Improving Usability, Safety and Patient Outcomes with Health Information Technology
F. Lau et al. (Eds.)

doi:10.3233/978-1-61499-951-5-229

A Hadoop/MapReduce Based Platform for Supporting Health Big Data Analytics

Alex KUO[a,1], Dillon CHRIMES[b], Pinle QIN[c], Hamid ZAMANI[a]
[a] *School of Health Information Science, University of Victoria, Victoria, BC, Canada*
[b] *Vancouver Island Health Authority, Victoria, BC, Canada*
[c] *Faculty of Big Data, North University of China, Shanxi, China*

Abstract. In this paper, we report our practical experience in designing and implementing a platform with Hadoop/MapReduce framework for supporting health Big Data Analytics. Three billion of emulated health raw data was constructed and cross-referenced with data profiles and metadata based on existing health data at the Island Health Authority, BC, Canada. The patient data was stored over a Hadoop Distributed File System to simulate a presentation of an entire health authority's information system. Then, a High Performance Computing platform called WestGrid was used to benchmark the performance of the platform via several data query tests. The work is important as very few implementation studies existed that tested a BDA platform applied to patient data of a health authority system.

Keywords. healthcare, big data analytics, Hadoop, MapReduce

1. Introduction

Big Data in healthcare is different from other disciplines such as social network or transactional business data in that it includes standardized structured, coded data (e.g. ICD, SNOMED CT), semi-structured data (e.g. HL7 messages), unstructured clinical notes, medical images (e.g. MRI, X-rays), genetic lab data, and other types of data (e.g. public health and mental or behavioral health). Huge volumes of very heterogeneous raw data are generated daily by a variety of hospital systems such as Electronic Health Records, Computerized Physician Order Entry, Picture Archiving and Communication Systems, Clinical Decision Support Systems, and Laboratory Information Systems. These information systems are utilized for functionalities in many healthcare settings such as physician offices and hospitals.

Several published studies have asserted that Big Data managed effeciently can improve care delivery while reducing healthcare costs [1-4]. A McKinsey Global Institute study suggests, "If US healthcare were to use big data creatively and effectively to drive efficiency and quality, the sector could create more than $300 billion in value every year" [5]. A number of published articles also reported using Big Data to improve population health with better policy decision making.

The process of extracting knowledge from sets of Big Data is called Big Data Analytics (BDA) [6]. Kuo et al. [7] and Chrimes et al. [8] further described the potential

[1] Corresponding Author: Dr. Alex Kuo, School of Health Information Science, University of Victoria, email: akuo@uvic.ca

process challenges of achieving full Big Data utilization in five distinct configuration stages: data aggregation, data maintenance, data integration, data analysis, and pattern interpretation. Among the analytic stages, data analysis over vast volumes is key for a successful BDA [9]. However, it is very difficult to efficiently analyze the data using traditional analytic software, such as IBM SPSS, Microsoft Excel or MathWorks MATLAB because Big Data is too large, too distributed, unstructured and heterogeneous. It can take several days, even months, to obtain a result over a very large data set (in terabytes and beyond). Moreover, for complex analyses, the computing time increases exponentially even with small amount of data growth. In the case of Bayesian Network, a popular algorithm for modeling knowledge in computational biology and bioinformatics, the computing time required to find the best network increases exponentially as the number of records rises incrementally. To address the analytical challenges, many recently published studies have suggested that using High Performance Computing (HPC), and parallelization of computing model can efficiently increase analysis performance for the computationally intense problems [10-14].

In this study, we described our practical experience among collaborations with Vancouver Island Health Authority (VIHA) funded research project for health Big Data Analytics. The main objective of this project was to collaborate with establishment of a BDA platform for application. A Hadoop/MapReduce framework formed the platform with noSQL database called HBase representing real hospital-specific metadata and file ingestion. Three billion of emulated patient data were generated and cross-referenced with inpatient profiles based on metadata dictionaries at VIHA.

2. Methods

The basic premise of the implementation of a BDA platform for use in healthcare was to construct the platform capable of compiling heterogeneous clinical data from diverse sources of the hospital system and querying large volumes quickly. Also, the applications must ensure patient data security/privacy. This section describes our approaches.

2.1. The Architecture of the Analytics Platform

The BDA platform harnesses the technical power and advanced programming to produce accessible front-end tools to end users that allow for analysis of large quantities of back-end data in an interactive enriching manner. All this must be accomplished at cost-effective expense for a successful platform to be deployed. Based on the design philosophy, we construct a dynamic platform with interfaced applications (i.e., Apache Phoenix, Spark, and Drill) linked to backend HBase over Hadoop Distributed File System (HDFS). With the Hadoop/MapReduce framework, the platform allowed users to easily analyze and visualize health Big Data [9, 15].
The platform included four components (see Figure 1):

(1). A clinical data warehouse stores healthcare data. Currently at VIHA there are over 1000 tables in its Admission, Discharge and Transfer (ADT) data from hospital system, and annually ca. one million patient encounters add to 50+ years archive (500 million at VIHA and 10 billion Provincially).
(2). High performance Linux clusters (WestGrid University System) were used to install software of big data technologies, build configurations, and run simulation queries (Hadoop ecosystems including Apache Phoenix, Spark and Drill).

(3). HBase noSQL database was used to store data from VIHA clinical data warehouse. HDFS distributes the data to indexed storage across the WestGrid clusters with backup, high availability and redundancy.

(4). A master deployment manager (DM) was used to access the clusters from sponsored accounts over the Portable Batch System (PBS) of the Resource Manager. The access to the DM is controlled by lightweight directory access protocol (LDAP) while accessing to worker nodes was restricted to only the user running the job. This architecture permitted an agile and stabilized access with system administrator that could be launched from any terminal for each PBS job.

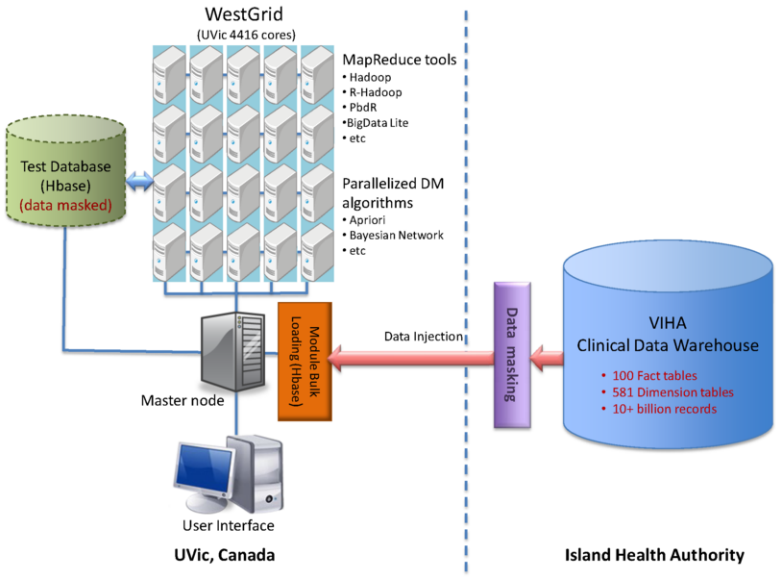

Figure 1. The Big Data Analytics Platform Architecture

2.2. The High Performance Computing Infrastructure

In this study, as described above, we relied on WestGrid's existing architecture as the computing infrastructure. WestGrid is a nationally Canadian funded program started in 2003, mainly used in western Canada while EastGrid and Ontario and Quebec grids are available. WestGrid installation at the University of Victoria (UVic) started in July 2010. The WestGrid computing facilities at UVic have 2 main clusters called Hermes and Nestor. The computing system of these two clusters share infrastructure such as resource management, job scheduling, networked storage, and service and interactive nodes. Hermes is a capacity cluster geared towards serial jobs with 84 nodes having 8 cores each and 120 nodes with 12 cores each, which gives a total of 2112 cores. Nestor is a large cluster consisting of 288 nodes (2304 cores) geared towards large parallel jobs. In this study, we use five dedicated worker nodes and one head node from Hermes cluster.

2.3. Data privacy protection

Ensuring patient data security and privacy was an important requirement in this study. The established platform used the following methods to protect data security and privacy:

(1). Data Masking – Typically this is carried out by database administrators thru rules and regulations set by business/security analysts based on current legislations of BC Ministry of Health. The goal was to generate a comprehensive list of sensitive elements specific to the organization and associated tables, columns, and relationships across the data warehouse and encryption of indexed key stores provided by HBase [8].
(2). Data replication – We worked in conjunction with Business Intelligence and Data warehouse, Clinical reporting, Application Platform Services, Database Administrators, and Physicians/Nurses groups to identify the masking or encryption required and optimal techniques to de-identify and restrict access to patient data. Once the data form distributed HBase data sets across working nodes, it was queried via Apache Phoenix, Spark and Drill only thru PBS held by WestGrid.
(3). Using HBase and WestGrid for Security/Privacy Mechanisms – HBase provided comprehensive security/privacy support thru its qualifiers and key-stores of data ingested. The access control to data stored in HBase was at table level, column family level and column level. HBase supports Kerberos authentication, Remote Procedure Call (RPC) and at-rest privacy protection. Data could not be queried without WestGrid for authentication.

3. Methods

3.1. Data Emulation and Modeling

A BDA platform was benchmarked its performance with clinical data warehouse utilization processes at the hospital level. Currently, huge volumes of health data are continuously generated and added into the archive. Within the archive of data warehouse, two of the largest data sets are the Admission, Discharge, Transfer (ADT) and the Discharge Abstract Database (DAD). ADT has over 1000 tables with 75 columns containing individual patient bed-tracking information, while the DAD is set by a data dictionary of 28 columns contains Canadian Institute for Health Information's (CIHI) diagnostic codes and discharge abstract metadata. These data sets are not system linked to form an all-encompassing database. Therefore, this study showed that these two data sets can be appropriately combined via big data technologies.

In a hospital system, the capacity to record patient data efficiently in the ADT is crucial to timely patient care and quality patient-care deliverables. Thus, the ADT system is often referred to as the source of truth for reporting operations of inpatient to outpatient and discharged [15]. In most Canadian hospitals, discharge records are subject to data standards set by the CIHI and entered into the Canada's national DAD. These two reporting systems, i.e., ADT and DAD, account for the majority of the patient data in hospitals, but they are seldom aggregated and integrated as a whole because of their complexity and large volume. A suitable analysis of ADT and DAD integrated data in this study shows many benefits of using big data technologies to produce high volumes while interactively applying new ways to query the data to find unknown correlations and trends.

Three billion of similated health raw data was constructed and cross-referenced with patient data profiles and metadata based on existing health data sets and elements in standardized data dictionaries. However, there are three main limiting factors in its deployment of using real patient data over application platform. The first reason is any

ethics approval for accessing the entire patient data of the health authority system will be very time consuming. Second, in the proposed analytic setting, the data will have to be migrated off the production database to avoid consuming network resources. This external architecture requires approval based on the VIHA's regulations for public disclosure. Finally, patient data must be masked/encrypted to standards set to pass privacy impact assements. Therefore, several teams are required to identify the sensitive data then scramble or mask data via optimal techniques to initalize the deployment with end users acceptance of its usability.

Over the span of twelve months in 2014-2015, several interviews were conducted with business intelligence data warehouse, clinical reporting, application platform, and health informatics architecture teams employed at VIHA [9]. During these interviews, an emulated health Big Data was generated from hospital admissions (based on encounter types) and a discharge system (based on diagnoses and procedures). In it, data profiles (including dependencies) and the importance of the metadata for the clinical reporting were confirmed and verified. Furthermore, current reporting limitations of the different combinations of the DAD and ADT data were recorded to form accurate simulation of the existing and future queries. To test the feasibility of the BDA platform and its performance, the emulated patient data had 90 columns that combined DAD and ADT metadata profiles. Thus, it was an accurate representation of the construct of real patient data based on encounter types, location and date/times.

3.2. Data Ingestion and Query Performance Evaluation

The pathway to running ingestions and queries over the BDA platform includes nine pipelined steps [9, 15]: (1) Generating *.csv* flat files, (2) Apache Phoenix Module Load, (3) HDFS Module and Ingestion of HFiles, (4) Bulkloading HFiles to HBase, (5) HBase Compression, (6) Phoenix SQL-like Queries, (7) Apache Spark and Drill Module Loads, (8) Notebook and Python/Pypark Module Loads, (9) Spark and Drill SQL-like Queries.

Thru this sequence, the Phoenix module loaded after Hadoop and HBase SQL code was directed and then iteratively run to ingest three billion rows to the existing HBase. Phoenix can run SQL-like queries against the HBase data. It was utilized to index and place schema over each *.csv* file bulkloaded to ingest using MapReduce. The queries via Apache Phoenix resided as a thin SQL-like layer on HBase. This allowed ingested data to form structured schema-based data in the noSQL database (i.e. HBase). The loads were 50 million each via the index and schema between HBase's RegionServers thru a functional SQL-like code of "salt bucket" that set the number of worker nodes in the cluster to five evenly distributed data. This additional code was deemed necessary as HDFS did not automatically distribute evenly and unbalanced data slowed performance [9]. Performance was measured with three main processes: HDFS ingestions, bulkloads to HBase, and query times. Three flat files (*.csv*) with different number of rows (50 million, 1 and 3 billon) were ingested to HDFS for testing (Table 1).

Table 1. HDFS Data Ingestions

Data Size	Ingestion Time
50 Million records (23GB)	~3-6 min
1 Billion records (451GB)	~60-120 min
3 Billion records (10TB)	~180-360 min

At an optimized iteration, Hadoop Distributed File System (HDFS) ingestion required three seconds but HBase required four to twelve hours to complete the Reducer of MapReduce. HBase bulkloads took a week for one billion and over two months for three billion (see Table 2).

Table 2. HBase Bulkload Durations

Data Size	Bulkloaded Time
50 Million records (0.5TB)	3-12 hrs
1 Billion records (10TB)	60-240 hrs
3 Billion records (30TB)	300-480 hrs

There were 22 test queries for different questions using the ADT and DAD combined data over 50 million, one three billion rows. All queries run on Zeppelin, Jupyter, Spark-terminal and *Pyspark*, as well as Drill took approximately the time of 50-120 seconds to load the data and query all 22 queries could be run at the same time. Spark was configured to run on specialized Yarn-client with 10 executors, four cores with 8 GB of RAM each; therefore, each node had two executors with a total of eight cores and 16 GB memory. However, Drill was faster with its configuration involving inherent ZooKeeper allocations via its *drillbit* components (see [8] for details).

4. Conclusion

In this study, Hadoop/MapReduce framework was proposed to implement the data-intensive distributed computing platform. Srirama et al. [16] indicated that Hadoop is suitable for simple iterative algorithms where they can be expressed as a sequential execution of constant MapReduce models (that could also be configured to be representative of the clinical event model of hospital systems). It is not well suited for complex statistical analysis or iterative problems. To amend the Hadoop's ecosystem weaknesses, we plan to engineer "R" to work over Hadoop (e.g. RHadoop*).* R provides a wide variety of statistical and graphical techniques, modeling, statistical tests, time-series analysis etc. R and Hadoop complement each other very well in BDA and in data visualizations [17].

This study comprised a constructed Hadoop/MapReduce framework to form a platform for Health Big Data. As indicated in the study [18], there are many analytical challenges to achieve full value of Big Data in Canadian healthcare systems because of information silos, various policies and regulations, and cultural *diversity* in *the healthcare systems*. These hinder patient data in different health care system to be fully integrated. However, our platform did allow for replication of patient data, which reduces architectural resource pressure while integrating data from different data sources to form one patient-encounter-centric database for ongoing analysis. Also, since HBase is linearly scalable and there were no differences in query durations; therefore, it is expected that query time will be a few milliseconds as the number of computing nodes increased to 100+.

Few studies have tested a variety of Big Data tools in Hadoop's ecosystem in healthcare. And even fewer studies have established a simulation of 3 billion patient records. Therefore, this study achieved the top three V's that define Big Data: high performance (or velocity) over its generator of detailed data (or variety) that formed

extremely large quantities (or volume) significantly contributed to ongoing development of Information Management and Information Technologies (IMIT) in healthcare [15].

References

[1] M.M. Hansen, T. Miron-Shatz, A.Y.S. Lau, C. Paton, Big data in science and healthcare: A review of recent literature and perspectives, *Yearbook of Medical Informatics* **9**:1 (2014), 21–26.

[2] J. Manyika, M. Chui, J. Bughin, B. Brown, R. Dobbs, C. Roxburgh, B. Hung, Big data: The next frontier for innovation, competition, and productivity, URL: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[3] Canada Health Infoway. Big Data Analytics in Health - White Paper 2013.

[4] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, *Health Information Science and Systems* **2**:3 (2014), 1-10.

[5] How Big Data Impacts Healthcare, *Harvard Business Review* (2014), 1-4.

[6] C.J. Saunders, N.A. Miller, S.E. Soden, D.L. Dinwiddie, A. Noll, N.A. Alnadi, et al., Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units, *Sci. Trans. Med.* **4**(154), 154ra135. http://doi.org/10.1126/scitranslmed.3004041, 2012.

[7] M.H. Kuo, T. Sahama, A.W. Kushniruk, E.M. Borycki, D. Grunwell, Health big data analytics: Current perspectives, challenges and potential solutions, *International Journal of Big Data Intelligence* **1**:4 (2014), 114–126.

[8] D. Chrimes, M.-H. Kuo, B. Moa, W. Hu, Towards a real-time big data analytics platform for health applications, *International Journal of Big Data Intelligence* **4**:2, (2017).

[9] D. Chrimes, B. Moa, M.-H. Kuo, A. Kushniruk, Operational efficiencies and simulated performance of big data analytics platform over billions of patient records of a hospital system, *Advances in Science, Technology and Engineering Systems Journal* **2**:1, 23-41 (2017).

[10] K. Deepthi, K. Anuradha, Big data mining using very-large-scale data processing platforms, *International journal of engineering research and applications,* **6**:2 (2016), 39-45.

[11] Y.P. Zhang, et al., i^2 MapReduce: Incremental MapReduce for mining evolving big data, *IEEE Transactions on Knowledge and Data Engineering* **27**:7 (2015), 1906-1919.

[12] D.P. Vaidya, S.P. Deshpande, Parallel data mining architecture for big data, *International journal of electronics, communication and soft computing science and engineering*, (2015), 208-213.

[13] X. Wu, X. Zhu, G.Q. Wu, W. Ding, Data mining with big data, *IEEE Transactions on Knowledge and Data Engineering* **26**:1 (2014), 97-107.

[14] E.A. Mohammed, B.H. Far, C. Naugler, Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends, *BioData Mining* **7**: 22 (2014), 1-23.

[15] D. Chrimes, *Towards a Big Data Analytics Platform with Hadoop/MapReduce Framework using Simulated Patient Data of a Hospital System*, Master thesis, 2016, School of Health Information Science, University of Victoria, BC, Canada.

[16] S.N. Srirama, P. Jakovits, E. Vainikko, Adapting scientific computing problems to clouds using MapReduce, *Future Generation Computer Systems* **28**:1 (2012), 184-192.

[17] S. Das, Y. Sismanis, K.S. Beyer, Ricardo: Integrating R and Hadoop*, Proceedings of the 2010 ACM SIGMOD/PODS Conference (SIGMOD'10). 2010, 987-998.*

[18] Y. Dufresne, S. Jeram, A. Pelletier, The True North Strong and Free Healthcare? Nationalism and Attitudes Towards Private Healthcare Options in Canada. *Canadian Journal of Political Science* **47**:3 (2014), 569–595.