SAFE: Policy Aware SPARQL Query Federation over RDF Data Cubes

Yasar Khan*, Muhammad Saleem~, Aftab Iqbal*,
Muntazir Mehdi* and Ratnesh Sahay*

***Insight Centre for Data Analytics, NUI Galway, Ireland**
~AKSW, University of Leipzig, Germany

Inspired by the publication of hundreds of Linked Datasets on the Web, researchers have been investigating federated querying techniques to enable access to this decentralised content. Query federation aims to offer clients a single-point-of-access through which distributed data sources can be queried in unison. In the context of Linked Data, various optimised query federation engines have been proposed that can federate multiple SPARQL interfaces. However, in the context of the Healthcare and Life Sciences (HCLS) domain – where data-integration is often vital – real world datasets contain sensitive information: strict ownership is granted to individuals working in hospitals, research labs, clinical trial organisers, etc. Therefore, the legal and ethical concerns on (i) preserving the anonymity of patients (or clinical subjects); and (ii) respecting data ownership through access control; are key challenges faced by the data analytics community working within the HCLS domain. However, our focus is on point (ii), i.e., develop a policy-based access control mechanism for user-restricted Linked Data resources residing at different locations.

The key challenges for federated querying are efficient source selection (i.e., determining which sources are (ir) relevant) and query planning (i.e., determining an efficient query execution strategy). Query federation engines often apply source selection at the level of endpoints, whereas in a controlled environment (e.g., healthcare), a user may only have access to certain information within an endpoint. Adding an access control layer to existing SPARQL query federation engines thus adds unique challenges: (i) source selection should be granular enough to enable effective access control, and (ii) it should be policy aware to avoid wasteful requests to unauthorised resources. Therefore, we developed SAFE [1], a SPARQL query federation engine that supports policy-based access to sensitive statistical data.

SAFE is motivated by the needs of three clinical organisations: University Hospital Lausanne (CHUV)[1], Cyprus Institute of Neurology and Genetics (CING)[2], and ZEINCRO[3], in the context of an EU project who wish to enable controlled federation over statistical clinical data – such as data from clinical trials – owned and hosted by multiple clinical sites, represented in the form of data cubes: multi-dimensional arrays of numeric data. These organisations wish to develop a platform for analysing clinical data across multiple clinical sites, which would allow for increasing the total number of patients that are included in each analysis, thus increasing the statistical power of conclusions related to biomarkers, effectiveness and/or side-effects of drugs or combinations of drugs, correlations between patient groups, etc. The ultimate goal is to enable the collaborative identification of new drugs and treatments while reducing the high costs associated with clinical trials. In order to employ data access restrictions on the anonymised multi-dimensional RDF data cubes, we thus require an access-control–based query federation approach that enforces and optimises for restricted user access over these RDF data cubes. To further illustrate and motivate, we now walk through an example.

[1] http://www.chuv.ch/
[2] http://www.cing.ac.cy/
[3] http://www.zeincro.com/

Figure 1 shows four sample data cubes published by three different clinical sites. Each observation represents the total number of patients exhibiting a particular adverse event. For example, the CS1-S1 observations describe the total number of patients (in the **Cases** column) that exhibit a particular combination of three adverse events: **Diabetes**, (Abnormal) **BMI_Abnormal** (Body Mass Index) and/or **Hypertension**. The value 0 or 1 indicates if the condition is present or not. For example, the second row in CS1-S1 shows that there are 26 cases presenting with both **Diabetes** and **Hypertension** but without **BMI_Abnormal**. Once the data are published by clinical sites, they should be accessible to clinical researchers. Figure 2 shows a sample SPARQL query specifying subject-selection criteria, asking for the counts of cases that involve some combination of diabetes, abnormal BMI, and hypertension. An answer returned by the query, i.e., number of cases, will play a major role in deciding the resources (i.e., number of subjects, location, etc.) required for conducting a clinical trial. However, answering such a query requires integrating RDF data cubes with three dimensions – **Diabetes**, **Hypertension**, **BMI_Abnormal** – and the respective counts originating from multiple clinical sites. Referring back to Figure 1, only three of the datasets (CS1-S1, CS2-S2 and CS3-S3) contain all required dimensions. An answer returned by the query (Figure 2) should list counts (i.e., cases) from these three RDF data cubes. However, assuming that the policy restrictions are applied to the user (say *Marie Stripolie*), who wants to execute the query and has access to CS1-S1 and CS2-S2 RDF data cubes only. Therefore, the query federation engine should retrieve results only from CS1-S1 and CS2-S2 and should not consider CS3-S3 for querying.

Diabetes	BMI_Abnormal	Hypertension	Cases
0	0	0	11
1	0	1	26

CS1 – S1

Diabetes	BMI_Abnormal	Hypertension	Cases
0	0	0	40
1	0	1	50

CS2 – S2

Diabetes	BMI_Abnormal	Hypertension	HIV	Cases
0	0	0	0	30
1	0	1	0	60

CS3 – S3

Diabetes	Smoking	Gender	Cases
0	0	0 (F)	90
1	0	1 (M)	120

CS1 – S4

Figure 1: Example data cubes published by three Clinical Sites (CS) - CS1, CS2 and CS3

```
1  PREFIX qb: <http://purl.org/linked-data/cube#>
2  PREFIX sehr: <http://hcls.deri.ie/l2s/sehr/1.0/>
3  SELECT ?diabetes ?bmi ?hypertension ?cases
4  WHERE { ?dataset a qb:DataSet.
5          ?observation qb:dataSet  ?dataset;
6          a qb:Observation;  sehr:Diabetes ?diabetes ;
7          sehr:BMI_Abnormal ?bmi ;
8          sehr:Hypertension ?hypertension ; sehr:Cases ?cases . }
```

Figure 2: Example subject selection criteria for clinical trials datasets

SAFE extends upon the FedX engine with two novel contributions: (i) GRAPH LEVEL SOURCE SELECTION in order to enable graph-based access-control and (ii) OPTIMISATIONS FOR FEDERATING QUERIES OVER STATISTICAL DATA that are represented using the RDF Data Cube Vocabulary. With these modifications, SAFE can (i) support more granular graph-level access control on top, and can (ii) efficiently reduce the query execution time when federating over RDF data cubes. It is important to note that no existing SPARQL query federation engine supports policy-aware access control over statistical datasets. Therefore, we argue that a specialised federation engine is required to address the specific challenges in combining statistical and distributed datasets with access restrictions.

SAFE's architecture is summarised in Figure 3, which shows its three main components: (i) Source Selection: performs multilevel source selection based on the capabilities of data sources; (ii) Policy Aware Query Planning: filters the selected data sources based on access rights defined for each user; and (iii) Query Execution: performs the execution of sub-queries against the selected sources and merge the results returned.

The demo for SAFE can be found at http://linked2safety.hcls.deri.org:8080/SAFE-Demo/. Please note that, for demo purposes we use fake user accounts and data due to ethical, legal and privacy concerns. Information on how to use the demo can be found on the demo site.

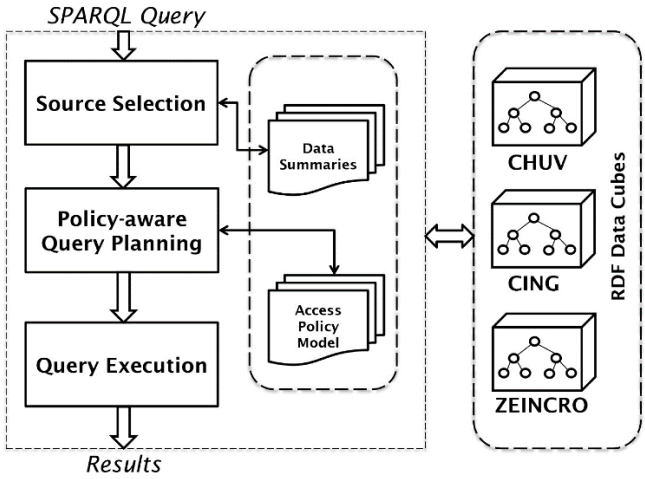

Figure 3: SAFE Architecture

Reference

[1] Yasar Khan, Muhammad Saleem, Aftab Iqbal, Muntazir Mehdi, Panagiotis Hasapi, Axel-Cyrille Ngonga Ngomo, Stefan Decker and Ratnesh Sahay. "SAFE: Policy Aware SPARQL Query Federation Over RDF Data Cubes". In Proceedings of the Semantic Web Applications and Tools for the LifeSciences (SWAT4LS), 2014, Germany.