PhD Lecture

Cassiopeia
House of Computer Science

In partial fulfillment of the terms for obtaining the PhD degree, Rudra Pratap Deb Nath will give a lecture on the following subject:

Aspects of Semantic ETL

on Thursday 1st of October 2020, 13:00 in room 02.13 at Selma Lagerlöfs Vej 300

Abstract:
Data Warehouse (DW) and On-Line Analytical Processing (OLAP) technologies perform efficiently when they are applied on data that are static in nature and well organized in structure. Nowadays, Semantic Web technologies and the Linked Data principles inspire organizations to publish their semantic data, which allow machines to understand the meaning of data, using the Resource Description Framework (RDF) model. In addition to traditional (non-semantic) data sources, the incorporation of semantic data sources into a DW raises the additional challenges of schema derivation, semantic heterogeneity, and schema and data management model over traditional ETL tools. Furthermore, most SW data provided by business, academic and governmental organizations include facts and figures, which raise new requirements for BI tools to enable OLAP-like analyses over those semantic (RDF) data.
In this thesis, we 1) propose a layer-based ETL framework for handling semantic and non-semantic data sources by addressing the challenges mentioned above, 2) propose a set of high-level ETL constructs for processing semantic data, and 3) implement appropriate environments (both programmable and GUI) to facilitate ETL processes and evaluate the proposed solutions. Our ETL framework is a semantic ETL framework because it integrates data semantically. We propose SETL, a unified framework for semantic ETL. The framework is divided into three layers: The Definition Layer, ETL Layer, and Data Warehouse Layer. In the Definition Layer, the semantic DW (SDW) schema, sources, and the mappings among the sources and the target are defined. In the ETL Layer, ETL processes to populate the SDW from sources are designed. The Data Warehouse Layer manages the storage of transformed semantic data. On top of SETL, we propose $SETL_{CONSTUCT}$ where we define a set of high-level ETL tasks/operations to process semantic data sources. We divide the integration process into two layers: The Definition Layer and Execution Layer. To create mappings among the sources and target constructs, we provide a mapping vocabulary called S2TMAP. In the Execution Layer, we propose a set of high-level ETL operations to process semantic data sources. Finally, we develop a GUI-based semantic BI system $SETL_{BI}$ to define, process, integrate, and query semantic and non-semantic data. In addition to the Definition Layer and the ETL Layer, $SETL_{BI}$ has the OLAP Layer, which provides an interactive interface to enable OLAP analysis over the semantic DW. $SETL_{BI}$ facilitates (1) DW designers with little/no SW knowledge to semantically integrate semantic and/or non-semantic data and analyze it in OLAP style, and (2) SW users with basic MD background to define MD views over semantic data for enabling OLAP-like analysis. We evaluate the framework by creating a multidimensional DW using real-world data. Taking the framework as a base point, researchers can aim to develop further interactive and automatic integration framework for SDWs. This project bridges the traditional BI technologies and SW technologies which in turn will open the door of further research opportunities like developing machine understandable ETL and warehousing techniques.

Members of the assessment committee are Professor Kristian Torp, Aalborg University, Denmark, Associate Professor Olaf Hartig, Linköping University, Sweden, and Associate Professor Panagiotis Vassiliadis, University of Ioannina, Greece. Professor Torben Bach Pedersen, Professor Katja Hose, and Professor Oscar Romero are Rudra Pratap Deb Nath's supervisors. The moderator is Associate Professor Gabriela Montoya.

All interested parties are welcome.