MODETL: A complete MODeling and ETL method for designing Data Warehouses from Semantic Databases

Selma Khouri
LIAS/ISAE-ENSMA
France
selma.khouri@ensma.fr

Ladjel Bellatreche
LIAS/ISAE-ENSMA
France
bellatreche@ensma.fr

Nabila Berkani
National High School for
Computer Science, Algeria
n_berkani@esi.dz

ABSTRACT

In last decades, Semantic DataBases ($\mathcal{SDB}$) have emerged and the major DBMS editors provide semantic support in their products. This is mainly due to the spectacular development of ontologies in several important domains like E-commerce, Engineering, Medicine, etc. Note that ontologies can be seen as a natural continuity of conceptual models. Contrary to traditional databases, where their instances are stored in a relational layout, $\mathcal{SDB}$ store ontological data according to one of three main storage layouts (horizontal, vertical, binary). Actually, $\mathcal{SDB}$ are serious candidates for business intelligence applications built around the Data Warehouse ($\mathcal{DW}$) technology. The important steps of the life-cycle warehouse design (user requirement analysis, conceptual design, logical design, ETL process, physical design) are usually managed in isolation way. This treatment is mainly due to the complexity of each phase. Actually, $\mathcal{DW}$ technology is quite mature for traditional data sources. As a consequence, leveraging its steps to deal with $\mathcal{SDB}$ becomes a necessity. In this paper, we propose a method that covers the most important steps of life-cycle of semantic $\mathcal{DW}$. To fitful our needs, four main objectives have been defined:

O_1: leveraging the integration framework by considering ontologies: a $\mathcal{DW}$ can be seen as a materialized data integration system, where data are viewed in a multi-dimensional way. Data integration systems are formally defined by a triple: $<\mathcal{G}, \mathcal{S}, \mathcal{M}>$, where $\mathcal{G}$ is the global schema, $\mathcal{S}$ is a set of local schemas that describes the structure of each source participating in the integration process, and $\mathcal{M}$ is a set of assertions relating elements of the global schema $\mathcal{G}$ with elements of local schemas $\mathcal{S}$. We defined an integration framework $<\mathcal{G}, \mathcal{S}, \mathcal{M}>$ adapted to $\mathcal{SDB}$ specificities, where schema $\mathcal{G}$ is represented by a domain ontology, the set of sources S considered are $\mathcal{SDB}$, and M represents the set of mapping assertions. A mathematical formalization of ontologies, $\mathcal{SDB}$ and semantic $\mathcal{DW}$ is given, based on the description logic formalism.

O_2: User requirements have to be expressed at the ontological level: the requirements model we proposed follows the *goal oriented paradigm*. After analyzing the major studies related to this paradigm, we proposed a goal model viewed as a pivot model, since it combines three widespread goal-oriented approaches: *KAOS, Tropos* and *iStar*. The goal model is then connected to the ontology meta-model in order to specify requirements at the ontological level. Requirements analysis allows the designer to construct the dictionary identifying the set of relevant concepts and properties used by the target application. The conceptual, logical and then physical model are defined based on that dictionary. The availability of the ontology allows exploiting its reasoning capabilities to correct the inconsistencies of the conceptual model, and to infer new facts.

O_3: The ETL process has to be defined at the ontological level and not at physical or conceptual levels: different ETL works proposed in the literature consider logical schemas of sources as inputs of the $\mathcal{DW}$ system, and make an implicit assumption that the $\mathcal{DW}$ model will be deployed using the same representation (usually using a relational representation). The third objective of our method ensures the definition of the ETL process at the ontological level independently of any implementation constraint. We defined a generic ETL algorithm, based on ten generic operators defined in the literature, that aims at populating the target $\mathcal{DW}$ schema, by data from $\mathcal{SDB}$.

O_4: the deployment process needs to consider the different storage layouts of semantic $\mathcal{DW}$: different deployment solutions are proposed and implemented using data access object design patterns. A prototype validating our proposal using the Lehigh University Benchmark ontology and Oracle $\mathcal{SDB}$ has been developed.

Categories and Subject Descriptors

H.2.7 [**Database management**]: [Data warehouse and repository]; D.2.10 [**Software engineering**]: Design—*Methodologies*

Keywords

Data warehouse design, Ontology, Semantic databases, ETL process

The 18th International Conference on Management of Data (COMAD), 14th-16th Dec, 2012 at Pune, India.