An ETL Process for OLAP Using RDF/OWL Ontologies

Marko Niinimäki and Tapio Niemi
Helsinki Institute of Physics, Technology Programme,
CERN, CH-1211 Geneva 23
Tel.: +41 22 7676179; Fax: +41 22 7673600
man@cern.ch, tapio.niemi@cern.ch

Abstract. In this paper, we present an advanced method for on-demand con-
struction of OLAP cubes for ROLAP systems. The method contains the steps
from cube design to ETL but focuses on ETL. Actual data analysis can then be
done using the tools and methods of the OLAP software at hand. The method is
based on RDF/OWL ontologies and design tools. The ontology serves as a basis
for designing and creating the OLAP schema, its corresponding database tables,
and finally populating the database.

Our starting point is heterogeneous and distributed data sources that are even-
tually used to populate the OLAP cubes. Mapping between the source data and
its OLAP form is done by converting the data first to RDF using ontology maps.
Then the data are extracted from its RDF form by queries that are generated us-
ing the ontology of the OLAP schema. Finally, the extracted data are stored in the
database tables and analysed using an OLAP software. Algorithms and examples
are provided for all these steps.

In our tests, we have used an open source OLAP implementation and a database
server. The performance of the system is found satisfactory when testing with a
data source of 450 000 RDF statements. We also propose an ontology based tool
that will work as a user interface to the system, from design to actual analysis.

Keywords: OLAP, ontology, ETL.

1 Introduction

The amount of available digital data is ever increasing mostly due to the World Wide
Web and intrawebs of organizations. In corporate environments, taking maximal benefit
of these data is essential for competitiveness and the success of a company. This is chal-
lenging since the data are located in various locations and stored in different formats. It
is obvious that only a very limited part of the data can be found and analysed manually.
Automated and powerful enough methods to fetch the right raw data and process it into
a form usable for analysis are thus needed.

To make this possible, the data must be made available in a machine readable format.
The Semantic Web technologies can be seen as a working solution, since they give a
standard method (Resource Description Framework, RDF) to store data items and their
relationships in a machine readable form. Data extraction can be done by using RDF
query languages. It is also possible (though not always trivial) to map XML or relational
data into the RDF form.
For actual data analysis there exist several techniques from statistical methods to spreadsheet tools. For analysing large amounts of data and trying to find cause and effect relationships, On-Line Analytic Processing (OLAP) has become popular during the past years. Briefly, OLAP is a decision making tool for analysts. It is designed to analyse large data warehouses in a multidimensional fashion and on different levels of details. The multidimensional data structure of OLAP is called an OLAP cube (for details, see e.g. [10]).

In theory, one can populate an OLAP cube instance using all the data in data sources. In many cases, this can lead to a massively sparse and inefficient OLAP cube. In some cases, such as using the World Wide Web as an information source, this approach would be totally impossible. Practically, the cube should be created on-demand, so that only required parts of the data sources are extracted and imported into the OLAP cube. To do so, we have to locate the data sources, convert them in a format that makes the semantics of the data explicit, and create an OLAP cube using parts of the data.

In this paper, we combine these two powerful technologies, Semantic Web and OLAP, resulting in a method of creating a fully functional tool for analysis. The method uses an ontology and mapping files that connect the data sources with the ontology. Extracting, transforming and loading the data (the so-called ETL process) into an OLAP system is done according to the ontology. We demonstrate our method by implementing it using open source tools.

Our starting point is a general, “meta-model” OLAP ontology that defines that there are dimensions and measures, and that the measure is a function of the dimensions. From this general OLAP ontology we form application specific sub-ontologies. These will be the starting point for the actual analysis, and the data will be formatted to conform with them. Resource Description Framework (RDF) and Ontology Web Language (OWL) are used to express these ontologies in a manner that is formal enough for analysis. We show that this model is analogous to a relational OLAP model. This gives us a clear theoretical background and a possibility to use relational OLAP servers in the implementation.

As an example in this paper, we use a Trade ontology containing information about trade of different products between countries. We demonstrate our method by combining this ontology with another one, describing country characteristics of the old editions of the World Fact Book[1]. The resulting OLAP cube can be used to analyse, for example, if countries with lots of telephones per 100 inhabitants actually export a great value in electronics products.

Since our use case is based on the idea of distributed, heterogeneous data, the data must be (i) converted into a form that conforms with our specific ontologies. This warrants that we use uniform names of dimensions and measures, as well as identifiers (like country names) that are declared by our ontology. Then (ii) parts of the data must be chosen and (iii) loaded into an OLAP system. Finally, the user can query and analyse the data using the methods offered by the OLAP system.

Phases (i) and (ii) have been discussed in our previous papers ([27], [29], [31]). Our paper about combining OLAP and RDF ontologies [29] was mainly related to data integration from distributed, heterogeneous sources, while a Grid implementation of our