International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 07
March 2018

An HDFS and Elastic Search Index Approach For Implementing Real-Time Or Near Real-Time Persisting Daily HealthCare Data

Mr P.Jagadeeswara Rao[1], K.Sai Ravali[2], K.Tejaswi[3], V.Yamini[4], L.Vamsi Krishna Vardhan[5]

[1]Assistant Professor, DEPT OF CSE, Dhanekula Institute of Engineering and Technology, A.P., India.

[2,3,4,5] B.Tech (CSE), Dhanekula Institute of Engineering and Technology, A.P., India.

Abstract: *Mayo Clinic (MC) healthcare generates a large number of HL7 V2 messages-0.7-1.1 million on weekends and 1.7-2.2 million on business days at present. With multiple RDBMS-based systems, such a large volume of HL7 messages still cannot be real-time or near-real-time stored, analyzed, and retrieved for enterprise-level clinic and nonclinic usage. To determine if Big Data technology coupled with Elastic Search technology can satisfy MC daily healthcare needs for HL7 message processing, a BigData platform was developed to contain two identical Hadoop clusters (TDH1.3.2 version)-each containing an ElasticSearch cluster and instances of a storm topology-MayoTopology for processing HL7 messages on MC ESB queues into an ElasticSearch index and the HDFS. The implemented BigData platform can process 62 ± 4 million HL7 messages per day while the ElasticSearch index can provide ultrafast free-text searching at a speed level of 0.2-s per query on an index containing a dataset of 25 million HL7-derived-JSON-documents. The results suggest that the implemented BigData platform exceeds MC enterprise-level patient-care needs.*

KEY WORDS: health-care analytics, big data, natural language processing, learning health-care system

INTRODUCTION

Mayo Clinic (MC) is the world's largest integrated not-for-profit healthcare system with hospitals and clinics located on the three main campuses (Rochester, MN, USA; Jacksonville, FL, USA; and Scottsdale/Phoenix, AZ, USA) and in more than 70 locations in southern Minnesota, western Wisconsin, and northeast Iowa [5]. Each day, it provides care for hundreds of thousands (~400 000 in 1999 [13] and > 1 million since 2014) of patients of all walks of life from all 50 US states and nearly 150 countries. Each day, MC generates a large number of digital (electronic) healthcare data or electronic health records (EHR) by three EHR instances (Rochester, MCHS, and FL/AZ) and more than 500 departmental clinical systems, which mainly use relational databases (RDB) or RDB management systems (RDBMS, mainly MsSQL, DB2, Oracle, and Sybase at MC) for storing transactional structured, and semi- or unstructured data in RDB tables. Data among EHR instances and departmental clinical systems are not integrated at the transactional level. At the enterprise level, MC has built three major RDBMS-based platforms or systems for integrating and effectively using the transactional healthcare data as follows:

a)MICS-Synthesis platform, pulls data from the three EHR instances and the departmental clinical systems, and allows a user to query and view all the data for a patient;

b)Amalga platform implemented by Microsoft Amalga Intelligence System, also pulls data from the three EHR instances and the departmental clinical systems, and allows a user to query and view all the patients matching specific criteria and their individual medical records; and

c)enterprise data trust (EDT), a semantically integrated enterprise data warehouse (EDW) [3], pulls data not only from the three EHR instances and the departmental clinical systems but also from education, research, and administrative RDBMS-based transactional systems in support of MC's analytic and decision-making processes.

BACKGROUND AND RELATED WORK

Figure 1 illustrates the learning cycle in an LHS: practice, data, research, and knowledge. With the rapid adoption of EHRs, clinical practice generates large amounts of clinical data.[11]Researchers have been extensively utilizing EHR data for secondary purposes including clinical decision support, outcomes improvement, biomedical research, and epidemiologic monitoring of the nation's health. Knowledge discovered through research can then be utilized to improve patient care. The most significant initiative related to the LHS is The National Patient-Centered Clinical Research Network (PCORnet) formed in 2013 by Patient-Centered Outcomes Research Institute (PCORI), which consists of 11 clinical data research networks (CDRNs) and 18 patient-powered research networks. These organizations have made significant progress toward analyzing the data within these networks focusing on common conditions, rare conditions, and genetic disorders. There are nationwide networks other than PCORnet that also play positive role in facilitating LHS. For instance the collaboration between Kaiser Permanente and Strategic Partners, Patient Outcomes Research To Advance Learning,[12] Scalable Collaborative Infrastructure for a Learning Health-care System,[13] PaTH (University of Pittsburgh/UPMC, Penn State, College of Medicine, Temple University Hospital, and Johns Hopkins University), leading four academic health centers,[14] and PEDSnet, another consortium of eight children's hospitals, are initiatives involving multiple institutions.[15] Large initiatives such as the PCORnet provide an infrastructure for a national LHS.

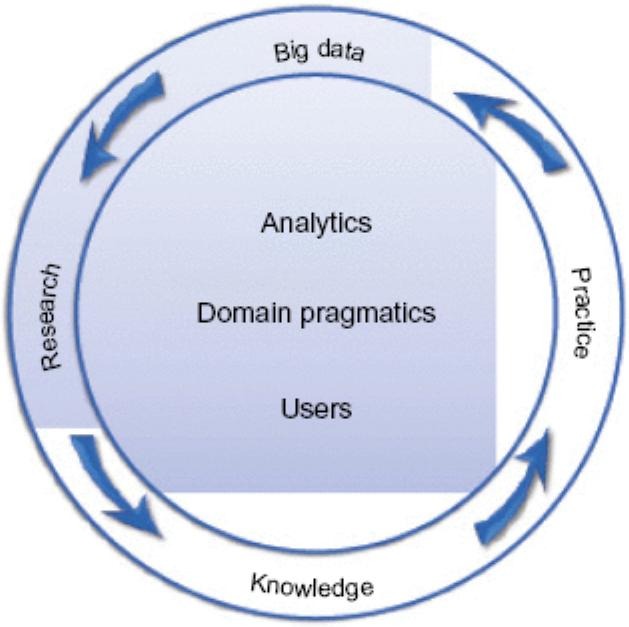


Figure 1. Learning cycle in an LHS. Analytics experts enable the cycle. Domain pragmatics provides the contextual information related to the domain, which is needed for discovering knowledge. Users are people who consume the knowledge.

NLP has been an integral component in the LHS, as evidenced by one of the review criteria in the recent CDRN phase II request for application, being the demonstration of NLP capability for phenotyping.[16]Figure 2 provides an overview of clinical NLP. At a high level, NLP generally consists of the following components: tokenization, syntactic parsing, semantic parsing, and pragmatic interpretation. It may also include upstream components such as speech recognition or optical character recognition or downstream components of data mining, text analytics, visualization, and summarization of the NLP results. In health care, a critical additional component is required – terminology mapping. This component takes the content that is clinically relevant and produces codes for unified semantic representations of clinical concepts. These codes are subsequently used in various applications such as billing, compliance, quality measurement, clinical decision support, and others.

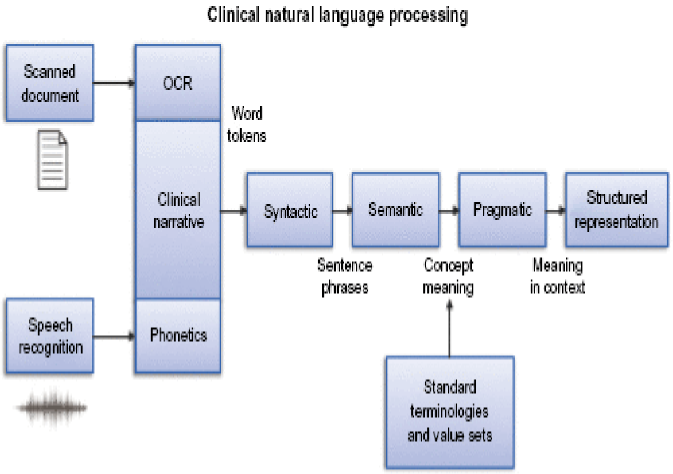


Figure 2. Generic clinical NLP process. Clinical NLP involves processing textual data obtained from clinical notes and voice dictated text. The process includes both syntactic and semantic processing. While syntactic components identify the grammatical structure of the text, the semantic components identify clinical concepts and its context such as experiencer, certainty, and negation.

Since the 1980s, NLP has been utilized to harness the information embedded in clinical narratives. One of the oldest and most studied clinical NLP systems is the Medical Language Extraction and Encoding System (MedLEE) developed by Friedman et al at the Columbia University in the mid-1990s.[17] MedLEE was initially developed on chest radiology reports[18] but further extended to work on any kind of clinical notes. An NIH-funded national center, Informatics for Integrating Biology, and the Bedside (i2b2), has organized both challenges and shared tasks focusing on problems less studied in clinical NLP and sharing annotated clinical notes that removed some of the barriers to the development of clinical NLP systems.[19—25]

However, one of the major bottlenecks in integrating NLP into clinical workflow has been the lack of computing infrastructure to implement real-time NLP solutions. With the recent advances in big data, it becomes apparent that the streaming and distributed computing capacity in the big data technology stack makes the implementation of NLP in the LHS possible.

One example of big data-empowered NLP solution is IBM Watson, a cognitive system developed by IBM Research Center with the capability of analyzing natural language content.[26] Watson incorporates multiple layers of NLP technologies including machine learning and a question answering system.[27—32] Recently, building upon the technologies behind Watson, IBM has invested in health-care analytics by improving clinical NLP capability. For example, Wang et al improved the performance of medical relation extraction in Watson.[33] IBM Watson is an independent analytical application that needs to be integrated into an EHR workflow for an effective use in clinical practice. In the recent past, EHRs do have some inbuilt NLP capabilities in their workflow. Cerner corporation has developed a sophisticated EHR that is not only safer and easier to use but smart enough to decipher the contextual meaning behind the descriptions in clinical text.[34] Chart Search, a search platform within Cerner has the capabilities to understand the intent of a query to perform semantic search.

Besides these big corporation initiatives, there are few efforts in academic institutions where big data-empowered NLP solutions have significantly advanced the clinical care by reducing the processing times of clinical data. Agerri et al (2015)[10] have demonstrated that the big data infrastructure can help scale NLP analytics to provide near real-time solutions.[35] On the other hand, Divita et al (2015)[9] explored an alternative approach for scaling NLP solutions through multithreading and running NLP modules concurrently. They took software engineering approach instead of assembling a robust hardware infrastructure for scaling the computing performance.

Essentially we have two models for performance scaling, as discussed above. One option is to have the right choice of robust hardware infrastructure, while the other is to engineer a robust software solution. At the Mayo Clinic, we took a middle path for scaling NLP applications, striking a fine balance between engineering a robust software solution and choosing a sophisticated big data infrastructure for deploying software. While big data-empowered

NLP offers the best hardware infrastructure, MedTagger is a suite of best-of-breed NLP modules developed based on rigorous software engineering models. We believe that this combination will help us realize the LHS as a possibility in the near future at Mayo Clinic.

Big Data Technologies Adopted

The big data implementation at Mayo has been designed for both analytical processing and new storage methodologies to facilitate faster retrieval. We chose Apache Hadoop as the big data platform, which includes components such as Apache Storm to provide real-time distributed computation environment, HBase for fast key-based data retrieval, and Elastic-search for efficient indexing and querying of information. In the following, we briefly describe these technologies.

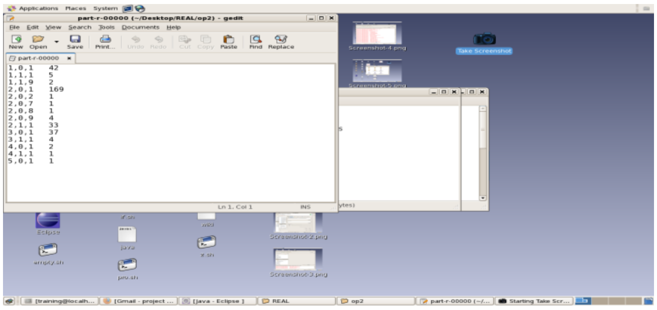
Apache Storm is a programming model agnostic stream-processing environment, which we used for streaming and scalable computing. Storm architecture consists of a cluster, where a master node distributes jobs to the slave nodes. The underlying structure of Storm is a graph topology, which consists of nodes that serve as the processing environment while the edges serve as the message broker communicating between the nodes. Nodes in the Storm topology essentially fall into two categories as follows: (i) Spouts to stream data from sources and (ii) Bolts to perform processing on data stream emitted by a Spout. A Bolt in turn emits a stream that can be utilized by other bolts. Apache Storm enables real-time analytics environment and data delivery through multiple processing streams of data effectively increasing the throughput.

Apache Hadoop is inherently designed for large-scale processing, predominantly in batch-processing mode across multiple, horizontally scaled server nodes built from commodity hardware.[44] Apache Hadoop allows the data to be processed faster and more efficiently than it would be in conventional supercomputer architecture. It relies on a parallel file system where computation and data are connected via high-speed networking. The Hadoop framework relies on map reduce formalism to deliver fault-tolerant scaling. In our big data implementation, we use MapReduce jobs to quickly search across clinical documentation to extract particular subsets of information that need to be processed through our Storm infrastructure. Currently, we are not using MapReduce jobs to scale the processing of any process in the big data implementation at the Mayo Clinic. We continue to investigate and develop more MapReduce and Spark capabilities in our infrastructure.

Apache HBase is an open-source distributed, nonrelational database modeled after Google's Big Table.[45]HBase does not support SQL as a query language, instead HBase provides a rich Java API. It is built on HDFS and hence can be deployed on commodity hardware. HBase is meant for a large amount of data in the range of billions of rows. An instance of HBase has a collection of tables. Each table contains rows with row-keys and arbitrary number of columns. These columns contain key–value pairs, which are versioned by timestamp by default. HBase was chosen to enable very fast key-based retrieval of documents stored in the big data environment.

Elasticsearch is a distributed full-text search engine that is built on Apache Lucene. Elasticsearch can handle large-scale real-time data to perform real-time analytics, which will enable the application layers to access the semantically enriched data in big data in near real time. Elastic search provides mechanisms to horizontally scale the retrieval by adding additional nodes for processing. It is fairly resilient that it can auto detect failure nodes and perform load balancing to ensure both data safety and accessibility. It also supports the notion of multitenancy where multiple indices can be housed on an instance of Elasticsearch.

Results:



Applications Places System
part-r-00000 (~/Desktop/REAL/op2) - gedit
File Edit View Search Tools Documents Help
New Open Save Print... Undo Redo Cut Copy Paste Find Replace
part-r-00000
Ln 1, Col 1
INS
Eclipse
empty.sh
java
pro.sh
wiki
x.sh
Screenshot-2.png
Screenshot-3.png

```
1,0,1    42
1,1,1    5
1,1,9    2
2,0,1    169
2,0,2    1
2,0,7    1
2,0,8    1
2,0,9    4
2,1,1    33
3,0,1    37
3,1,1    4
4,0,1    2
4,1,1    1
5,0,1    1
```

Fig: 1 Analysis Result

CONCLUSION

The Big Data work displayed in this paper, MC has deliberately chosen to construct a rationale and physical information stockpiling and handling engineering—UDP, which will utilize the present actualized Big Data stage as its inside or on the other hand center part while set up RDBMS-based frameworks and web applications—using the current RDBMS-based replication what's more, information distribution center condition—around the center so that the venture coordinated framework can give better information administrations to venture level clinical and nonclinical utilization at MC. The objective of the Big Data stage in MC UDP is to obtain and store endeavor information, and administration endeavor information for perception, investigation, furthermore, investigation that are required by big business level clinical utilization (determination, treatment, counteractive action, or clinical announcing) furthermore, nonclinical utilization (logical research, business insight, or on the other hand wellbeing data trade) at MC. What's more, the information continued in the stage can be any sorts of big business medicinal services information, regardless of whether they are organized, semi structured or on the other hand unstructured information.

REFERENCES

[1] Ross MK, Wei W, Ohno-Machado L. Big data and the electronic health record. Yearb Med Inform. 2014;9(1):97–104.

[2] McGlynn EA, Lieu TA, Durham ML, et al. Developing a data infrastructure for a learning health system: the PORTAL network. J Am Med Inform Assoc. 2014;21:596–601

[3] The Foundation for Continuous Improvement in Health and Health Care . Digital Infrastructure for the Learning Health System: Institute of Medicine. 2011.

[4] Fernandes L, O'Connor M, Weaver V. Big data, bigger outcomes: healthcare is embracing the big data movement, hoping to revolutionize HIM by distilling vast collection of data for specific analysis. JAHIMA. 2012;83(10):38–43. quiz 44.

[5] K. N. Eggleston et al., "The net value of health care for patients with type 2 diabetes, 1997 to 2005," Ann. Internal Med., vol. 151, no. 6, pp. 386–393, Sep. 2009.