PhD Dissertation Presentation

Semantic Preservation of Standardized Health Documents in Big Data

Mr. Shujaat Hussain

Department of Computer Science and Engineering

Kyung Hee University, Republic of Korea

Email: shujaat.hussain@oslab.khu.ac.kr

Date: November 6, 2019

Advisor: Prof. Sungyoung Lee, PhD

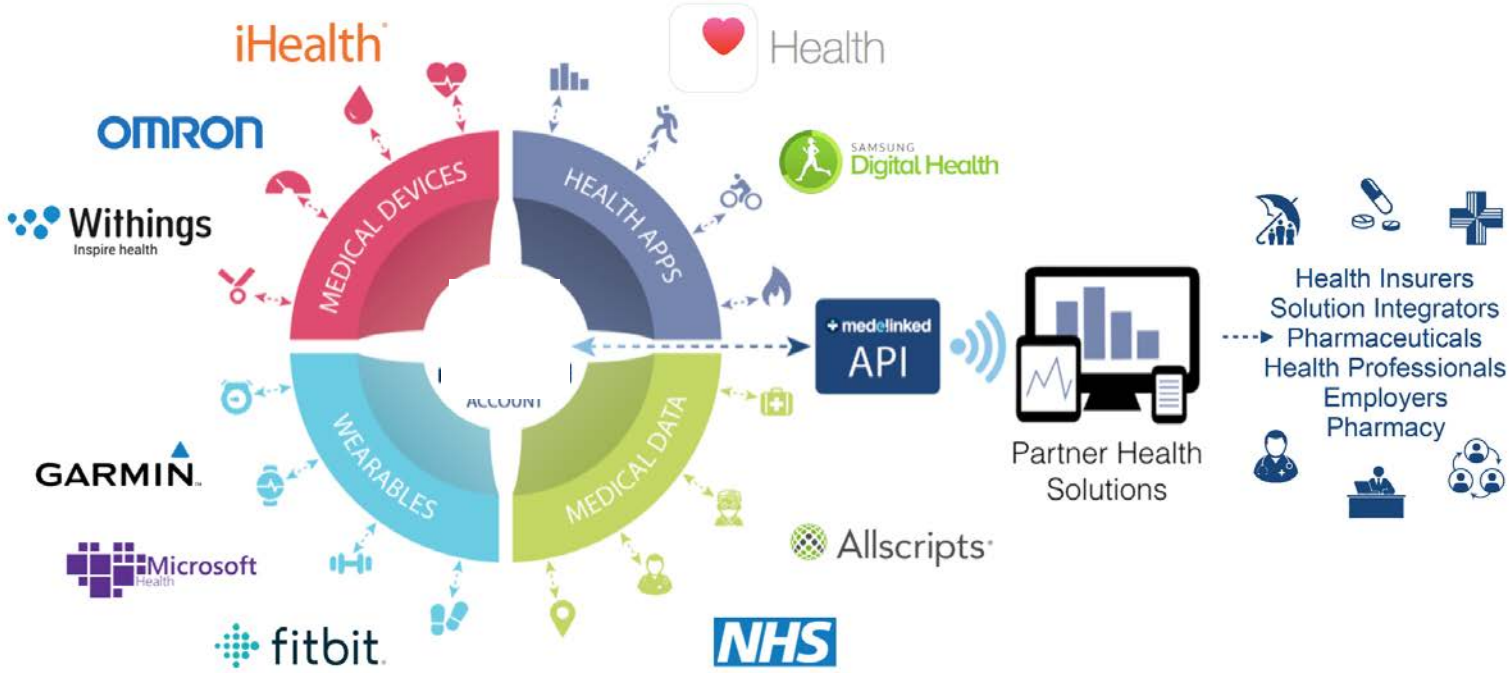
Presentation Agenda

- Introduction
  - Background and Motivation
  - Problem Statement
- Related Works
- Proposed Methodology
- Experiments and Results
- Uniqueness and Contributions
- Conclusion and Future Research

Background

Multiple Data Sources

Complexity in Schema and Knowledge



Knowledge Representation

- Complex metadata
- Linked Data
- Semantic Enriched dataset

Large Complex Dataset

- 30% of data is healthcare
- 97% hospitals use EHR (US)
- 80% data unstructured
- Critical to understand

Complex Health Standards

Important Vs of Big Data – Volume and Variety

Standardized Health Documents have critical information and the semantics of the document has to be protected.

Motivation

Semantic Preservance

Means lossless data which is critical in healthcare

Adaptable to Standardized schema(Open EHR, HL7 CDA)

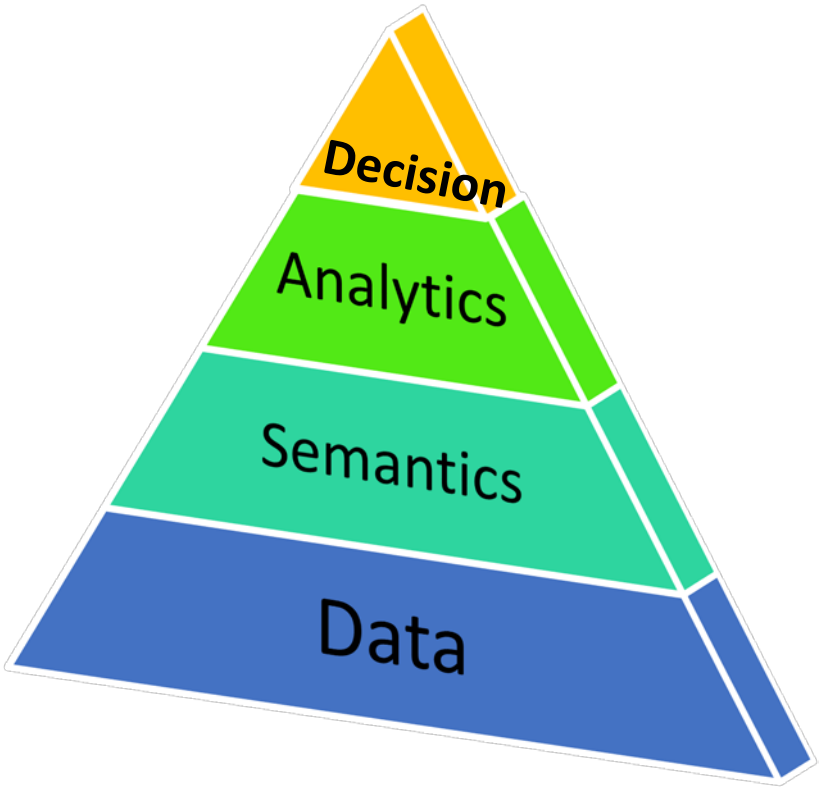
Complete Document Knowledge on disposal

Data continuity is important in standardized health documents

HL7 CDA standard is universally acceptable*

Healthcare Data

- Variety of Data Formats
  - Unstructured
  - Semi structured
  - Flat file
  - XML



Enabling Better Decision making

* Slavov, Vasil, et al. "A new tool for sharing and querying of clinical documents modeled using HL7 Version 3 standard." *Computer methods and programs in biomedicine* 112.3 (2013): 529-552.

Problem Statement

Problem statement

Semantic loss occurs in standardized and complex health care documents during partitioning big data framework and results in compromised decision making ability[1][2]

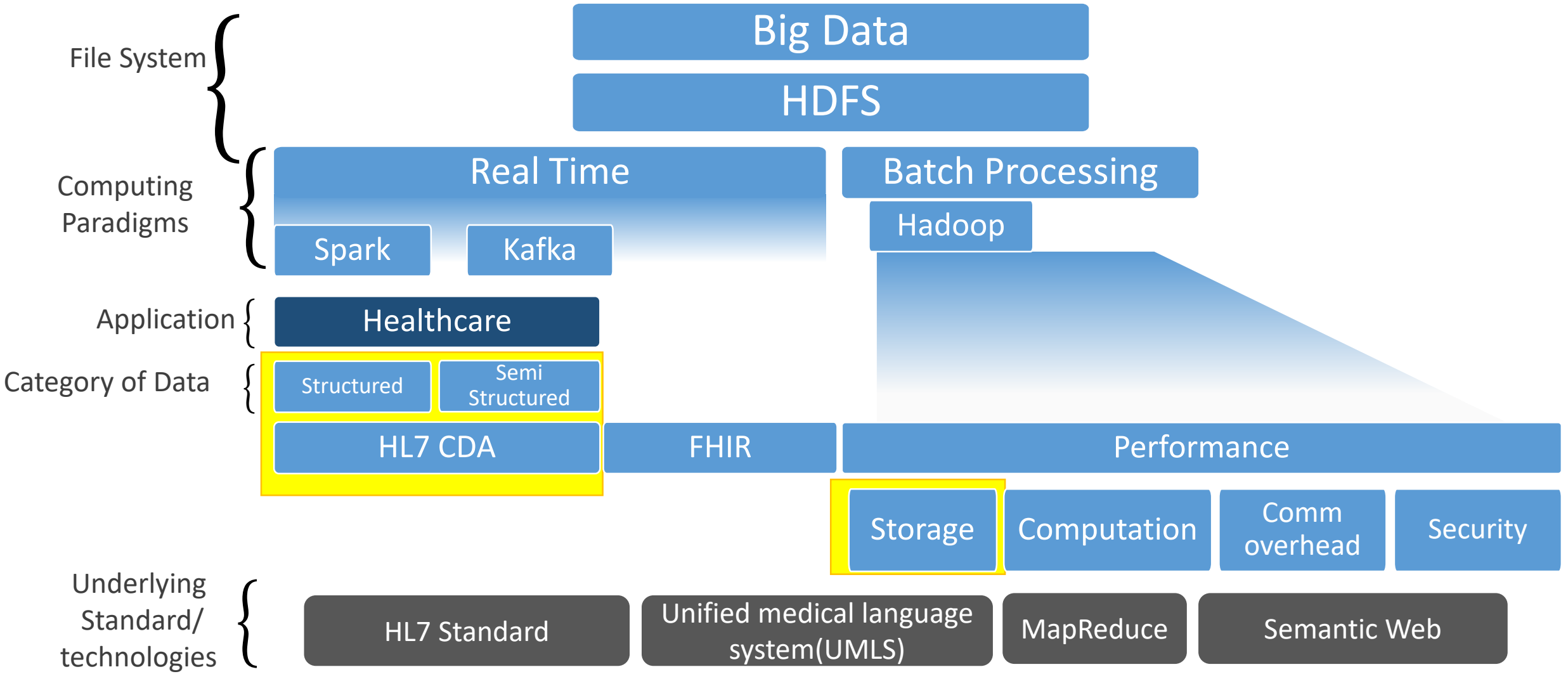
Goal

Semantic preservance in big data framework for standardized healthcare documents and ensure all the semantics from big data is extracted.

Challenges

- Challenge 1: There exists no data correlation with conventional Big Data framework partition
- Challenge 2: Identifying documents with semantic loss based on the application scenario
- Challenge 3: Resolving conflicts in partitioned data through semantic matching

Taxonomy


File System
Big Data
HDFS
Computing Paradigms
Real Time
Batch Processing
Spark
Kafka
Hadoop
Application
Healthcare
Category of Data
Structured
Semi Structured
HL7 CDA
FHIR
Performance
Storage
Computation
Comm overhead
Security
Underlying Standard/ technologies
HL7 Standard
Unified medical language system(UMLS)
MapReduce
Semantic Web

Related work and limitations

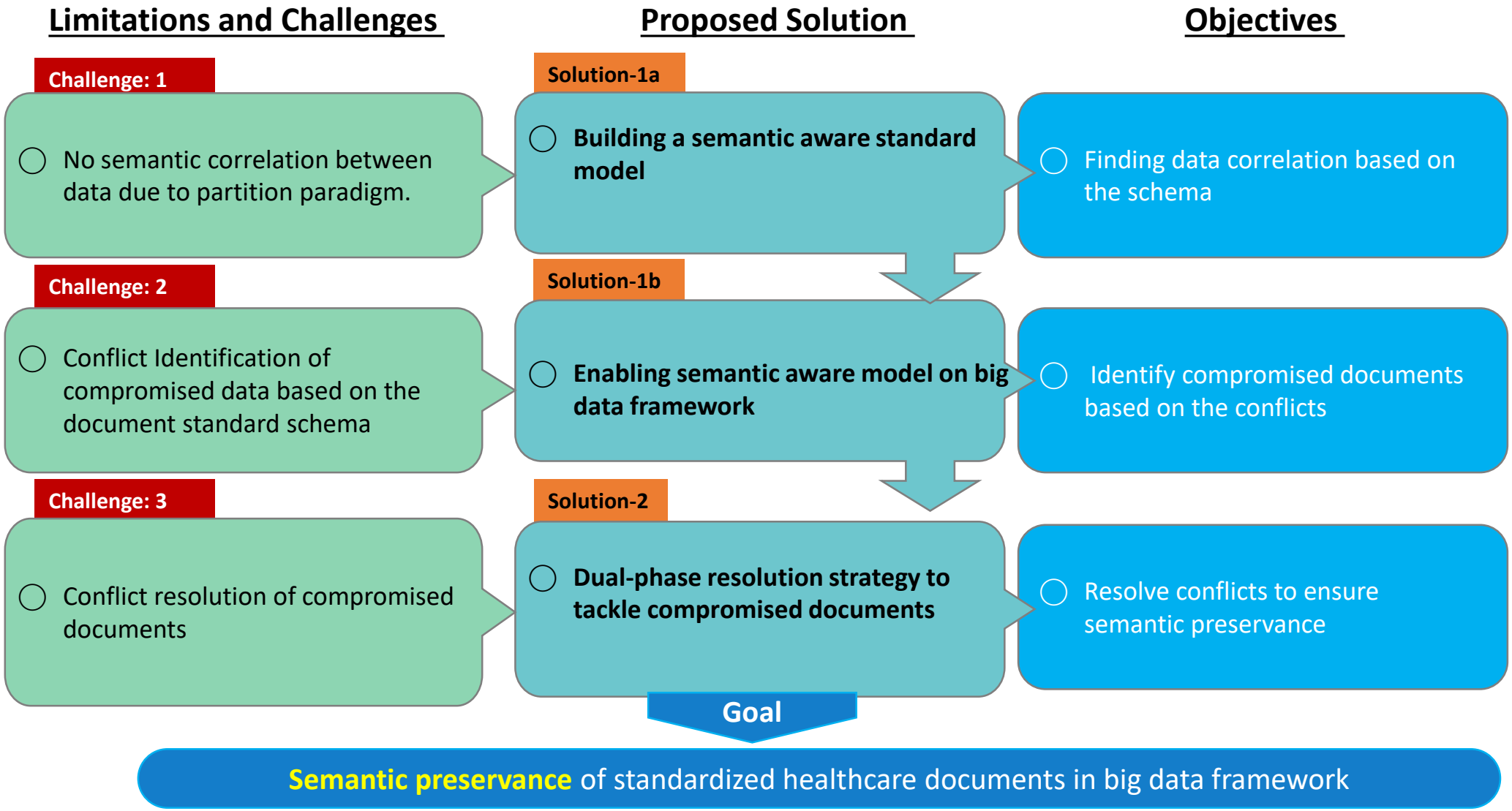
	Reference	Data Handling	Multiple Data Sources	Standard Validation	Data preprocessing (schema)	Partition	Data Distribution	Time Complexity
Challenge: 1	Silvestri et al (2019)	Spark	○	●	Dependent	Line by Line	Default	(n^2)
	Nunziato et al (2016)	Spark	●	○	Dependent	Not Applied	Default	(n^2)
Challenge: 2	HL7-CDA clinical documents management (2013)	MongoDB	○	●	Dependent	Not Applied	Default	NA
	Ko *et al.* (2014)	Pseudo Distributed	○	○	Not Applied	Line by Line	NA	NA
Challenge: 3	Hiromasa *et al.* (2012)	Hadoop	●	○	Dependent	Line by Line	Default	(n^2)
	Medoop (2013)	Hadoop	○	○	Dependent	Schema Based	Default	(n^2)

● Yes
○ No

Overall Limitation

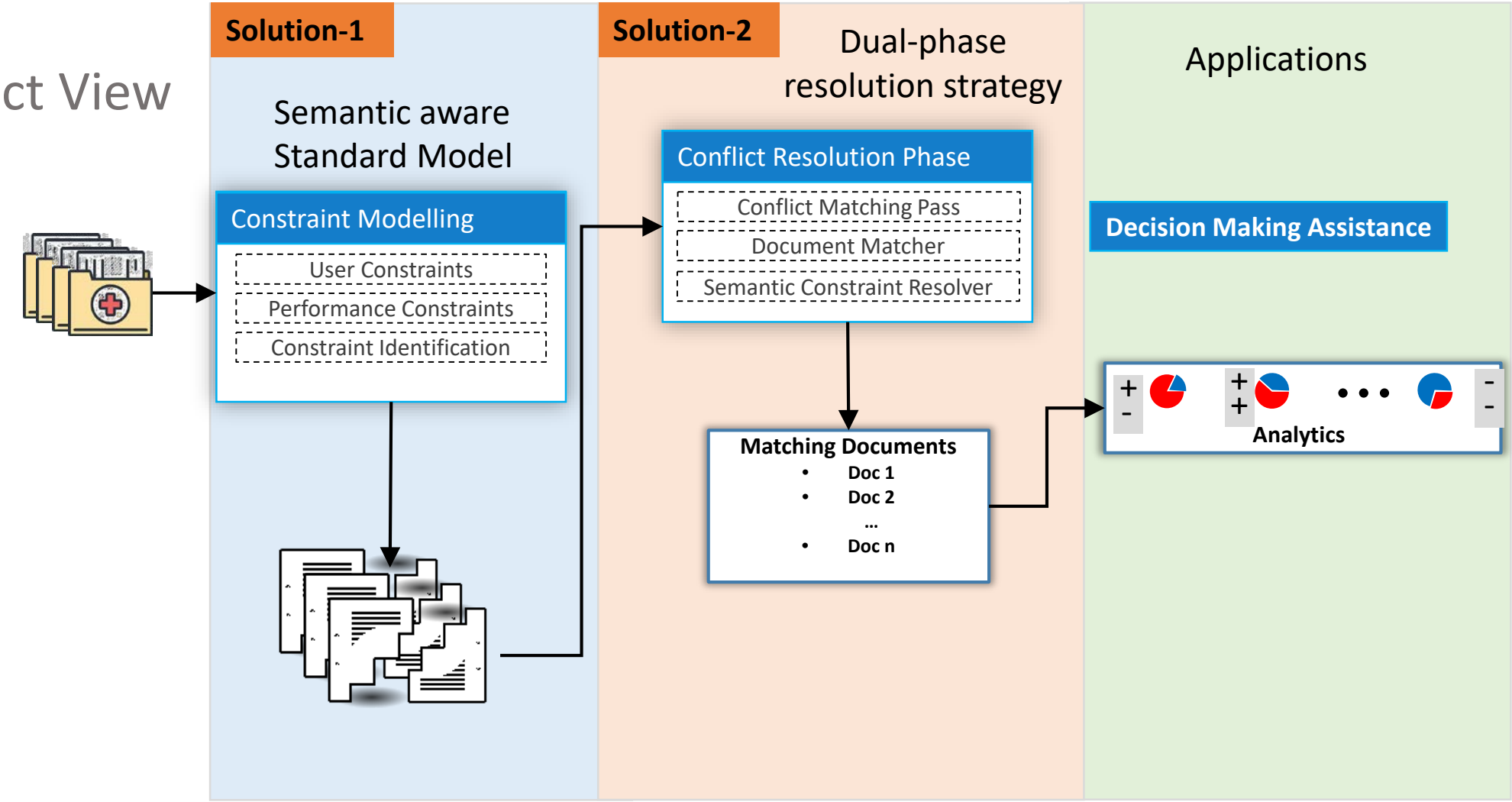
- No semantic or data schema connected with the dataset
- Number of documents compromised are ignored or partially used
- No semantic resolution for the compromised documents

Limitation, Objective, and Proposed Solution

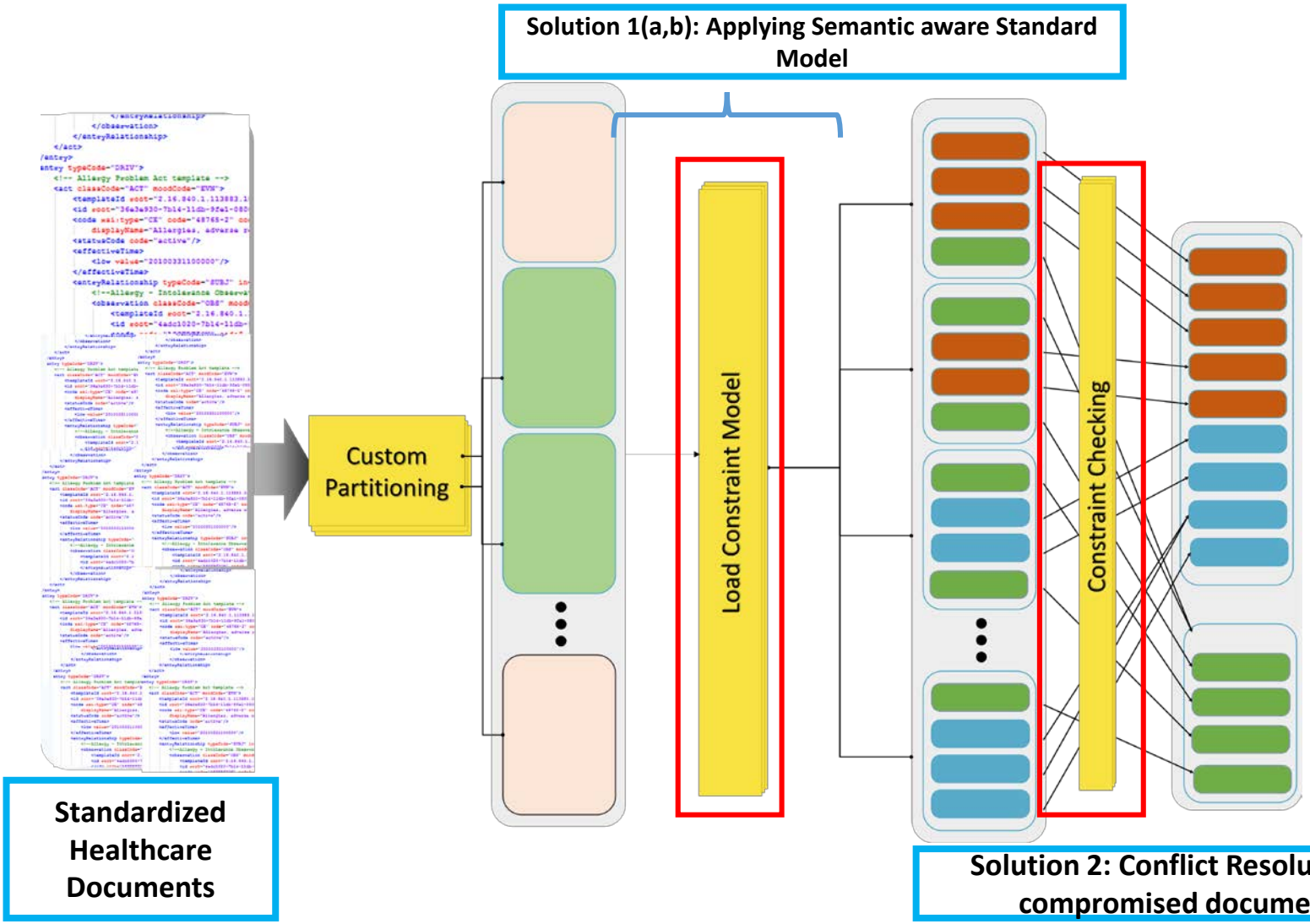


Proposed Methodology

- Abstract View



Overall Workflow : Semantic aware Standard Model, conflict identification and resolution



Three main steps

- Identify constraints and applying on the dataset
- Conflict Identification of the documents partitioned
- Conflict Resolution of the documents

Compromised Document

Regular Document

Regular Document

Shujaat Hussain, 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 2014

Problem during data processing

Allergies on Node 1

Data Node 1

Substance	Reaction	Status
Penicillin	Hives	Active
Aspirin	Wheezing	Active
Codeine	Nausea	Active

Medications on Node 2

Data Node 2

Medication	Instructions
Lisinopril 5 mg	1 tablet a day
Atenolol 25mg	1 tablet a day

Issues

- Incomplete information on two nodes for one patient
- No node has complete picture
- Partial Semantics at best

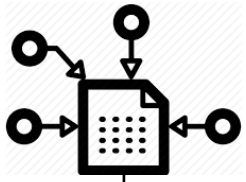
```
<entry typeCode="DRIV">
    <act classCode="ACT" moodCode="EVN">
        <templateId root="2.16.840.1.113883.10.20.1.27"/>
        <!-- Problem act template -->
        <id root="36e3e930-7b14-11db-9fe1-0800200c9a66"/>
        <code nullFlavor="NA"/>
        <entryRelationship typeCode="SUBJ">
            <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>
```

```
<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="cdbd5b05-6cde-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>
```

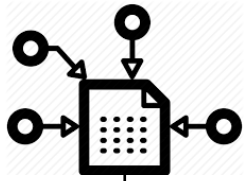
Proposed Methodology: Semantic Standard Model-As-is To be

Existing Method



- **Partitioner**
  - Input Record Formatter
  - Size Splitter
- **Data Transformation**
  - Data processing
  - Data Extraction
- **Generate Analytics**
  - User Parameters
  - Data Facts
- **Show Results**

Proposed Idea



- **Partitioner**
  - Input Record Formatter
  - Size Splitter
- **Constraint Modelling**
  - Performance Constraints
  - User Constraints
- **Applying Constraint Model**
  - Conflict Identification
  - Data Classification
- **Dual Phase Resolution**
  - Semantic Matching
  - UMLS Resolution
- **Data Transformation**
  - Data processing
  - Data Extraction
- **Generate Analytics**
  - User Parameters
  - Data Facts
- **Show Results**

Existing Approach Limitations

- ❑ After the partition, documents are passed to data transformation and analytics in broken parts and full picture is not shown due to semantic loss

Solution for challenge 1

- ❑ Creation and application of semantic standard model to find compromised documents and the resolution of the documents.

Solution 1a: Semantic Aware Standard Model

- Semantic Standard model is based on constraints, in which the output of a decision is constrained by certain parameters.

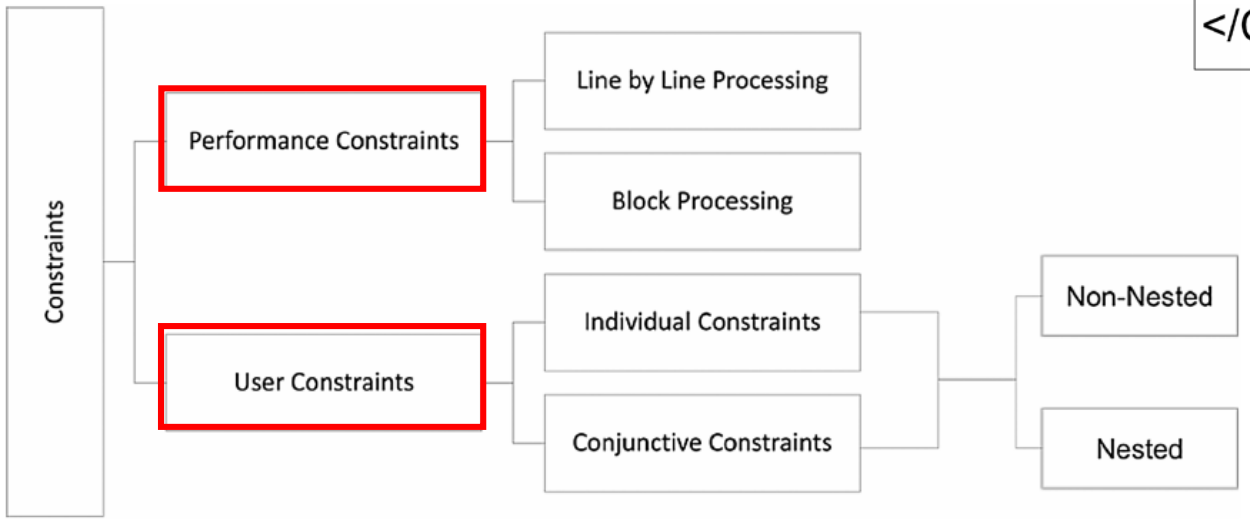
Clinical Document Architecture

<ClinicalDocument>…**CDA Header**…
<structuredBody><section><text> …**Single Narrative Block**
…</text>
<observation>…</observation>
<substanceAdministration>
<supply>…</supply>
</substanceAdministration><observation>
<externalObservation>…
</externalObservation></observation>
</section><section><section>…..</section>
</section> </structuredBody>
</ClinicalDocument>

Human Readable Part

Complex Machine Processable Parts

Constraint Hierarchy



Constraint in CDA Document

User constraint table.

Constraint name	Type	Nested
Clinical statements	Conjunctive	Yes
entryRelationship	Individual	Yes
Entry	Individual	No
Section	Individual	Yes
Whole document	Individual	No

Solution 1: Semantic Aware Model

Inside CDA Document

```
- <!--
    ************************************************************

    CDA Body

    ************************************************************
  -->
- <component>
  - <structuredBody>
      <!--  **********************  -->
    - <!--
        ************************************************************

        Allergies, Adverse Reactions, Alerts

        ************************************************************

      -->
    - <component>
      - <section>
          <templateId root="2.16.840.1.113883.10.20.22.2.6.1" />
          <!-- Alerts section template  -->
          <code code="48765-2" codeSystem="2.16.840.1.113883.6.1" />
          <title>Allergies, Adverse Reactions, Alerts</title>
        - <text>
          - <table border="1" width="100%">
            + <thead>
            - <tbody>
              - <tr>
                  <td>Penicillin</td>
                - <td>
                    <content ID="reaction1">Hives</content>
                  </td>
                  <td>Active</td>
```

Table of Contents

Allergies, Adverse Reactions, Alerts

Substance	Reaction	Status
Penicillin	Hives	Active
Aspirin	Wheezing	Active
Codeine	Nausea	Active

Hospital Course

The patient was admitted and started on Lovenox and nitroglycerin paste. The patient had serial cardiac enzymes and was ruled out for myocardial infarction. The patient underwent a dual isotope stress test. There was no evidence of reversible ischemia on the Cardiolite scan. The patient has been ambulated. The patient had a Holter monitor placed but the report is not available at this time. The patient has remained hemodynamically stable. Will discharge.

Hospital Discharge Diagnosis

Unspecified chest pain

Solution 1: Semantic Aware Model

Individual Constraint

- **Section**
  - Section title
  - Description and explanatory narrative
  - LOINC section code
  - Requirements for a text element

```
<section>
    <templateId root="2.16.840.1.113883.10.20.22.2.58" />
    <code code="75310-3" displayName="Health Concerns Document'
codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC" />
    <title>Health Concerns Section</title>
    <text>
      ...
    </text>
    <entry>
        <!-- Health Status Observation -->
    </entry>
    <entry>
        <!-- Health Concern Act -->
    </entry>
</section>
```

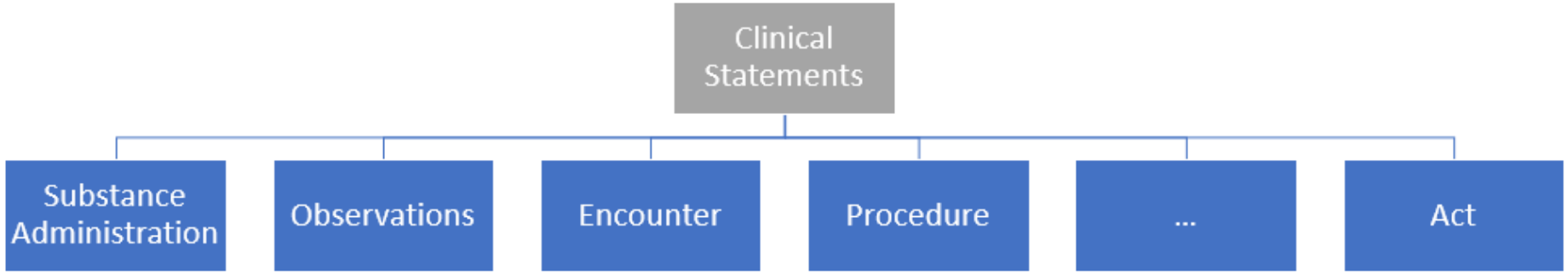
Individual Constraints are those which are simple and are not dependent on other constraints or vice versa. They do not have any subtypes.

Solution 1: Semantic Aware Model

Conjunctive constraints

- Clinical Statements
  - Key template metadata (e.g., template identifier, etc.)
  - Description and explanatory narrative.
  - Required CDA acts, participants and vocabularies.
  - Optional CDA acts, participants and vocabularies.

Conjunctive Constraints are nested and contain subtypes as well as multiple individual constraints which increases the complexity.



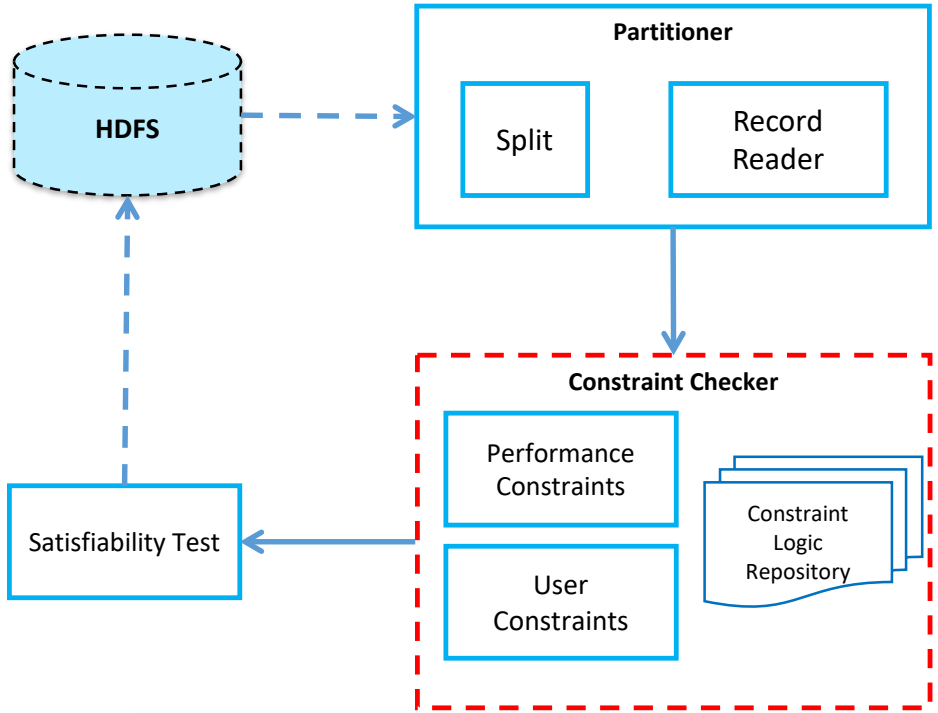
Solution 1b: Semantic Standard Model

Algorithm

Input: Standardized Documents
Output: Conflicted Document

Output: $documents\{key, conflicteddocument\}$: 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 ∀ md_i ∈ middledocs do
18 |  applyuserscenario(md_i) ;
19 end
```



Contributions

- Conflict Identification of head of the documents partitioned
- Conflict Identification of tail of the documents partitioned'
- Documents with no conflicts sent for processing and analytics

Shujaat Hussain, Maqbool Hussain, Muhammad Afzal, …Sungyoung Lee, "Semantic Preservation of standardized healthcare documents in big data", International Journal of Medical Informatics 2019

Identification of Compromised Documents

- After applying Semantic aware standard model the output is conflicted documents and the different markers identified are
  - Number of constraints violated
  - Individual Constraints
  - Conjunctive Constraints

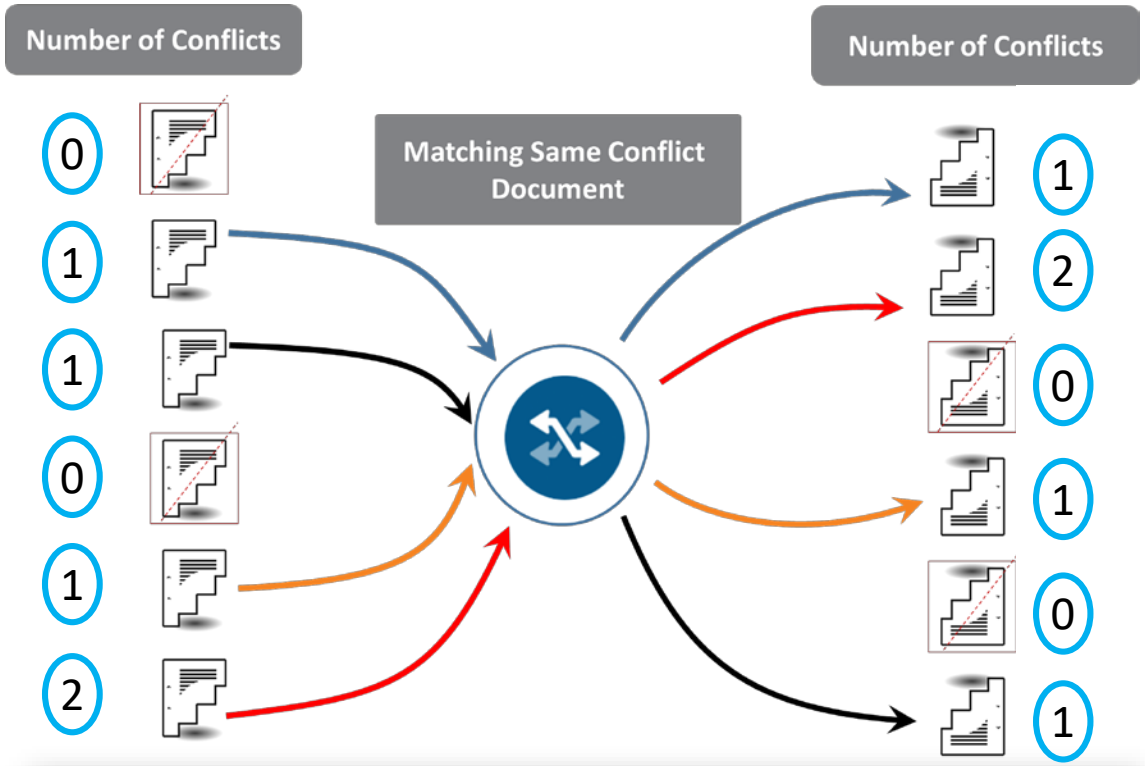
- Based on the conflict markers, we need a resolution strategy to make the document whole so that no conflicts remain.
- A Dual Phase resolution strategy was introduced.

Solution 2: Dual-phase resolution strategy

Input: Conflicted Documents
Output: Conflict free Document

Input : $UD = \{ud_1, ud_2, ud_3 \ldots, ud_n\}$: Incomplete Upper Documents
$LD = \{ld_1, ld_2, ld_3 \ldots, ld_n\}$: Incomplete Lower Documents
Output: Complete CDA® documents

for $\forall\ ud_i \in UD$ **do**
  /* Conflicts are scanned from the upper document and if conflicts are one or greater, we try to resolve */;
  $Conflictmarker_i \leftarrow checkconflict(ud_i)$;
  if $Conflictmarker \geqslant 1$ **then**
    for $\forall\ ld_j \in LD$ **do**
      /* Conflicts are scanned from the lower document and if conflicts of upper and lower documents are same, we append and validate */;
      $Conflictmarker_j \leftarrow checkconflict(ud_j)$;
      if $Conflictmarker_i == Conflictmarker_j$ **then**
        $fulldoc \leftarrow appendDocument(ud_i, ld_j)$;
        $flag \leftarrow validateDocument(fulldoc)$;
        /* If the document is validated, then we create a pair of indexes of upper and lower documents for duplicate checking which triggers second pass if any */;
        if $flag = 1$ **then**
          $Pair < Integer, Integer > MapperRecord \leftarrow list.add(i, j)$;
        end
      end
    end
  end
end



First Pass of Conflict resolving

- The documents are matched to the head and tail of the documents.
- The documents are validated through the schema.
- Some documents may not match or match with two heads/tails which then requires second phase.

Shujaat Hussain, Maqbool Hussain, Muhammad Afzal, …Sungyoung Lee, "Semantic Preservation of standardized healthcare documents in big data", International Journal of Medical Informatics 2019

Orphan and Duplicate Matches

- 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.

- Second Phase
  - 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.

Solution 2: UMLS Querying Workflow-Example

```
<component>
    <section>
        <templateId root="2.16.840.1.113883.10.20.1.4"/>
        <code code="10157-6" codeSystem="2.16.840.1.113883.6.1"/>
        <title>Family history</title>
        <text>
            <paragraph>Father (deceased)</paragraph>
            <table border="1" width="100%">
                <thead>
                    <tr>
                        <th>Diagnosis</th>
                        <th>Age At Onset</th>
                    </tr>
                </thead>
                <tbody>
                    <tr>
                        <td>Myocardial Infarction (cause of death)</td>
                        <td>57</td>
                    </tr>
                    <tr>
                        <td>Hypertension</td>
                        <td>40</td>
                    </tr>
                </tbody>
            </table>
```

2.16.840.1.113883.6.1

Family history

Unified Medical language System (UMLS)

2.16.840.1.113883.10.20.1.42

Family history

```
<entry typeCode="DRIV">
    <organizer moodCode="EVN" classCode="CLUSTER">
        <templateId root="2.16.840.1.113883.10.20.1.23"/>
        <!-- Family history organizer template -->
        <statusCode code="completed"/>
        <subject>
            <relatedSubject classCode="PRS">
                <code code="9947008" codeSystem="2.16.840.1.113883.6.96"
                    displayName="Biological father"/>
                <subject>
                    <administrativeGenderCode code="M"
                        codeSystem="2.16.840.1.113883.5.1" displayName="Male"/>
                    <birthTime value="1912"/>
                </subject>
            </relatedSubject>
        </subject>
        <component>
            <observation classCode="OBS" moodCode="EVN">
                <templateId root="2.16.840.1.113883.10.20.1.42"/>
                <!-- Family history cause of death observation template -->
                <id root="d42ebf70-5c89-11db-b0de-0800200c9a66"/>
                <code code="ASSERTION" codeSystem="2.16.840.1.113883.5.4"/>
                <statusCode code="completed"/>
                <effectiveTime value="20001020"/>
                <value xsi:type="CD" code="22298006"
                    codeSystem="2.16.840.1.113883.6.96" displayName="MI"/>
                <entryRelationship typeCode="CAUS">
                    <observation classCode="OBS" moodCode="EVN">
                        <id root="6898fae0-5c8a-11db-b0de-0800200c9a66"/>
                        <code code="ASSERTION"
```

Family history

Concept Matching of Upper and Lower Document

Contributions

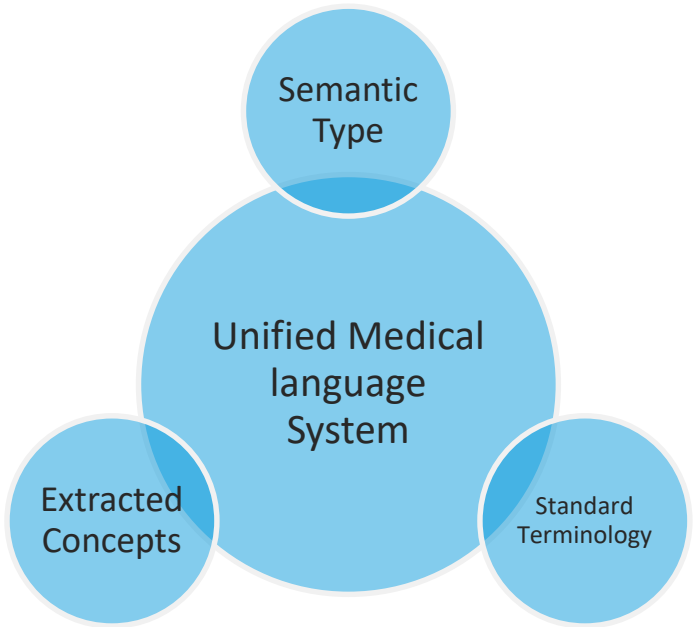
- A methodology for autonomous resolution of compromised document based on their conflicts and semantic concepts based on the classified documents

Solution 2: Dual-phase resolution strategy

Input: Orphan and Duplicated Documents
Output: Conflict free Document

Input : $UD = \{ud_1, ud_2, ud_3 \ldots, ud_n\}$: Orphan and Duplicate Upper Documents
$LD = \{ld_1, ld_2, ld_3 \ldots, ld_n\}$:Orphan and Duplicate Lower Documents
Output: $CD = \{cda_1, cda_2, cda_3 \ldots, cda_n\}$ 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 ($conceptud_i$) and first constraint of lower document ($conceptld_j$) */;
4 $conceptud_i \leftarrow getLastConstraint(ud_i)$;
5 $conceptld_j \leftarrow getFirstConstraint(ld_j)$;
6 /* Get semantic type of concept of $conceptud_i$ and $conceptld_j$ */;
7 $semtypeud_i \leftarrow getSemanticConstraint(conceptud_i)$;
8 $semtypeld_j \leftarrow getSemanticConstraint(conceptld_j)$;
9 /* Match Semantic types of lower and upper document concepts, join them and validate for full CDA® document */;
10 **if** $semtypeud_i == semtypeld_j$ **then**
11 $fulldoc_i \leftarrow appendDocument(ud_i, ld_j)$;
12 $flag \leftarrow validateDocument(fulldoc_i)$;
13 **end**
14 **end**
15 **end**



2nd Pass of conflict resolution

- UMLS repository is used for second phase of resolution.
- It has a vocabulary of enriched concepts of biomedical terms.
- The documents remaining compromised after phase 1 go through phase 2

Experimental Setup

Experiments

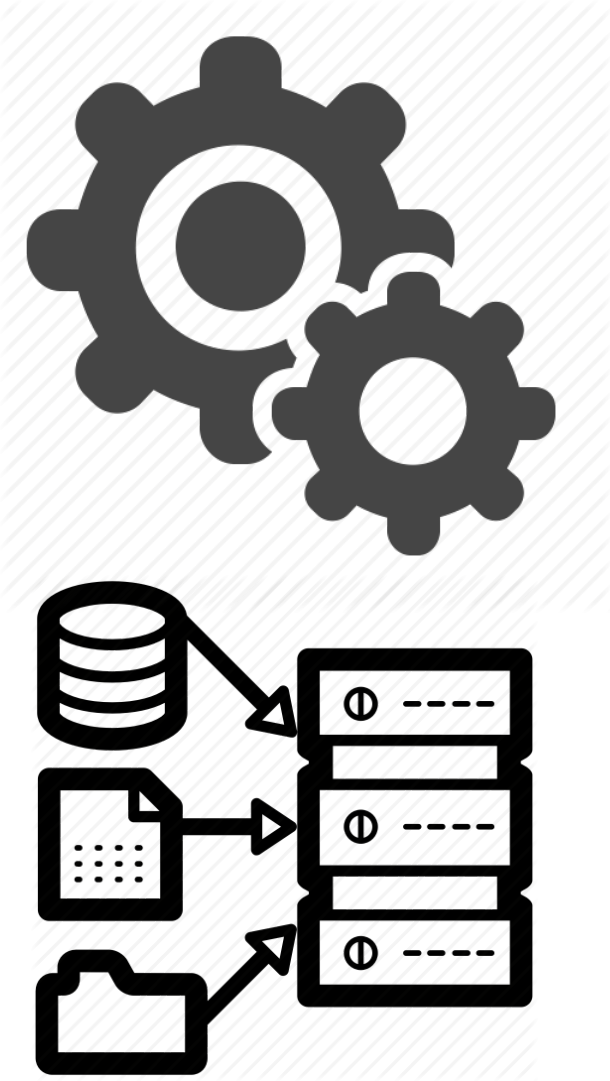
- **Experiment 1**: Constraint Conflicts against Healthcare Documents
- **Experiment 2**: Dual Phase Resolution strategy

Experiment Setup and Big Data Framework

- Ubuntu 16.04
- HDFS (Hadoop and Spark)
- 6 Machine Cluster

Dataset

- Unique 700 C-CDA documents*
- Overall Dataset: 87000 documents through bootstrapping and sampling
- Block Sizes: 32 MB, 64MB, 128 MB, 256 MB

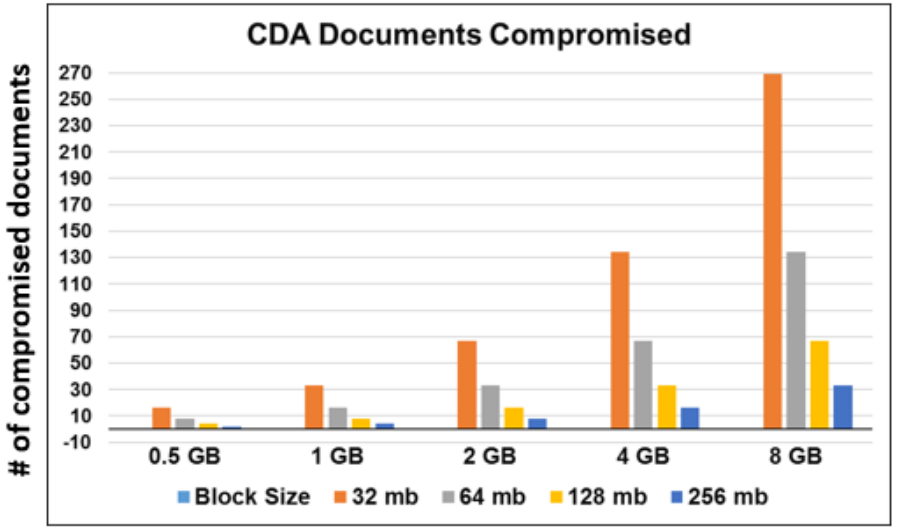


* https://github.com/jmandel/sample_ccdas

Solution 1: Individual Constraints

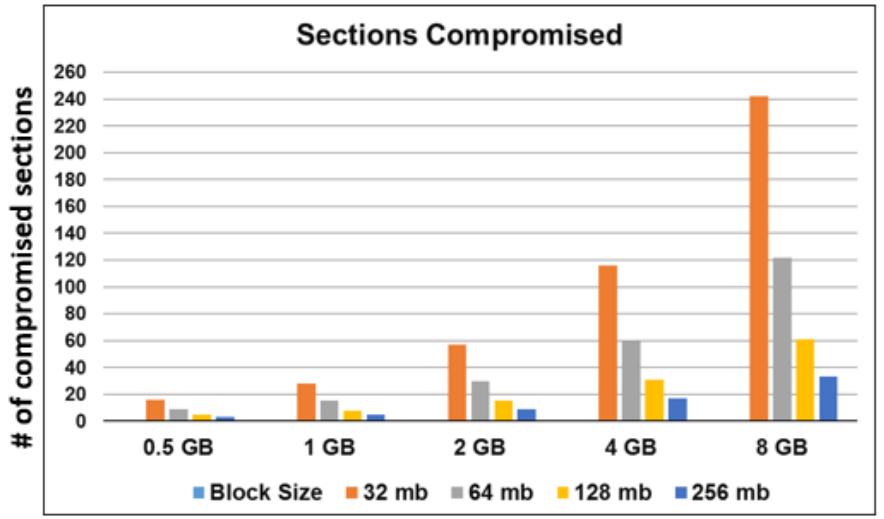
Results Discussion

- Overall the number of documents compromised got higher with the dataset without any constraints applied.
- Higher level individual constraints resulted in more conflicted documents.



Higher value is bad which means more conflicts and compromised documents

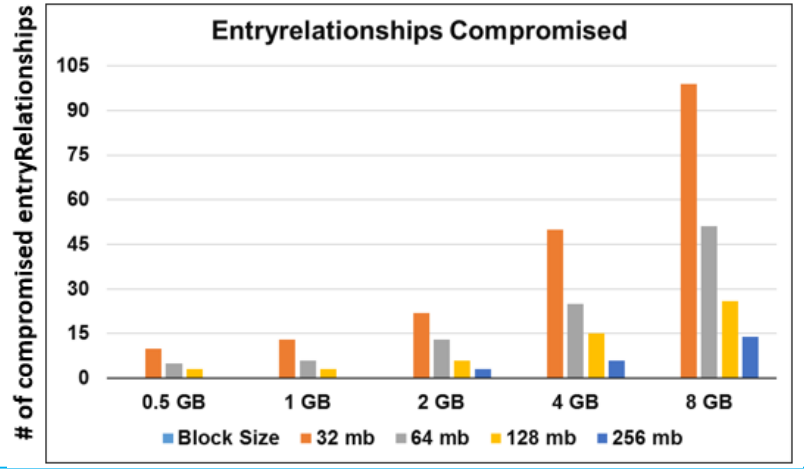
Block Size	Individual Constraints									
	Whole Documents					Sections				
	.5 GB	1 GB	2 GB	4 GB	8 GB	.5 GB	1 GB	2 GB	4 GB	8 GB
32 mb	16	33	67	134	269	16	28	57	116	242
64 mb	8	16	33	67	134	9	15	30	60	122
128 mb	4	8	16	33	67	5	8	15	31	61
256 mb	2	4	8	16	33	3	5	9	17	33



Solution 1: Conjunctive constraints

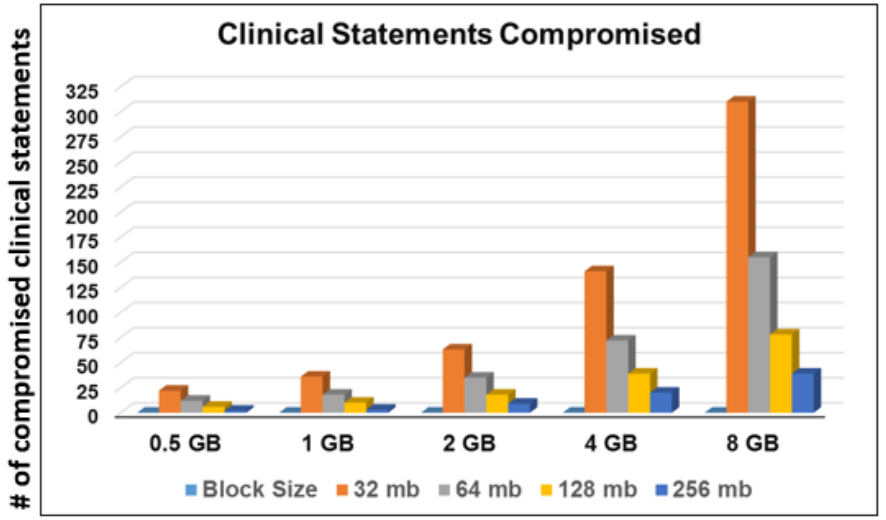
Results Discussion

- Clinical Statements (CS) is a superset of Entryrelationships (ER).
- CS is the most complex constraint and it has been compromised the most
- ER is an individual constraint but present inside CS constraints



Higher value is bad which means more conflicts and compromised documents

Block Size	Conjunctive Constraints									
	Clinical Statements					EntryRelationships				
	.5 GB	1 GB	2 GB	4 GB	8 GB	.5 GB	1 GB	2 GB	4 GB	8 GB
32 mb	22	36	63	141	310	10	13	22	50	99
64 mb	12	18	35	72	155	5	6	13	25	51
128 mb	6	10	18	39	78	3	3	6	15	26
256 mb	2	3	9	20	39	0	0	3	6	14



Results And Evaluation: Solution 2 Dual Phase resolution

Experimental Details

- Two phases of conflict resolution
  - **Matching documents**
  - UMLS Querying

Results Discussion

- The conflicted documents are matched depending on the number of conflicts identified.
- The table shown are the conflict for the 1GB dataset.

Conflicted Documents in Phase 1

Resolution 1st Pass 32 MB Block

Upper part	Lower part	Upper part	Lower part
Doc. 1	Doc. 18	Doc. 18	Doc. 24
Doc. 2	Doc. 11	Doc. 19	Doc. 13
Doc. 3	Doc. 34	Doc. 20	Doc. 21
Doc. 4	Doc. 28	Doc. 21	Doc. 34
Doc. 5	Doc. 23	Doc. 22	Doc. 21
Doc. 6	Doc. 4	Doc. 23	Doc. 14
Doc. 7	Doc. 20	Doc. 24	Doc. 33
Doc. 8	Doc. 19	Doc. 25	Doc. 16
Doc. 9	Doc. 29	Doc. 26	Doc. 6
Doc. 10	Doc. 1	Doc. 27	Doc. 25
Doc. 11	Doc. 26	Doc. 28	Doc. 17
Doc. 12	Doc. 30	Doc. 29	Doc. 32
Doc. 13	Doc. 21	Doc. 30	Doc. 22
Doc. 14	Doc. 8	Doc. 31	Doc. 34
Doc. 15	Doc. 31	Doc. 33	Doc. 27
Doc. 17	Doc. 3	Doc. 34	Doc. 9
Doc. 16	Doc. 5		

Results And Evaluation: Solution 2 Dual Phase resolution

Experimental Details

- Two phases of conflict resolution
  - Matching documents
  - **UMLS Querying**

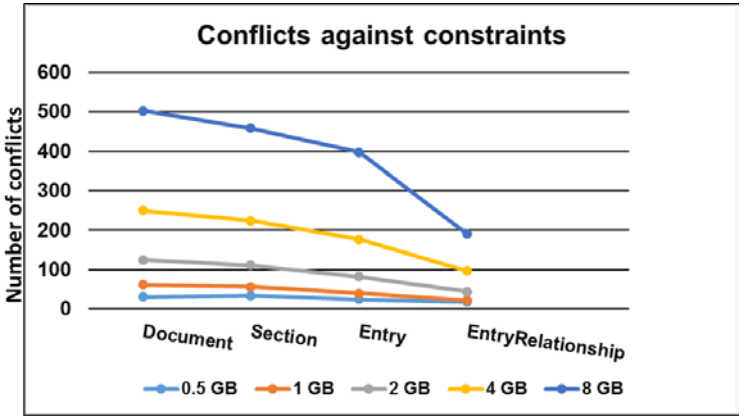
Results Discussion

- The orphan and duplicated documents are matched through semantic type
- All the document conflicts were resolved during the second phase of the resolution.

Second pass of resolution.

Orphan and duplicate documents in 32 MB block

Document	Document type
Upper Doc. 3	Duplicate
Upper Doc. 20	Duplicate
Upper Doc. 21	Duplicate
Upper Doc. 22	Duplicate
Upper Doc. 31	Duplicate
Upper Doc. 13	Duplicate
Lower Doc. 2	Orphan
Lower Doc. 7	Orphan
Lower Doc. 10	Orphan
Lower Doc. 12	Orphan
Lower Doc. 15	Orphan
Lower Doc. 21	Duplicate
Lower Doc. 34	Duplicate



The constraints against size is shown accumulating all the blocks

Contributions and Uniqueness

The compromised documents were successfully identified through semantic aware standard model

- ❑ Compromised Healthcare documents processed through Big Data Frameworks were identified

Dual-phase resolution are applied to compromised documents

- ❑ The healthcare documents in the dataset are semantically preserved for meaningful extraction though conflict resolution technique.

- Potential Applications
  - Document Interoperability
  - Predictive Analytics
  - Decision Support Systems

Conclusion and Future Work

Identifying compromised documents based on constraint model

- Identifying constraints based on schema standard and user scenario as well as performance.
- Classifying documents into three categories conflicts

Conflict Resolution of the documents compromised

- All the documents were semantically preserved and the meaningful data can be extracted from them to better decision making.

- Future Research
  - In future, we will focus on performance, as the first goal was semantic preservance in this study.

Publications

- SCI/ SCIE Journals (8)
  - First Author- Two Published
  - Co-Author- Six Published
- Local Journal (2)
  - Co-Author-Two Published
- International Conferences (9)
  - First Author – Five Publications
  - Co-Author- Nine Publications
- Local Conferences (3)
  - First Author – Three Publications
- Domestic Patents (3)
  - 2 Registered
  - 1 Applied

Total Publications = 25

Selected References

- [1] Panahiazar, Maryam, et al. "Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases." 2014 IEEE International Conference on Big Data (Big Data). IEEE, 2014.
- [2] Bansal, Srividya K. "Towards a semantic extract-transform-load (ETL) framework for big data integration." 2014 IEEE International Congress on Big Data. IEEE, 2014.
- [3] Mohanpurkar, Arti, and Prasad kumar Kale. "Big data analysis using partition technique." International Journal of Computer Science and Information Technologies (IJCSIT) 6.3 (2015): 2871-2875.
- [4] Dhar, Vasant. "Data science and prediction." Communications of the ACM56.12 (2013): 64-73.
- [5] Chawla, Nitesh V., and Darcy A. Davis. "Bringing big data to personalized healthcare: a patient-centered framework." Journal of general internal medicine 28.3 (2013): 660-665.
- [6] Tsai, Chun-Wei, et al. "Big data analytics: a survey." Journal of Big Data 2.1 (2015): 1-32.
- [7] Ko, Kyung Dae, et al. "Predicting the severity of motor neuron disease progression using electronic health record data with a cloud computing Big Data approach." Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on. IEEE, 2014.
- [8] Naive Bayesian Classification Approach in Healthcare Applications. R. Bhuvaneswari and K. Kalaiselvi Department of CSE, Saveetha Engineering College Chennai, India (2012)
- [9] Decision Support in Heart Disease Prediction System using Naive Bayes. n Journal of Computer Science and Engineering (IJCSE) (2012)

Selected References

- [10] Yu, Hongyong, and Deshuai Wang. "Research and Implementation of Massive Health Care Data Management and Analysis Based on Hadoop." Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on. IEEE, 2012.
- [11] Wang, Ye, et al. "Large-Scale Clinical Data Management and Analysis System Based on Cloud Computing." Frontier and Future Development of Information Technology in Medicine and Education. Springer Netherlands, 2014. 1575-1583.
- [12] Arian, et al. "Assessing the Predictability of Hospital Readmission Using Machine Learning." Twenty-Fifth IAAI Conference. 2013.
- [13] Danning, et al.."Mining high-dimensional administrative claims data to predict early hospital readmissions."JAMIA (2013)
- [14] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica et al., "A view of cloud computing," Communications of the ACM, vol. 53, no. 4, pp. 50–58, 2010.
- [15] J. Zhou, N. Bruno, M.-C. Wu, P.-A. Larson, R. Chaiken, and D. Shakib. Scope: parallel databases meet mapreduce. The VLDB Journal-The International Journal on Very Large Data Bases, 21(5):611{636, 2012.
- [16] T. White. Hadoop: the definitive guide. O'Reilly, 2012.
- [17] A. Hadoop. Apache hadoop nextgen mapreduce (yarn). 2013.
- [18] www.predixionsoftware.com
- [19] Horiguchi, Hiromasa, et al. "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 12.1 (2012): 151.
- [20] Arshdeep Bahga, et al. "A Cloud-based Approach for Interoperable Electronic Health Records (EHRs)." IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS Sep 2013

THANK YOU!

Any questions or comments?

Conjunctive Constraints

- Clinical Statements
  - Key template metadata (e.g., template identifier, etc.)
  - Description and explanatory narrative.
  - Required CDA acts, participants and vocabularies.
  - Optional CDA acts, participants and vocabularies.