



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Thesis for the Degree of Doctor of Philosophy

**SEMANTIC PRESERVATION OF STANDARDIZED
HEALTHCARE DOCUMENTS IN BIG DATA**

Shujaat Hussain

**Department of Computer Science and Engineering
Graduate School
Kyung Hee University
South Korea**

February 2020

SEMANTIC PRESERVATION OF STANDARDIZED HEALTHCARE DOCUMENTS IN BIG DATA

Shujaat Hussain

**Department of Computer Science and Engineering
Graduate School
Kyung Hee University
South Korea**

February 2020

SEMANTIC PRESERVATION OF STANDARDIZED HEALTHCARE DOCUMENTS IN BIG DATA

by

Shujaat Hussain Kausar

Supervised by

Prof. Sungyoung Lee

Submitted to the Department of Computer Science and Engineering and the
Faculty of Graduate School of Kyung Hee University in partial fulfillment of the
requirements of the degree of Doctor of Philosophy

Dissertation Committee:

Prof. Tae-Seong Kim _____

Prof. Sung-Ho Bae _____

Prof. Lok-Won Kim _____

Prof. Hyon Woo Seung _____

Prof. Sungyoung Lee _____

Achievement of each challenging goal is possible due to self-efforts and the guidance of elders, especially those who were close to our heart.

My humble efforts, I dedicated to my beloved

Father (may ALLAH rest his soul in peace) & *Mother,*

Brothers & Sisters,

whose endless love, encouragement, support, and prayers make me able to achieve such success and honor.

Along with my all respected, hardworking, and supportive

Teachers.

Abstract

Standardized healthcare documents have a high adoption rate in today's hospital setup. This brings several challenges as processing the documents on a large scale takes a toll on the infrastructure. The complexity of these documents compounds the issue of handling them which is why applying big data techniques is necessary. The nature of big data techniques can trigger accuracy/semantic loss in health documents when they are partitioned for processing. This semantic loss is critical with respect to clinical use as well as insurance, or medical education.

In this study a novel technique is proposed to avoid any semantic loss that happens during the conventional partitioning of healthcare documents in big data through a constraint model based on the conformance of clinical document standard and user based use cases. The study used clinical document architecture (CDA) datasets on Hadoop Distributed File System (HDFS) through uniquely configured setup and identified the affected documents with respect to semantic loss after partitioning and separated them into two sets: conflict free documents and conflicted documents. The resolution for conflicted documents was done based on different resolution strategies that were mapped according to CDA specification. The first part of the technique is focused in identifying the type of conflict in the blocks that arises after partitioning. The second part focuses on the resolution mapping of the conflicts based on the constraints applied depending on the validation and user scenario. The semantic aware standard model is created through constraint modeling for detection of the compromised documents. In the next stage a dual phase resolution strategy is used to prevent the semantic loss in the compromised documents.

This study used a publicly available dataset of CDA documents, identified all conflicted documents and resolved all the them successfully to avoid any semantic loss. In the experiments up to 87000 CDA documents were tested and successfully identified the conflicts and resolved the

semantic issues. The study has presented a novel study that focuses on the semantics of big data which did not compromise the performance and resolved the semantic issues risen during the processing of clinical documents. For the future work, more focus will be on the performance aspect of the technique and the applications for lossless data as the focus of current study was semantic preservice.



Acknowledgement

First and foremost, I render my humble and sincere thanks to the *Almighty Allah* for showering HIS blessings upon me. The Almighty gave me the strength, courage, and patience during my doctoral study.

Special thanks to my advisor Prof. Sungyoung Lee who provided me guidance, strength, support, and courage in overcoming the difficult challenges throughout my time as his student. I have learned a lot from him in becoming a productive person in diverse situations. He inspired me with his dynamic personality and his unreserved help and guidance lead me to finish my thesis. He has great role in polishing my skills such as thinking, creativity and technical soundness which are key ingredients for high quality research.

I am grateful to my dissertation evaluation committee for their insight comments and valuable suggestions during the dissertation defense. Their valuable comments improved the presentation and contents of this dissertation.

I am extremely grateful to some of my colleagues who have always provided me time, expertise, and encouragement in my course of research. They were always present to guide me in difficult situations of my PhD duration. I would like to thank Bilal, Wajahat, Maqbool, Afzal, Asif, Asad, Fahim, Bilal Amin, Taqdir, Jamil, Rahman Ali, Tae Ho, Kifayat, Idris, Fahad and Zeeshan. They have contributed enormously in successfully performing various academic and personal tasks that confronted me during my stay at South Korea.

I am very thankful to all of my current and former Ubiquitous computing lab fellows and colleagues for their kind support to my personal and academic life at Kyung Hee University. I am highly obliged to brilliant researchers and fellows- Mrs. Seoungae Kim, Aamir, Jae Hun Bang, Abdul Qadir, Khalil, Usman, Imran, Musarrat, Sadiq, Omar Farooq, Mahmood, Saeed Ullah,

Waqas, Ahsan Kazmi, Waseem, Junaid, Ahsan Raza, Saad, Shoaib, Aunas, Saba, Piran, Dildar, Aftab, Jawad, Maqbool Ali, Max, Anees Talha, Ammad, Ahmed, Asad, Fakhar, Latif, Zain Abbas Shehreyar, Jalal, Amjad Adnan, Qasim, Kamran and Faraz. This journey would not have been possible without their support. They contributed a lot in to my personal and academic life to polish myself. Also, I appreciate all my Korean and international friends who worked as a team with me and developed my team work skills and provided me wonderful memories during my stay in South Korea.

Last but not the least, I would like to express my sincere gratitude to my mother, sisters, and brothers for their endless love, support, prayers, and encouragement. Their support and encouragement has made this dissertation possible. Finally, I would like to leave the remaining space in memory of my father for giving me a privileged life where I was encouraged in all of my pursuits and inspiration to follow my dreams.



Shujaat Hussain

Feb, 2020

Table of Contents

Abstract	i
Table of Contents	v
List of Figures	viii
List of Tables	xi
Chapter 1 Introduction	1
1.1 Background	1
1.2 Motivation	5
1.3 Problem Statement	6
1.4 Contributions	9
1.4.1 Semantic Aware Standard Model	9
1.4.2 Two Phase Resolution Strategy	10
1.4.3 Accelerated Similarity Computations	10
1.5 Thesis Organization	10
Chapter 2 Related Work	12
2.1 Preliminaries	12
2.1.1 Big Data framework and concepts	12
2.1.2 Standardized Clinical Documents	14
2.1.2.1 Clinical Document Architecture	15

2.1.2.2	Continuity of Care Record (CCR) and Continuity of Care Document (CCD)	17
2.1.2.3	FHIR [®] (Fast Healthcare Interoperability Resources)	17
2.2	Background and Related Work	19
2.2.1	Hadoop Based Health Frameworks	19
2.2.2	General Health Frameworks	21
2.3	Summary of Related Work	29
Chapter 3	Proposed Methodology	30
3.1	Uniqueness	30
3.2	Abstract view of proposed methodology	31
3.3	Semantic Aware Standard Model	34
3.3.1	Constraint Modeling	34
3.3.1.1	Performance Constraints	35
3.3.1.2	User constraints	35
3.3.2	Conflict identification phase	37
3.4	Dual Phase Resolution strategy	40
3.4.1	First pass	41
3.4.2	Second pass	43
3.5	Accelerated Similarity Computation	48
3.6	Case study for Lossless data	54
3.6.1	Disease Based Analytics	55
Chapter 4	Simulation Results and Evaluation	58
4.1	Experimental evaluation	58
4.1.1	Experimental Setup	58
4.1.2	Dataset description	58
4.1.3	CDA [®] preservation conflicts	59
4.1.3.1	Entry, section and whole documents	59
4.1.3.2	Clinical Statements and EntryRelationship	60

4.1.3.3	Conflicts against size and constraints	61
4.1.4	Resolution results	63
4.2	Discussion	66
4.2.1	Conflict Marking in Compromised Documents	66
4.2.2	Number of compromised documents for resolution	66
4.2.3	Health Interoperability	66
4.2.4	Big data file systems	67
4.2.5	Conventional behavior of Hadoop	67
4.2.6	Additional passes in resolution phase	67
4.2.7	Limitations of this work	68
Chapter 5	Conclusion and Future Directions	69
5.1	Conclusion	69
5.2	Future Directions	70
5.2.1	Future work: Health Imaging Data	70
5.2.2	Future work: Precision Medicine	71
Chapter A	Healthcare standards and Big Data framework	72
A.1	Healthcare standards	73
A.1.1	FHIR	76
A.1.2	OPENEHR	77
A.1.3	HL7 CDA	78
A.2	Big Data Frameworks	83
A.2.1	Spark	84
A.2.2	Hadoop (extensions) YARN	85
A.3	Relationship between Healthcare standards and Big Data framework	86
Bibliography		86
A List of Publications		99

List of Figures

1.1	Split documents blocks	4
1.2	Data Pyramid	5
1.3	Key Standards, Categories and application	6
1.4	Existing System Work-flow	7
1.5	Proposed Solution Workflow	9
2.1	Hadoop Basic architecture	13
2.2	Complete Flow of MapReduce.	14
2.3	CDA [®] Document Sample.	15
2.4	CDA [®] Structure and Example	16
2.5	CCR to CCD Conversion.	17
2.6	Patient Resource Example.	18
2.7	Big Data Classification [1]	20
2.8	Medoop Workflow [2]	21
2.9	Semantic ETL framework [3]	22
2.10	Framework of the Clinical Diagnosis and Treatment System [4]	23
2.11	Big Data Architecture for personalized medicine [5]	23
2.12	Big Data Analytics in Healthcare [6]	24
2.13	Personalized Big Data Analysis Framework [7]	26
2.14	North America IoT in Healthcare Market Growth, by Component, 2012–2022(USD Billion) [8]	27
2.15	Architectural elements of healthcare IoT systems [8].	28

2.16 Big data analytics-enabled transformation model [9]	28
3.1 Class Diagram for abstract View of the Document	31
3.2 Layers of proposed methodology	32
3.3 Complete flow of the conflict identification and resolution strategy phases	33
3.4 Complete flow of the conflict identification and resolution strategy phases	34
3.5 Constraint Hierarchy	35
3.6 Clinical Statement Types	37
3.7 Conflict Identification	39
3.8 Conflict Resolution	41
3.9 Second Pass Workflow	44
3.10 UMLS Integrating subdomains	45
3.11 Unified Medical Language System	47
3.12 Unified Medical Language System Work Flow	48
3.13 Creation of complex data objects	49
3.14 Similar Data Computation Workflow	51
3.15 Calculate Attribute Probability	53
3.16 Stakeholders of the lossless data	54
3.17 Abstract Workflow of proposed methodology for Applications	55
3.18 Allergies Code	56
3.19 medications Code	56
3.20 General case study of analytics and visualization	57
4.1 CDA Documents Compromised	59
4.2 CDA Sections Compromised	60
4.3 CDA Entries Compromised	60
4.4 Entryrelationships Compromised	61
4.5 clinical statements Compromised	61
4.6 Conflicts against constraints	62
4.7 Conflicts against size.	62

A.1 Healthcare Data Sources	73
A.2 Example of a DeviceObservationReport (JSON-format) [10]	76
A.3 openEHR specification components [11]	77
A.4 CDA Header	79
A.5 CDA Body	80
A.6 CDA Entries	82
A.7 Major Components of Clinical Document Architecture [12]	83
A.8 Spark System Overview [13]	84
A.9 YARN architecture [14]	86



List of Tables

1.1	Most common healthcare Sources (adapted from [15])	3
2.1	CDA Discharge Summary Sections and Details	16
2.2	Related Work Comparison	29
3.1	User Constraint Table	36
3.2	Records having ML-Care attributes	53
4.1	Conflicts remaining in First pass on different datasets	63
4.2	First Pass of Resolution	64
4.3	Second Pass of Resolution	65
A.1	Most common standards in healthcare Adapted from [16]	76

1.1 Background

Clinical data includes electronic health records, patient demographics and images, prescriptions, discharge summaries, insurance information and data from sensory devices. Big clinical repositories containing clinical and biological data are increasingly becoming available for research and enhanced analytics [17].

There is an opportunity to transform and associate the clinical data that could detect patterns and support in advanced health analytics. This can lead to cost-effective solutions and impact patient life positively due to a better understanding of the data. There are long term potential benefits as many complex scenarios can be better understood like surgeries and patient readmission. Moreover, 1.2 billion clinical documents are generated in the U.S each year and about 60 percent of them contain patient data in an unstructured or semi-structured format [18]. Better decisions and quicker processing can be done from the health analytics extracted from the transformed clinical documents [1].

In the past few years, health and wellness applications have emerged as a fast growing category of mobile applications. This increasing trend is considered as a prompt and useful resource for collecting users' data which are used for generating recommendations for a healthy lifestyle. Using smart phone features, applications like Microsoft Health, Apple Healthkit, Samsung S Health, and Google Fit collect users data by monitoring their daily activities, e.g., eating habits, sleeping patterns, and workout routines to generate certain recommendations which are helpful in maintaining a healthy lifestyle. The adaptation rate of such applications is on the rise with downloads in millions [19–23]. Healthcare data is mostly in semi-structured or unstructured form. The additional thing about the data is the complexity, dynamicity and heterogeneous nature of the data [24–26]

makes it challenging to extract useful information using conventional data processing tools & techniques [27]. Without effective decision support systems, it is difficult to comprehend or process the data [28].

This creates the demands for Big Data Analytics into healthcare. Big Data analytics empowers us to generate valuable insights from complex data which would have been very difficult to get. When applied to the healthcare data, patterns can be identified and thus lead to cost reduction, improved healthcare quality and enable decision-making in a timely fashion [25, 27, 29, 30]. A McKinsey Global Institute report [31] mentions that by utilizing Big Data effectively, reduction of healthcare expenditure can create a value of more than \$300 billion yearly, . Hidden knowledge can be uncovered using automation and better decisions using Big Data frameworks and technology [32].

The healthcare outcomes using Big Data still fall short [33–36] and there is still a lot to be done to realizing the potential of Big Data analytics. Another major issue is lack of user intervention in processing information using Big Data analytics in Healthcare and this can lead to erroneous results [37, 38].

Studies show that the main technical issues in Big Data analytics include partitioned and fragmented data [15, 29, 39], which limits the observational data [30, 40], validation, standardization issues [29, 40–43], inconsistencies [15, 41, 44, 45], reliability, and semantic standard interoperability [46–48].

The amount of data generated by smart phones and supportive need to include data from other resources make data volume enormous and its structure more complex. Although smart phones are sufficiently equipped with large memory size and computational resources for on-device storage and processing ability, however, to achieve increased battery life, data backup, centralized data storage, and to fulfill data-cross-sharing, there is another approach gaining momentum in a majority of applications which is the adoption of cloud services. The Big Data analytics has the ability to identify diseases allows swift and accurate diagnosis and evaluation of therapies [33, 49–51]. Data linkage from different sources and identifying patterns, the prediction through Big Data analytics can also be used for transformation of real-time data into valuable insights. This is of real importance where quick decision making and emergency medical situations arise as it can mean

the difference between life and death [29]. In the healthcare industry, stakeholders like physicians, healthcare decision makers and insurance companies generate data in diverse healthcare standards which are mostly structured and semi-structured. One of the challenges in healthcare is data processing due to the complex schema and structure of the standards [52]. Nowadays data for all patients is recorded from admission to discharge, like previous history, imaging, reports, tests, doctor recommendations and constant monitoring of patients through different devices [53] as shown in Table 1.1. The data deluge is a norm in healthcare, however extracting analytics from the data is still a big challenge. When using multiple data sources patient privacy is of utmost importance in healthcare as well. Data sharing between stakeholders helps in deriving insights but can also cause a lot of concern for privacy [9, 27, 36, 38, 40, 54] but privacy is not in the scope of this study.

Type	Description	Source
Clinical	Electronic Medical Records (EMRs)	Hospitals and Clinics
	Diagnostic	Laboratories
	Biomarkers	Diagnostic Companies
	Ancillary	Hospitals & Clinics
Calims	Medical Claims	Payers
	Prescription Claims	
Clinical Research	Clinical Trails	Pharma Companies
Patient-generated Data	Social Media	Web Health Portals and Social Media Websites
	Wearable & Sensors	Device Data Systems

Table 1.1: Most common healthcare Sources (adapted from [15])

The exponential growth in the amount of data generated during the last few years have greatly changed ideas about the value, management, and expertise of such data [55]. As of 2012, about 2.5 Exabyte data are created each day, and this doubles about every three years. The current amount

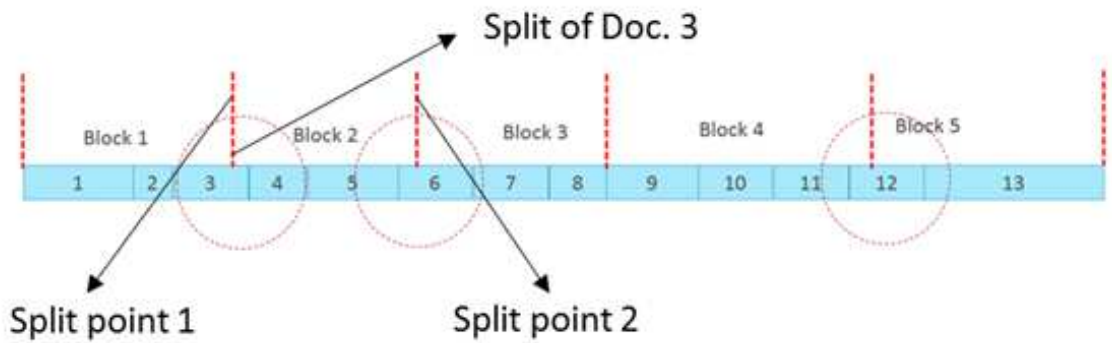


Figure 1.1: Split documents blocks

of data generated each second is more than all of the collective data from the previous 20 years. In 2011, the human digital universe contained 1.7 Zbytes, and this dataset is expected to increase by almost five times by 2015 (7.9 Zbytes) [56]. Similarly, the speed of data generation is more important than its volume. Real-time or nearly real-time data streaming [57, 58] makes system more agile. Big Data sources generate data in various formats such as images, audio GPS signals, text, sensory data, and huge amount of healthcare data through monitoring devices and mobile applications [59]. Healthcare in big data refers to the clinical and administrative datasets and they are complex, huge and cannot be handled with traditional infrastructure. Big Data in healthcare has the potential to change the way data is processed and reduce the operational structure of IT in healthcare. Big data is changing all aspects of modern day life specially in healthcare [60]. As much as 30% of the entire stored data is related to health and a normal patient will generate 80 MB of electronic medical data every year [61]. The complex nature of data sets usually results in inefficient and difficult management of conventional databases. Big data has four key characteristics i.e. volume, variety, veracity and velocity [62]. It means that big data can play a key role not only where size matters but it also helps in handling the complex nature of the data sets.

As seen in Figure 1.1, one block is represented as a health document and dotted lines as partitioning markers. Document 3 is chopped during the partitioning which results in a semantic loss. The case is same with the document 6 and 12. In Figure 1.1, three out of thirteen documents have semantic loss due to partition which is usually based on size and results in the documents being split on two different places for processing. These documents are almost impossible to

process for extracting any meaningful information as they have complex sections, relationships, and data types conforming to the semantics of standard specification. Cutting them in different places can result in total loss of the data in that document.

1.2 Motivation

The main motivation of this study is semantic preservation of all the standardized health/clinical documents due to their critical nature. Some documents are separated into two parts in a different part of the cluster due to the partitioning nature of the big data frameworks as seen in Figure 1.1. Better semantics can only be achieved from the complete data which is then translated in analytics and later enabling in better decision making.

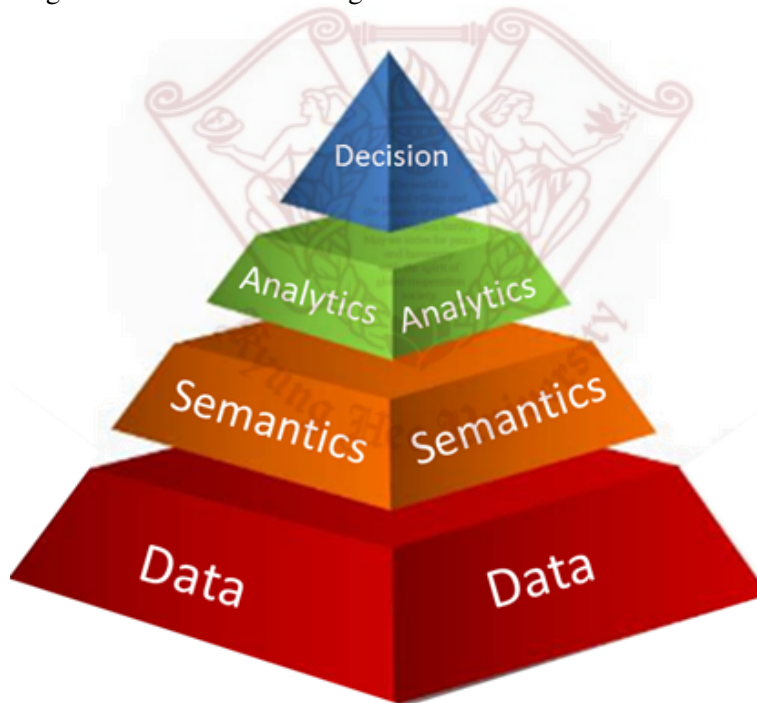


Figure 1.2: Data Pyramid

As shown in Figure 1.2, if the data is compromised during partition, all the subsequent actions to be done from that data are also compromised. An additional challenge is to get adaptable to standardized schema like (Open EHR, HL7 CDA) so that complete document knowledge is on our disposal. Data continuity is important in standardized health documents and in this study

HL7 CDA standard is used due to its high adoption rate in hospitals and other medical related institutes [63].

1.3 Problem Statement

Semantic loss occurs in standardized and complex health care documents during partitioning in big data framework and results in compromised decision making ability [3,5]. The Key standards, categories and application are shown in figure 1.3.

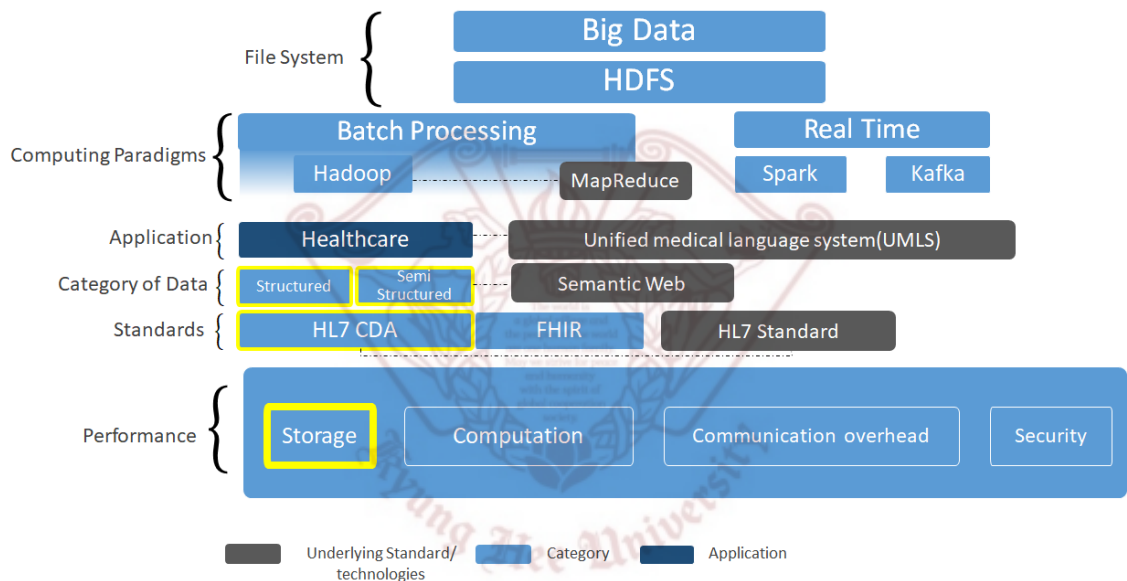


Figure 1.3: Key Standards, Categories and application

The taxonomy highlights the focus on structured and semi structured data. There are images, unstructured doctor notes and but many other kinds of health and wellness data which also adds to the complexity but is not used in this study. The two most common standards are CDA and FHIR. FHIR is a fast interoperable standard which is becoming popular in developers very fast due to its ease of use. CDA is one of the most common EHR guidelines being used in the hospitals.

One of the problems that occur with big medical records and repository is the way big data frameworks partition the data for processing. The big data frameworks depend on divide and conquer strategy for huge datasets in partitioning. This strategy ends in an inconsistent division

of the records and results in a semantic loss. The generic work-flow to depict existing systems is presented in Figure 1.4.

In the existing systems, partitioner splits the dataset in different blocks with respect to a pre-defined size and default input record formatter. Input record formatter usually do a line by line parsing but it could be overridden to user preferences. This triggers semantic loss in documents due to same document being dispersed in geographically different places to be processed. This data is transformed according to the user application ignoring the lost documents in process and passing it for analytics and decision making.

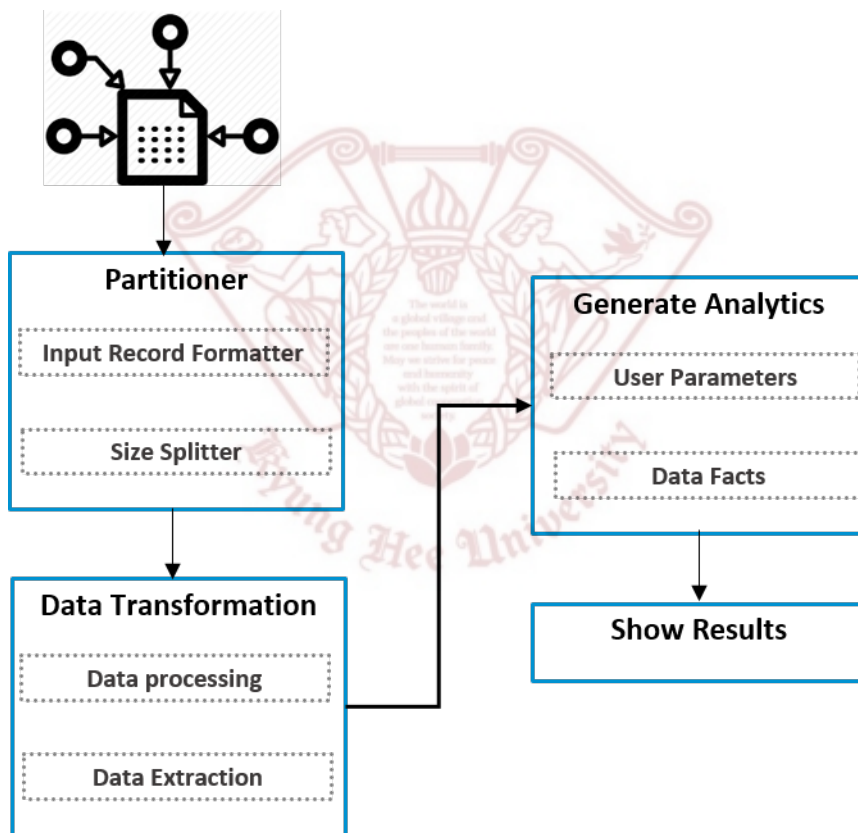


Figure 1.4: Existing System Work-flow

Semantic loss occurs when a part of the health document is chopped during the data partitioning and disseminated to two different nodes of the big data cluster for processing. Semantic preservation is needed to fully understand the meaning of a chopped document. The goal is semantic preservation in big data framework for standardized healthcare documents and ensure all

the semantics from big data is extracted. One solution is to specifically check that the document is not cut in between and resulting in 100% accuracy in preserving the semantics. The issue with this approach is that it results in massive performance bottlenecks in preprocessing as the conventional partitioning does not take the data continuity into consideration. The conventional partitioning divides the data into small predefined sizes/blocks and the primary goal is division of data by size which is processable by commodity nodes. No regard for data format and semantic loss is considered and due to this high performance is achieved at the cost of semantic loss. This approach is a major hindrance in performance perspective/semantic loss when the health documents are processed. So this study proposed a configurable and optimized document partitioning approach for health documents based on the solution in which the data blocks are verified based on constraint model. The additional overhead of semantic preservation is also addressed to make the solution more compatible with respect to performance.

The proposed solution workflow is shown in Figure 1.5. The proposed workflow shows three additional models highlighted for semantic preservation of the documents being used in the Big Data frameworks. In this study, constraint modelling creates the constraint from schema of health standard which are classified as user constraints. The performance constraints are dependent on the type of big data framework being used for data processing. The constraints are applied on the documents after partitioning in the big data framework to identify the compromised documents. After detection of compromised documents, the resolution of documents is done for semantic preservation. After complete semantic preservation, data processing is done and then only passing it for analytics and decision making.

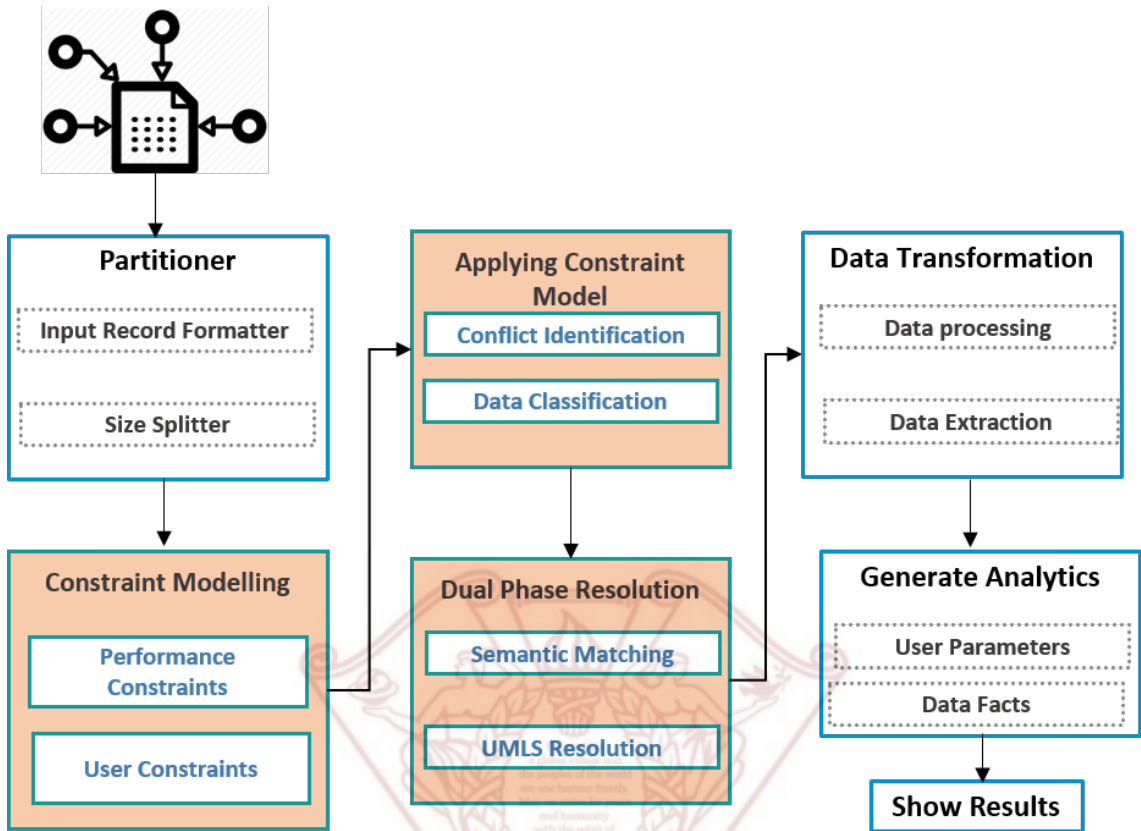


Figure 1.5: Proposed Solution Workflow

1.4 Contributions

We summarize the main contributions of this thesis as below:

1.4.1 Semantic Aware Standard Model

The semantic aware standard model constructs a constraint model which covers up the CDA[®] standard conformance and also user derived criteria. The constraint model finds correlation in the data based on the schema. There are two major constraints types. Individual constraints are those which are simple and not dependent on other constraints or vice versa. They do not have any subtypes. Conjunctive constraints are nested and contain subtypes as well as multiple individual constraints which increases the complexity. The constraints assist in detection of conflicted

documents in terms of semantic loss.

1.4.2 Two Phase Resolution Strategy

After applying Semantic aware standard model the output is conflicted documents and the different conflict markers identified are number of constraints violated, individual and conjunctive constraints. Based on the conflict markers, a resolution strategy is needed to make the document complete so that no conflicts remain. A Dual Phase resolution strategy was introduced in this study. In the first phase, the duplicate documents occur when two or more documents are conflicted with similar markers i.e. same constraints. This results in some documents being orphan which means their other half matched with someone else. The second phase was introduced and semantic concept type matching is used for resolving orphan and duplicate documents. For semantic concept matching, Unified Medical Language System (UMLS) is used. The motivation for using UMLS is the large biomedical concepts from over 100 source vocabularies. A methodology for autonomous resolution of compromised document based on their conflicts and semantic concepts based on the classified documents

1.4.3 Accelerated Similarity Computations

Semantic preservation operations takes additional time which adds to the overall time and makes it a high time complexity solution. To make semantic preservation a feasible solution, more performance gains are needed. Data complexity among intermediate partitioned data can make it a very time intensive process. Naive record and data matching gives quadratic complexity. So if the data has one million records, there will be 1 trillion record matches. So a sub linear time complexity was introduced through min-max hashing. Similar record identification for faster processing is proposed and weighted attributes in the dataset are set for hashing in domain dependent scenario. It was fast and accurate estimation using locality sensitive hashing and using weighted attributes

1.5 Thesis Organization

This dissertation is organized into chapters as following.

- **Chapter 1: Introduction.** Chapter 1 provides brief introduction of the research work on semantic preservation for standardized healthcare documents and in particular the HL7 clinical document architecture (CDA). It focuses on the problems in the areas, the goals to achieve these problems, and finally the objectives achieved in this research work.
- **Chapter 2: Related Work.** A background detail is provided in this chapter about the Big Data partition and discuss different matching techniques and approaches, for achieving semantic preservation. This chapter also provides the state-of-the-art literature for the multiple data sources. Finally, it provides comparison of these systems with the proposed system of the research thesis to reflect the limitations of current systems addressed by the proposed system.
- **Chapter 3 Proposed Methodology.** A proposed solution in the form of a framework for achieving semantic preservation is presented in this chapter to overcome the limitations of current approaches. This chapter also provides overview of the concepts used in the thesis related to the proposed approach. It defines the scope of thesis in achieving the semantic preservation through conflict identification and conflict resolution. A constraint model is proposed in the first phase and the way it will assist in conflict identification. The second part of this chapter focus on conflict resolution through two passes.
- **Chapter 4: Case study for lossless data.** Different healthcare standards and the big data framework survey and details are provided in this chapter and the healthcare standards and Big Data framework can be interlinked to benefit the overall field.
- **Chapter 5: Results and Evaluation.** The results and evaluation of different techniques used in the proposed framework are highlighted in this chapter. It explains two types of results and evaluation. Firstly, it describes the results of compromised documents with respect to size and constraints. Secondly, the compromised documents are shown with respect to the resolution passes and the preservation happening in the two steps.
- **Chapter 6: Conclusion and Future Directions.** This chapter concludes the thesis and also provides future directions in this research area. The main contribution of the thesis is also highlighted in this chapter.

2.1 Preliminaries

Below we discuss some preliminary terms and concepts that need to be understood before understanding the general concept of MapReduce [64] and HDFS (Hadoop Distributed File System) [65]. We are using CDA[®] documents as the case study for clinical documents as CDA[®] standard is very comprehensive and semantically rich.

2.1.1 Big Data framework and concepts

Hadoop has the best fault-tolerant, high-throughput, and server-failure survival mechanisms [66]. Like Google File System, Hadoop also maintain replicas of its data splits across different machines to provide data locality and reliability. Default chunk size of HDFS is 64 MB, and these chunks are once write-multiple-read chunks. Hadoop's master-slave mechanism is shown in Figure 2.1. It consists of NameNode and DataNodes. A NameNode also called MasterNode, is responsible for controlling the whole MapReduce job through a JobTracker and controlling tasks in a job through a Tasktracker while working as DataNode (e.g., single node cluster). A DataNode, also called SlaveNode, consists of DataNode and Tasktracker. A SecondaryNameNode in large clusters is used to generate snaps of NameNode to avoid loss and works as standby NameNode. NameNode stores meta-data about the files/data chunks stored in DataNodes, while DataNodes store the actual data chunks (64 MB). By default, each data chunk is replicated by a factor of 3. *MapReduce*: The MapReduce framework is a platform which enables parallelism seamlessly. The input is partitioned, key/value pairs from each portion of the input is calculated and grouped by key, and reduced as shown in figure 2.2. MapReduce also has inherent features like network performance,

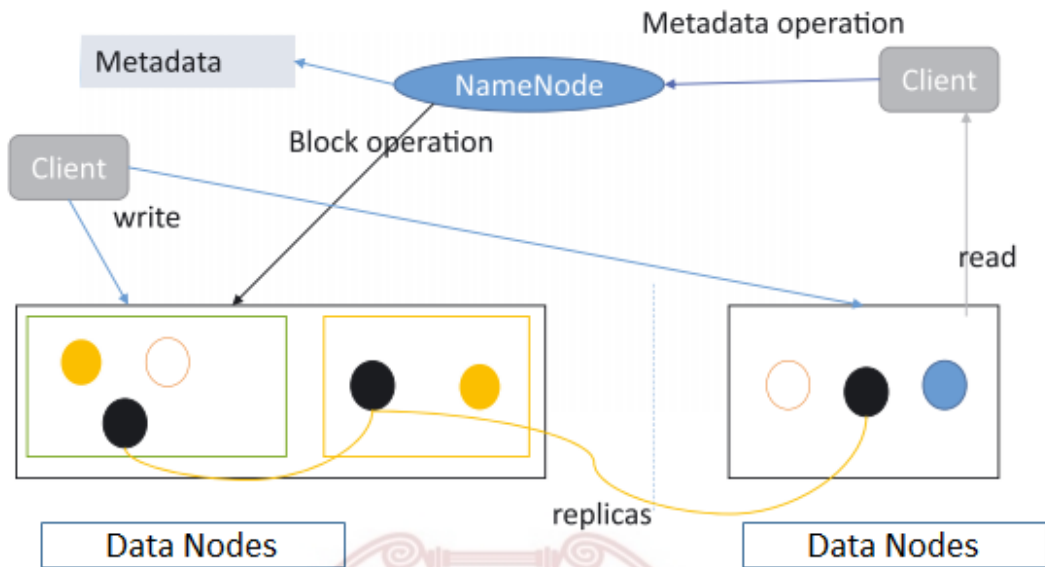


Figure 2.1: Hadoop Basic architecture

load balancing and fault tolerance. The Apache Hadoop [67] is the widely used open source application of MapReduce for distributed programming in Java.

Map: In MapReduce, the map function is basically used to transform the input which is being processed in parallel on the Hadoop cluster.

Intermediate Data: The output of the map function is called an intermediate data. It is in key value format and accumulated after map operation and communication between the data nodes. It is temporary and serves as an input to the reduce function.

Reduce: Reduce function shapes the intermediate data as a single entity usually as summarization step. The reduce phase summarizes the intermediate data based on the unique keys.

Example: Word count is a common example in MapReduce and it is intended to find the number of occurrences of each word in the input files. In mapper phase, the input is given line by line and the line is tokenized into words and a key value pair is formed where the key is word and value is 1. In the reducer phase, the keys are grouped and values for the same keys are added to find the number of occurrences [68].

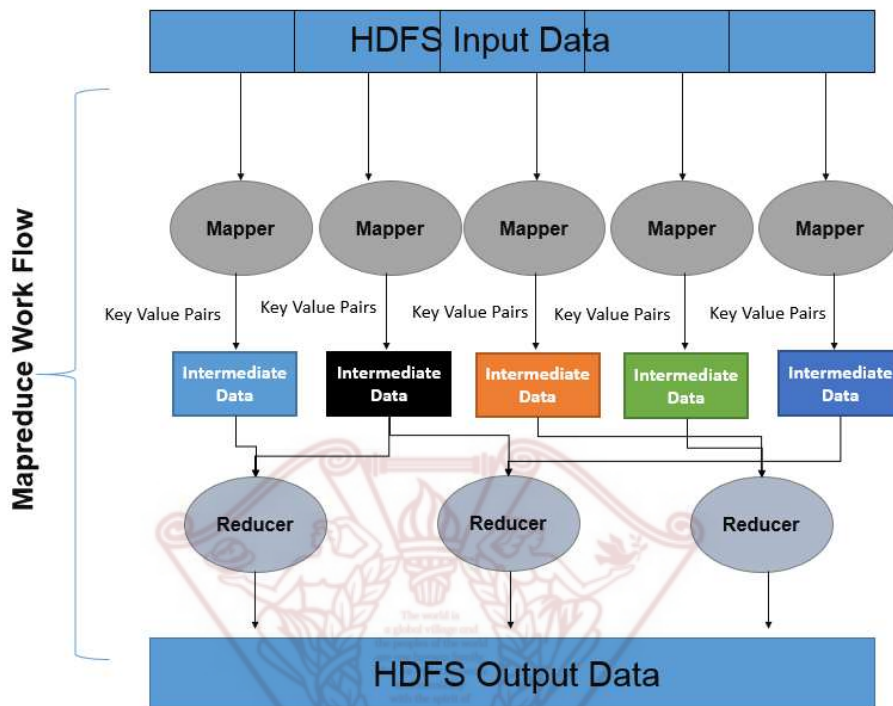
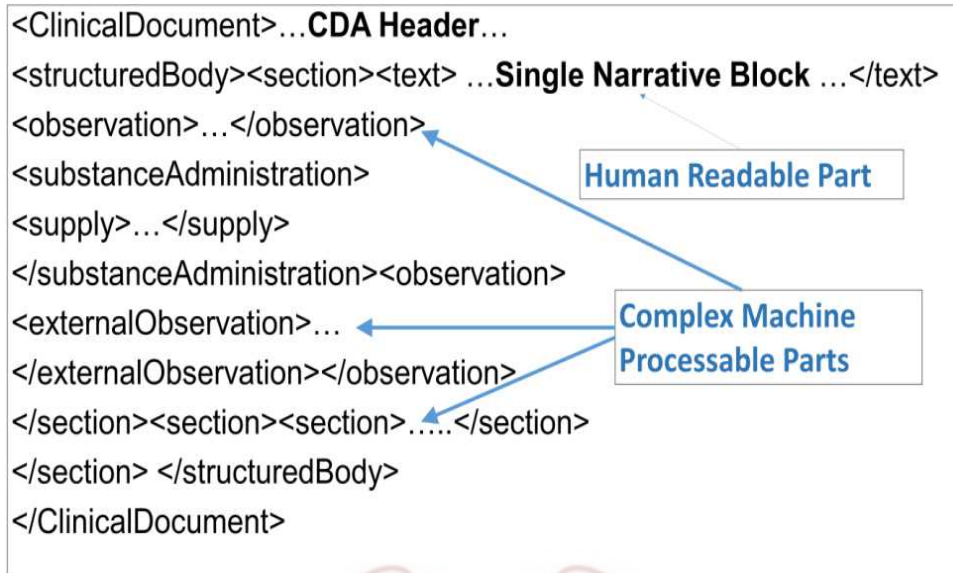


Figure 2.2: Complete Flow of MapReduce.

2.1.2 Standardized Clinical Documents

Over the recent past, certain standards have been constructed equipped with modalities using digital technology (Ultrasound, CT Scan, MR etc.) and peripheral devices like printers. Imaging devices and printers are dealt with a standard Digital Imaging and Communications in Medicine (DICOM) [69]. In a hospital-based scenario, for management of non-imaging data, there are multiple standards for medical documents like openEHR, HL7[®] CDA[®] and FHIR[®] are used. They provide protocols for messaging, management and integration of clinical documents. Our focus in this study is the semantic preservation of clinical documents and enabling in better health interoperability.

Figure 2.3: CDA[®] Document Sample.

2.1.2.1 Clinical Document Architecture

Clinical Document Architecture (CDA[®]) [12] is a standard which provides a concrete basis for common architecture, coding, semantic understanding, and markup language for clinical documents. CDA[®] documents are coded in XML, which is composed of a header for identification of the patient, encounter details, care provider and the other part is the main body which consists of some clinical observations in compulsory and optional sections. For rendering of the CDA[®] document, a human-readable part is included to derive narrative of the document as shown in figure 2.3. The structured part depends on coding systems like LOINC [70] and SNOMED [71] to represent concepts. Most importantly, CDA[®] structure is derived from standard reference model called HL7[®] RIM (Reference Information Model) [72] as it is the root of all the information models and provides clinical or administrative context and expresses how different pieces data are connected [73].

There are many parts in the CDA[®] document body which is composed of clinical statements such as procedures, acts, observations and their respective codes. Clinical statements have complex data types like code with equivalents (CE), coded value (CV) and concept descriptor (CD). A sample section is shown in figure 2.4 where different entries, clinical statements, and entryRela-

tionships are a challenge to parse or query in a meaningful time. The codes in figure 2.4 are taken from medical coding classification systems and meaningful use is a term used to define minimum requirements from the U.S. government for exchange of clinical patient data between healthcare providers and other stakeholders like insurers and patients [74]. The details can contain diagnosis, plan of care, family history, allergies, vital signs etc. A more detailed view as shown in Table 2.1.

Section	Contains
Allergies, Adverse, Alerts	Substance, Reaction, Status
Hospital Discharge Medications	Medication and Instructions
Plan of Care	Planned Activity and Planned Date
Family history	Parents Diagnosis and Age At Onset
Functional Status	Functional Condition, Effective Dates, Condition
Immunizations	Vaccine, Date, Status
Procedures	Procedure and Date
Problems	Condition, Effective Dates and Condition Status
Vital Signs	date/Time, Height, Weight, BP etc
Social History	Social History Element, Description and Dates

Table 2.1: CDA Discharge Summary Sections and Details



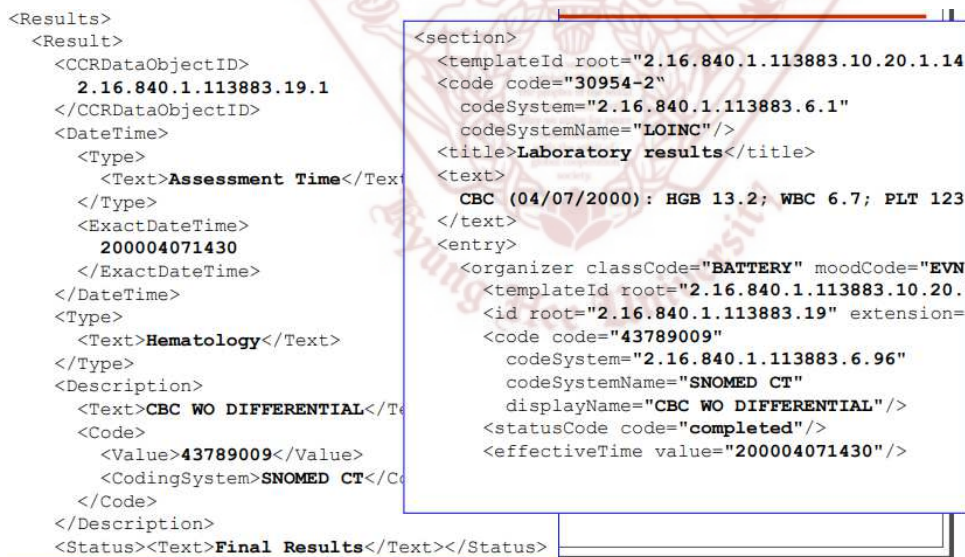
Figure 2.4: CDA[®] Structure and Example

2.1.2.2 Continuity of Care Record (CCR) and Continuity of Care Document (CCD)

The primary use case for the CCR is to provide a summary of the patients' health status i.e. problems, medications etc [75].

Continuity of Care Document (CCD) goal is to apply CCR content to the CDA[®] framework and provide the necessary clinical information for the continuation of care and assist in reducing medical errors [76]. CCD uses a detailed set of constraints for CDA[®] elements and the templates define how to use CDA[®] elements to communicate the clinical data, but the scope of the data within the templates is determined by CCR dataset. In some sense, the constraints applied in CCD is the subset of constraints in CDA[®].

In figure 2.5, a basic conversion is shown from CCR to CCD which makes it a CDA[®] compliant document and tightly constrained compared to C-CDA[®] document. This is due to additional constraints applied on CCD document.



```

<Results>
  <Result>
    <CCRDataObjectID>
      2.16.840.1.113883.19.1
    </CCRDataObjectID>
    <DateTime>
      <Type>
        <Text>Assessment Time</Text>
      </Type>
      <ExactDateTime>
        200004071430
      </ExactDateTime>
    </DateTime>
    <Type>
      <Text>Hematology</Text>
    </Type>
    <Description>
      <Text>CBC WO DIFFERENTIAL</Text>
      <Code>
        <Value>43789009</Value>
        <CodingSystem>SNOMED CT</CodingSystem>
      </Code>
    </Description>
    <Status><Text>Final Results</Text></Status>
  </Result>
</Results>

```

```

<section>
  <templateId root="2.16.840.1.113883.10.20.1.14"
    <code code="30954-2"
      codeSystem="2.16.840.1.113883.6.1"
      codeSystemName="LOINC"/>
    <title>Laboratory results</title>
    <text>
      CBC (04/07/2000): HGB 13.2; WBC 6.7; PLT 123
    </text>
    <entry>
      <organizer classCode="BATTERY" moodCode="EVN"
        <templateId root="2.16.840.1.113883.10.20.1.14"
          <id root="2.16.840.1.113883.19" extension="1"
            <code code="43789009"
              codeSystem="2.16.840.1.113883.6.96"
              codeSystemName="SNOMED CT"
              displayName="CBC WO DIFFERENTIAL"/>
            <statusCode code="completed"/>
            <effectiveTime value="200004071430"/>
          </id>
        </organizer>
      </entry>
    </section>

```

Figure 2.5: CCR to CCD Conversion.

2.1.2.3 FHIR[®] (Fast Healthcare Interoperability Resources)

FHIR[®] is the latest specification based on emerging industry trends, approaches and lessons learned through defining and implementing HL7[®] v2, HL7[®] v3 and the RIM, and CDA[®] [77].

FHIR[®] provides granular access to data, provides a streamlined approach to interoperabil-

ity [78] and focuses on implementation whereas CDA[®] addresses interoperability for clinical documents, mixing narrative and structured data. CDA does not provide granular data access and added additional challenges to the implementers due to its complexity [79]. The basic building block in FHIR[®] document is a resource [80] and it represents the clinical data and is a well defined and meaningful expression. An example of a patient resource [80] can be found in figure 2.6.

```

<Patient xmlns="http://hl7.org/fhir">
  <id value="glossy"/>
  <meta>
    <lastUpdated value="2014-11-13T11:41:00+11:00"/>
  </meta>
  <text>
    <status value="generated"/>
    <div xmlns="http://www.w3.org/1999/xhtml">
      <p>Henry Levin the 7th</p>
      <p>MRN: 123456. Male, 24-Sept 1932</p>
    </div>
  </text>
  <extension url="http://example.org/StructureDefinition/trials">
    <valueCode value="renal"/>
  </extension>
  <identifier>
    <use value="usual"/>
    <type>
      <coding>
        <system value="http://hl7.org/fhir/v2/0203"/>
        <code value="MR"/>
      </coding>
    </type>
    <system value="http://www.goodhealth.org/identifiers/mrn"/>
    <value value="123456"/>
  </identifier>
  <active value="true"/>
  <name>
    <family value="Levin"/>
    <given value="Henry"/>
    <suffix value="The 7th"/>
  </name>
  <gender value="male"/>
  <birthDate value="1932-09-24"/>
  <careProvider>
    <reference value="Organization/2"/>
    <display value="Good Health Clinic"/>
  </careProvider>
</Patient>

```

Figure 2.6: Patient Resource Example.

Resources have an extensive range i.e. from clinical content such as care plans and patient

medical reports as well as Message Header and capability statements [80]. FHIR[®] resources definitions do not churn out semantically consistent data as it serves many distinct contexts in addition to clinical data such as workflows, guidelines, health reporting, wearable devices etc. In this study, the focus is on clinical data which has concrete concepts such as MedicationPrescription, AdverseReaction, Procedure and Condition.

A single CDA[®] document can be decomposed into various FHIR[®] resources and can be made it much simple to comprehend. FHIR[®] is potentially the future of CDA[®], after a certain level of maturity [78].

2.2 Background and Related Work

During this study, I found out research involving both big data and healthcare but most of the studies are focused on getting an inherent benefit on big data frameworks like scalability and replication. Our work is based on loss of semantics in clinical data in big data and to preserve semantics without compromising significant performance. This is a unique approach as most of the work I found is based on efficient retrieval in big data based on high level APIs.

2.2.1 Hadoop Based Health Frameworks

Hadoop [67] is used for discharge summaries by Horiguchi et al [81]. The data set consists of log files and was artificially generated and conformed to a schema and grouped through different parameters like date, quantity and drugs.

In one study [82], Big Data Analysis using Balanced Partition technique which gives better performance with help of PIG and generate a histogram for the respective partition. This Histogram is generating according to the specified column by user in interface. They claimed that Hadoop is providing the best performance for big data on respective data set, it has uncontrolled chunks which are balanced in this project with the help of a balanced partition algorithm. They control the uncontrolled chunks of Hadoop Framework using a balance partition algorithm as they discovered that Hadoop's simple partitioning method does not preserve correlation between data chunks. So they devised partitioning Framework partitioning to help for balancing data chunks

into respective partitions.

The semantic transformation model is proposed in [1] as shown in Figure 2.7 .The storage layer is using Hadoop Distributed File System (HDFS) for storing the data. This layer has a semantic partitioner to extract the complete healthcare document.The query formulator has two parts i.e. the extractor and the builder. This query is based on the health analytics scenario and it maps the concepts of this query to HDFS repository

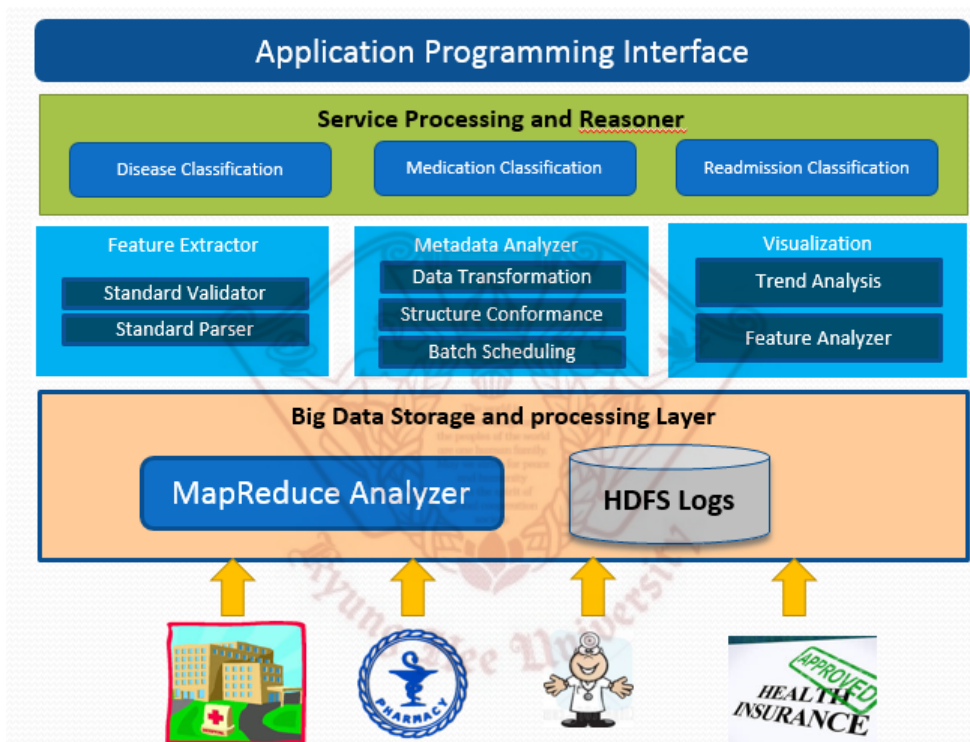


Figure 2.7: Big Data Classification [1]

A cloud based approach is used for data aggregation management [83] and the big data framework used is Hadoop. For querying and storing relations, the authors have used Hive [84] and HBase [85]. Hybrid XML database and HBase framework are used to handle heart disease clinical data analysis online [86]. They use a mix of conventional DBMS and big data which is challenging due to different data access web services and techniques. Medoop [2] is a medical platform which is developed for supporting a centralized health information exchange (HIE) in China because the size of such data can grow massively. Medoop uses Hadoop and HBase as its

underlining framework. Medoop merges all files to a large one and creates an indexing file containing the information of all merged files. Medoop stores both the indexing file and the merged file on HDFS as shown in Figure 2.8. Any frequently used data, however, is stored separately on HBase.

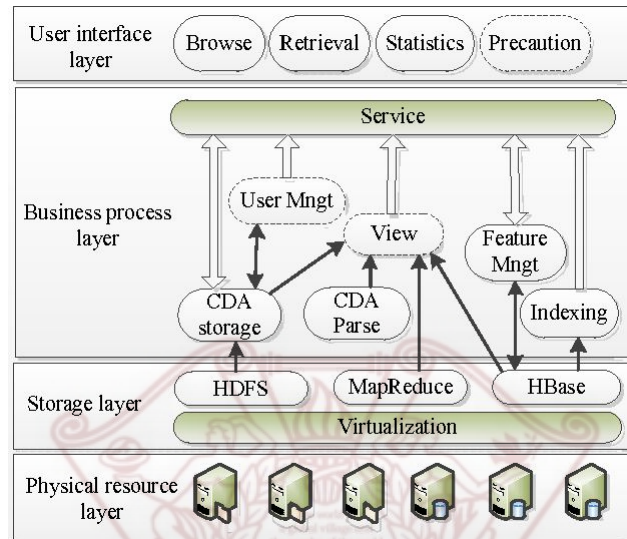


Figure 2.8: Medoop Workflow [2]

2.2.2 General Health Frameworks

The mobile applications are used for purposes of monitoring healthcare is used for minimizing costs of traditional health care treatment. Monitoring patients and accessing medical records easily at all times is a clear advantage. The concept of the Health cloud [87], is a prototype which utilizes the public Amazon cloud to manage patient records and relevant medical images. The Project has developed an android application for viewing JPEG2000 standard images with image annotation exploiting the multi-touch functions of the Android OS. The mobile device is now an essential part of the distributed architecture [88] and analysis of sensory data to determine human activities are done using MapReduce and many studies are now using big data technology for extracting context out of sensory data.

The Hospital data is used to build a hospital-specific Predixion model [89]. The prediction model is then used to risk stratify patients upon admission. Risk scores are updated throughout the

patient's stay. Readmission risk scores are used by care givers to target appropriate patient care paths.

Big Data scientists are dealing with the Variety of data that includes various formats such as structured, numeric, unstructured text data, image, video, and audio. The authors in [3] proposed Semantic Extract-Transform-Load (ETL) framework that uses semantic technologies to integrate and publish data from multiple sources as open linked data provides an extensible solution for effective data integration, facilitating the creation of smart urban apps for smarter living. The semantic ETL framework is shown in Figure 2.9

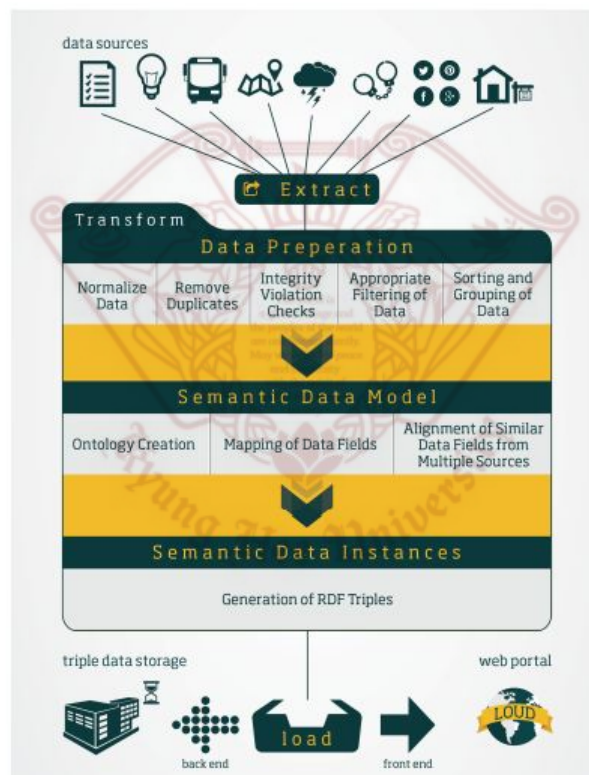


Figure 2.9: Semantic ETL framework [3]

The authors in [4] have proposed a framework of the Clinical Diagnosis and Treatment System shown in Figure 2.10. In CDTS, a new clinical tabular document model is provided as a standard for clinical document representation. The critical component from the perspective of the doctor or hospital is a semantic inference mechanism that consists of two stages: knowledge extraction and reasoning.

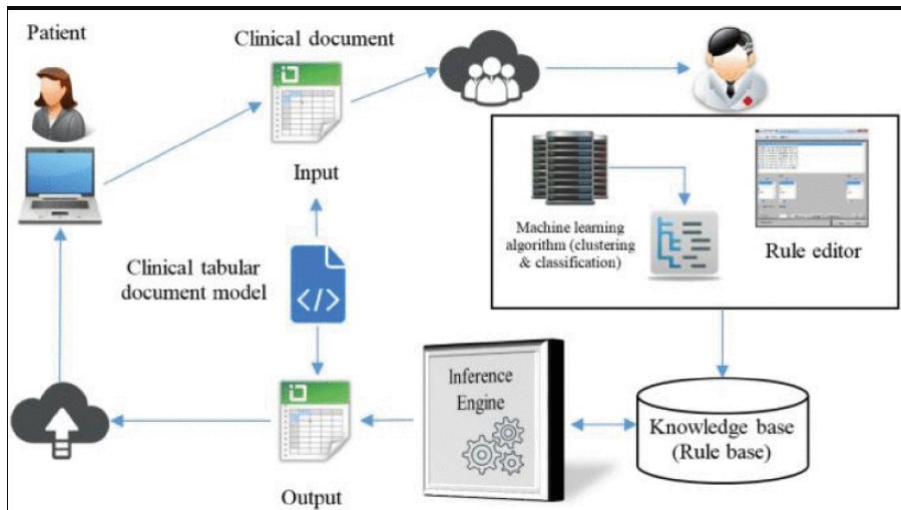


Figure 2.10: Framework of the Clinical Diagnosis and Treatment System [4]

The authors in [5] focus on the need for a semantic data-driven environment to address the big data issue. They discuss a path for empowering personalized medicine using big data and semantic web technology and proposed a framework shown in Figure 2.11. Based on the architecture, they stored datasets from different resources including EHRs, Genomics, and Medical Imaging into the Hortonwork repository and then use scripting tools like Pig and Hive to clean and prepare our data. One of their main application is data retrieval.

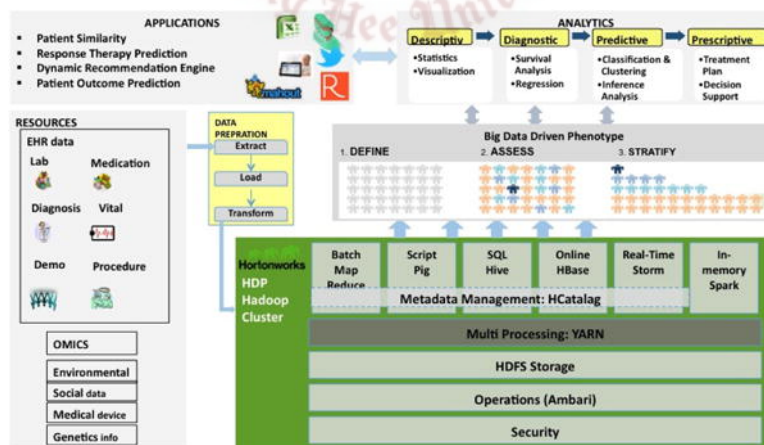


Figure 2.11: Big Data Architecture for personalized medicine [5]

In this paper [6], dataset alike Electronic Medical Records (EMR) produced from numerous

medical devices and mobile applications are induced into MongoDB using Hadoop framework with Improved processing technique to improve outcome of processing patient records as shown in Figure 2.12.

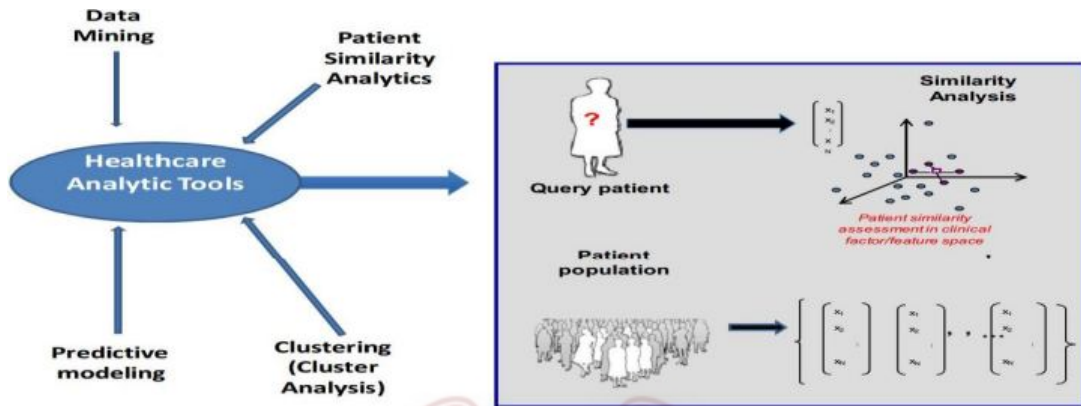


Figure 2.12: Big Data Analytics in Healthcare [6]

The review here in [90] explains the role of healthcare and medical informatics and the limitations in the current research. The complexity and heterogeneity is highlighted in data since the literature contains different kinds of sources of information on Big Data definition, Big Data Analytics techniques and their application and challenges in healthcare. The technical details related to implementation and results are not taken into consideration in this study.

The healthcare data also includes wellness data which is collected through different health monitoring devices and smartphones. The smartphones nowadays have built-in sensors which are highly effective for activity and context identification. In the research work presented in [91], a smartphone was used to identify walking and running [92] whereas GPS data was used to recognize transport and commuting.

‘Nike + iPod’ [93] initiated by Nike, logs user well-being activities such as running, jogging, and gym activities via the Nike+ hardware device paired with Apple’s iPhone or iPod. Activity data is subsequently published over Nike’s portal [94], which provides data visualization services and data persistence services. ‘MapMyRun’ [95], is a similar application that keeps track of user’s workout activities and nutrition intake with intuitive visualizations and track mapping services. Activity data gathered from a smartphone can be stored over the MapMyRun Portal [95] or

exported as log files to be sync with cloud storage services like dropbox [96].

Maintaining log files for activities is turning out to be very important as it contains vital information about our wellbeing. These logs can vary from our daily life activities to our workout and exercise activities. Most of the available life-logging applications are focused on wellbeing and workout tracking. Life-log data recorded by the smartphones provide improved activity tracking by utilizing the built-in sensors and GPS capabilities of the phone. A novel feature selection algorithm is used for accelerometer classification [97] and it utilizes multimodal sensor data from accelerometer, audio, GPS, and Wi-Fi. Another approach based on this technique takes the context information and prompts the user for an activity label [98]. This label and the sensory data is saved and stored in the cloud. Another smartphone based hierarchical approach is used for activity modeling and real time activity recognition [99].

These applications utilize cloud and web for the persistence of activity data. This data is used as the basis for improved visualization over the web and smartphone, and can also be used for expert analysis such as physicians and trainers. Cloud computing has introduced a new revolution in the development of the internet. The rapid rise of cloud computing and mobile computing has started a new computing paradigm that is mobile cloud computing. Mobile cloud computing has, however, a set of challenges once integrated into a mobile application with a cloud service. There have been many elastic models for mobile applications as the mobile application is launched inside the mobile device, however, later the processing or data is migrated to the cloud. Research in mobile cloud computing has ranges from topics considering energy saving, data management to migration, social networks, and healthcare. The potential of applying mobile cloud computing and Big Data for purposes of monitoring healthcare has the potential of minimizing costs of traditional health care treatment. Monitoring patients and accessing medical records easily at all times is a clear advantage. In addition, taking action with some intelligent emergency management system when the patient has been identified as being in distress is a further advantage. The concept of the Health cloud [87], is a prototype which utilizes the public Amazon cloud to manage patient records and relevant medical images. The Project has developed an android application for viewing JPEG2000 standard images with image annotation exploiting the multi-touch functions of the Android OS. The mobile device is now an essential part of the distributed architecture [88] and

analysis of sensory data to determine human activities are done using MapReduce and many studies are now using big data technology for extracting context out of sensory data. In [7], a big data analysis model is proposed which updates the knowledge base and gives users a personalized recommendation based on the analysis of the data and is shown in Figure 2.13. They have designed a personalized adaptive analysis technique for data handling and transformation and responds to information utilization APIs in a real time manner.

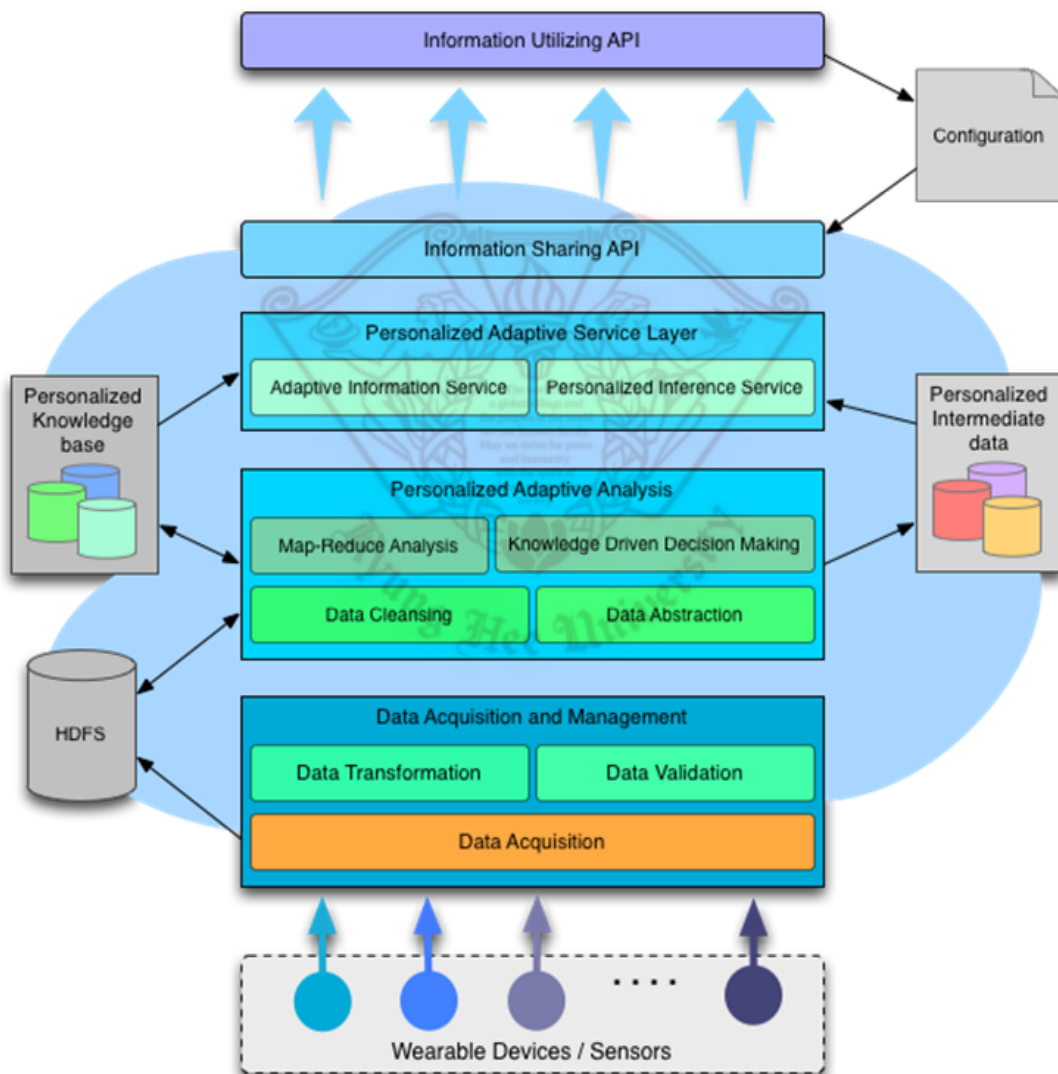


Figure 2.13: Personalized Big Data Analysis Framework [7]

Internet of Things (IoT) and Big Data Analytics are increasingly gaining popularity for the next generation of eHealth and mHealth services [8]. IoT in healthcare covering the markets of medical devices, systems, software, and services is expected to grow to a market size of 300B dollars by 2022 according to the market analyst, Grand View Research, as shown in Figure 2.14.

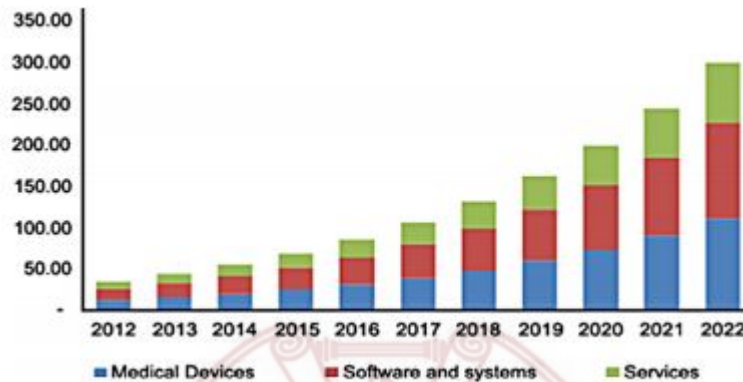


Figure 2.14: North America IoT in Healthcare Market Growth, by Component, 2012–2022(USD Billion) [8]

Figure 2.15 outlines the general architectural elements required for healthcare IoT systems (Health-IoT) , which includes three main components: (i) body area sensor network, (ii) Internet connected smart gateways, also known as Fog layer [8], or a local access network, and (iii) cloud and big data support. Various applications provide services to different stakeholders in the system through this platform.



Figure 2.15: Architectural elements of healthcare IoT systems [8].

A big data analytics-enabled transformation model [9] based on practice-based view is developed, which reveals the causal relationships among big data analytics capabilities, IT-enabled transformation practices, benefit dimensions, and business values. This model was tested in healthcare setting and path-to-value chains were identified for healthcare organizations to provide additional insights.

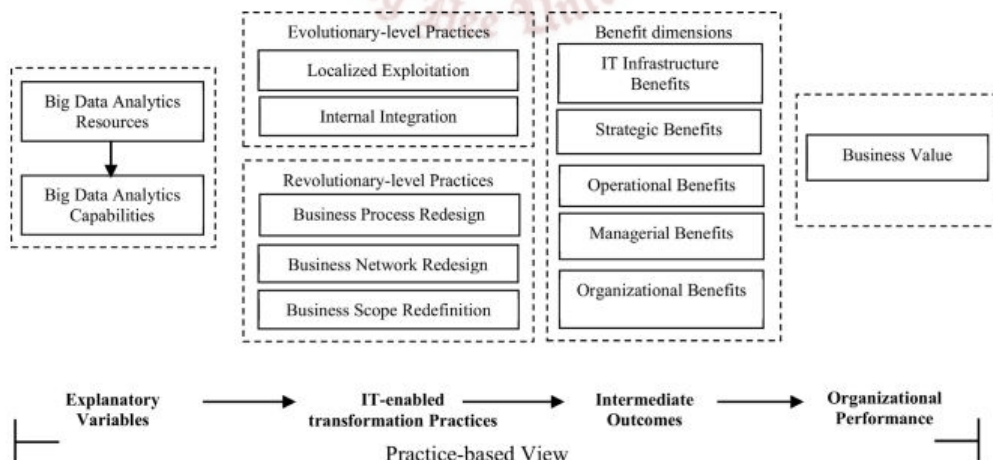


Figure 2.16: Big data analytics-enabled transformation model [9]

2.3 Summary of Related Work

Most of the studies have used big data frameworks for storing and querying but have not focused on semantic loss during the big data partition. They extract partial or no semantics from any bad or damaged documents in the framework. Our motivation is to remove this semantic loss before the processing has started to give better and complete results. The Table 2.2 shows a benchmark which includes the time complexity of the record matching, the type of big data framework, variety of data, standard data and schema based processing.

Reference	Data Handling	Multiple Data Sources	Standard Validation	Data preprocessing (schema)	Partition	Data Distribution	Time Complexity
Silvestri et al (2019)	Spark	No	Yes	Dependent	Line by Line	Default	Quadratic
Nunziato et al (2016)	Spark	Yes	No	Dependent	Not Applied	Default	Quadratic
HL7-CDA clinical documents management (2013)	MongoDB	No	Yes	Dependent	Not Applied	Default	NA
Ko et al. (2014)	Pseudo Distributed	No	No	Not Applied	Line by Line	NA	NA
Hiromasa et al. (2012)	Hadoop	Yes	No	Dependent	Line by Line	Default	Quadratic
Medoop (2013)	Hadoop	No	No	Dependent	Schema Based	Default	Quadratic

Table 2.2: Related Work Comparison

Hadoop and its ecosystem is still popular and used for processing of healthcare data [47, 100–104] , but for real time systems newer tools like Spark, Storm and Kafka are also used nowadays [47].

In the background section, this study has discussed different types of standardized clinical documents which vary from tightly constrained (CCD) to loosely constrained document (FHIR[®]). FHIR[®] is mentioned as a loosely constrained relative to CDA[®] as multiple FHIR[®] resources can be created from one CDA[®] document. The schema is flexible in nature and inherits constraints of the resource (CDA[®] document).

C-CDA[®] documents are selected for this study as it has strong constraints / structure and the constraint model devised will inherently also satisfy based on CDA[®] schema.

3.1 Uniqueness

Generally there is no semantic correlation between data due to partition paradigm of big data frameworks, so to tackle that a semantic aware standard model was built. This model identifies the constraints in the schema which are most influential in keeping the semantics of document. This will assist in finding data correlation based on the schema. The conflict identification of compromised data is based on the document standard schema and needs to be addressed so the semantic aware standard model is applied in the big data framework to identify compromised documents based on the conflicts. After identification, conflict resolution of compromised documents needs to be done. This study introduced dual-phase resolution strategy to tackle compromised documents and resolve conflicts to ensure semantic preservation.

3.2 Abstract view of proposed methodology

The semantic transformation is needed so that the CDA document can be used for processing. For this purpose a model is needed which can maintain the basic relationships but remove the complexity of sections, complex data types and relationships. A generalized object model is needed to create a simple and understandable reduced document. The class diagram in Figure 3.1 shows the reduced document which contains the patient id and encounter information. It also contain the list of reduced sections that the health analytic query may need. The reduced section contains the section code, title and collection of clinical statements. The clinical statement is an atomic object which contains the code and value (description). The remaining classes' i.e. driver, XMLInputFormat [105], mapper and reducer are supporting the distributed environment of the system and is explained in resolution strategy.

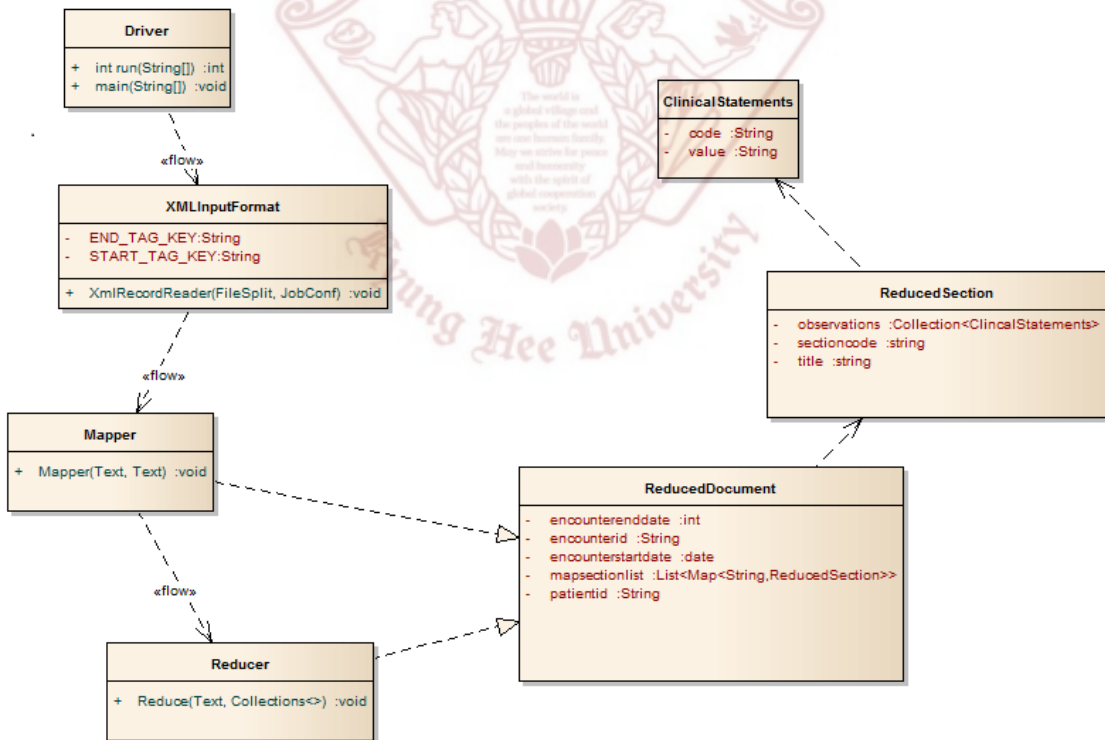


Figure 3.1: Class Diagram for abstract View of the Document

The proposed methodology consists of four main phases i.e. the constraint modeling, con-

conflict identification, the resolution strategy and the accelerated similarity computation as shown in Figure 3.2.

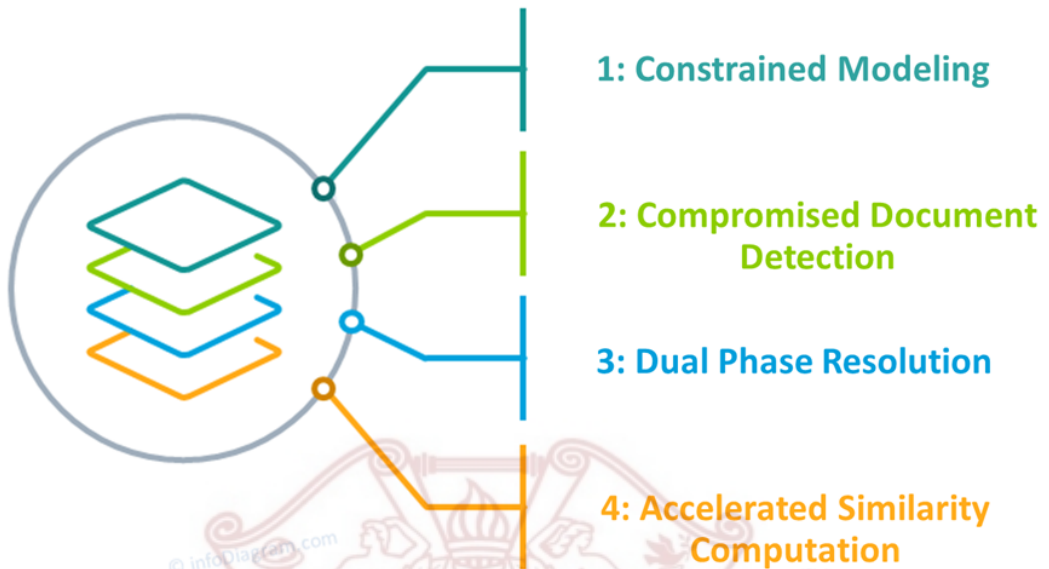


Figure 3.2: Layers of proposed methodology

The semantic preservation methodology consists of three main phases i.e. the constraint modeling, conflict identification and the resolution strategy. Constraint modeling is based on the schema of the document which is the first phase. The second phase is the conflict identification and the third phase is a resolution strategy in which documents are completed and validated. These three phases ensure the completeness of all documents in the dataset as shown in Figure 3.3.

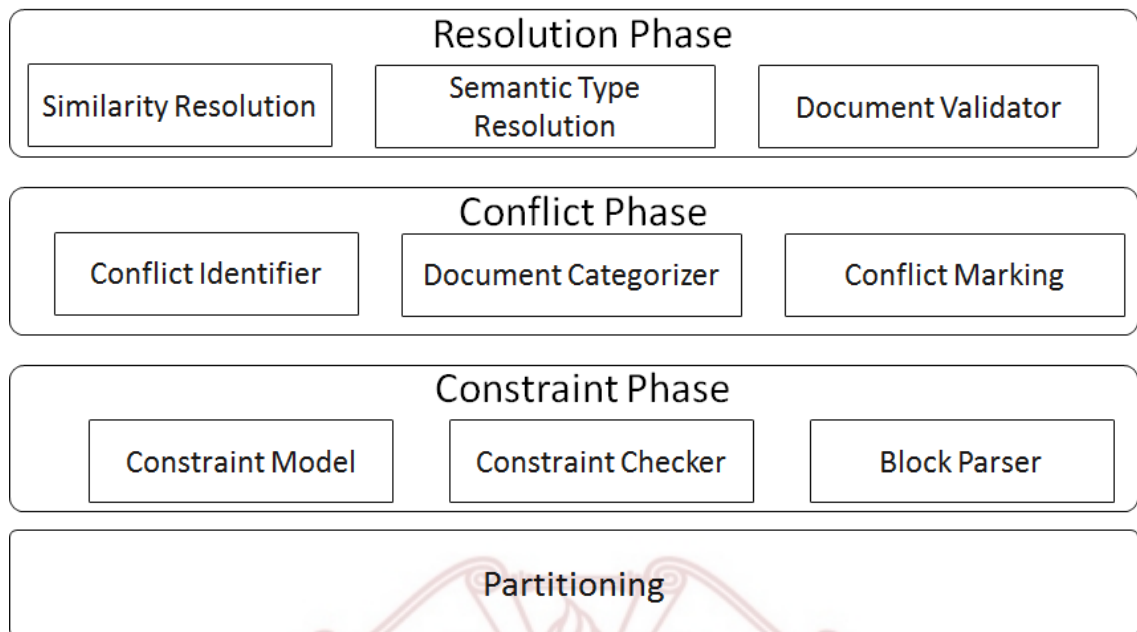
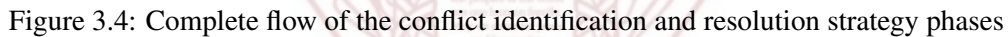


Figure 3.3: Complete flow of the conflict identification and resolution strategy phases

The dataset is fed to the custom partitioner which depending on the dataset schema customized. The complete flow of the conflict identification and resolution strategy phases is shown in the figure 3.4.



3.3.1 Constraint Modeling

Collection @ khu

framework like the replication factor of the data, the block size of the partitions and memory allocations.

The user constraints are based on a user scenario or an application requirement. User constraints are further divided into individual constraints and conjunctive constraints with the nested and non-nested division.

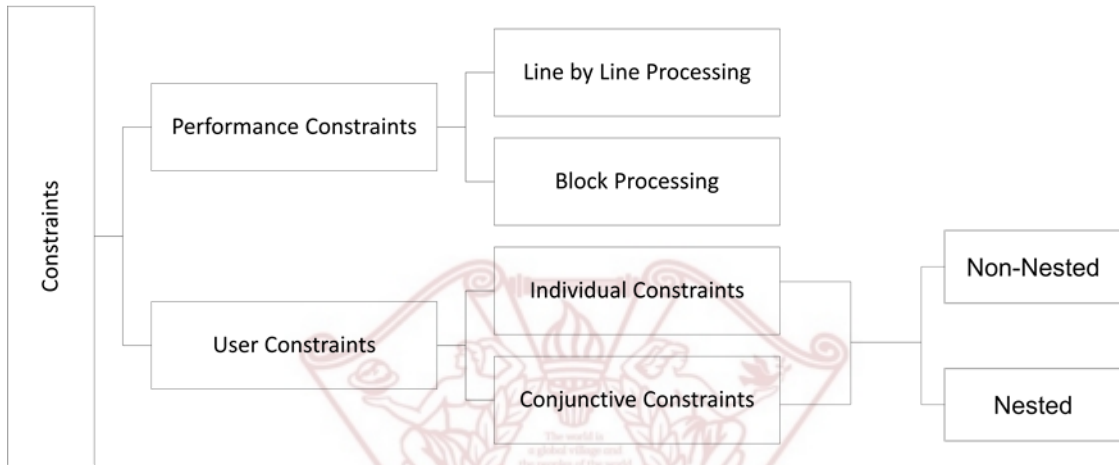


Figure 3.5: Constraint Hierarchy

3.3.1.1 Performance Constraints

Performance constraints are function based constraints and indicate the kind of setting/configuration the blocks will be processed. The number of maximum and minimum mappers can be tweaked for more performance and similarly with the reducers. In this study, block processing was used instead of line by line processing. The motivation behind that was due to the fact that in line by line processing semantics cannot be understood. Other performance parameters were unchanged as semantic preservation is our main goal.

3.3.1.2 User constraints

The user constraints are adapted in the overall scenario as it is based on the user requirements. In Table 3.1, different types of user constraints and their nature are shown. The conjunctive constraints consist of two or more individual constraints and additionally nested and non-nested

constraints as seen in Figure 3.5. Nested constraints are those which encapsulates itself in their body whereas non-nested constraints have multiple categories which also needs to be checked for the constraint to be fully applicable.

Constraint Name	Type	Nested
Clinical Statements	Conjunctive	Yes
entryRelationship	Individual	Yes
Entry	Individual	No
Section	Individual	Yes
Whole Document	Individual	No

Table 3.1: User Constraint Table

Sections and Entries

The sections refer to primarily coded multiple entries and narrative block which contains human readable part [106]. There is different medical information in the sections like procedures, allergies, vital signs as shown in Figure 2.4.

For example, if an user scenario requires all the users with immunizations to be grouped for analytics or research, the section constraint will be applied from Table 3.1.

Entries encompass structured content for computer processing within a document section. The entry part of the CDA[®] documents contains clinical statements and is often critical for the preservation of semantics. If a user scenario is further narrowed down to finding a specific type of immunizations like influenza, then the entry constraint is applied.

Sections can have multiple sections within them but there are no multiple categories when searching it, so a section is a nested individual constraint. Entry is a non-nested individual constraint as it is not nested.

Clinical statements and entryRelationship

There are nine different kinds of clinical statements which makes it a conjunctive constraint. There are multiple constraints to be checked such as Observation, Act and SubstanceAdministration. The clinical statement is a conjunctive constraint as it has nine types and it is also nested

If the expert scenario requires immunizations of influenza in a particular location or time period, then the clinical statements constraint will be applied. Clinical statements generally have

a lot of machine-processable information.

In CDA[®] documents, the entryRelationship is used to connect multiple clinical statements. Structure wise it can have a parent or child relationship with clinical statements which makes it a complex constraint as shown in Figure 3.6 .



Figure 3.6: Clinical Statement Types

3.3.2 Conflict identification phase

During the creation of blocks from the dataset, some documents are chopped in between as the blocks are based on size. This results in a semantic loss for these documents that are cut on the boundaries of the block. There were $2(n-1)$ CDA[®] documents that were cut off from boundaries of the block where n is the number of blocks. In every block, there are at least two compromised documents except the first one and the last one.

The default behavior of partitioner in Hadoop is to give the data line by line to the mapper from the block. This strategy does not constitute any semantics as the CDA[®] documents need a lot of context and a single line can provide none. The custom partitioner in our methodology sends full blocks to the mapper as input so that maximum context can be extracted from it. All blocks are of the same size except the last.

Algorithm 1: Conflict Identification in Map Phase

Input : Full Blocks of the CDA[®] documents

Output: *documents{key, conflictdocument}*: Key Value pairs of conflicted documents from the blocks

```

1  /* Find conflicted document in the upper part of the block and added in
   key value pair. key 1 is for upper documents to be differentiated in
   later stages for processing */;
2  upperdocumentconflict ← scanBlock(block);
3  if upperdocumentconflict == TRUE then
4      upperdoc ← extractDocument(block);
5      documents.key- > 1;
6      documents.value- > upperdoc;
7  end
8  /* Find conflicted document in the lower part of the block and added in
   key value pair. key 2 is for lower documents to be differentiated in
   later stages for processing */;
9  lowerdocumentconflict ← scanBlock(block);
10 if lowerdocumentconflict == TRUE then
11     lowerdoc ← extractDocument(block);
12     documents.key- > 2;
13     documents.value- > lowerdoc;
14 end
15 /* All the documents in the middle are processed as they are complete
   (no semantic loss) according to userscenario defined */;
16 middledocs ← extractDocuments(block);
17 for  $\forall md_i \in middledocs$  do
18     applyuserscenario( $md_i$ );
19 end

```

The constraints are applied in the mapper phase into three different categories base on the block

division into three parts: upper, lower, and middle. The upper part is checked and the incomplete document is extracted from it and is assigned a unique key to it. This key will indicate that it is an incomplete document and the upper part is missing (except for the first block) as shown in algorithm 1. Similarly, at the end of the block, the incomplete document is removed and assigned a unique key which indicates that the lower part of the document is missing (except for the last block). The remaining part of the block which is in the middle has no semantic loss as it has no compromised documents. This part of the block is ready for analytics or any data extraction based on the user or scenario requirements.

The satisfiability test is to check whether all the constraints are applied from the constraint repository as shown in Figure 3.7.

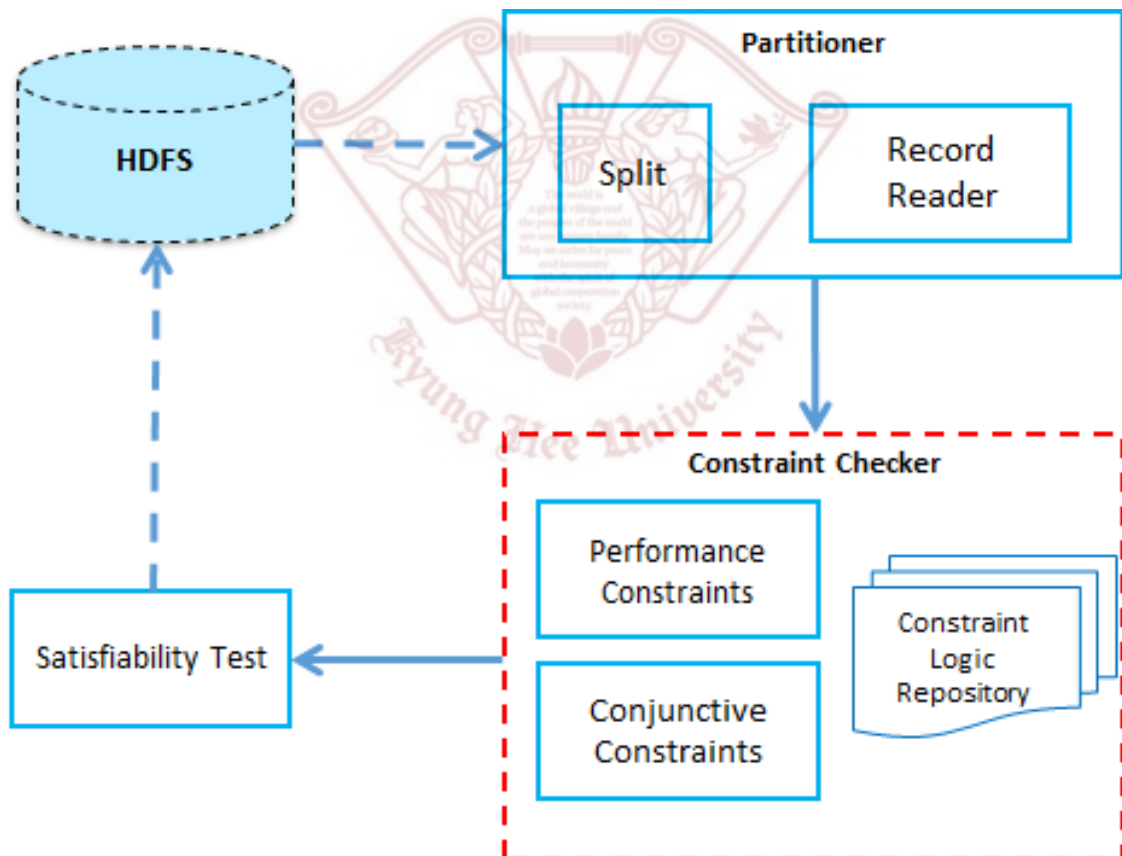


Figure 3.7: Conflict Identification

The two sets of incomplete documents are sent as values with the unique key to the reducer.

All the mappers will send these key value pairs as intermediate data. This data is received in the reducer for resolution of the incomplete documents so that their data can be extracted without any semantic loss.

3.4 Dual Phase Resolution strategy

In the reducer phase, all the compromised documents are processed and completed so that there is virtually no semantic loss. There are two types of key value pairs received by the reducer i.e. the upper part of the document and lower part.

There are two sets of incomplete documents i.e. the ones whose lower part is missing and the others are missing their upper part. Both the sets can have one or more conflicts in most cases and every lower part and upper part joins to make a complete and unique document. These incomplete parts are joined based on the applicable constraint.

When the constraints are applied, the possible matching combinations reduce significantly. For example, if the constraint applied is of the entryRelationship and that part of the document is nested, I will only match the nested open tag documents and close tag documents as shown in Figure 3.8.

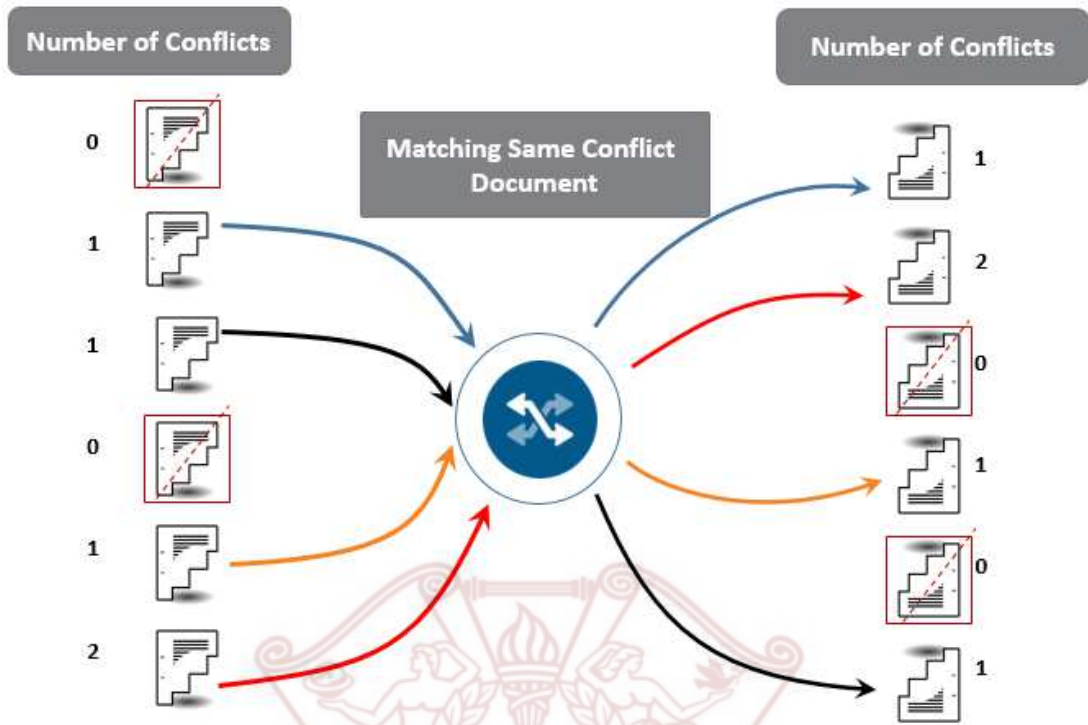


Figure 3.8: Conflict Resolution

During resolution strategy, the parts with zero conflicts are ignored as their incompleteness does not affect the overall scenario or the use case. The incomplete parts will be only matched if the number of conflicts is 1 or more than 1 and the same for each side. This reduces the comparison process where the worst case is 2^n .

For complete resolution, I devised a two pass method. The first pass resolves the process by matching the number of conflicts and validation. While most of the documents got resolved, some duplication occurred when one upper document matched two lower documents and vice versa. The second pass was introduced to eliminate the duplication by matching the semantic type of the documents. Below are the details for both passes of the resolution phase.

3.4.1 First pass

Algorithm 2 is matching the incomplete documents where one or multiple conflicts are identified. For example in Figure 3.8, first and fourth document on the left side do not have any conflicts when

the constraints were applied. The upper part with one missing tag/constraint is matched with all the lower parts in the other set. After joining them, it forms a complete CDA[®] document. This document is then checked for validation through the CDA[®] schema. If the document violates the schema, next lower part document in the list is appended with the upper document until matched and validated. When validated, indexes of upper and lower part are saved.

Algorithm 2: First Pass of Conflict Resolution

Input : $UD = \{ud_1, ud_2, ud_3 \dots, ud_n\}$: Incomplete Upper Documents

$LD = \{ld_1, ld_2, ld_3 \dots, ld_n\}$: Incomplete Lower Documents

Output: Complete CDA[®] documents

```

1  for  $\forall ud_i \in UD$  do
2      /* Conflicts are scanned from the upper document and if conflicts are
        one or greater, then resolution happens */;
3       $Conflictmarker_i \leftarrow checkconflict(ud_i)$ ;
4      if  $Conflictmarker \geq 1$  then
5          for  $\forall ld_j \in LD$  do
6              /* Conflicts are scanned from the lower document and if
                conflicts of upper and lower documents are same, append and
                validate */;
7               $Conflictmarker_j \leftarrow checkconflict(ld_j)$ ;
8              if  $Conflictmarker_i == Conflictmarker_j$  then
9                   $fulldoc \leftarrow appendDocument(ud_i, ld_j)$ ;
10                  $flag \leftarrow validateDocument(fulldoc)$ ;
11                 /* If the document is validated, then create a pair of
                    indexes of upper and lower documents for duplicate checking
                    which triggers second pass if any */;
12                 if  $flag = 1$  then
13                      $Pair < Integer, Integer > MapperRecord \leftarrow list.add(i, j)$ ;
14                 end
15             end
16         end
17     end
18 end
  
```

This phase sometimes results in duplication or zero matches for some parts. This happens when the upper part of the document is matched with two or more lower parts and still comes

out as a validated document. This is also true for the vice versa scenario when a lower part matches multiple upper parts of the document and results in a valid document. This behavior is erroneous and needs further processing to correctly join every document completely. In one of the cases during experiments, the upper part of the vital signs section was joined by a lower part of a procedure section. Upon deeper inspection, this combination resulted in a validated document due to the cutting point as it was a human readable part. The human readable part does not breach the reference information model (RIM) constraints. In view of this, a second pass was introduced to match the duplicate and orphan documents based on additional semantics.

3.4.2 Second pass

In the second pass, all the duplication are removed from the CDA[®] documents. Duplications are found through the pair record saved in algorithm 2. Any upper or lower document index that has occurred more than once indicates a duplication and no index entry indicates an orphan. These partial documents are input to the second pass as shown in Figure 3.9 which indicates the work flow of second phase.

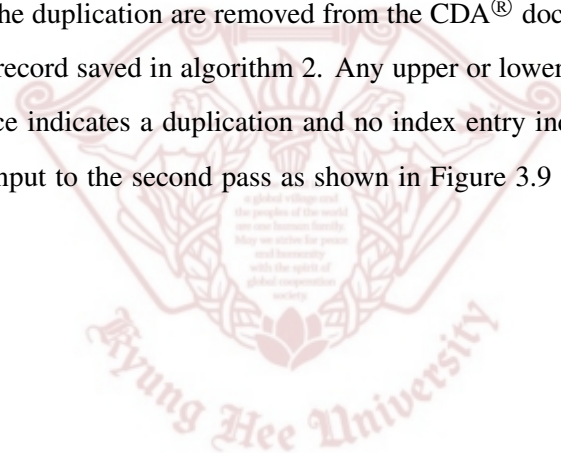




Figure 3.9: Second Pass Workflow

For this study use Unified Medical Language System (UMLS) [107] repository. It is large resource with 1.5 million concepts, 6 million terms and over 20 million relations. It has a complex structure including a metathesaurus and semantic Network. UMLS is a terminology integration system that helps bridge across namespaces and integrate information sources and the subdomains can be shown in figure 3.10

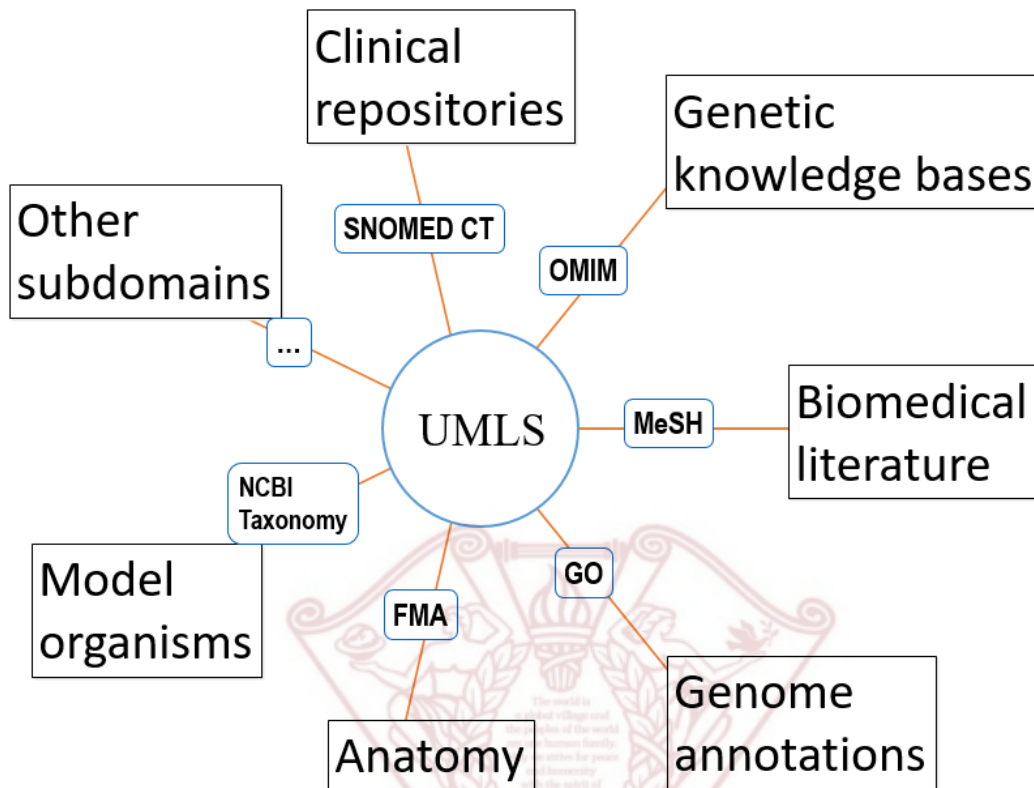


Figure 3.10: UMLS Integrating subdomains

This pass focuses on solving more complex scenarios where additional semantics is needed. This behavior is highly probable when the descriptive part of the CDA[®] document i.e. the human readable part was the cutting point in figure 2.4.

Algorithm 3: Second Pass of Conflict Resolution**Input :** $UD = \{ud_1, ud_2, ud_3 \dots, ud_n\}$: Orphan and Duplicate Upper Documents $LD = \{ld_1, ld_2, ld_3 \dots, ld_n\}$: Orphan and Duplicate Lower Documents**Output:** $CD = \{cda_1, cda_2, cda_3 \dots, cda_n\}$ Complete CDA documents

```

1  for  $\forall ud_i \in UD$  do
2      for  $\forall ld_j \in LD$  do
3          /* Get medical concept of the last constraint of upper document
              (conceptudi) and first constraint of lower document (conceptldj) */;
4          conceptudi  $\leftarrow$  getLastConstraint(udi);
5          conceptldj  $\leftarrow$  getFirstConstraint(ldj);
6          /* Get semantic type of concept of conceptudi and conceptldj */;
7          semtypudi  $\leftarrow$  getSemanticConstraint(conceptudi);
8          semtypldj  $\leftarrow$  getSemanticConstraint(conceptldj);
9          /* Match Semantic types of lower and upper document concepts, join
              them and validate for full CDA® document */;
10         if semtypudi == semtypldj then
11             fulldoci  $\leftarrow$  appendDocument(udi, ldj);
12             flag  $\leftarrow$  validateDocument(fulldoci);
13         end
14     end
15 end

```

For example, if an incomplete section is describing vital signs and the other half that should be found has to be definitely contains vital signs related entries and clinical statements. But during the first pass, the vital signs section was matched to two sections i.e. procedure section and vital signs. This is a duplication of the CDA[®] document and erroneous. To eliminate this, the second pass is executed.

The input in this pass are duplicated and orphan documents in the first pass. The medical/clinical concepts are extracted from the last section/entry/clinical statement of the upper documents. Similarly, medical/clinical concepts are extracted from the first section/entry/clinical statement of the lower documents as shown in line 4 and 5 in algorithm 3.

Then the semantic type of the medical concepts of the upper document and lower document are matched in line 7 and 8. This is done by the terminology services of UMLS (Unified Medical Language System) [107]. UMLS repository is used for second phase of resolution. It has a

vocabulary of enriched concepts of biomedical terms and can be used for semantic type matching as shown in Figure 3.11.

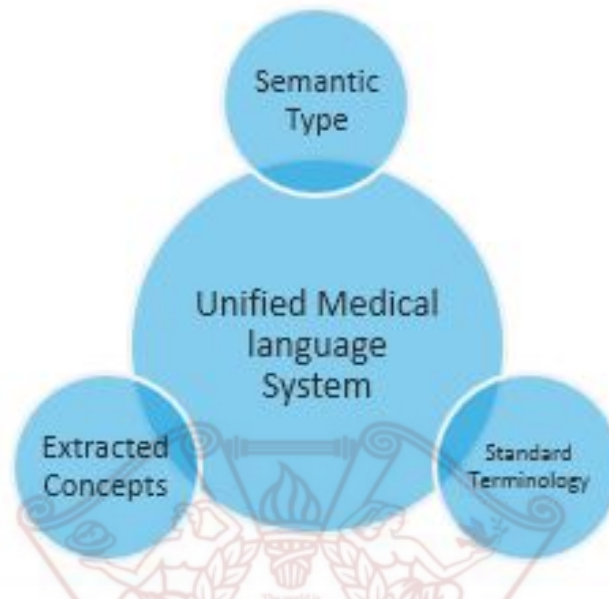


Figure 3.11: Unified Medical Language System

This study cross checked the semantic types of the extracted concepts and if they match, the complete document is validated again through the first pass again shown in algorithm 3 (line 10 to 12). In figure 3.12 two compromised documents can be seen send their codes and the semantic concept is extracted . This concept is then matched to make a full document of two compromised documents.

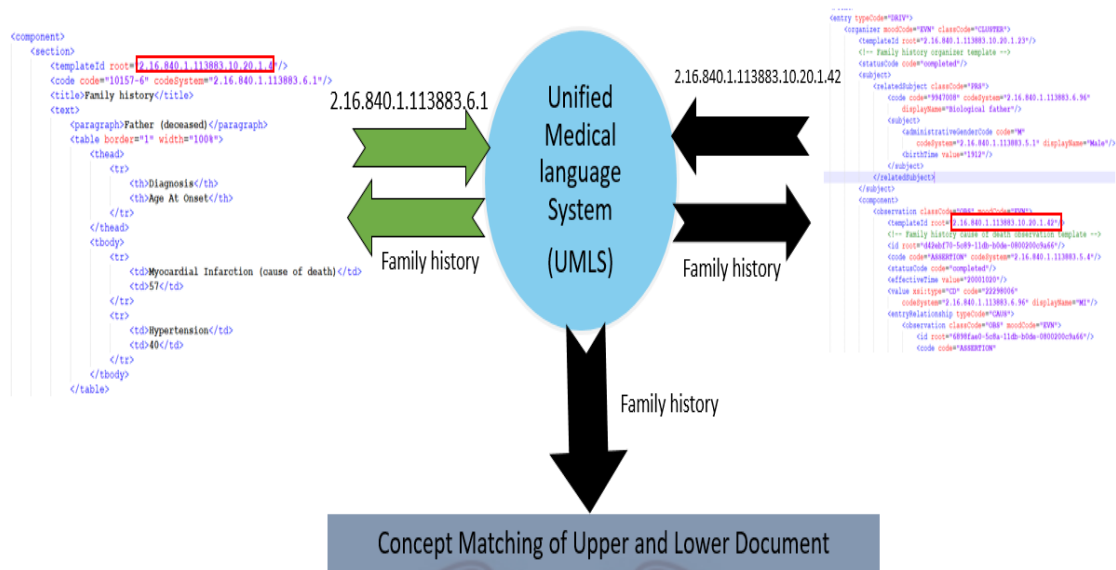


Figure 3.12: Unified Medical Language System Work Flow

The semantic type of the extracted concept helps in removing duplication in the second pass and help us correctly join and validate all compromised CDA[®] documents in our dataset.

3.5 Accelerated Similarity Computation

To make semantic preservation a feasible solution, more performance gains are required. Complex data among intermediate data make it a very time intensive process. Naive record and data matching gives quadratic complexity. So if the data has one million records, there will be 1 trillion record matches. The input is the complex data objects created from intermediate data and help of the schema of the dataset. The workflow of complex data object creation is shown in 3.13.

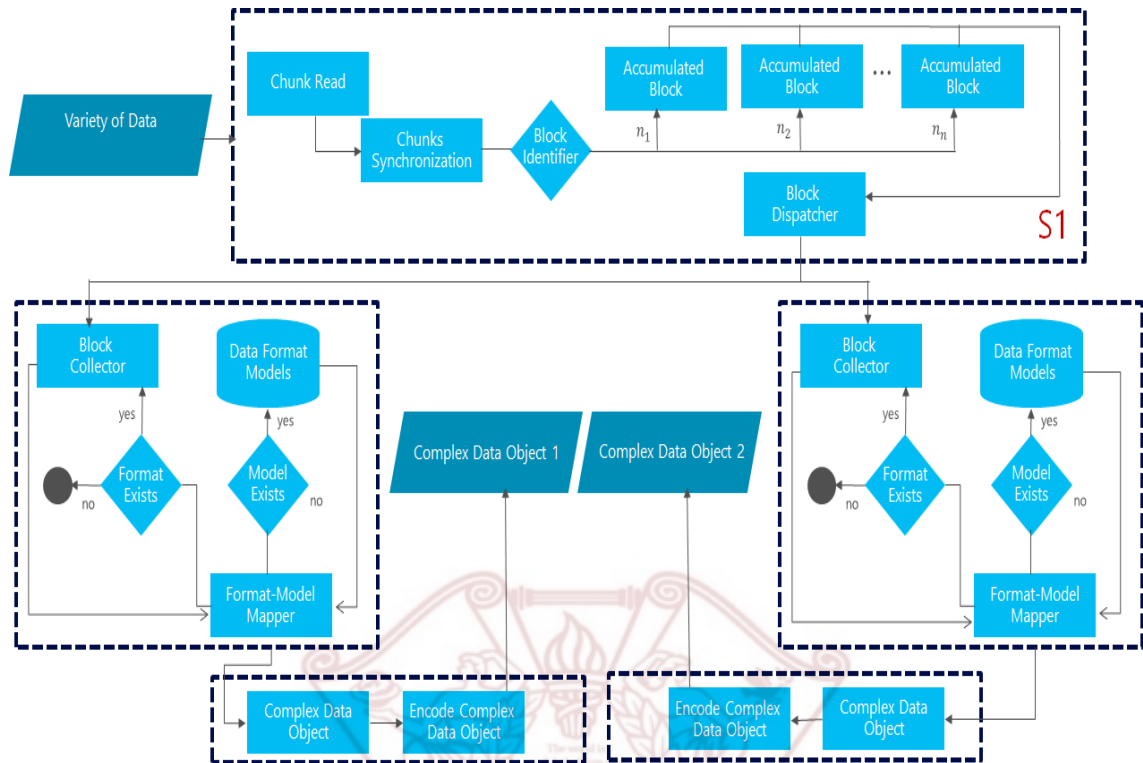


Figure 3.13: Creation of complex data objects

The goal is similar data computations for partitioned data on the same processing data node. The advantage is significant performance gains due to processing of similar data in one location.

Algorithm 4: Similar Data Computation

Input : $A = \{o_1, o_2, o_3, \dots, o_n\}$: Complex Mapped Objects

Output: S : Hash tables of candidate similar records

```

1      /*      [Numerical Encoding] where, E is encoded list      */;
2   $E \leftarrow encodeAttributes(attributes, code)$ ;
3  /* [Create the adjacency list of records ] where, A is object list */;
4   $AL \leftarrow createAdjacencyMatrix(E, A)$ ;
5      /*      [Hash Function Generation ] where, H is hash table      */;
6   $H \leftarrow hashGeneration(AL)$ ;
7      /*      [Create MinHash Matrix ] where, M is minhash matrix      */;
8   $M \leftarrow buildMinhashMatrix(AL, H)$ ;
9      /*      [Band Splitting] where b is the number of bands      */;
10  $C \leftarrow bandSplitting(AL, b)$ ;
11      /*      [Compute Hash Codes] where S is the Hash Table      */;
12  $S \leftarrow computehashCodes(b, C)$ ;
13  $ComputeSimilarlity(S)$ ;

```

In the algorithm 4, the mapped objects are the inputs and for them a hash function is generated and a matrix is created. The similarity is found through jaccard similarity. The signature column generation is done through Minhash matrix as shown in Figure 3.14.

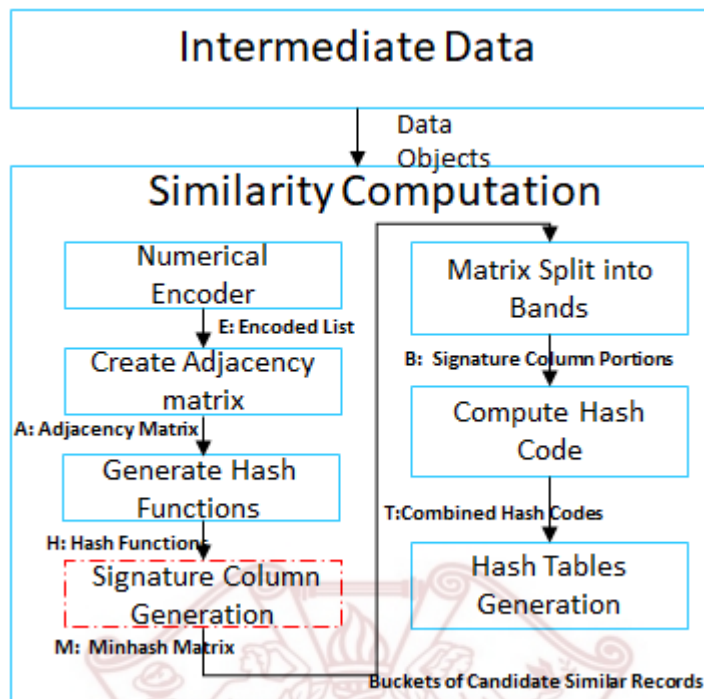


Figure 3.14: Similar Data Computation Workflow

There are two kinds of weighted attributes i.e. must care attributes and least care attributes. The must care attributes are needed for those which are more influential with respect to the user scenario. The must care attributes are identified by the domain expert is dependent on the dataset being processed. For example, if the analytics needs to generate with respect to the diagnosis and gender, then they become the most significant attributes and the similarity is found with respect to them.

Algorithm 5: Signature Column Generation**Input :** Adjacency Matrix $MC = \{mc_1, mc_2, mc_3, \dots, mc_n\}$; $LC = \{lc_1, lc_2, lc_3, \dots, lc_n\}$; $k = \text{sizeofmatrix}$ **Output:** : Hash values of weighted attributes

```

1  for  $\forall i \in n$  do
2      /* All the must care attributes are checked and their signature is
        generated */;
3      if  $mc_1 == TRUE$  then
4          if  $S(i) == 1$  then
5              for  $\forall j \in k$  do
6                  if  $h_j(i) == c_j$  then
7                       $c_j = h_j(i)$ ;
8                  end
9              end
10         end
11     end
12     /* All the least care attributes are checked and their signature is
        generated */;
13     if  $lc_1 == TRUE$  then
14         if  $S(i) == 1$  then
15             for  $\forall j \in k$  do
16                 if  $h_j(i) == c_j$  then
17                      $c_j = h_j(i)$ ;
18                 end
19             end
20         end
21     end
22 end

```

The algorithm 5 describes to generate the Signature Column Generation of weighted attributes. The weighted attributes' hash are later on used for approximation matching. Table 3.2 shows must-care attributes in red colors and least care attributes in blue.

$$\forall av_l \in ML_{P_i}, Pr\left(\min\left(\pi(ML_{P_i})\right) = av_l\right) = \frac{1}{|ML_{P_i}|}$$

$$\text{Example: } Pr\left(\min\left(\pi(ML_{P_1})\right) = Gender\right) = Pr\left(\min\left(\pi(ML_{P_1})\right) = Income\right) = \frac{1}{6}$$

Figure 3.15: Calculate Attribute Probability

The probability of the weighted attributes can be calculated by the Figure 3.15. Income is the least-care attribute and gender is must-care attribute while the other parameters/fields are ignored during matching

Patient #	Diagnosis	Location	Gender	Income	Blood Group	Insurance
P1	Asthma	Seoul	Male	Medium	A	INS-1234
P2	Asthma	Busan	Male	High	B+	INS-3467
P3	Flu	Seoul	Female	Low	A	INS-1222
P4	Flu	Seoul	Male	Low	A	INS-9575
P5	Flu	Seoul	Male	Medium	B+	INS-4673
P6	Asthma	Seoul	Male	High	B+	INS-9801

Table 3.2: Records having ML-Care attributes

3.6 Case study for Lossless data

Lossless data is important in many fields but it is one of the most important trait in healthcare industry due to its critical nature. There are many stakeholders in the health industry like doctors and physicians, health insurance companies, pharmaceutical companies and systems which are based on prediction and are decisions enablers as shown in Figure 3.16.



Figure 3.16: Stakeholders of the lossless data

Physicians need data for evaluating similar patient symptoms in a broad view of finding out about an outbreak. Pharmaceuticals need lossless data for prediction of what medicines could be needed in the upcoming season. Health Insurers need lossless data for latest medication prices and their ratios for coverage of patients. Predictive analytics enables in decision support systems which in turn requires lossless data. The abstract workflow is shown on how the different applications based on this system will work is shown in Figure 3.17.

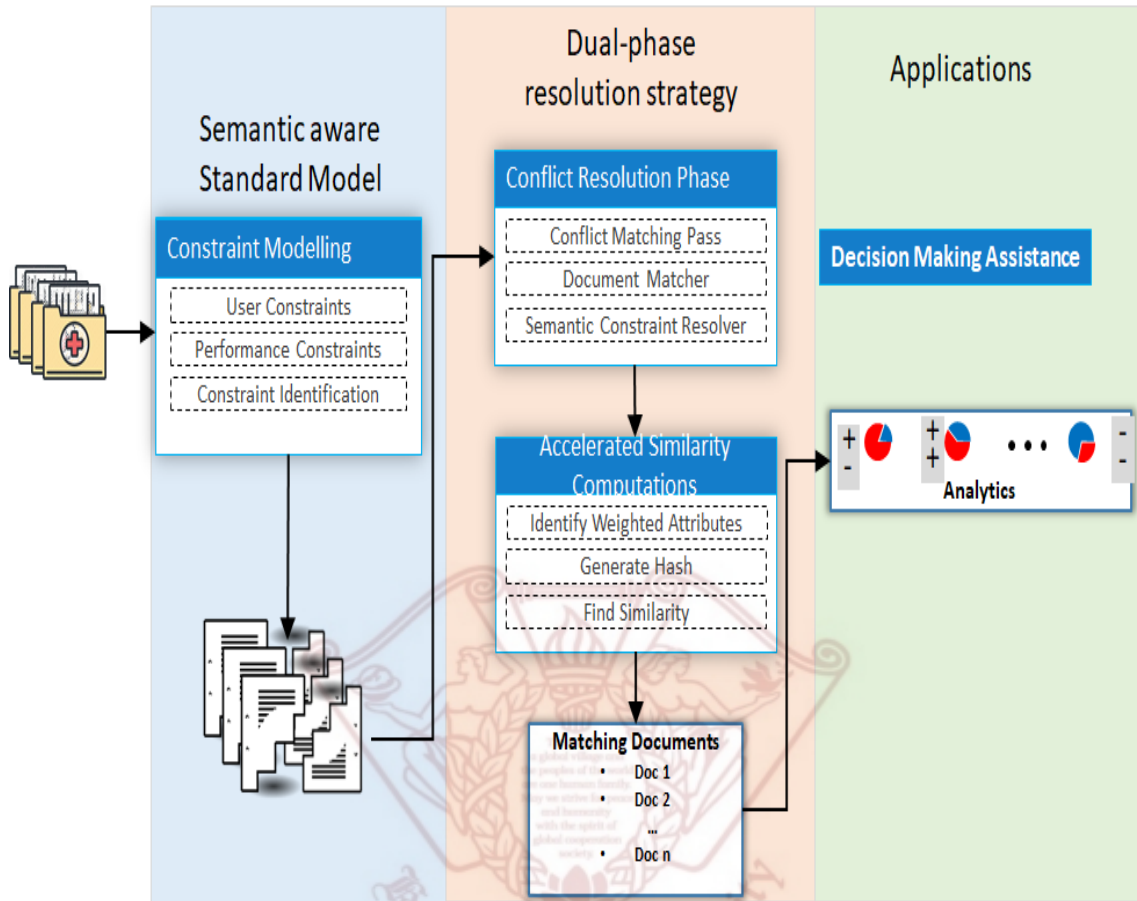


Figure 3.17: Abstract Workflow of proposed methodology for Applications

3.6.1 Disease Based Analytics

The issue with data processing using big data framework is showed below as allergies and medications can go to different nodes. For example if the allergies go to node 1 shown in Figure 3.18 and medications go to node 2 shown in Figure 3.19


```

<entry typeCode="DRIV">
  <act classCode="ACT" moodCode="EVN">
    <templateId root="2.16.840.1.113883.10.20.1.27"/>
    <!-- Problem act template -->
    <id root="36e3e930-7b14-11db-9fe1-0800200c9a66"/>
    <code nullFlavor="NA"/>
    <entryRelationship typeCode="SUBJ">
      <observation classCode="OBS" moodCode="EVN">
        <templateId root="2.16.840.1.113883.10.20.1.18"/>
        <!-- Alert observation template -->
        <id root="4adc1020-7b14-11db-9fe1-0800200c9a66"/>
        <code code="ASSERTION" codeSystem="2.16.840.1.113883.5.4"/>
        <statusCode code="completed"/>
        <value xsi:type="CD" code="282100009"
          codeSystem="2.16.840.1.113883.6.96"
          displayName="Adverse reaction to substance"/>
        <participant typeCode="CSM">
          <participantRole classCode="MANU">
            <playingEntity classCode="MMAT">
              <code code="70618"
                codeSystem="2.16.840.1.113883.6.88"
                displayName="Penicillin"/>
            </playingEntity>
          </participantRole>
        </participant>
      </entryRelationship>
    </act>
  </entry>

```

Figure 3.18: Allergies Code

The allergies and medications going to different nodes hampers the overall meaningful facts extraction for patients as no node has complete picture. This will create incomplete facts for allergy of Hives.

```

<substanceAdministration classCode="SBADM" moodCode="EVN">
  <templateId root="2.16.840.1.113883.10.20.1.24"/>
  <!-- CCD Medication activity template -->
  <templateId root="1.3.6.1.4.1.19376.1.5.3.1.4.7"/>
  <!-- IHE Medications Template -->
  <id root="0dbd5b05-6cde-11db-9fe1-0800200c7a26"/>
  <statusCode code="active"/>
  <effectiveTime xsi:type="PIVL_TS">
    <period value="24" unit="h"/>
  </effectiveTime>
  <routeCode code="PO" codeSystem="2.16.840.1.113883.5.112"
    codeSystemName="RouteOfAdministration"/>
  <doseQuantity value="1"/>
  <consumable>
    <manufacturedProduct>
      <templateId root="2.16.840.1.113883.10.20.1.53"/>
      <!-- Product template -->
      <manufacturedMaterial>
        <code code="370619" codeSystem="2.16.840.1.113883.6.88"
          codeSystemName="RX NORM"
          displayName="Atenolol 25 MG Oral Tablet"/>
        <originalText>Atenolol 25 MG Oral Tablet</originalText>
      </code>
      <name>Tenormin</name>
    </manufacturedMaterial>
  </manufacturedProduct>
</consumable>
</substanceAdministration>

```

Figure 3.19: medications Code

Second case study is for readmission rates for Congestive Heart Failure (CHF). Readmission of patients with chronic diseases is a significant and growing problem in the United States and an increasing burden on the healthcare system. Preventable patient readmissions cost the U.S. healthcare system about \$25 billion every year, according to a study by PricewaterhouseCoopers (2010). Experts believe that high readmission rates, when patients are readmitted within 30 days of discharge, indicate that the nation's hospitals are not adequately addressing patient health issues. To tackle this problem, the U.S. Centers for Medicare and Medicaid Services (CMS) has imposed penalties on hospitals for preventable readmissions related to chronic conditions such as heart failure, starting in 2012. One if five people have CHF, and if any of the sections containing readmission information is compromised, the overall goal of cost reduction can be hampered.

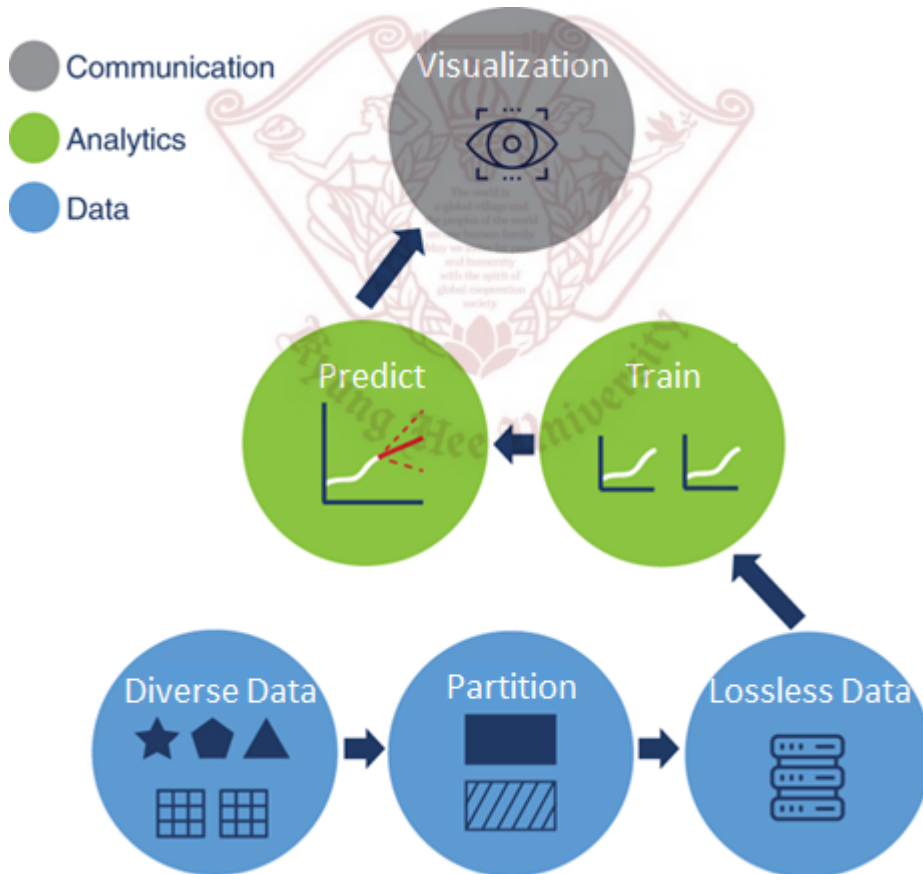


Figure 3.20: General case study of analytics and visualization

4.1 Experimental evaluation

In this section, the conflict identification algorithm based on the constraint model is implemented, and the results are explained. In all experiments, highlighting the conflicts based on the constraint level and resolution of these conflicts are discussed and analyzed.

4.1.1 Experimental Setup

In this study, Hadoop was used as a primary big data framework due to it being the most common framework in academia. The file system is HDFS which is used by both Hadoop and Spark. The operating system used is Ubuntu 16.05 and the data cluster consists of six data nodes.

4.1.2 Dataset description

For the experiment, this study used a publicly available dataset containing 700 unique consolidated CDA[®] documents [108]. The data is enlarged by replicating the originally collected documents in a randomized order through sampling and bootstrapping. The Consolidated CDA implementation guide has nine subtypes of commonly used CDA documents, and each of the nine types have a document template defined in the Consolidated CDA guide, which is the source for implementing these CDA documents. The study experimented with the enlarged dataset ranging from 512 MB (5464 documents) to 8GB (87000 documents).

4.1.3 CDA[®] preservation conflicts

The experiments were done on datasets ranging from 512 MB to 8 GB. The conflicts in the experiment were directly related to the size of the dataset and the block size during the MapReduce cycle. The block sizes were 32 MB, 64 MB, 128 MB and 256 MB. The block size impacted the conflicts inversely as the bigger size of the block decreased the number of conflicts. This made sense as fewer blocks meant fewer conflicts but bigger blocks result in severe performance deficiency as every map function has to process bigger blocks.

4.1.3.1 Entry, section and whole documents

The number of CDA[®] documents compromised are shown in Figure 4.1. 269 documents in an 8 GB dataset are compromised. This just shows the number of documents that are split during the processing phase.

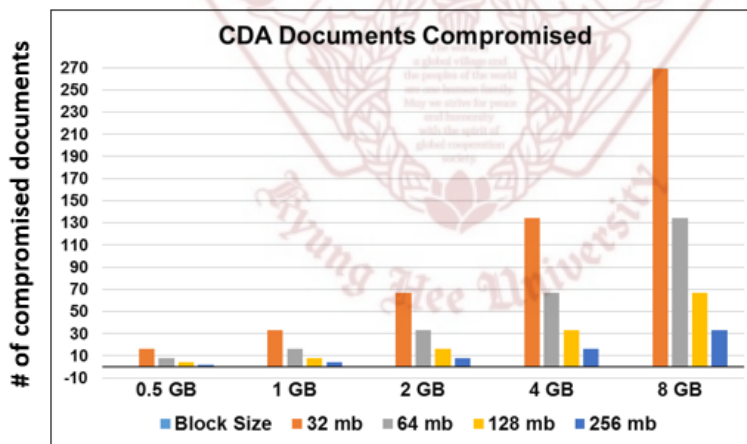


Figure 4.1: CDA Documents Compromised

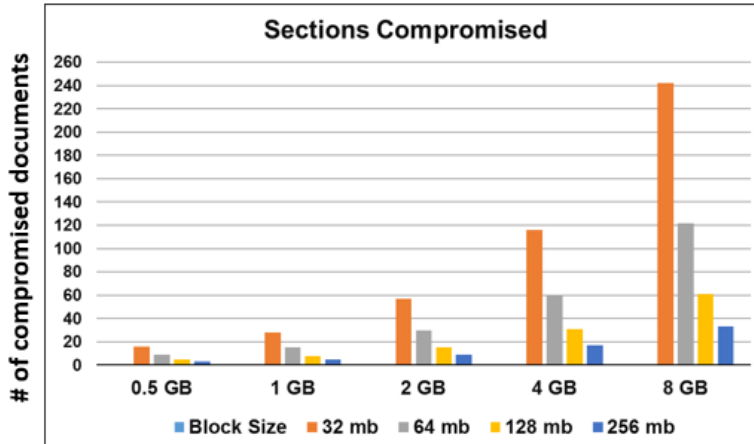


Figure 4.2: CDA Sections Compromised

Similar behavior is displayed in the sections constraint in Figure 4.2 as it has higher conflicts than entry (4.3) with respect to the block size and dataset. This is due to the fact that it is high level constraints and there is more chance of it being compromised being a parent tag.

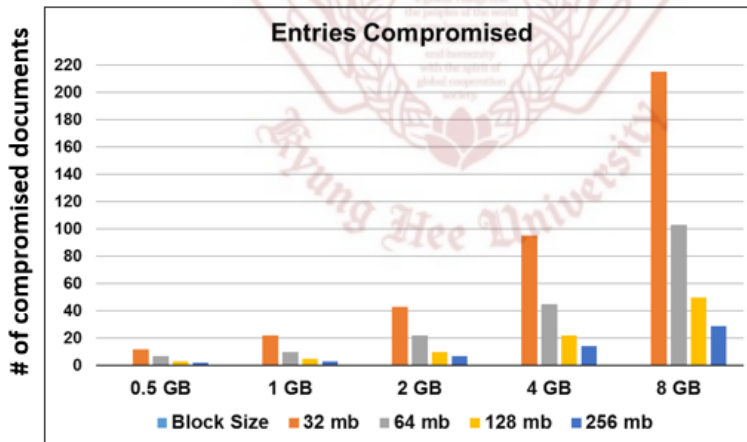


Figure 4.3: CDA Entries Compromised

4.1.3.2 Clinical Statements and EntryRelationship

In this experiment, the number of entryRelationships and clinical statements that were compromised semantically in the data set are shown in second row of Figure 4.4.

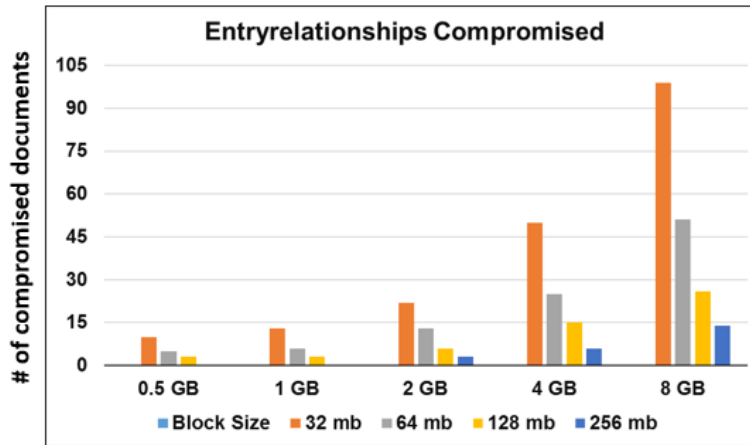


Figure 4.4: Entryrelationships Compromised

There are nine types of clinical statements in CDA[®] schema. All the 9 clinical statements types were accumulated in Figure 4.5.

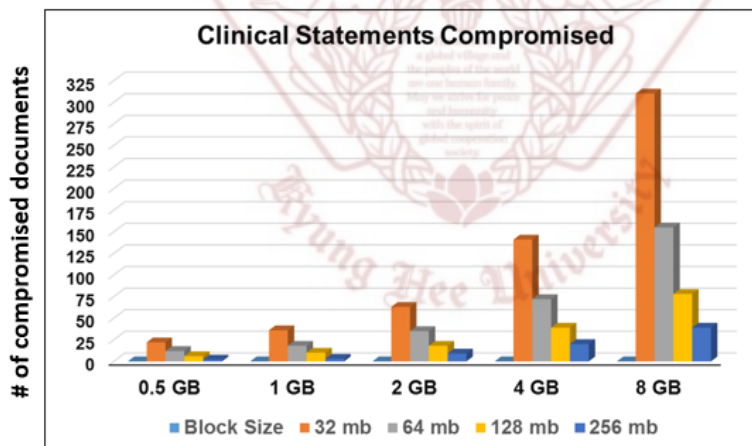


Figure 4.5: clinical statements Compromised

4.1.3.3 Conflicts against size and constraints

The conflicts against the constraints are shown in Figure 4.6. It is clear that a constraint is up in the hierarchy of schema, more conflicts will arise. EntryRelationship has the least conflicts as it is child tag of all the other constraints. The clinical statements have been omitted from this result as they have nine different types.

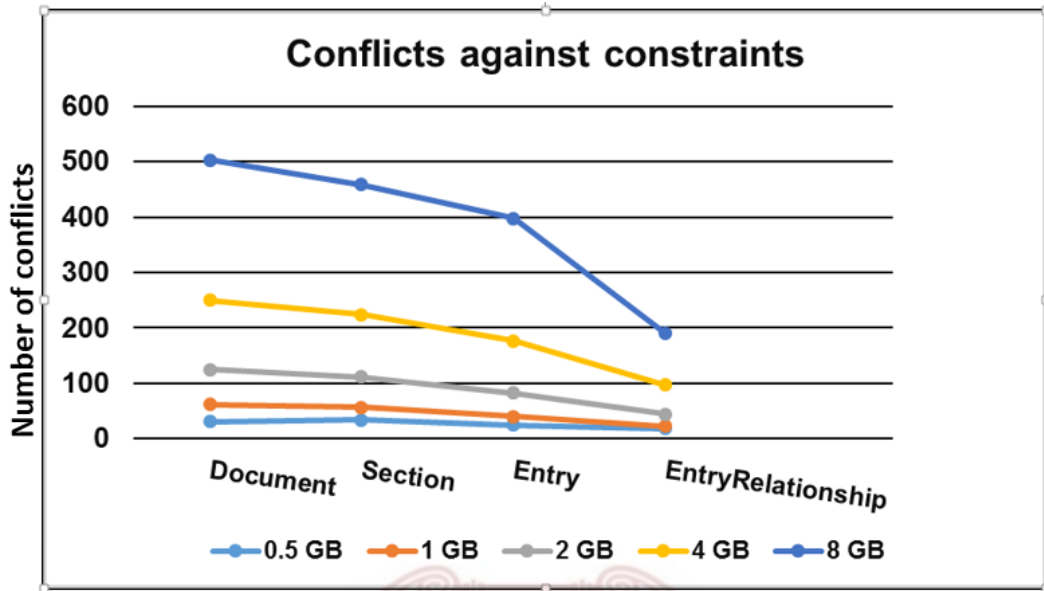


Figure 4.6: Conflicts against constraints

The conflicts against the size is shown in Figure 4.7. The increase of conflicts in big data is polynomial. It further shows that conflicts will increase when the dataset increases.

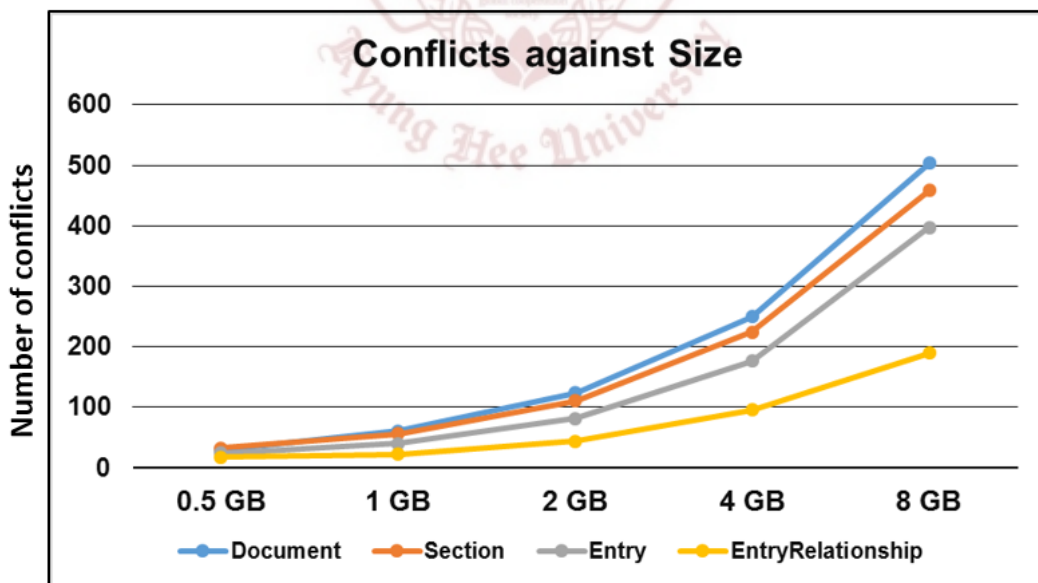


Figure 4.7: Conflicts against size.

4.1.4 Resolution results

The resolution phase has two passes. The first pass matches the upper and lower part of the documents based on the number of conflicts and validates them. If they are validated based on HL7[®] RIM [72], they are mapped and the corresponding index of the document is saved.

In the first phase, some conflicts remained unresolved on different block sizes. 32 MB block data had more remaining conflicts after the first pass due to its size as bigger block size meant few or no conflicts in the first pass as shown in Table 4.1. In the first phase, for the block of 256 MB, all the split documents were resolved through the validation process except 8 GB data set.

Number of documents needed for second pass				
Block Size	1 GB	2 GB	4 GB	8 GB
32 MB	6	14	40	101
64 MB	4	5	19	46
128 MB	0	2	4	13
256 MB	0	0	0	4

Table 4.1: Conflicts remaining in First pass on different datasets

In Table 4.2 for 1 GB dataset only due to space issues for 32 MB and 64 MB. All the document numbers are sequentially assigned when identifying conflicts and does not have any significance in regard to resolution phase.

Resolution 1st Pass 32 MB Block				Resolution 1st Pass 64 MB Block	
Upper Part	Lower Part	Upper Part	Lower Part	Upper Part	Lower Part
Doc. 1	Doc. 18	Doc. 18	Doc. 24	Doc. 1	Doc. 12
Doc. 2	Doc. 11	Doc. 19	Doc. 13	Doc. 2	Doc. 15
Doc. 3	Doc. 34	Doc. 20	Doc. 21	Doc. 3	Doc. 7
Doc. 4	Doc. 28	Doc. 21	Doc. 34	Doc. 4	Doc. 14
Doc. 5	Doc. 23	Doc. 22	Doc. 21	Doc. 5	Doc. 8
Doc. 6	Doc. 4	Doc. 23	Doc. 14	Doc. 6	Doc. 17
Doc. 7	Doc. 20	Doc. 24	Doc. 33	Doc. 7	Doc. 2
Doc. 8	Doc. 19	Doc. 25	Doc. 16	Doc. 8	Doc. 13
Doc. 9	Doc. 29	Doc. 26	Doc. 6	Doc. 9	Doc. 4
Doc. 10	Doc. 1	Doc. 27	Doc. 25	Doc. 10	Doc. 11
Doc. 11	Doc. 26	Doc. 28	Doc. 17	Doc. 11	Doc. 4
Doc. 12	Doc. 30	Doc. 29	Doc. 32	Doc. 12	Doc. 9
Doc. 13	Doc. 21	Doc. 30	Doc. 22	Doc. 13	Doc. 6
Doc. 14	Doc. 8	Doc. 31	Doc. 34	Doc. 14	Doc. 5
Doc. 15	Doc. 31	Doc. 33	Doc. 27	Doc. 15	no Doc.
Doc. 17	Doc. 3	Doc. 34	Doc. 9	Doc. 16	Doc. 1
Doc. 16	Doc. 5			Doc. 17	Doc. 11

Table 4.2: First Pass of Resolution

As seen in Table 4.2 the highlighted documents are mapped multiple times like lower documents 34 and 21 in 32 MB block. Upper documents 20 and 21 are also mapped erroneously to the same lower document 21 in 32 MB block. There can only be one to one mapping.

These documents are filtered in Table 4.3 as duplicate or orphan documents and are inputs for the second pass of the resolution phase. All the documents in Table 4.3 are resolved successfully through semantic type matching.

Orphan and Duplicate Documents in 32 MB Block		Orphan and Duplicate Documents in 64 MB Block	
Document	Document Type	Document	Document Type
Upper Doc. 3	Duplicate	Upper Doc. 15	Orphan
Upper Doc. 20	Duplicate	Upper Doc. 17	Duplicate
Upper Doc. 21	Duplicate	Lower Doc. 4	Duplicate
Upper Doc. 22	Duplicate	Lower Doc. 11	Duplicate
Upper Doc. 31	Duplicate	Lower Doc. 10	Orphan
Upper Doc. 13	Duplicate	Upper Doc. 10	Duplicate
Lower Doc. 2	Orphan	Upper Doc. 9	Duplicate
Lower Doc. 7	Orphan	Upper Doc. 11	Duplicate
Lower Doc. 10	Orphan	Lower Doc. 16	Orphan
Lower Doc. 12	Orphan		
Lower Doc. 15	Orphan		
Lower Doc. 21	Duplicate		
Lower Doc. 34	Duplicate		

Table 4.3: Second Pass of Resolution

UMLS repository is used for second phase of resolution due to its vocabulary of enriched concepts of biomedical terms. UMLS has the large biomedical concepts from over 100 source vocabularies and millions of concepts and relationships.

In Figure 3.12, the upper document on the right side sends a last code which indicates the cutting point was at family history section at the end. In the lower part of the document, the first code is extracted from the template id and given to the web service to give a semantic concept. In Figure 3.12, the lower document also sends family history which confirms that both the upper and lower part are actually one document and were chopped during partitioning. The semantic concept is matched and resolution is managed for all the documents in Table 4.3. During phase two of the resolution, all the document were successfully resolved. Although there might be a possibility that two different documents chopped might return same semantic type. For this possibility, the documents are also checked if they are CDA compliant.

4.2 Discussion

Based on the experimental results in this study, we deduced some important observations to consider during processing of standardized healthcare documents in Big Data.

4.2.1 Conflict Marking in Compromised Documents

The conflict markers are based on the constraints and they are based on the standard that is being followed in the dataset. This makes the constraint identification critical and important in the overall process. This needs to be done by the domain expert in the dataset standard and knows how the semantics are compromised on the data.

4.2.2 Number of compromised documents for resolution

In this study, the focus was on semantic preservation of the clinical documents in the big data storage and to ensure that every document is interpretable after partitioning and execution of the framework. The Table 4.2 indicate the number of documents that are compromised which were found through conflict markers. This was only for 1 GB dataset and the compromised documents were large in numbers for bigger dataset in the experiments.

4.2.3 Health Interoperability

Modern healthcare depends on successful communication between different stakeholders. Interoperability is needed to provide information, enable better decision making, reduce redundancy and improve safety [77]. Semantic loss occurs when the documents being mapped/translated are incomplete which reduces the ability for data queries and business rules. Health interoperability cannot be achieved if the health documents lose their semantics in the partitioning and processing phases. In this regard, semantic preservation is very important for health interoperability as it needs a complete document mapping from one format to other. These documents should have high level semantic preservation so that health interoperability can be achieved.

4.2.4 Big data file systems

This study has focused on HDFS due to its widespread use in the big data. It is the primary file system in Hadoop and Spark which are the most extensively used framework. Partitioning is a big part of big data frameworks irrespective of the technology and will remain an integral part of the batch processing and real time processing systems. Amazon simple service storage (S3) manages data as objects and each object is identified by a unique user assigned key. The HDFS and Hadoop framework can interact with data in Amazon S3 which further shows the flexibility and usability of HDFS. Oracle uses an Oracle Database File System (DBFS) that creates a standard file system interface and the files are stored in database tables. It is more in line with RDBMS than the conventional file system and is tailored to serve the purpose of the Oracle. Kudu is columnar store manager in Apache Hadoop platform and can be connected to other Hadoop ecosystem components like Spark, Impala It uses tables and has a SQL like a schema and is generally collocated with HDFS. The motivation for choosing HDFS for our study and experimental setup was due to its compatibility with diverse storage systems both commercial and open source.

4.2.5 Conventional behavior of Hadoop

We deviated from the default Hadoop implementation practices as Hadoop by default has the line by line processing in mapper phase. There are also additional input format processing modules like multiple line processing and XML processing but these also hinders the ability to extract the semantic concepts and values of the clinical data because of its comprehensive nature. In our implementation, we parsed the whole block in the mapper phase and identified the compromised documents in the first phase.

4.2.6 Additional passes in resolution phase

In the resolution phase in the methods section, there are two passes which ensured the semantic preservice of the documents. The dataset used did not reveal any compromised documents but we imagined another issue that could arise in the resolution process. The second pass of the resolution phase compares semantic types of each end of the document to join it. In a rare case, the semantic types of one upper document can match two lower documents and vice versa. For

this unique case, an additional 3rd pass could be introduced in the future work.

4.2.7 Limitations of this work

The proposed semantic preservation method requires a data schema for creating a semantic aware standard model. The constraints are created from the schema which is then used for detection of the compromised documents. This limitation will be addressed in our future work as this study focuses on semi structured and structured data.



5.1 Conclusion

Obtaining lossless data for healthcare and medical fields from diverse healthcare standards and data formats is a very important and critical task. The main goal of this thesis was to achieve lossless data while being processed in the big data frameworks as in today's world, big data frameworks are the most effective way to handle complex data in large volumes. The thesis achieved the lossless data by creation of a semantic aware standard model and dual phase resolution strategy.

Standardized healthcare documents maintains a lot of information due to the comprehensiveness of their schema. Every clinical document has critical potential and losing its semantics can have adverse effects. In default behavior of the partitioning in the big data frameworks, many clinical standardized documents loose their semantics. It results in incomplete information to the stakeholders and reduce the decisions making ability. Using our proposed technique, affected documents were identified through constraint modeling depending on the scenario and then the resolution of incomplete documents was done to avoid any semantic loss. Most of the work is focused on the schema of the standard being followed. This study choose HL7 CDA for its high adoption rate in the hospitals and medical centers.

Most of the studies have used big data frameworks for storing and querying but have not focused on semantic loss during the big data partition. Our motivation was to remove this semantic loss before the processing has started to give better and complete results. The overall goal in this critical dataset is always maximum extraction of the concepts to get meaningful understanding of the dataset to enable better decision making.

This work is beneficial as it does not compromise the performance of the big data framework significantly and extract the information from all the clinical documents which are semantically

compromised during the process. One objective of this study apart from semantic preservation is the accelerated similarity computations. The similarity computations were done by assigning weights to the features in the dataset and approximate matching was done to get intermediate on similar locations. The overall overhead during semantic preservation is offset by the accelerated similarity computations in the later phase of the methodology. The presented methodology has the potential to produce several benefits in term of practical implementation in data analytics and health interoperability field.

5.2 Future Directions

Further extensions will be done evaluate the system for computation and complexity of the overall execution. For future work, the focus will be on the performance aspect of the technique as almost 30% of the data in Big Data is health related so the performance aspect is very critical and important. Additional case studies will be worked on further cementing the effectiveness of semantic preservation. Fast response time will be the goal for us in the future work specially with respect to the accelerated similarity computations (ASC) in the data cluster. The exhaustive search takes a lot of time and has quadratic complexity which can bottleneck the overall cluster. The hashing technique is ASC helps in approximate hashing and the results for ASC are under process and in the future will be focused on to the complete the process and offset the semantic preservation overhead.

5.2.1 Future work: Health Imaging Data

In current work, the standardized textual and XML data is taken into account. Figures and images carry an important piece of information in healthcare data. Adding non-textual data in the lossless data may increase the chance of better analytics as they are a good source for the quick presentation of the complex contents.

5.2.2 Future work: Precision Medicine

Precision medicine is one of emerging field in disease prevention and treatment and it takes people's individual variations in genes, environment, and lifestyle into account. It is a hot topic and different countries have started working to contribute to precision medicine initiative. The lossless data can be used for precision medicine.



Appendix A

Healthcare standards and Big Data framework

While big data on Public Health are growing with the diffusion of telemedicine and e-health and more generally with that of Internet of Things (IoT) sensors and networking digital platforms, their relationship with healthcare infrastructural investments is still pioneering [109]. Digitalized data already provide many benefits to healthcare organizations through disease prediction and surveillance, population health management and patient care improvement. Moreover, big data can stimulate innovation, cost and risk reduction and productivity gains [110]. Big data are useful, not only for standard mobile-health (m-health) operations, but also for healthcare investments [29]. Those investments need to match growing expenses, due to aging population trends, with public budget constraints: Hence the importance of big data-driven cost savings [111]. Big data represent an essential source of information for healthcare Project Finance (PF) investments and their data-driven business plans, whose input data increasingly depend on timely and massive information [112].

The description about the healthcare data sources can be observed in Figure A.1. In today's world multiple data sources are used for healthcare records which increases complexity in data and creates more knowledge. It results in complex health standards and big data specifically in terms of volume and variety. In healthcare it is critical and all the semantics needs to be extracted for a complete picture. Standardized Health Documents have critical information and the semantics of the document has to be protected.

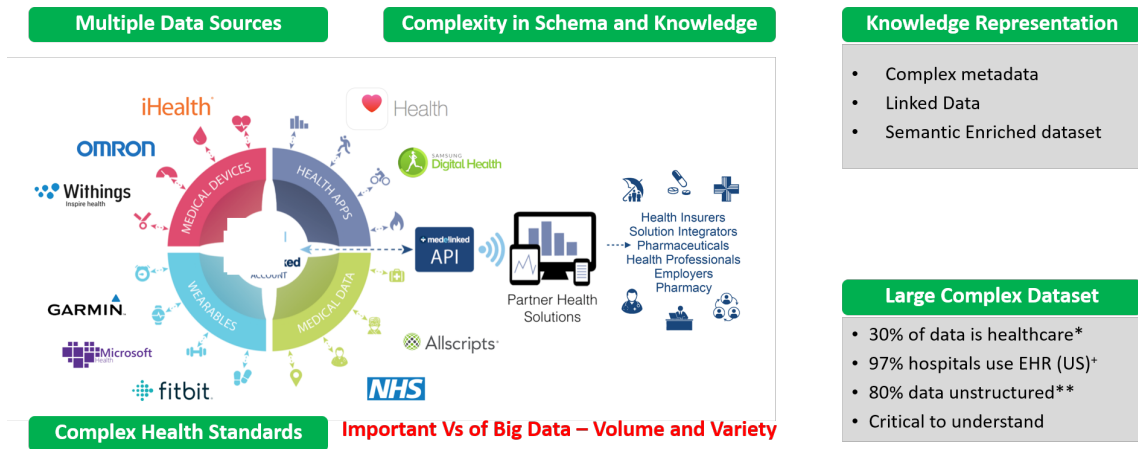


Figure A.1: Healthcare Data Sources

Telemedicine can contribute to reshaping infrastructure, referring to smart facilities (like “intelligent” hospitals) and Telecommunications (TLC) networks [113]. For example, healthcare policymakers can conveniently use networked big data to enrich their infrastructural feasibility plans, whereas private managers may extract valuable information from public databases [41].

A.1 Healthcare standards

Clinical data is very diverse and complex as diseases, observations, operations, drugs, vital signs and lab results etc all come under its umbrella.

Due to this diversity in information systems (electronic health records, disease registries, clinical trial documentations, mortality databases), the data is sometimes incomplete, incorrect and context dependent [16]. Clinical data are usually based or shaped according to the requirements of the systems for which they are collected, like mining, billing, and communicating. The most popular standards were classified in Table A.1. The credit for this table belongs to [16].

Standard Development Organization	Standard	Scope
Federative Committee on Anatomical Terminology (FCAT)	Terminologia Anatomica (TA)	Anatomy terms in English and Latin
Health Level Seven (HL7)	v2	Messaging protocol; several of the chapters of this standard cover clinical content
	v3 (RIM)	Information ontology; especially the “Clinical Statement” work aims to create reusable clinical data standards
	CDA Level 1–3	Information model for clinical documents (embedding of terminology standards in level 2 and 3); especially the Continuity of Care Document (CCD) specifications and the Consolidated CDA (C-CDA) specifications add detail to standards for clinical documents
	FHIR	Information and Document model; several parts of the core specification deal with clinical content
Integrating the Healthcare Enterprise (IHE)	Several Integration profiles	Clinical workflows including references to clinical data standards to be used
International Organization for Standardization (ISO)	TS22220:2011	Identification of subjects of care
	21090:2011	Harmonized data types for information exchange
	13606	High-level description of clinical information models

	23940 (ContSys)	Health care processes for continuity of care
	14155	Clinical investigations
	IDMP	Medicinal products
National Electrical Manufacturers Association (NEMA)	DICOM	
openEHR foundation	openEHR	Clinical information model specification
Regenstrief Institute	LOINC	Terminology for lab and other observables
	UCUM	Standardised representation of units of measure according to the SI units (ISO 80000)
PCHAlliance (Personal Connected Health Alliance)	Continua Design Guidelines	Collecting data from personal health devices
SNOMED International, formerly knowns as the International Health Terminology Standards Development Organisation	SNOMED CT	Terminology / Ontology for representing the electronic health record (“context model” = Information model for SNOMED CT)
World Health Organization (WHO)	ICD-10 / ICD-11	Disease classification
	ICF	Classification of functioning, disability and health
	ICHI	Health procedure classification
	INN	Generic names for pharmaceutical substances
	ATC	Drug ingredient classification
World Organization of Family Doctors (WONCA)	ICPC	Primary care classification

Table A.1: Most common standards in healthcare Adapted from [16]

A.1.1 FHIR

Representational State Transfer (REST) systems as described by Fielding [114] (often referred to as RESTful architectures) have recently been widely adopted as the dominant information abstraction of the World Wide Web. The practical advantages of RESTful architectures include light-weight interfaces that allow for faster transmission and processing of data structures, more suitable for mobile phones and tablet devices [80].



```
{
  "resourceType": "DeviceObservationReport",
  "text": {
    "status": "generated",
    "div": "Device observation report"
  },
  "contained": [
    {
      "resourceType": "Observation",
      "text": {
        "status": "generated",
        "div": "<div>Jan 30 2014: Body Weight = 80 kg</div>"
      },
      "name": {
        "coding": [
          {
            "system": "http://loinc.org",
            "code": "3141-9",
            "display": "MDC_MASS_BODY_ACTUAL"
          }
        ]
      },
      "valueQuantity": {
        "value": 80,
        "units": "kg",
        "system": "http://unitsofmeasure.org",
        "code": "kg"
      },
      "status": "final",
      "reliability": "ok"
    },
    {
      "instant": "2008-12-11T14:45:00",
      "source": {
        "reference": "device/d1"
      },
      "subject": {
        "reference": "Patient/ihe-pcd"
      },
      "virtualDevice": [
        {
          "channel": [
            {
              "metric": [
                {
                  "observation": {
                    "reference": "#o1"
                  }
                }
              ]
            }
          ]
        }
      ]
    }
  ]
}
```

Figure A.2: Example of a DeviceObservationReport (JSON-format) [10]

The FHIR effort aims to simplify and accelerate HL7 adoption by being easily consumable

but robust, and by using open Internet standards where possible like JSON as shown in Figure A.2. Using an easily consumable format for the standard avoids the need for complex custom tooling [80].

FHIR aims to define the key entities involved in healthcare information exchange as resources. Each resource is a distinct identifiable entity. Example resources include: Patient, Device and Document. At the time of this writing there are 32 resources defined with many more under consideration. The development team estimates that there will be approximately 150 resources defined in total. As a resource oriented environment FHIR allows for very simple implementation of base artifacts, their transmission and persistence.

A.1.2 OPENEHR

openEHR is an open standard specification in health informatics that describes the management and storage, retrieval and exchange of health data in electronic health records (EHRs) [115]. In openEHR, all health data for a person is stored in a "one lifetime", vendor-independent, person-centred EHR. The openEHR specifications include an EHR Extract specification [11] but are otherwise not primarily concerned with the exchange of data between EHR-systems as this is the focus of other standards such as EN 13606 and HL7. The openEHR specification components are shown in Figure A.3.

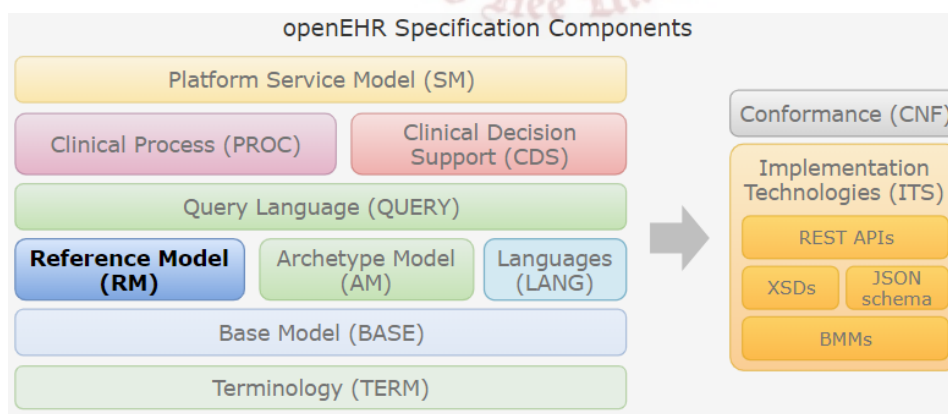


Figure A.3: openEHR specification components [11]

A.1.3 HL7 CDA

The need for a clinical document standard stemmed from the desire to unlock the considerable clinical content currently stored in free-text clinical notes and to enable comparison of content from documents created on information systems of widely varying characteristics. Given the variability in clinical notes, including structure, underlying information models, degree of semantic encoding, use of standard healthcare terminologies, and platform and vendor-specific features, it is currently difficult to store and exchange documents with retention of standardized semantics over both time and distance.

The CDA is a document markup standard that specifies the structure and semantics of “clinical documents.” A clinical document [116] is a documentation of observations and services and has the following defining characteristics:

Persistence. A clinical document continues to exist in an unaltered state, for a time period defined by local and regulatory requirements.

Stewardship. A clinical document is maintained by a person or organization entrusted with its care. Potential for authentication. A clinical document is an assemblage of information that is intended to be legally authenticated.

Wholeness. Authentication of a clinical document applies to the whole and does not apply to portions of the document without full context of the document.

Human readability. A clinical document is human readable.

The header of the CDA document gives the context as shown in A.4. The CDA header enables clinical document exchange across and within institutions for the stakeholders and management of clinical document. It also facilitate ID, category type, title, date, version fields are present there for identification of the document. The header also includes confidentiality status to assist systems in managing access to sensitive data. Confidentiality status can also be applied to specific segments or sections of the document. It can include participants and authors in it. An author can be a person or a device [117]. The CDA header is very comprehensive as the schema shows and handles information on authentication, patient demographics, encounters and other involved parties.

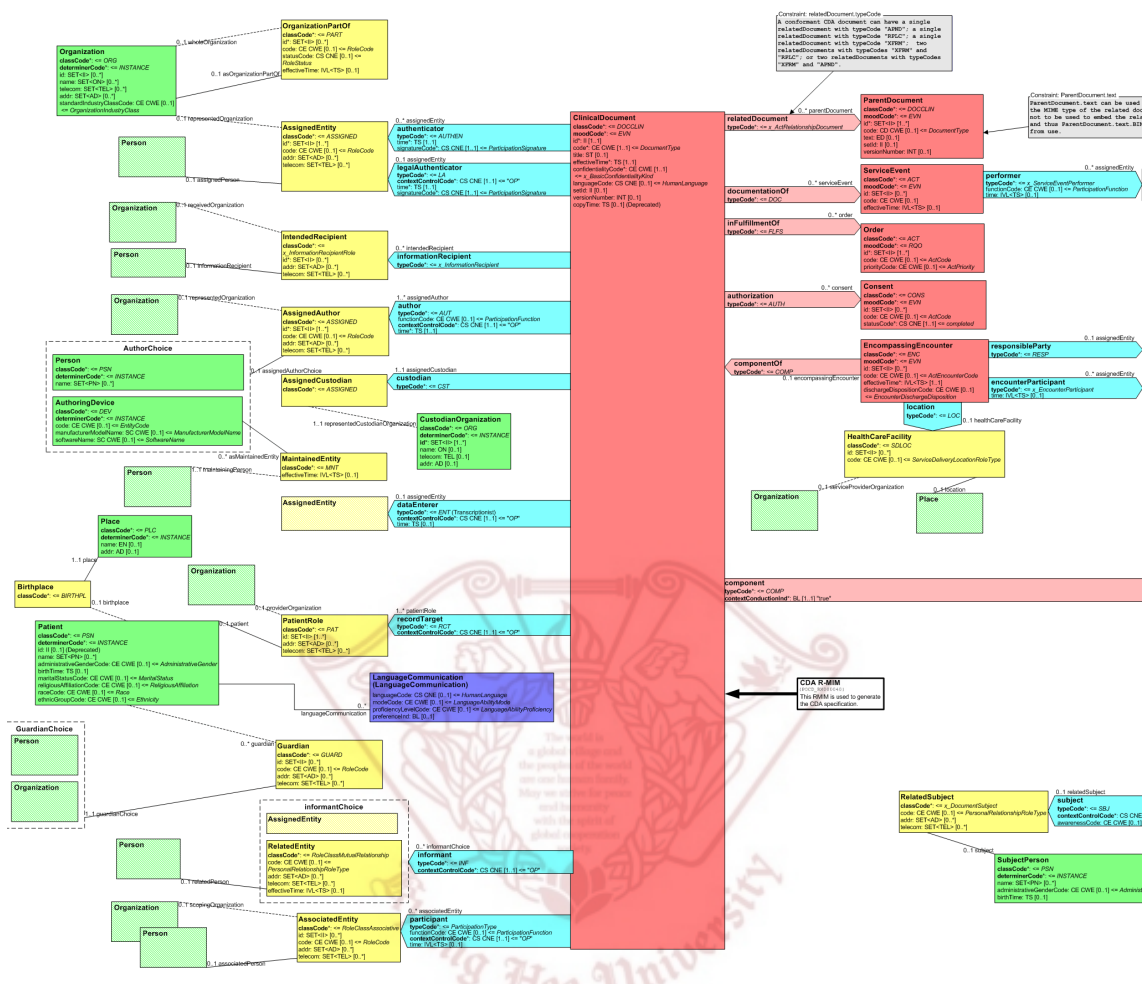
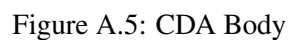


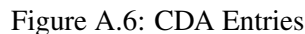
Figure A.4: CDA Header

The CDA body contains the clinical information, and can have structured or unstructured information. Figure A.5 shows a schema for a structured body, which is wrapped by the `<StructuredBody>` element, and which is divided up into recursively nestable document sections which adds to the overall comprehensiveness and complexity. In the CDA body and the document section specifically, the narrative block comprises of content to be rendered, whereas CDA entries represents machine processable structured content. In the narrative block of the section, CDA entries encode clinical contents.



Document entries in a CDA occur in structures and include coded entries(<coded_>, entry>; a recursively nesting wrapper from the nine type of clinical statements. There can be a link between two or more Clinical Statements also allows a relationship to be stated independently of the related Clinical Statements. One kind of clinical statement is observationMedia and the observationMedia Entry provides an in-line graphic depiction of the section findings. Another type of entry is RegionofInterest for referencing image-related spatial coordinates as observations. The clinical statement SubstanceAdministration is used for medication history and medication administration orders. Its consumable participant is played by a LabeledDrug or Material entity in the role of a ManufacturedProduct. The entry relationship is used where the narrative is fully derived from clinical statements.





Collection @ khu


```

<ClinicalDocument>
  ... CDA Header ...
  <structuredBody>
    <section>
      <text>(a.k.a. "narrative block")</text>
      <observation>...</observation>
      <substanceAdministration>
        <supply>...</supply>
      </substanceAdministration>
      <observation>
        <externalObservation>...
      </externalObservation>
    </section>
    <section>
      <section>...</section>
    </section>
  </structuredBody>
</ClinicalDocument>

```

Figure A.7: Major Components of Clinical Document Architecture [12]

Figure A.7 shows two `<observation>`, CDA entries and `<substanceAdministration>`. entry containing a nested `<supply>` entry, although several other CDA entries are defined. These entries are derived from classes in the RIM and enable formal representation of clinical statements in the narrative.

A.2 Big Data Frameworks

Huge amount of data is generated from multiple sources in recent times. Almost two exabytes of data is being generated on the internet everyday [118]. In one minute, three days worth of videos are uploaded to Youtube, 30,000 new posts are created on the Tumblr blog platform, 100,000 Tweets are shared on Twitter and more than 200,000 pictures are posted on Facebook [119]. The key features for big data frameworks are (1) the programming model, (2) the supported programming languages, (3) the type of data sources, (4) the compatibility of the framework with existing machine learning libraries, and (5) the fault tolerance strategy.

A.2.1 Spark

Apache Spark is a powerful processing framework that provides an ease of use tool for efficient analytics of heterogeneous data. It was originally developed at UC Berkeley in 2009 [13]. Spark has several advantages compared to other Big Data frameworks like Hadoop and storm. Spark is used by many companies such as Yahoo, Baidu, and Tencent. A key concept of Spark is Resilient Distributed Datasets (RDDs). An RDD is basically an immutable collection of objects spread across a Spark cluster. In Spark, there are two types of operations on RDDs: (1) transformations and (2) actions. Transformations consist in the creation of new RDDs from existing ones using functions like map, filter, union and join.

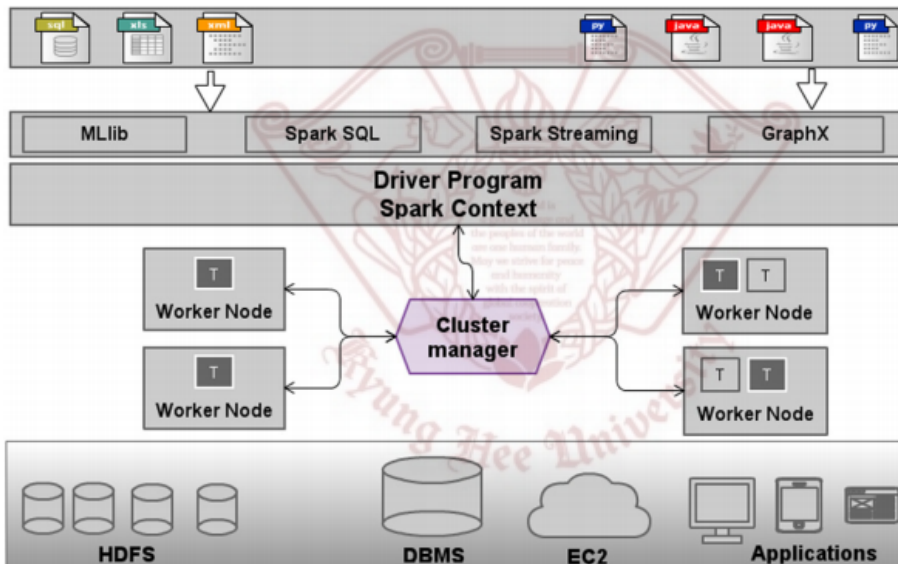


Figure A.8: Spark System Overview [13]

Actions consist of final result of RDD computations. In Figure A.8, a Spark architecture overview is shown. There are three main components in a spark cluster which is based on master/slave architecture:

The slave node is represented by the Driver Program. It maintains an object called Spark-Context that manages and supervises running applications. The cluster manager is responsible for managing the application workflow assigned by Driver Program to workers. All the resources in

clusters are controlled and supervised by the cluster manager and returns their state to the Driver Program. Worker Nodes: During the execution of a Spark program each Worker Node represents a container of one operation.

A.2.2 Hadoop (extensions) YARN

Hadoop is an Apache project founded in 2008 by Doug Cutting at Yahoo and Mike Cafarella at the University of Michigan [64]. Hadoop consists of two main components: (1) Hadoop Distributed File System (HDFS) for data storage and (2) Hadoop MapReduce, an implementation of the MapReduce programming model [120]. Hadoop MapReduce has two main versions. In the first version Hadoop MapReduce has two main components i.e. the task tracker and job tracker. The Task Tracker mainly manages the execution of the Map and Reduce functions whereas the Job Tracker represents the master and allows resource management and job scheduling/monitoring. It manages the Task Trackers [67]. The second version of Hadoop is called YARN and Job Tracker has two major features that have been split into separate daemons i.e. a global Resource Manager and per-application Application Master. In Figure A.9, an illustration is shown for the overall architecture of YARN. As shown in Figure A.9, the MapReduce jobs are received and run by the Resource Manager. The ResourceManager allocates resources to the Application Master which then works with the Node Manager(s) to execute and monitor the tasks. In YARN, the Resource Manager (respectively the Node Manager) replaces the Job Tracker (respectively the Task Tracker) [14].

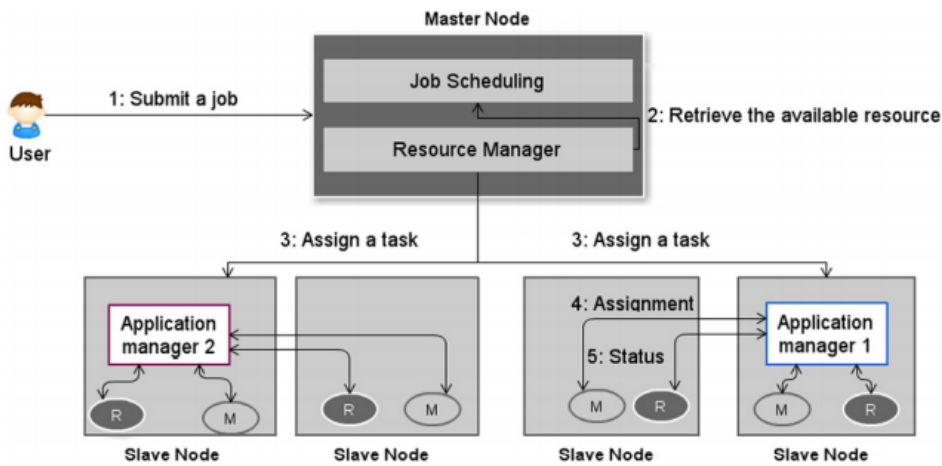


Figure A.9: YARN architecture [14]

A.3 Relationship between Healthcare standards and Big Data framework

The challenges in clinical data is not only growing volume but the diverse nature of data that is being collected. It is collected in different formats and stored in various databases as it is very hard to get the stakeholder organizations to agree on a data format (structured or unstructured). A literature research was performed and the most common definitions of big data are summarized. Most of the literature focuses on the four ‘V’ concepts: volume, variety, velocity, and veracity [121]. Clinical data is still stored in conventional RDBMS and spreadsheets but now data can come under the form of free text (electronic report) or images (patients’ scans). This kind of data can be classified as structured or semi-structured (missing values or inconsistencies). Different sources: variety is also used to mean that data can come from different sources and they don’t usually come from the same institution. Due to advancements in the healthcare machinery, a large amount of images are produced in a short time

Bibliography

- [1] S. Hussain and S. Lee, "Semantic transformation model for clinical documents in big data to support healthcare analytics," in *Digital Information Management (ICDIM), 2015 Tenth International Conference on*. IEEE, 2015, pp. 99–102.
- [2] W. Lijun, H. Yongfeng, C. Ji, Z. Ke, and L. Chunhua, "Medoop: A medical information platform based on hadoop," in *e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on*. IEEE, 2013, pp. 1–6.
- [3] S. K. Bansal, "Towards a semantic extract-transform-load (etl) framework for big data integration," in *2014 IEEE International Congress on Big Data*. IEEE, 2014, pp. 522–529.
- [4] S. Yang, R. Wei, J. Guo, and L. Xu, "Semantic inference on clinical documents: combining machine learning algorithms with an inference engine for effective clinical diagnosis and treatment," *IEEE Access*, vol. 5, pp. 3529–3546, 2017.
- [5] M. Panahiazar, V. Taslimitehrani, A. Jadhav, and J. Pathak, "Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases," in *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, 2014, pp. 790–795.
- [6] J. A. Basco and N. C. Senthilkumar, "Real-time analysis of healthcare using big data analytics," *IOP Conference Series: Materials Science and Engineering*, vol. 263, p. 042056, nov 2017.
- [7] S. Hussain, B. H. Kang, and S. Lee, "A wearable device-based personalized big data analysis model," in *International Conference on Ubiquitous Computing and Ambient Intelligence*. Springer, 2014, pp. 236–242.

- [8] F. Firouzi, A. M. Rahmani, K. Mankodiya, M. Badaroglu, G. V. Merrett, P. Wong, and B. Farahani, "Internet-of-things and big data for smarter healthcare: from device to architecture, applications and analytics," 2018.
- [9] Y. Wang, L. Kung, W. Y. C. Wang, and C. G. Cegielski, "An integrated big data analytics-enabled transformation model: Application to health care," *Information & Management*, vol. 55, no. 1, pp. 64–79, 2018.
- [10] B. Franz, A. Schuler, and O. Krauss, "Applying fhir in an integrated health monitoring system," *EJBI*, vol. 11, no. 2, pp. 51–56, 2015.
- [11] https://specifications.openehr.org/releases/RM/latest/ehr_extract.html, [Online; accessed 19-April-2019].
- [12] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, and A. Shabo, "HL7 clinical document architecture, release 2," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 30–39, 2006.
- [13] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets." *HotCloud*, vol. 10, no. 10-10, p. 95, 2010.
- [14] R. Li, H. Hu, H. Li, Y. Wu, and J. Yang, "Mapreduce parallel programming model: a state-of-the-art survey," *International Journal of Parallel Programming*, vol. 44, no. 4, pp. 832–866, 2016.
- [15] N. Szlezak, M. Evers, J. Wang, and L. Pérez, "The role of big data and advanced analytics in drug discovery, development, and commercialization," *Clinical Pharmacology & Therapeutics*, vol. 95, no. 5, pp. 492–495, 2014.
- [16] W. R. Hersh, M. G. Weiner, P. J. Embi, J. R. Logan, P. R. Payne, E. V. Bernstam, H. P. Lehmann, G. Hripcsak, T. H. Hartzog, J. J. Cimino *et al.*, "Caveats for the use of operational electronic health record data in comparative effectiveness research," *Medical care*, vol. 51, no. 8 0 3, p. S30, 2013.

- [17] I. M. Mullins, M. S. Siadaty, J. Lyman, K. Scully, C. T. Garrett, W. G. Miller, R. Muller, B. Robson, C. Apte, S. Weiss *et al.*, “Data mining and clinical data repositories: Insights from a 667,000 patient data set,” *Computers in biology and medicine*, vol. 36, no. 12, pp. 1351–1377, 2006.
- [18] “Healthcare information and management systems society (himss),” *Retrieved March 2019 from <https://www.himss.org/>*, 2014.
- [19] <http://runkeeper.com/>, [Online; accessed October2019].
- [20] <https://www.noom.com/>, [Online; accessed October2019].
- [21] <http://www.azumio.com/s/argus/index.html>, [Online; accessed October2019].
- [22] <https://www.myfitnesspal.com/>, [Online; accessed October2019].
- [23] <https://www.fitocracy.com>, [Online; accessed October2019].
- [24] F. Martin-Sanchez, V. Aguiar-Pulido, G. Lopez-Campos, N. Peek, and L. Sacchi, “Secondary use and analysis of big data collected for patient care,” *Yearbook of medical informatics*, vol. 26, no. 01, pp. 28–37, 2017.
- [25] Y. Wang and N. Hajli, “Exploring the path to big data analytics success in healthcare,” *Journal of Business Research*, vol. 70, pp. 287–299, 2017.
- [26] B. Cyganek, M. Graña, B. Krawczyk, A. Kasprzak, P. Porwik, K. Walkowiak, and M. Woźniak, “A survey of big data issues in electronic health record analysis,” *Applied Artificial Intelligence*, vol. 30, no. 6, pp. 497–520, 2016.
- [27] F. F. Costa, “Big data in biomedicine,” *Drug discovery today*, vol. 19, no. 4, pp. 433–440, 2014.
- [28] D. R. Leff and G.-Z. Yang, “Big data for precision medicine,” *Engineering*, vol. 1, no. 3, pp. 277–279, 2015.
- [29] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” *Health information science and systems*, vol. 2, no. 1, p. 3, 2014.

- [30] J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, “Big data analytics to improve cardiovascular care: promise and challenges,” *Nature Reviews Cardiology*, vol. 13, no. 6, p. 350, 2016.
- [31] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, “Big data: The next frontier for innovation, competition,” *Washington, DC: McKinsey Global Institute*, 2011.
- [32] L. Haar, “Big data expected to have big impact on diagnostic imaging,” 2014.
- [33] A. Asante-Korang and J. P. Jacobs, “Big data and paediatric cardiovascular disease in the era of transparency in healthcare,” *Cardiology in the Young*, vol. 26, no. 8, pp. 1597–1602, 2016.
- [34] C. Auffray, R. Balling, I. Barroso, L. Bencze, M. Benson, J. Bergeron, E. Bernal-Delgado, N. Blomberg, C. Bock, A. Conesa *et al.*, “Making sense of big data in health research: towards an eu action plan,” *Genome medicine*, vol. 8, no. 1, p. 71, 2016.
- [35] S. Fodeh and Q. Zeng, “Mining big data in biomedicine and health care.” *Journal of biomedical informatics*, vol. 63, p. 400, 2016.
- [36] J. Wu, H. Li, S. Cheng, and Z. Lin, “The promising future of healthcare services: When big data analytics meets wearable technology,” *Information & Management*, vol. 53, no. 8, pp. 1020–1033, 2016.
- [37] G. Asokan and V. Asokan, “Leveraging “big data” to enhance the effectiveness of “one health” in an era of health informatics,” *Journal of epidemiology and global health*, vol. 5, no. 4, pp. 311–314, 2015.
- [38] M. Grossglauser and H. Saner, “Data-driven healthcare: from patterns to actions,” *European journal of preventive cardiology*, vol. 21, no. 2_suppl, pp. 14–17, 2014.
- [39] D. V. Dimitrov, “Medical internet of things and big data in healthcare,” *Healthcare informatics research*, vol. 22, no. 3, pp. 156–163, 2016.

- [40] B. E. Huang, W. Mulyasasmita, and G. Rajagopal, "The path from big data to precision medicine," *Expert Review of Precision Medicine and Drug Development*, vol. 1, no. 2, pp. 129–143, 2016.
- [41] C. S. Kruse, R. Goswamy, Y. J. Raval, and S. Marawi, "Challenges and opportunities of big data in health care: a systematic review," *JMIR medical informatics*, vol. 4, no. 4, p. e38, 2016.
- [42] H. Geerts, P. A. Dacks, V. Devanarayan, M. Haas, Z. S. Khachaturian, M. F. Gordon, S. Maudsley, K. Romero, D. Stephenson, B. H. M. Initiative *et al.*, "Big data to smart data in alzheimer's disease: The brain health modeling initiative to foster actionable knowledge," *Alzheimer's & Dementia*, vol. 12, no. 9, pp. 1014–1021, 2016.
- [43] R. Budhiraja, R. Thomas, M. Kim, and S. Redline, "The role of big data in the management of sleep-disordered breathing," *Sleep medicine clinics*, vol. 11, no. 2, pp. 241–255, 2016.
- [44] M. Cox and D. Ellsworth, "Application-controlled demand paging for out-of-core visualization," in *Proceedings. Visualization'97 (Cat. No. 97CB36155)*. IEEE, 1997, pp. 235–244.
- [45] S. R. Sukumar, R. Natarajan, and R. K. Ferrell, "Quality of big data in health care," *International journal of health care quality assurance*, vol. 28, no. 6, pp. 621–634, 2015.
- [46] I. D. Dinov, "Volume and value of big healthcare data," *Journal of medical statistics and informatics*, vol. 4, 2016.
- [47] N. Peek, J. Holmes, and J. Sun, "Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics," *Yearbook of medical informatics*, vol. 23, no. 01, pp. 42–47, 2014.
- [48] S. Salas-Vega, A. Haimann, and E. Mossialos, "Big data and health care: challenges and opportunities for coordinated policy development in the eu," *Health Systems & Reform*, vol. 1, no. 4, pp. 285–300, 2015.
- [49] E. Capobianco, "Systems and precision medicine approaches to diabetes heterogeneity: a big data perspective," *Clinical and translational medicine*, vol. 6, no. 1, p. 23, 2017.

- [50] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.
- [51] H. M. Krumholz, "Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system," *Health Affairs*, vol. 33, no. 7, pp. 1163–1170, 2014.
- [52] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "Maui: making smartphones last longer with code offload," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, 2010, pp. 49–62.
- [53] A. Bell, P. Rogers, C. Farnell, B. Sparkman, and S. C. Smith, "Wireless patient monitoring system," in *Healthcare Innovation Conference (HIC), 2014 IEEE*. IEEE, 2014, pp. 149–152.
- [54] E. A. Mohammed, B. H. Far, and C. Naugler, "Applications of the mapreduce programming framework to clinical big data analysis: current landscape and future trends," *BioData mining*, vol. 7, no. 1, p. 22, 2014.
- [55] H. W.C, ". big data management, technologies, and applications," *IGI Global*, 2013.
- [56] S. Mohanty, M. Jagadeesh, and H. Srivatsa, *Big data imperatives: Enterprise 'Big Data'warehouse, 'BI'implementations and analytics*. Apress, 2013.
- [57] H. U. Buhl, M. Röglinger, F. Moser, and J. Heidemann, "Big data," *WIRTSCHAFTSINFORMATIK*, vol. 55, no. 2, pp. 63–68, Apr 2013. [Online]. Available: <https://doi.org/10.1007/s11576-013-0350-x>
- [58] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, "Big data: the management revolution," *Harvard business review*, vol. 90, no. 10, pp. 60–68, 2012.

- [59] M. Idris, S. Hussain, M. H. Siddiqi, W. Hassan, H. S. M. Bilal, and S. Lee, "Mrpack: Multi-algorithm execution using compute-intensive approach in mapreduce," *PloS one*, vol. 10, no. 8, p. e0136259, 2015.
- [60] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018.
- [61] M. D. Huesch and T. J. Mosher, "Using it or losing it? the case for data scientists inside health care," *Retrieved April 2019 from <https://catalyst.nejm.org/case-data-scientists-inside-health-care/>*, 2017.
- [62] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," *International Journal of Production Economics*, vol. 165, pp. 234–246, 2015.
- [63] V. Slavov, P. Rao, S. Paturi, T. K. Swami, M. Barnes, D. Rao, and R. Palvai, "A new tool for sharing and querying of clinical documents modeled using hl7 version 3 standard," *Computer methods and programs in biomedicine*, vol. 112, no. 3, pp. 529–552, 2013.
- [64] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.
- [65] D. Borthakur *et al.*, "Hdfs architecture guide," *Hadoop Apache Project*, vol. 53, 2008.
- [66] M. Idris, S. Hussain, M. Ali, A. Abdulali, M. H. Siddiqi, B. H. Kang, and S. Lee, "Context-aware scheduling in mapreduce: a compact review," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 17, pp. 5332–5349, 2015.
- [67] T. White, *Hadoop: The definitive guide*. "O'Reilly Media, Inc.", 2012.
- [68] O. O'Malley, "Programming with hadoop's map/reduce," *ApacheCon EU*, 2008.
- [69] I. Indrajit and B. Verma, "Dicom, hl7 and ihe: A basic primer on healthcare standards for radiologists," *Indian Journal of Radiology and Imaging*, vol. 17, no. 2, p. 66, 2007.

- [70] C. J. McDonald, S. M. Huff, J. G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook *et al.*, “Loinc, a universal standard for identifying laboratory observations: a 5-year update,” *Clinical chemistry*, vol. 49, no. 4, pp. 624–633, 2003.
- [71] K. A. Spackman, K. E. Campbell, and R. A. Côté, “Snomed rt: a reference terminology for health care.” in *Proceedings of the AMIA annual fall symposium*. American Medical Informatics Association, 1997, p. 640.
- [72] A. Hasman *et al.*, “HI7 rim: an incoherent standard,” in *Ubiquity: Technologies for Better Health in Aging Societies, Proceedings of Mie2006*, vol. 124, 2006, p. 133.
- [73] R. Margaret, “Clinical document architecture (cda),” Retrieved April 2019 from <https://searchhealthit.techtarget.com/definition/Clinical-Document-Architecture-CDA>.
- [74] —, “Meaningful clinical document architecture (cda),” Retrieved April 2019 from <https://searchhealthit.techtarget.com/definition/meaningful-use>.
- [75] J. M. Ferranti, R. C. Musser, K. Kawamoto, and W. E. Hammond, “The clinical document architecture and the continuity of care record: a critical analysis,” *Journal of the American Medical Informatics Association*, vol. 13, no. 3, pp. 245–252, 2006.
- [76] HL7, “Ccd-continuity of care document,” Retrieved April 2019 from <https://corepointthealth.com/resource-center/hl7-resources/ccd/>.
- [77] T. Benson and G. Grieve, *Principles of health interoperability: SNOMED CT, HL7 and FHIR*. Springer, 2016.
- [78] G. Greive, “Fhir cda position statement and roadmap joint statement with lantana. health intersections blog,” Retrieved April 2019 from <http://www.healthintersections.com.au/?p=2202>, 2014.
- [79] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni, “Smart on fhir: a standards-based, interoperable apps platform for electronic health records,” *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 899–908, 2016.

- [80] D. Bender and K. Sartipi, "HI7 fhir: An agile and restful approach to healthcare information exchange," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. IEEE, 2013, pp. 326–331.
- [81] H. Horiguchi, H. Yasunaga, H. Hashimoto, and K. Ohe, "A user-friendly tool to transform large scale administrative data into wide table format using a mapreduce program with a pig latin based script," *BMC medical informatics and decision making*, vol. 12, no. 1, p. 151, 2012.
- [82] P. Kale and A. Mohanpurkar, "Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases," in *International Journal of Computer Science and Information Technologies*,), 2015, pp. 2871–2875.
- [83] A. Bahga and V. K. Madiseti, "A cloud-based approach for interoperable electronic health records (ehrs)," *Biomedical and Health Informatics, IEEE Journal of*, vol. 17, no. 5, pp. 894–906, 2013.
- [84] D. Du, *Apache Hive Essentials*. Packt Publishing Ltd, 2015.
- [85] A. H. Team, "Apache hbase reference guide," *Apache, version*, vol. 2, no. 0, 2016.
- [86] Y. Wang, L. Wang, H. Liu, and C. Lei, "Large-scale clinical data management and analysis system based on cloud computing," in *Frontier and Future Development of Information Technology in Medicine and Education*. Springer, 2014, pp. 1575–1583.
- [87] C. Doukas, T. Pliakas, and I. Maglogiannis, "Mobile healthcare information management utilizing cloud computing and android os," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 1037–1040.
- [88] C. Paniagua, H. Flores, and S. N. Srirama, "Mobile sensor data classification for human activity recognition using mapreduce on cloud," *Procedia Computer Science*, vol. 10, pp. 585–592, 2012.
- [89] <https://www.iotone.com/software/predixion-insight/s293/>, [Online; accessed October2019].

- [90] N. Mehta and A. Pandit, "Concurrence of big data analytics and healthcare: A systematic review," *International journal of medical informatics*, vol. 114, pp. 57–65, 2018.
- [91] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.
- [92] M. J. Duncan, H. M. Badland, and W. K. Mummery, "Applying gps to enhance understanding of transport-related physical activity," *Journal of Science and Medicine in Sport*, vol. 12, no. 5, pp. 549–556, 2009.
- [93] <http://nikeplus.nike.com/plus>, [Online; accessed 19-April-2019].
- [94] <http://www.apple.com/ipod/nike/run.html>, [Online; accessed 19-April-2019].
- [95] <http://www.mapmyrun.com/>, [Online; accessed 19-April-2019].
- [96] <http://www.dropbox.com/>, [Online; accessed 19-April-2019].
- [97] M. Han, Y.-K. Lee, S. Lee *et al.*, "Comprehensive context recognizer based on multimodal sensors in a smartphone," *Sensors*, vol. 12, no. 9, pp. 12 588–12 605, 2012.
- [98] I. Cleland, M. Han, C. Nugent, H. Lee, S. Zhang, S. McClean, and S. Lee, "Mobile based prompted labeling of large scale activity data," in *Ambient Assisted Living and Active Aging*. Springer, 2013, pp. 9–17.
- [99] M. Han, J. H. Bang, C. Nugent, S. McClean, and S. Lee, "Harf: A hierarchical activity recognition framework using smartphone sensors," in *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction*. Springer, 2013, pp. 159–166.
- [100] W. Wang, K. Haerian, H. Salmasian, R. Harpaz, H. Chase, and C. Friedman, "A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from pubmed citations," in *AMIA annual symposium proceedings*, vol. 2011. American Medical Informatics Association, 2011, p. 1464.

- [101] C.-L. Hung and Y.-L. Lin, "Implementation of a parallel protein structure alignment service on cloud," *International journal of genomics*, vol. 2013, 2013.
- [102] L. Wang, D. Chen, R. Ranjan, S. U. Khan, J. Kolodziej, and J. Wang, "Parallel processing of massive eeg data with mapreduce," in *2012 IEEE 18th International Conference on Parallel and Distributed Systems*. Ieee, 2012, pp. 164–171.
- [103] B. Meng, G. Pratz, and L. Xing, "Ultrafast and scalable cone-beam ct reconstruction using mapreduce in a cloud computing environment," *Medical physics*, vol. 38, no. 12, pp. 6603–6609, 2011.
- [104] D. Markonis, R. Schaer, I. Eggel, H. Müller, and A. Depeursinge, "Using mapreduce for large-scale medical image analysis," in *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*. IEEE, 2012, pp. 1–1.
- [105] H. Choi, K.-H. Lee, and Y.-J. Lee, "Parallel labeling of massive xml data with mapreduce," *The Journal of Supercomputing*, vol. 67, no. 2, pp. 408–437, 2014.
- [106] R. Kernan, "Clinical document architecture (cda),consolidated-cda (ccda) and their role in meaningful use (mu)," *Retrieved April 2019 from <https://www.healthit.gov/>*, 2012.
- [107] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.
- [108] J. Mandel, "Repository for sample ccda documents," *Retrieved April 2019 from <https://github.com/jmandel>*.
- [109] J. Li, W. Ding, H. Cheng, P. Chen, D. Di, and W. Huang, "A comprehensive literature review on big data in healthcare," 2016.
- [110] I. de la Torre Díez, H. M. Cosgaya, B. Garcia-Zapirain, and M. López-Coronado, "Big data in health: a literature review from the year 2005," *Journal of medical systems*, vol. 40, no. 9, p. 209, 2016.
- [111] J. Archenaa and E. M. Anita, "A survey of big data analytics in healthcare and government," *Procedia Computer Science*, vol. 50, pp. 408–413, 2015.

- [112] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information sciences*, vol. 275, pp. 314–347, 2014.
- [113] R. Moro Visconti and D. Morea, "Big data for the sustainability of healthcare project financing," *Sustainability*, vol. 11, no. 13, p. 3748, 2019.
- [114] R. T. Fielding and R. N. Taylor, *Architectural styles and the design of network-based software architectures*. University of California, Irvine Doctoral dissertation, 2000, vol. 7.
- [115] <https://en.wikipedia.org/wiki/OpenEHR>, [Online; accessed 19-April-2019].
- [116] R. H. Dolin, L. Alschuler, C. Beebe, P. V. Biron, S. L. Boyer, D. Essin, E. Kimber, T. Lincoln, and J. E. Mattison, "The hl7 clinical document architecture," *Journal of the American Medical Informatics Association*, vol. 8, no. 6, pp. 552–569, 2001.
- [117]
- [118] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International journal of information management*, vol. 35, no. 2, pp. 137–144, 2015.
- [119] W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, and E. M. Nguifo, "An experimental survey on big data frameworks," *Future Generation Computer Systems*, vol. 86, pp. 546–564, 2018.
- [120] I. Polato, R. Ré, A. Goldman, and F. Kon, "A comprehensive view of hadoop research—a systematic literature review," *Journal of Network and Computer Applications*, vol. 46, pp. 1 – 25, 2014.
- [121] P. Kubben, M. Dumontier, and A. Dekker, *Fundamentals of Clinical Data Science*. Springer, 2019.

International Journal Papers:

- [1] **Shujaat Hussain**, Maqbool Hussain, Muhammad Afzal, Jamil Hussain, Jaehun Bang, Hyonwoo Seung and Sungyoung Lee, "Semantic Preservation of standardized healthcare documents in big data", International Journal of Medical Informatics (SCI, IF:2.731), Vol. 129, pp.133-145, 2019
- [2] **Shujaat Hussain**, Jae Hun Bang, Manhyung Han, Muhammad Idris Ahmed, Muhammad Bilal Amin, Sungyoung Lee, Chris Nugent, Sally McClean, Bryan Scotney and Gerard Parr, "Behavior Life Style analysis for mobile sensory data in cloud computing through MapReduce", Sensors (SCIE, IF:2.048), Vol.14, No.11, pp.22001-22020, 2014
- [3] Jamil Hussain, Anees Ul Hassan, Hafiz Syed Muhammad Bilal, Muhammad Afzal, **Shujaat Hussain**, Jaehun Bang, Oresti Banos and Sungyoung Lee, "Model-based adaptive user interface based on context and user experience evaluation", Journal on Multimodal User Interfaces (SCIE, IF: 1.031), Vol.12, Issue 1, pp.1-16, 2018
- [4] Dinh-Mao Bui, **Shujaat Hussain**, Eui-Nam Huh and Sungyoung Lee, "Adaptive Replication Management in HDFS based on Supervised Learning", IEEE Transactions on Knowledge and Data Engineering (SCI, IF:2.067), Vol.28, Issue 6, pp.1369-1382, 2016
- [5] Mahmood Ahmad, Muhammad Bilal Amin, **Shujaat Hussain**, Byeong Ho Kang, Taechoong Cheong and Sungyoung Lee, "Health Fog: a novel framework for health and wellness applications", Journal of Supercomputing, (SCI, IF:0.858), Vol.72, Issue 10, pp.3677-3695, 2016

- [6] Muhammad Bilal Amin, Oresti Banos, Wajahat Ali Khan, Hafiz Syed Muhammad Bilal, Jinhyuk Gong, Dinh-Mao Bui, Soung Ho Cho, **Shujaat Hussain**, Taqdir Ali, Usman Akhtar, Tae Choong Chung and Sungyoung Lee, "On Curating Multimodal Sensory Data for Health and Wellness Platforms", *Sensors* (SCIE, IF: 2.033), vol. 16,no. 7, doi:10.3390/s16070980 , 2016
- [7] Muhammad Bilal Amin, Wajahat Ali Khan, **Shujaat Hussain**,Dinh-Mao Bui,Oresti Bano,Byeong Ho Kang and Sungyoung Lee, "Evaluating Large-Scale Biomedical Ontologies Over Parallel Platforms", *IETE Technical Reviews*(SCIE, IF:0.88), DOI:10.1080/02564602.2015.1117399, 2015
- [8] Muhammad Idris, **Shujaat Hussain**, Muhammad Hameed Siddiqi, Waseem Hassan, Hafiz Syed Muhammad Bilal and Sungyoung Lee, "MRPack: Multi-Algorithm Execution Using Compute-Intensive Approach in MapReduce", *PLoS One*(SCIE, IF:3.234), Vol.10, No.8, DOI: 10.1371/journal.pone.0136259, 2015.
- [9] Muhammad Idris, **Shujaat Hussain**, Maqbool Ali, Arsen Abdulali, Muhammad Hameed Siddiqi, Byeong Ho Kang and Sungyoung Lee, "Context-aware scheduling in MapReduce: a compact review", *Concurrency and Computation: Practice and Experience* (SCIE, IF: 0.997) Published online in Wiley Online Library, DOI: 10.1002/cpe.3578, 2015
- [10] Rahman Ali, Muhammad Hameed Siddiqi, Muhammad Idris Ahmed,Taqdir Ali, **Shujaat Hussain**, Eui-Nam Huh, Byeong Ho Kang and Sungyoung Lee, "GUDM: Automatic Generation of Unified Datasets for Learning and Reasoning in Healthcare", *Sensors* (SCIE, IF: 2.245), Vol.15, No.7, pp.15772-15798, 2015.

International Conference Papers:

- [11] **Shujaat Hussain** and Sungyoung Lee, "Visualization and descriptive analytics of wellness data through Big Data", *The Tenth International Conference on Digital Information Management (ICDIM 2015)*, Jeju, Korea, Oct 21-23, 2015

- [12] **Shujaat Hussain** and Sungyoung Lee, "Semantic transformation model for clinical documents in big data to support healthcare analytics", The Tenth International Conference on Digital Information Management (ICDIM 2015), Jeju, Korea, Oct 21-23, 2015
- [13] **Shujaat Hussain**, Byeong Ho Kang and Sungyoung Lee, "A wearable device-based personalized big data analysis model", UCAMI 2014, Belfast, Ireland, Dec 2-5, 2014
- [14] **Shujaat Hussain**, Manhyung Han, Jae Hun Bang, Chris Nugent, Sally McClean and Bryan Scotney, "Activity recognition and resource optimization in mobile cloud through MapReduce", 15th International Conference on E-Health Networking, application & service, Oct 9-12, 2013
- [15] **Shujaat Hussain**, Muhammad Bilal Amin, Zeeshan Pervez, Ammar Ahmad Awan, Sungyoung Lee. "A hybrid cloud based smart home". AAL SUMMIT 2012.
- [16] Wajahat Ali Khan, Muhammad Bilal Amin, Oresti Banos, Taqdir Ali, Maqbool Hussain, Muhammad Afzal, **Shujaat Hussain**, Jamil Hussain, Rahman Ali, Maqbool Ali, Dongwook Kang, Jaehun Bang, Tae Ho Hur, Bilal Ali, Muhammad Idris, Asif Razzaq, Sungyoung Lee and Byeong Ho Kang, "Mining Minds: Journey of Evolutionary Platform for Ubiquitous Wellness", 12th International Conference on Ubiquitous Healthcare (u-Healthcare 2015), Osaka, Japan, Nov 30- Dec 2, 2015
- [17] Muhammad Idris, **Shujaat Hussain**, Mahmood Ahmad and Sungyoung Lee, "Big Data Service Engine (BISE): Integration of Big Data Technologies for Human Centric Wellness Data", Second International Conference on Big Data and Smart Computing, Jeju, Korea, Feb 9-12, 2015
- [18] Oresti Banos, Muhammad Bilal Amin, Wajahat Ali Khan, Muhammad Afzel, Mahmood Ahmad, Maqbool Ali, Taqdir Ali, Rahman Ali, Muhammad Bilal, Manhyung Han, Jamil Hussain, Maqbool Hussain, **Shujaat Hussain**, Tae Ho Hur, Jae Hun Bang, Thien Huynh-The, Muhammad Idris, Dong Wook Kang, Sang Beom Park, Hameed Siddiqui, Le-Ba Vui, Muhammad Fahim, Asad Masood Khattak, Byeong Ho Kang and Sungyoung Lee, "An

Innovative Platform for Person-Centric Health and Wellness Support”, 3rd International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2015), Granada, Spain, April 15-17, 2015

- [19] Muhammad Idris, **Shujaat Hussain** and Sungyoung Lee, ”In-Map/In-Reduce: Concurrent Job Execution in MapReduce”, 2014 IEEE TrustCom, Beijing, China, Sep 24-26, 2014
- [20] Muhammad Idris, **Shujaat Hussain**, Taqdir Ali, Byeong Ho Kang and Sungyoung Lee, ”Semantics Based Intelligent Search in Large Digital Repositories using Hadoop MapReduce”, UCAmI 2014, Belfast, Ireland, Dec 2-5, 2014
- [21] **Shujaat Hussain**, Byeong Ho Kang and Sungyoung Lee, ”A wearable device-based personalized big data analysis model”, UCAmI 2014, Belfast, Ireland, Dec 2-5, 2014
- [22] Ammar Ahmad Awan, Muhammad Bilal Amin, **Shujaat Hussain**, Aamir Shafi and Sungyoung Lee, ”An MPI-IO Compliant Java based Parallel I/O Library”, 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid2013), Delft, Netherlands, May 13-16, 2013
- Domestic Journal Papers:**
- [23] Muhammad Bilal Amin, Wajahat Ali Khan, Bilal Ali Rizvi, Jaehun Bang, Muhammad Sadiq, Taqdir Ali, Taeho Hur, **Shujaat Hussain**, Imran Ali, Dohyung Kim and Sungyoung Lee., ”Health and Wellness platforms: A Survey on Services and Enabling Technologies”, The Journal of The Korean Institute of Communication Sciences, Vol.35, pp.9-25, 2018.
- [24] Wajahat Ali Khan, Syed Imran Ali, Muhammad Bilal Ali, **Shujaat Hussain**, Huh Tae Ho, Lee Seongyoung, ”Healthcare Standard Interoperability”, The Journal of The Korean Institute of Communication Sciences, Vol.35, pp.73-84, 2018.

Domestic Patents:

- [25] Sungyoung Lee, and **Shujaat Hussain**, ”A wearable device-based personalized big data analysis model”, Korean Intellectual Property Office, Registration No. 10-1595057-0000, Date: 2016. 02. 17.

- [26] Sungyoung Lee, and **Shujaat Hussain**, “A hybrid cloud gateway for smart home”, Korean Intellectual Property Office, Registration No. 10-1501731-0000, Date: 2015. 03.12

