SAFE
Policy Aware SPARQL Query Federation
Over RDF Data Cubes

Dr. Ratnesh Sahay
Semantics in eHealth & Life Sciences (SeLS)
Insight Centre for Data Analytics
NUI Galway, Ireland

10-12-2014
SWAT4LS-2014, Berlin
Germany
1. **Showcase #1 – Phase III Clinical Trial: Subject Selection Criteria:**
   - the unbiased randomised selection of subjects in phase III clinical trials
   - e.g. return subjects with diabetes, weight >80 and hasCancer

2. **Showcase #2 – Phase IV Post Marketing Surveillance trial:**
   - the pharmacovigilance of a drug after it receives permission to be sold
   - e.g. Test DrugX association with headaches

3. **Showcase #3 – Chemoinformatics:**
   - identification of relations between molecular fragments and specific adverse side effect categories.
   - e.g. Test chemicalFragmentX(of DrugX) with rash
The Problem

return number of patients that have been administered the drug Insulin and exhibit BMI > 25 and Hypertension and Diabetes as adverse events

Switzerland

Cyprus

Greece
Patients’ anonymity

- Anonymised Clinical Data cubes
- Insensitive clinical parameters without personal information

Data Ownership & Privacy

- Access-Control Based Query Federation
SAFE - Secure SPARQL Query Federation

**SPARQL Query**

- **Source Selection**
- **Access Policy Filter**
- **Query re-Writer**

**Results**

<table>
<thead>
<tr>
<th>RDF Data Cubes</th>
<th>CHUV (Data Cubes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CING (Data Cubes)</td>
</tr>
<tr>
<td></td>
<td>ZEINCRO (Data Cubes)</td>
</tr>
</tbody>
</table>

**SAFE**

- RDF Data Cubes
- Index
- Access Policy Model

**Sources**

- CHUV
- CING
- ZEINCRO
SELECT ?diabetes ?bmi ?hypertension ?cases
WHERE {
?observation a qb:Observation .
?observation sdmx-measure:Cases ?cases.
}
SAFE - Source Selection

- SPARQL Query + User Info
- Source Selection
- Access Policy Filtering
- Query Rewriting

Diagram:
- For each query triple \( t(s, p, o) \in T \)
- Level 1: Data Sources
  - CHUV
  - CING
  - ZEINCRO
- Level 2: Named Graphs
  - g1
  - g2
  - g3
Enabling networked knowledge

**SAFE - Source Selection**

**SPARQL Query + User Info**

**Source Selection**

**Access Policy Filtering**

**Query Rewriting**

**Triples Patterns**


?observation sdmx-measure:Cases ?cases.

**Capable Sources**

{S1, S2, S3, S4}  
{S1, S2, S3}  
{S1, S2, S3}  
{S1, S2, S3, S4}

**Capable Sources**

S1={ Diabetes, BMI, Hypertension, Cases }

S2={ Diabetes, BMI, Hypertension, Cases }

S3={ Diabetes, BMI, Hypertension, Cases, HIV }

S4={ Diabetes, Smoking, Gender, Cases }

**Index**

Join Awareness
SAFE - Access Policy

- SPARQL Query + User Info
- Triple Pattern-based Source Selection
- Access Policy Filtering
- Query Rewriting

Access Policy Framework

Requested Data: S1 S2 S3

Oya
Clinical Researcher
Expertise - Diabetes

Input

Grants Access
Denies Access

S1
S2
S3
Enabling networked knowledge

**Example Access Policy**

- **AP1**
  - `type`: Access_Policy
  - `applies_to`: {S1, S2}
  - `grants_access`: Read
  - `assigned_to`: Oya

- **Oya**
  - `type`: User
  - `haslocation`: Galway
  - `hasPurpose`: Perform p-value analysis
  - `hasRole`: Clinical Researcher
  - `hasDomain`: Diabetes

**SPARQL Query**

```
ASK WHERE {
?accessPolicy a AccessPolicy.
?accessPolicy appliesToNamedGraph S1.
?accessPolicy grantsAccess acl:Read_l2s,
?accessPolicy hasUser Oya.
}
```
SAFE – Query Rewriting

- **Graph Information will be added to the query triples**
  - `SELECT …. WHERE { GRAPH <S1> { …. } }`
  - `SELECT …. WHERE { GRAPH <S2> { …. } }`

- **Sub queries sent to relevant sources**
  - `S1`
  - `S2`

- **Integration of results obtained from each sources**

<table>
<thead>
<tr>
<th></th>
<th>Diabetes</th>
<th>BMI</th>
<th>Hypertension</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td><strong>S2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>120</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>90</td>
</tr>
</tbody>
</table>
## Evaluation - DataSets

<table>
<thead>
<tr>
<th>Dataset</th>
<th># triples</th>
<th># obs</th>
<th># sub</th>
<th># pred</th>
<th># obj</th>
<th># size</th>
<th># index size</th>
<th># index generation time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal Dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHUV</td>
<td>0.8 M</td>
<td>96 K</td>
<td>96 K</td>
<td>36</td>
<td>88</td>
<td>31 MB</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CING</td>
<td>0.1 M</td>
<td>17 K</td>
<td>17 K</td>
<td>21</td>
<td>51</td>
<td>5 MB</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ZEINCRO</td>
<td>0.4 M</td>
<td>49 K</td>
<td>49 K</td>
<td>24</td>
<td>59</td>
<td>15 MB</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1.3 M</td>
<td>162 K</td>
<td>162 K</td>
<td>81</td>
<td>198</td>
<td>51 MB</td>
<td>8 KB</td>
<td>10 sec</td>
</tr>
<tr>
<td><strong>External Dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>World Bank</td>
<td>77 M</td>
<td>10 M</td>
<td>10 M</td>
<td>58</td>
<td>40 K</td>
<td>19 GB</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IMF</td>
<td>18 M</td>
<td>1.8 M</td>
<td>1.8 M</td>
<td>30</td>
<td>3151</td>
<td>3.51 GB</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Eurostat</td>
<td>0.3 M</td>
<td>38 K</td>
<td>44 K</td>
<td>31</td>
<td>5717</td>
<td>205 MB</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Trans. Int.</td>
<td>43 K</td>
<td>3939</td>
<td>4286</td>
<td>64</td>
<td>5290</td>
<td>9.2 MB</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>95 M</td>
<td>12 M</td>
<td>2 M</td>
<td>183</td>
<td>54 K</td>
<td>23 GB</td>
<td>12 KB</td>
<td>571 sec</td>
</tr>
</tbody>
</table>
## Evaluation - Berlin SPARQL Benchmark

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Q11</th>
<th>Q12</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Triple Patterns</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>16</td>
<td>7</td>
<td>8</td>
<td>11</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td># of Sources</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td># of Results</td>
<td>41</td>
<td>50</td>
<td>348</td>
<td>41</td>
<td>62</td>
<td>1983</td>
<td>5</td>
<td>10</td>
<td>1701</td>
<td>19656</td>
<td>570</td>
<td>41</td>
</tr>
<tr>
<td>Filters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 9 Patterns</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIMIT Modifier</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order By Modifier</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISTINCT Modifier</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>REGEX Operator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>UNION Operator</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation - Source Selection

- Sum of triple-pattern-wise sources selected for each query

<table>
<thead>
<tr>
<th>Systems</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Q11</th>
<th>Q12</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAFE</td>
<td>8</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>15</td>
<td>13</td>
<td>15</td>
<td>16</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>FedEx</td>
<td>9</td>
<td>13</td>
<td>16</td>
<td>24</td>
<td>20</td>
<td>14</td>
<td>16</td>
<td>19</td>
<td>15</td>
<td>17</td>
<td>9</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

- Number of SPARQL ASK requests used for source selection

<table>
<thead>
<tr>
<th>Systems</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Q11</th>
<th>Q12</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAFE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FedEx</td>
<td>36</td>
<td>28</td>
<td>40</td>
<td>64</td>
<td>48</td>
<td>40</td>
<td>44</td>
<td>40</td>
<td>21</td>
<td>21</td>
<td>9</td>
<td>21</td>
<td>35</td>
</tr>
</tbody>
</table>
Evaluation - Source Selection Time

- Source Selection Time

![Graph showing evaluation results for source selection time. The graph compares SAFE (S.S.) and FedEx (S.S.) with query times in milliseconds on a log scale.](image-url)
Evaluation - Query Execution Time

- Query Execution Time

### SAFE vs. FedEx

<table>
<thead>
<tr>
<th>Query</th>
<th>SAFE</th>
<th>FedEx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>227</td>
<td>336</td>
</tr>
<tr>
<td>Q2</td>
<td>36,266</td>
<td>8,656</td>
</tr>
<tr>
<td>Q3</td>
<td>34,161</td>
<td>3,851</td>
</tr>
<tr>
<td>Q4</td>
<td>3,583</td>
<td>501</td>
</tr>
<tr>
<td>Q5</td>
<td>89,168</td>
<td>860</td>
</tr>
<tr>
<td>Q6</td>
<td>45,031</td>
<td>938,722</td>
</tr>
<tr>
<td>Q7</td>
<td>36,024</td>
<td>2407</td>
</tr>
<tr>
<td>Q8</td>
<td>41,154</td>
<td>19,648</td>
</tr>
<tr>
<td>Q9</td>
<td>14,888</td>
<td>88,968</td>
</tr>
<tr>
<td>Q10</td>
<td>13,844</td>
<td>185,392</td>
</tr>
<tr>
<td>Q11</td>
<td>223</td>
<td>368</td>
</tr>
<tr>
<td>Q12</td>
<td>32,763</td>
<td>0.21</td>
</tr>
<tr>
<td>Avg</td>
<td>32,763</td>
<td>0.21</td>
</tr>
</tbody>
</table>

*Query times-out for FedEx*
SAFE - Highlights

- **Source Selection**
  - SPARQL SERVICE
  - Using SPARQL ASK queries
  - Using a catalog/index
  - SAFE - Hybrid (catalog/index + ASK)

- **Lightweight Cache**
  - RDF Cube Data Structure
  - AccessPolicy

- **Join Aware**
  - Excludes ineligible sources before actual query join

- **Provenance – via RDF Named Graphs**
  - Self-contained Data Cubes
  - Creator
  - Location
  - Date
  - Access rights
Conclusion & Future Work

✓ Efficient source selection with a lightweight indexing
✓ Policy aware query execution
✓ Evaluated against internal and external sets
✓ Performance is significantly improved compared to FedX

■ Cooking at the moment!

□ Evaluation extended to federation engines (ANAPSID, HiBISCuS)
□ Benchmarking for query federation over statistical data cubes
□ SAFE extension for normal RDF data
SAFE - Team

- Yasar Khan
- Muhammad Saleem
- Aftab Iqbal
- Muntazir Mehdi
- Aidan Hogan
- Panagiotis Hasapis
- Axel-Cyrille Ngonga Ngomo
- Stefan Decker
- Ratnesh Sahay

http://linked2safety.hcls.deri.org:8080/SAFE-Demo/

Thank You