with 8GB of main memory, 750 GB of storage. We have purposely used commodity hardware to exploit the features of Hadoop. Out of these four machines one machine is configured as Name Node, two machines for data storage and on the fourth machine the application is configured. This system is set with default replication factor (3 replicas) that is configured in the Hadoop environment. We have experimented this setup with operations like uploading of data and querying of data. Data on the aforementioned health parameters for about 75 patients is uploaded to the system. The system is simulated (based on the sample data) to produce as many records as it requires for conducting the comparison. Heart beat rate of a patient for a day is initially queried, also we queried the number of days when the heart beat rate exceeds certain threshold value. We ended up this experiment analyzing the age group of patients who are likely to get higher heart beat rates. The results of this experiments is tabulated below.

Table 1. Hadoop vs RDBMS

Data Size (No of Patient Records)	Time taken by proposed frame-work (sec)	Time taken be RDBMS (sec)	Operation
5000	0.823	4.213	Upload
5 Million	1.145	8.987	Upload
50 Million	3.652	47.598	Upload
5000	12.467	75.912	Query
5 Million	22.445	140.876	Query
50 Million	35.098	270.045	Query

From the table above it is evident that the performance of proposed framework is stable even when the number of record increased drastically whereas the performance of the RDBMS based system is lagging as the data size increases. The sample dataset used for this application consists of data such as blood pressure level, heart beat rate,

age and the disease along with other information. In order to infer knowledge from such datasets, queries are constructed and these queries are execued both on the RDBMS also on the proposed framework. Hadoop based framework converts the the given task into Map-Reduce tasks and executes these tasks on the nodes. We have inferred knowledge (finding out the age group and variation in the heart beat rate due to a disease) in the form of a triplet, {*Age Group,Heart Beat Rate,Disease}*. The more is the details of the inferrence so is the complexity of queries to be constructed. For inferring such knowledge we have used the Select, Aggregation and Join queries.

It is observed that the RDBMS based system failed as we increase the size of datasets also the execution involved in in may sub queris and joins. The next experiment that we performed is to evaluate the performance of MapReduce and that of Hive. Basically this involves quering the massive database thereby to arrive at knowledge from the datastore. Aforementioned health data is again used for this purpose. One intersting phenomenon observed out of this experiment is that the performance of Hive proved to be better than that of the MapReduce. When the same experiment is repeated on the archived data of various intervals, and as the data size increases Hive had offered better performance than that of MapReduce jobs. RDBMS cannot be applied in situations like analyzing the archived data of massive data size. Such analysis is required to infer knowledge as we have indicated few usecases in section 2.

4 Conclusion

In this paper capabilities of exisitng healthcare informatics systems is addressed. Also the need for decentralized parallel architecture for processing and extracting knowledge out of the datasets is highlighted. For this purpose a modern framework for processing and inferring of voluminous multivariate data is proposed. The proposed framework is fully based on open source tools which provides options for customization and really suits the existing distributed architecture. Also we highlighted the performance evaluation of this framework against RDBMS based data storage. This framework can be configureable on any commodity hardware because of the inherent fault tolerance mechanisms provided by the Hadoop environment.

References

1. PWC on Clinical Informatics, http://www.pwc.com/us/en/press-releases/2012/clinical-informatics-full-report-press-releases.jhtml
2. TRAI, http://www.trai.gov.in/WriteReadData/PressRealease/Document/PR-TSD-03JULY2013.pdf
3. Darwish, A., Hassanien, A.E.: Wearable and Implantable Wireless Sensor Network Solutions for Healthcare Monitoring. Sensors 11(6), 5561–5595 (2011)

4. Bao., Y., Ren., L., Zhang., L., Zhang., X., Luo, Y.: Massive sensor data management framework in Cloud manufacturing based on Hadoop. In: 10th IEEE International Conference on Industrial Informatics (INDIN), pp. 397–401 (2012)
5. Apache Hadoop, http://hadoop.apache.org/
6. White, T.: Hadoop: The Definitive Guide, pp. 9–72. O'Reilly Media Inc., CA (2012)
7. Apache Hive, http://hive.apache.org/
8. Apache Sqoop, http://sqoop.apache.org/
9. Ting, K., Cecho, J.J.: Apache Sqoop Cookbook, pp. 1–23. O'Reilly Media Inc., CA (2013)
10. SpagoBI, http://en.wikipedia.org/wiki/SpagoBI
11. Spago BigData, http://www.spagoworld.org/xwiki/bin/view/SpagoBI/BigData
12. Dobre., C., Xhafa, F.: Parallel Programming Paradigms and Frameworks in Big Data Era. International Journal of Parallel Programming (2013)
13. McKinsey healthcare Report, http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care
14. Ullah, S., Higgins, H., Braem, B., Latre, B., Blondia, C., Moerman, I., Saleem, S., Rahman, Z., Kwak, K.S.: A Comprehensive Survey of Wireless Body Area Networks. Journal of Medical Systems 36(3), 1065–1094 (2012)
15. Postgresql Wiki, http://wiki.postgresql.org/wiki/FAQ
16. Python data analytics library, http://pandas.pydata.org/