Efficient Database Model to Represent Numerical Research Data of Material Flow Analysis.

Master Thesis - Computer Science

06 June 2018
Mohammad Mahadi Hasan
Outline

- **Introduction**
  - Problem Statement
  - Approach

- **Background**
  - Relational Database
  - The Semantic Web
  - Triple stores

- **Construction of Relational Database**
  - MySQL architecture and design process
  - Implementation of web interface

- **Construction of RDF Triple Store**
  - Mapping relational database to RDF triples
  - Apache Jena TDB architecture

- **Database Performance Analysis**

- **Discussion**
Introduction

- **Problem Statement**

Usually, the research data of Industrial Ecology (IE) is stored in .mat or .csv files on a local machine which imposes challenges like:

- Lack of an efficient data storing mechanism.
- Lack of an online platform to query and retrieve data publicly.
- Data sharing and re-using terms has not been discussed for such data.

- **Approach**

- Study on Relational Database (RDB)
  - MySQL database.
  - Implementation of web interface.

- Study on Semantic Web and its components
  - RDF - Data Model.
  - Apache Jena TDB triple store.
Relational Database (RDB)

- A collection of data sets organized and stored in relational tables.
- Each of these tables has one primary key column and shares at least one column (referred as foreign key) with another table to establish relationship.
- Data can be extracted by querying with query language like SQL.
- RDBMS is a database management system that is used to create and maintain relational databases.
- Popular RDBMS examples include Microsoft Access, SQL Server, Oracle Database, MySQL, PostgreSQL etc.
MySQL Database Architecture

- **The application layer**: client-side layer which handles the connection string, authentication and most importantly security.
- **The server layer**: the brain of the overall architecture. Any kind of query statement is executed in this layer.
- **The storage engine layer**: offers different storage engines.

**Indexes:**

- Indexes are used to find a data entry quickly.
- Let us consider following example:
  
  ```sql
  SELECT * FROM table WHERE id=1;
  ```

- Without the indexes, the query will go through every row and column.
- The optimization here is to add an index, for example, primary key. It runs the query only against the table indexes rather than all the column data.
The Semantic Web

- "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries".
- Data is structured and published following semantic web standards: RDF.

RDF- The Data Model

Resource Description Framework is a "framework for representing information in the Web".

- The basic elements of RDF are triples.
- A triple is a set of three entities that form a statement in the form of subject-predicate-object expressions.

*Harry Potter → has author → J. K. Rowling*
Background

URIs, Vocabularies & Ontologies

- Uniform Resource Identifiers (URIs) are short strings that identify resources in the web.

\[
<\text{http://dbpedia.org/resource/Berlin}> \quad (1) \\
<\text{http://www.w3.org/1999/02/22-rdf-syntax-ns#type}> \quad (2) \\
<\text{http://dbpedia.org/class/yago/CapitalsInEurope}> \quad (3)
\]

\( subject \rightarrow predicate \rightarrow object \)

\( (1) \rightarrow (2) \rightarrow (3) \)

This triple refers that resource (1) has a relationship (2) with the resource (3).

- The “vocabularies” on the Semantic Web is used for data integration between data sets.

- The term “ontology” is for more complex, and more formal collection of terms, where “vocabulary” is used for more basic use.
SPARQL – Query Language

- SPARQL is the standardized query language for RDF, just like SQL is the standardized query language for relational databases.
- Like SQL, SPARQL also follow the same “SELECT...FROM...WHERE...” query structure.

```
:id1 foaf:name "André Schürrle"
:id1 foaf: based_near : Dortmund
:id2 foaf:name "Nils Petersen"
:id2 foaf:based_near : Freiburg

SELECT ?name
WHERE {
  ?x foaf: name ?name .
  ?x foaf:based_near : Freiburg .
}
```

The predicate has a constant value of foaf:based_near and the object has a constant value of : Freiburg match to one of RDF triples. The result is Nils Petersen.
Triple Stores

- Triple Store (RDF triple store) is a specialized DBMS for RDF triples.
- RDF data can be stored in two ways:
  - in files, triples are stored following one of the serialization formats, or
  - in special kind of databases for triples, called triple stores.

- Triple stores have three possible architectures:
  - In-memory: stores the triples in main memory. It is fast but expensive.
  - Native Store: storage systems with own database. For example- Jena TDB, Sesame Native, Virtuoso, AllegroGraph, Oracle 11g etc.
  - Non-native Store: storage using a third-party RDBMS. For example- Jena SDB backed by MySQL database.
MySQL Design Process:

- Six different CSV files containing the data of country, unit specification, dataset (list of datasets), process list, stock data and flow data between processes.

- Database Schema, Tables, and Constraints
  - Database Schema => database name.
  - 6 tables.
  - 6 primary keys and 11 foreign keys; are said to be database constraints.
  - 589165 data rows.
  - Roughly 55mb of SQL dump.
## Construction of Relational Database

### stocks
- `seq_id`
- `system_id`
- `stock_dataset_number`
- `process_id`
- `ISO_code`
- ...
- `Unit_id`
- `comment`

### countries
- `seq_id`
- `system_id`
- `ISO_code`
- `name`
- ...
- `alternative_name7`

### flows
- `seq_id`
- `system_id`
- `flow_dataset_number`
- `process_id_source`
- `process_id_target`
- `region_source`
- `region_target`
- ...
- `Unit_id`
- `comment`

### dataset
- `system_id`
- `system_definition`
- `dataset_name`
- ...
- `Unit_id`
- `comment`

### process_list
- `system_id`
- `process_id`
- `process_name`
- ...
- `YPos`

### Unit_classification
- `unit_id`
- `si_unit_id`
- `unit_code`
- ...
- `factor`

---

**Green** columns = primary keys.
**Purple** columns = foreign keys.
Implementation of web interface

- Back-End Development
  - Mechanisms to communicate with server-side database.

- Front-End Development
  - Mechanisms to present data in client-side.

Figure: web interface architecture
Data Table with the catalogue of datasets:

- **Query 1:** Select all the dataset available in the database from dataset table;

![Catalogue of Data Sets](image)

Figure: DataGrid with list of datasets available in the database
Data Table with material stocks:

- **Query 2:** Select all the stocks within the dataset (identified by the id) and filter the data with year (2007, 2008) and country (Austria, Germany, Switzerland and Italy) from stock table;

Figure: Data Grid with stock data between processes
Construction of Relational Database

Data Table with material flows:

- **Query 3:** Select all the flows within the dataset (identified by the id) and filter the data with year (2005) and source process ‘scrap market’ and target process ‘foundries’ from flow table;

Figure: Data Grid with flow data between processes
Construction of Triple Store

- The majority of data on the current Web is stored in relational databases.
- Semantic web is useful, especially if data from different sources has to be shared or integrated.
- Therefore, it is important to introduce mapping technologies between relational database and RDF.

- **D2RQ**: A mapping mechanism to use RDB as RDF graphs (in addition, it exposes RDB to RDF triples)
D2RQ architecture

- A D2RQ Engine accesses a Non-RDF database.
- The mapping language describes the relationship between an ontology (or vocabulary) and a relational data model.
- RDF APIs can be embedded with Java applications via Jena or Sesame APIs.
- The RDF dump file can be stored in any triple store.
- D2RQ server provides an HTML view to explore the mapped database.

Figure: Architecture of D2RQ
Construction of Triple Store

D2RQ: Overview and Features

- The D2RQ platform serves different tools to offer RDF-based access to the content of relational databases.

  - The *generate-mapping* tool creates a D2RQ mapping file by analysing the relational database schema.

    ```
    >generate-mapping -u username -p password -o mapping.ttl jdbc:mysql:///ief
    ```

  - The *dump-rdf* tool is used to dump the contents of the entire database into a single RDF file.

    ```
    ```

  - The *D2R-server* takes the mapping file as input and provides a web interface where RDF data can be browsed.

    ```
    >d2r-server mapping.ttl
    ```
Construction of Triple Store

Database table to RDF triples with D2RQ:

```
<"primary key column value" "tableName_columnName" "columnValue"">

<table>
<thead>
<tr>
<th>id</th>
<th>country</th>
<th>aspect_of_dataset</th>
<th>indicator_element</th>
<th>process</th>
<th>system_id</th>
<th>unit_id</th>
<th>value</th>
<th>year</th>
</tr>
</thead>
<tbody>
<tr>
<td>17011</td>
<td>Germany</td>
<td>Stock</td>
<td>slag</td>
<td>Scrap market</td>
<td>1</td>
<td>kilo ton</td>
<td>444.090</td>
<td>1906</td>
</tr>
</tbody>
</table>

"stocks17011" "stocks_country" "Germany">
"stocks17011" "stocks_aspect_of_dataset" "Stock">
"stocks17011" "stocks_indicator_element" "slag">
"stocks17011" "stocks_process" "Scrap market">
"stocks17011" "stocks_system_id" "1">
"stocks17011" "stocks_unit_id" "kilo ton">
"stocks17011" "stocks_value" "444.0907">
"stocks17011" "stocks_year" "1906">
```
Construction of Triple Store

### Apache Jena TDB

- TDB is a component of Jena that used as native RDF storage.
- TDB is stored in a single directory in the filing system backed by a dataset.
- A complete TDB dataset consists of:
  - **The node table**: stores the representation of RDF terms. It consists of two mappings: Node to Nodeld and Nodeld to Node.
  - **Triple and Quad indexes**: used for the default graph. Triples are held as 3-tuples of Nodelds in triple indexes where quads are held as 4-tuples of Nodelds.
  - **The prefixes table**: stores index for Graph->Prefix->URI mapping. It provides the mechanism for Jena API to serialize the triples in RDF/XML or Turtle.
Experimental Settings:

- Two training datasets:
  - **IEF dataset:** consisting the material stock and flow data of Industrial Ecology Freiburg research group.
    - 6 tables, 589165 rows.
    - D2RQ generates roughly 10 million triples.
    - 7596007 final triples (ignoring the empty “ ” column values).
  - **ISWC dataset:** sample dataset with the information about conferences, papers, authors and topics from ISWC 2002 conference¹.
    - 9 tables, 96 rows.
    - 322 triples after mapping.

¹http://iswc2002.semanticweb.org/
### Results Evaluation:

<table>
<thead>
<tr>
<th>SQL Queries on MySQL</th>
<th>SPARQL Queries on Jena TDB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query Execution Time (ms)</strong> (average of 10 run)</td>
<td><strong>Query Execution Time (ms)</strong> (average of 10 run)</td>
</tr>
<tr>
<td><strong>Query 1</strong></td>
<td>9</td>
</tr>
<tr>
<td><strong>Query 2</strong></td>
<td>127</td>
</tr>
<tr>
<td><strong>Query 3</strong></td>
<td>80</td>
</tr>
<tr>
<td><strong>Query 4</strong></td>
<td>16</td>
</tr>
</tbody>
</table>

Table 1: Query execution time comparisons for ief dataset
Database Performance Analysis

Results Evaluation:

<table>
<thead>
<tr>
<th>SQL Queries on MySQL</th>
<th>SPARQL Queries on Jena TDB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Query Execution Time (ms)</td>
</tr>
<tr>
<td></td>
<td>(average of 10 run)</td>
</tr>
<tr>
<td>Query 9</td>
<td>3</td>
</tr>
<tr>
<td>Query 10</td>
<td>3</td>
</tr>
<tr>
<td>Query 11</td>
<td>2</td>
</tr>
<tr>
<td>Query 12</td>
<td>2</td>
</tr>
<tr>
<td>Query 13</td>
<td>132</td>
</tr>
<tr>
<td>Query 14</td>
<td>7</td>
</tr>
<tr>
<td>Query 15</td>
<td>5</td>
</tr>
<tr>
<td>Query 16</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 2: Query execution time comparisons for iswc dataset
Results Evaluation:

- MySQL works faster than TDB
  - MySQL storage engines are faster architecturally.
  - Use of indexing to identify records.

- TDB works slower because of enumeration
  - TDB query execution method does not finish until results are fully enumerated.
  - Enumeration process takes longer time depending on number of predicates (number of columns in case of MySQL) and triples.

```java
ResultSet results = qexec.execSelect();
long numResults = ResultSetFormatter.consume(results);
```

The variation of query execution time in TDB (when triples are available in main memory!).
Discussions

- Most of the web contents are backed by RDB.
- RDB’s are easy to implement; become complex with the number of tables increases.
- RDB works as “closed loop”.
- Data can not be integrated from different sources if the database schemas are not in same structure.

- Triple Stores allow data to be extended across different data sources.
- Data can be integrated with multiple data sources by adding properties.
- A small data source can be enlarged into a bigger and richer data source.
Discussions

Future Work

- Query optimization for SQL joins.
- Allow filtering data with a range of years (1901-2000 instead of using 1901, 1902, ..., 2000).
- Introduce filtering based on a collection of regions (not only country wise, also continents wise like “Europe“).
- Improvement of default RDF vocabulary terms within D2RQ (use of more meaningful and publicly well-known vocabulary).
- Build a domain-specific complete ontology from the dataset and make it public.
Thank You!