Towards a Programmable Semantic Extract-Transform-Load Framework for Semantic Data Warehouses

Rudra Pratap Deb Nath
Aalborg University
Aalborg, Denmark
rudra@cs.aau.dk

Katja Hose
Aalborg University
Aalborg, Denmark
khose@cs.aau.dk

Torben Bach Pedersen
Aalborg University
Aalborg, Denmark
tbp@cs.aau.dk

ABSTRACT

In order to create better decisions for business analytics, organizations increasingly use external data, structured, semi-structured and unstructured, in addition to the (mostly structured) internal data. Current Extract-Transform-Load (ETL) tools are not suitable for this "open world scenario" because they do not consider semantic issues in the integration process. Also, current ETL tools neither support processing semantic-aware data nor create a Semantic Data Warehouse (DW) as a semantic repository of semantically integrated data. This paper describes SETL: a (Python-based) programmable Semantic ETL framework. SETL builds on Semantic Web (SW) standards and tools and supports developers by offering a number of powerful modules, classes and methods for (dimensional and semantic) DW constructs and tasks. Thus it supports semantic-aware data sources, semantic integration, and creating a semantic DW, composed of an ontology and its instances. A comprehensive experimental evaluation comparing SETL to a solution made with traditional tools (requiring much more hand-coding) on a concrete use case, shows that SETL provides better performance, knowledge base quality and porgrammer productivity.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration—*Data warehouse and repository*

Keywords

Semantic ETL Framework; RDF; Semantic Integration; Semantic Data Warehouse; Knowledge Base

1. INTRODUCTION

Business Intelligence (BI) tools support intelligent business decisions by analyzing available organizational data. Data Warehouses (DWs) are used to store the large data volumes from different operational databases in enterprises and On-Line Analytical Processing (OLAP) queries are applied on DWs to answer business questions. Extract-Transform-Load (ETL) is the backbone process of a DW. The strength of a DW depends on how good its ETL process is. Extraction retrieves data from appropriate data sources. Transformation converts the source data according to the target schema of the DW, typically either a star or snowflake schema. Loading stores the transformed data into the DW. DW/OLAP technologies perform efficiently when they are applied on data that are static in nature and well-organized in structure.

Nowadays, the Web is also an important source of business information. Moreover, Semantic Web (SW) technologies and the Linked Data (LD) principles inspire organizations to publish their business-related data using the Resource Description Framework (RDF) [29]. As a result, besides analyzing internal data available in a DW, it is often desirable to incorporate external data from various (semantic) sources into the DW to derive the needed business knowledge.

The inclusion of external data, especially RDF data, however, raises several challenges for integration and transformation in comparison to the traditional ETL process. One of the drawbacks of using RDF data in the corporate analysis process is that the data sometimes do not have any schema, or only have a poor or complex schema. Moreover, different sources describe the data in their own way, introducing heterogeneity problems. Therefore, to build a successful data warehouse system with these heterogeneous data, the integration process should emphasize the semantic relationship of the data. Traditional ETL tools are unable to process such external data because they (1) do not support semantic-aware data, (2) are entirely schema-dependent, (3) do not focus on meaningful semantic relationships to integrate data from disparate sources [4], and (4) do not support deriving new information by active inference and reasoning on the data. Thus, a DW with both internal and external (semantic) data requires more powerful tools to define, integrate and transform data semantically.

SW technology was introduced with the vision of converting the 'Web of Documents' to the 'Web of Data' where data are presented and exchanged in a machine-readable and understandable format, and data are integrated with each other semantically. The journey towards this vision is quite successful, and a number of standards, languages, and tools have been developed to express semantic-based

DOLAP'15, October 23 2015, Melbourne, VIC, Australia

DOI: http://dx.doi.org/10.1145/2811222.2811229.