QB4OLAP: A New Vocabulary for OLAP Cubes on the Semantic Web

Lorena Etcheverry\textsuperscript{1}  Alejando A. Vaisman\textsuperscript{2}

\textsuperscript{1}lorenae@fing.edu.uy
Instituto de Computación
Universidad de la República, Uruguay

\textsuperscript{2}avaisman@ulb.ac.be
Université Libre de Bruxelles, Belgium

COLD 2012 - 11th International Semantic Web Conference,
Boston, 2012
Outline

1. Motivation and Contribution
2. Preliminaries
   - Multidimensional Model
   - OLAP Operators
3. The QB4OLAP Vocabulary
   - Representing the Model in QB4OLAP
   - Implementing OLAP Operators in QB4OLAP
   - Analyzing a QB cube using QB4OLAP
4. Conclusion and Future Work
   - Conclusion
   - Future Work
Outline

1. Motivation and Contribution

2. Preliminaries
   - Multidimensional Model
   - OLAP Operators

3. The QB4OLAP Vocabulary
   - Representing the Model in QB4OLAP
   - Implementing OLAP Operators in QB4OLAP
   - Analyzing a QB cube using QB4OLAP

4. Conclusion and Future Work
   - Conclusion
   - Future Work
Motivation

- OLAP (On-line Analytical Processing) allows analyzing huge amounts of data for decision-making.
- Multidimensional data are seen as data cubes (DC).
- ETL (Extract, Transform, Load) process initially loads the DW; then, data is refreshed periodically.
- ETL is costly and resource-consuming.

DW Architecture, Malinowski & Zimányi, 2008
A Possible Architecture

A Possible Architecture

The Story so Far

- The RDF Data Cube Vocabulary (QB) [Cyganiak et al. 2012] (W3C Working Draft) does not directly support the classical multidimensional model for OLAP.
  - Oriented to statistical data analysis.
  - Does not represent dimension structure.
  - Does not bind measures to aggregate functions.
  - Dimension hierarchies not accounted for directly.
  - **Consequence:** OLAP operators are difficult to define over QB (see Kämpgen et al., ILD, ESWC 2012).

- The OpenCube vocabulary (OC) [Etcheverry and Vaisman. ESWC 2012] fully represents OLAP model, but ...
  - no relationship is provided among concepts in OC and QB
  - **Consequence:** is not possible to use OC operators over data already published in QB. Data must be rewritten.
An RDF vocabulary (QB4OLAP), that fully represents the classical multidimensional model, based on OC and QB.

A set of OLAP operators implemented as SPARQL queries.

Algorithms that:

- automatically build the SPARQL queries that implement OLAP operators.
- build QB4OLAP cubes from QB cubes, allowing the reuse of published data without rewriting observations.

Contribution
Outline

1. Motivation and Contribution

2. Preliminaries
   - Multidimensional Model
   - OLAP Operators

3. The QB4OLAP Vocabulary
   - Representing the Model in QB4OLAP
   - Implementing OLAP Operators in QB4OLAP
   - Analyzing a QB cube using QB4OLAP

4. Conclusion and Future Work
   - Conclusion
   - Future Work
Multidimensional Model

DATE dimension schema

year \rightarrow \text{month}

year_1 \leftarrow \text{month}_1
\text{month}_1
\text{month}_i

DATE dimension instance

Measures

#People

Aggregate Function: sum

Each measure is associated with an AGGREGATE FUNCTION that is used to compute measure values while traversing dimension hierarchies

GEO_TO dimension schema

region \rightarrow \text{country}

region_1
region_2
\ldots
\text{country}_1
\text{country}_j
\text{country}_k

GEO_TO dimension instance

GEO_FROM dimension schema

region \rightarrow \text{country}

GEO_FROM dimension instance

Also suppose a GENDER dimension with only one level (GENDER) and two level members (F,M)
Instance Example

Number of people that migrated between regions in Spain

<table>
<thead>
<tr>
<th>FROM</th>
<th>TO</th>
<th>Fem 2006</th>
<th>Masc 2006</th>
<th>Fem 2007</th>
<th>Masc 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galicia</td>
<td>Com. Valenciana</td>
<td>–</td>
<td>–</td>
<td>678</td>
<td>–</td>
</tr>
<tr>
<td>Navarra</td>
<td>Com. Valenciana</td>
<td>242</td>
<td>266</td>
<td>318</td>
<td>427</td>
</tr>
<tr>
<td>Com. Valenciana</td>
<td>Galicia</td>
<td>780</td>
<td>–</td>
<td>930</td>
<td>–</td>
</tr>
</tbody>
</table>
OLAP Operators

(a) Cube $C$

(b) $Slice(C, Gender)$

(c) $RollUp(C, From, country)$

(d) $Dice(C, \#people > 500)$
Outline

1 Motivation and Contribution

2 Preliminaries
   - Multidimensional Model
   - OLAP Operators

3 The QB4OLAP Vocabulary
   - Representing the Model in QB4OLAP
   - Implementing OLAP Operators in QB4OLAP
   - Analyzing a QB cube using QB4OLAP

4 Conclusion and Future Work
   - Conclusion
   - Future Work
The QB4OLAP Vocabulary
Schema Examples

**Dimension Schema**

```
eg:geoFrom a qb:DimensionProperty.
eg:country a qb4o:LevelProperty;
  qb4o:inDimension eg:geoFrom.
eg:region a qb4o:LevelProperty;
  qb4o:inDimension eg:geoFrom;
  qb4o:parentLevel eg:country.
```

**Fact Schema**

```
eg:migr_2es a qb:DataStructureDefinition ;
  qb:component [ qb:dimension eg:date] ;
  qb:component [ qb:dimension eg:geoFrom] ;
  qb:component [ qb:dimension eg:geoTo] ;
  qb:component [ qb:dimension eg:gender] ;
  qb:component [qb:measure eg:numPeople;
    qb4o:hasAggregateFunction qb4o:sum] .

eg:dataset_migr qb:structure eg:migr_2es.
```
Instance Examples

Dimension Instance$^{ab}$

geo:ES qb4o:inLevel eg:country.
eg:ES11 qb4o:inLevel eg:region;
rdf:label "Galicia";
skos:broader geo:ES .
eg:ES22 qb4o:inLevel eg:region;
rdf:label "Navarra";
skos:broader geo:ES .
eg:ES52 qb4o:inLevel eg:region;
rdf:label "Comunidad Valenciana";
skos:broader geo:ES .

Fact Instance

eg:migr_r_2es1 a qb:Observation;
qb:dataSet eg:dataset_migr;
eg:date 2007;
eg:geoFrom geo:ES11;
eg:geoTo geo:ES52;
eg:gender sex:F;
eg:numPeople 678.0.

$^a$prefix geo: http://eurostat.linked-statistics.org/dic/geo
$^b$prefix sex: http://eurostat.linked-statistics.org/dic/sex
Roll-Up Example

RollUp(Migr2es,From,country)

    ?id eg:geoFrom ?countryFrom . ?id eg:geoTo ?regionTo .
WHERE { 
    SELECT ?date ?countryFrom ?regionTo ?gender (SUM(?numPeople) AS ?numPCountry) 
    (iri(f(?date,?from,?regionTo,?gender)) AS ?id) 
WHERE { 
    ?o eg:numPeople ?numPCountry .
    ?regionFrom skos:broader ?countryFrom . ?countryFrom qb4o:inLevel eg:country
}GROUP BY ?date ?countryFrom ?regionTo ?gender}
Creating a QB4OLAP schema from a QB schema

Given a schema inQB, a schema in QB4OLAP can be built, that allows to apply OLAP operators over existent QB observations. The algorithm does not modify the set of observations.

INPUT:

- a schema in QB
- for each measure in the schema, an aggregate function
- for each dimension in the schema, a set of levels and a partial order among them (parent level)
- for each dimension level member, its correspondent dimension level and its corresponding level member in the parent level
Creating a QB4OLAP schema from a QB schema II

1: for all $d_i \in D_1$ ($d_i$ a qb:DimensionProperty) do
2: Create a new dimension $d_j \in D_2$ ($d_j$ a qb:DimensionProperty)
3: Add triples ($d_i$ a qb4o:LevelProperty) and ($d_i$ qb4o:inDimension $d_j$)
4: Add a triple ($dsd_2$ qb:component [qb4o:level]$d_i$])
5: Obtain a hierarchy of levels $hl_i$ and a hierarchy of level members $hm_i$ for $d_i$
6: for all $l_i \in hl_i$ do
7: Add triples ($l_i$ a qb4o:LevelProperty) and ($l_i$ qb4o:inDimension $d_j$)
8: for all $lm_i \in hm_i$ such that $lm_i$ belongs to level $l_i$ do
9: Add a triple ($lm_i$ qb4o:inLevel $l_i$).
10: end for
11: end for
12: for all $(l_j, l_k) \in hl_i$ such that $l_j \rightarrow l_k$ do
13: Add a triple ($l_j$ qb:parentLevel $l_k$)
14: end for
15: end for
16: for all $m_i$ such that ($dsd_1$ qb:component [qb:measure $m_i$]) do
17: Add a triple ($dsd_2$ qb:component [qb:measure $m_i$;qb:hasAggregateFunction $ag_i$])
18: end for

COMPLEXITY: the main source of complexity is the size of the set of dimension level members, which is usually small, compared to the set of observations (facts).
Outline

1 Motivation and Contribution

2 Preliminaries
   - Multidimensional Model
   - OLAP Operators

3 The QB4OLAP Vocabulary
   - Representing the Model in QB4OLAP
   - Implementing OLAP Operators in QB4OLAP
   - Analyzing a QB cube using QB4OLAP

4 Conclusion and Future Work
   - Conclusion
   - Future Work
Conclusion

- An RDF vocabulary for representing the classical multidimensional model s.t.
  - ANSI architecture is supported (conceptual, logical and physical levels clearly identified).
  - OLAP applications and operators can be implemented naturally and easily maintained and extended.
- A set of OLAP operators implemented as SPARQL queries.
- Algorithms that automatically build the SPARQL queries that implement such OLAP operators.
- Preliminary tests over proof-of-concept prototype.
- More info: https://code.google.com/p/publishing-multidimensional-data/
Future Work

- Extend the operator set (e.g., Drill-Across).
- Perform stress tests with real data
- Query processing and optimization.
- Incorporate all of these into the general framework.
Thanks for your attention.
Questions?
Contact:

- Lorena Etcheverry lorenae@fing.edu.uy
- Alejandro A. Vaisman avaisman@ulb.ac.be

R. Cyganiak and D. Reynolds.
The RDF Data Cube Vocabulary, March 2012.

L. Etcheverry and A. A. Vaisman.
Enhancing OLAP Analysis with Web Cubes.
