Web-Enabler: Transformation of Conventional HIMS Data to Semantics Structure Using Hadoop MapReduce

Muhammad Idris*, Sungyoung Lee*
{idris,sylee}@oslab.khu.ac.kr
*Dept. of Computer Engineering, Kyung Hee University

Abstract

Objective: Data exchange, interoperability, and access as a service in healthcare information management systems (HIMS) is the basic need to provision health-services. Data existing in various HIMS not only differ in the basic underlying structure but also in data processing systems. Data interoperability can only be achieved when following a common structure or standard which is shareable such as semantics based structures. We propose web-enabler: A Hadoop MapReduce based distributed approach to transform the existing huge variety data in variety formats to a conformed and flexible ontological format that enables easy access to data, sharing, and providing various healthcare services. Results: For proof of concept, we present a case study of general patient record in conventional system to be enabled for analysis on the web by transforming to semantics based structure. Conclusion: This work achieves transformation of stale as well as future data to be web-enabled and easily available for analytics in healthcare systems.

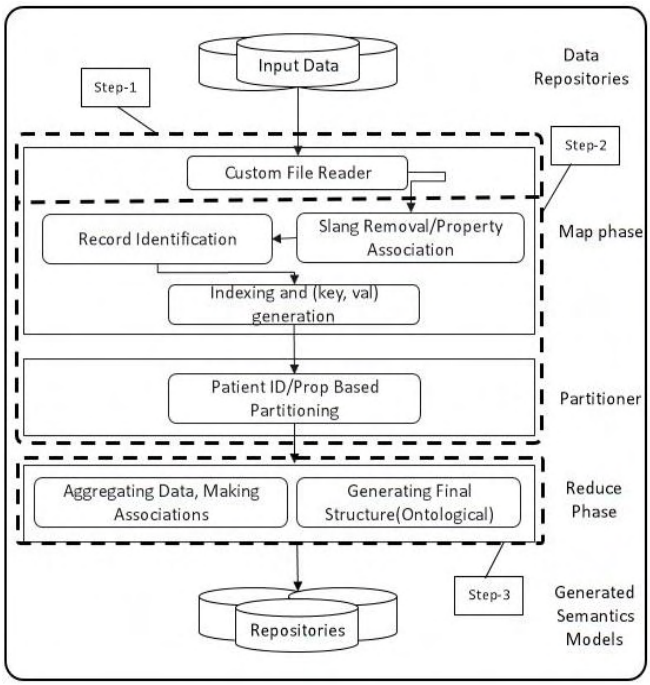
1. INTRODUCTION

Future internet is contributing ominously in various fields of science and technology [1]. In the recent few years, where the need for interoperable data and services has increased, medical systems envision enabling its stale and future data to the web for better patient health management through analytics and providing services. Hospital Information (HIM) systems face major problems in enabling their large stale and future data in their databases and repositories in various formats. The major challenge faced arises from difficulty in transforming data to semantics based structure such as ontological structure which can be easily shared and web-enabled in the current and future Internet of Things (IoT)[2]. This objective can be achieved by using a distributed approach of processing huge stale datasets to transform into a semantically connected structure. Hadoop MapReduce is a distributed processing model to process huge datasets on a clusters of commodity machines.

Hadoop MapReduce's [3] effectiveness can be observed from the data that exist in the existing systems in variety formats. MapReduce has the ability to process various formats of data at one place and generate a commonly agreed data based on a pre-defined model/structure. Conventional data processing systems i.e. sequential systems need single input format and preprocessing. However, in MapReduce this task can be easily performed. Existing systems process the data for specific purposes such as PMR/HER systems and are not suitable for generating a common ontological structure.

Semantics based structures are currently the main focus of the research community in IoT as future web is envisioned to be the web of services and things. Semantics web has the ability to express complex data structure and perform reasoning on it. In healthcare, HL7[4] provides standards which are currently used/followed by majority of hospitals internationally. HL7 provides standards such as Clinical Document Architecture (CDA) [5], a XML-based markup standard intended to specify structure and semantics of clinical documents. It has the characteristics of wholeness, persistence, readability, and context.

Existing works in the literature mainly achieve interoperability between the data using same standard by different systems. LinkEHR-Ed [6] by Maldonado is proposed to build, process, and validate using various reference models. Similarly, Web services-based Artemis architecture [7] by Dogac et al. works for mediation among healthcare systems. In summary, most existing systems target data interoperability. However, there is lack of a system that can efficiently work to transform existing BigData to the newly format and structures. The process artifacts and structure of our proposed distributed approach is shown in Figure 1.



2. PROPOSED SYSTEM WORKING MODEL

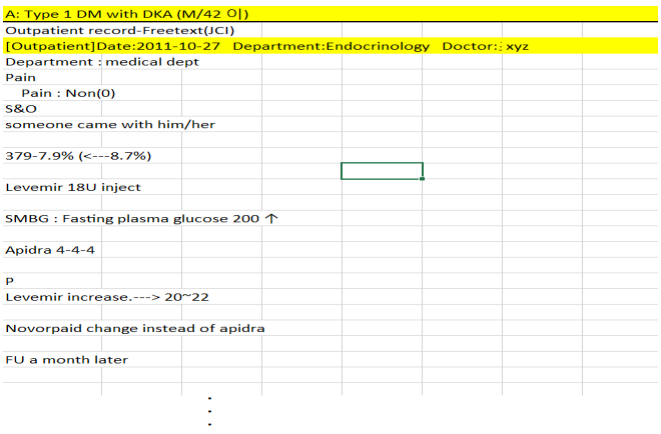
This section explains the working model of proposed approach. Development of HIMS and building such systems requires understanding of the existing patient medical records and standards by the HL7 community. Based on these constraints, medical systems can be made interoperable and made available to the web. As the proposed approach follows MapReduce based distributed model hence works in three steps. Steps include data-load, map, and reduce as shown in the Figure 1. In the following subsections, we explain each step separately.

2.1 Custom File Reader

Custom file reader handles managing variety input data formats by various systems such as excel format, CSV, XML, etc. Patient records generated in different countries exist in different formats and contents. This component deals to read each file format by recognizing the file format and parsing it to be processed by the MapReduce's mapper components. A basic medical record has he least possible fields which are required to be filled. In general case, a single file consists of single patient's record, therefore it works to read whole file as single input to the mapper.

2.2 Map Phase (Data Generation)

This phase of the proposed approach works to process the data in three sub-phases. Each phase works to conform the data to MapReduce <key, value> pairs. It also maintains relationship between the properties of various fields in the semantic structure. The basic work for transformation from conventional structure to a conformed structure is performed here. Each patient medical record (PMR) or electronic medical record (EMR) is parsed to extract various fields necessary for building the CDA document (XML-based). The basic structure of Excel-based PMR can be seen in the Figure 2.



Each student has a unique identifier (ID) which distinguishes it from other patients and as well as its records. Based on these various properties, the data is processed as described below.

a. Slang Removal; This phase works to remove any inconsistency/slang-words in the text and applies Parts-Of-Speech (POS) technique to process the data, and associate specific properties to it. The data parsed is managed in a structured format and passed to the next step to relate to a specific identifier.

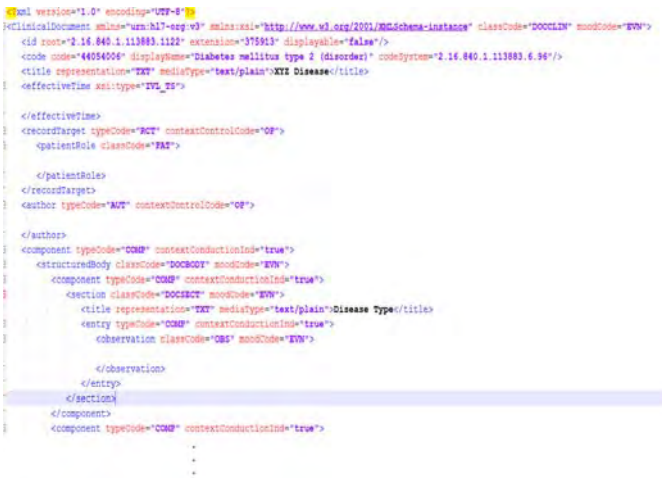
b. Record Identification: As each document represents a single patient's record, this component generates and creates a whole model of patient record based on the associated properties with specific identifications (ID) which are used in reducer for reduction. In case, a patient has many visit history or has visit many hospitals, their data can be connected using the patient identifier. Identifier can also further be integrated to investigate the user whole history background.

c. Indexing and <k,v> generation: This step works to generate (key, value) pairs as basic need of MapReduce programming model. The model of (key, value) pairs is described by the model and is used to transfer data to the reduce step. Record's identification is used as key while the whole record (properties and values) are passed as single serialized object i.e. value part of map step output.

d. Combiner: This component is not explicitly included in the architecture. However, in case of combining the user/patient data after the map out when processing data from multiple branches or hospitals would require a combiner optimization facility to mend the performance of the system.

2.3 Reduce Step

The reducer of the web-enabler works to collect and aggregate the Map step output based on keys (patient IDs) and generates XML-based RDFS structure. Ontology structure follows XML-based structure also called semantics based structure. This XML-liked structure generated in Reduce step follows CDA standard provided by HL7 as shown in Figure 3.



The methods to generate such CDA document is stored as in-memory rules. The output is a single document representing a single patient record as XML-based schema with values which can be easily shared, and provided as services. Associations in patients' medical records are also made including various observations, visits by a single patient. The final semantically connected data is called an ontology which is interoperable for HL7 standards, and can be used for wider purposes such as disease analytics, reasoning, and predictions.

3. DISCUSSION

Currently in hospitals, majority of them use the conventional approach of generating patient records, and processing them for their own purposes. Semantic web-based approaches i.e. HL7 standards such as clinical document architecture (CDA), SNOMED CT, MEDLINE is used by some of hospitals in limited countries. As the semantic web technology itself is evolving, therefore most of hospital systems currently have not shifted. However, it is envisioned by HL7 community and other stakeholders that data located in various countries, locations, Databases need to be conformed to an interoperable format so that it can be used for the better healthcare of patients.

The proposed approach not only transforms the existing data to a web-enabler format, but also new data could be directly linked or stored in ontological format. For new data, many systems exist that store data in ontological format, however, to connect that data to the existing, it needs some matching and mapping techniques. Therefore, any new system that is being built should conform to the new technology i.e. Hl7 CDA ontological format which can then be easily integrated with stale data and used for futuristic analytics and predictions. The use of distributed system for generating ontologies opens new way of introducing parallel processing for ontologies to boost their performance in data processing.

4. CONCLUSION

The proposed approach is based on BigData technology i.e. Hadoop MapReduce and semantic web technology, which play an important role in both aspects including performance of the system and making the data available for sharing and easy access. It is cost effective, scalable, and web-enabled with strong emphasis on the importance of IoT and transforming the existing traditional approaches to conform to the newly discovered approaches in the HIMS field. It emphasizes on using the medical records for better treatment based on its historical data, and performing effective data analytics to extract trends from patients' data. Trend analytics can become easy when we have data in a single agreed format as by HL7. Semantic structures are easily shareable, have expressiveness, and cost effective in processing. These benefits can lead us to better analytics on patients' data. The distributed solution of using Hadoop MapReduce is helpful in processing large data in distributed places in variety formats which cannot be processed by a conventional system easily and cost-effectively.

Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2014-(H0301-14-1003)

REFERENCES

[1] Semantic web health care and life sciences (hcls) interest group. http://www.w3.org/blog/hcls. (Last visited in May 2010)

[2] Kopetz, Hermann. "Internet of things." *Real-Time Systems*. Springer US, 2011. 307-323.

[3] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.

[4] HL7 Standards: http://www.hl7.org/implement/standards/

[5] Dolin, Robert H., et al. "HL7 clinical document architecture, release 2." *Journal of the American Medical Informatics Association* 13.1 (2006): 30-39.

[6] Maldonado JA, et al. LinkEHR-Ed: A multi-reference model archetype editor based on formal semantics. Int J Med Inform 2009; 78:559–570.

[7] Dogac A, et al. Artemis: Deploying semantically enriched Web services in the healthcare domain. Information Systems 2006; 31:321–339.