BVG Trust

J. Sci. Trans. Environ. Technov. 2018, 12(1) : 6-11

Scientific Transactions in Environment and Technovation

A Study on the Relationship of Semantic Web Components using OLAP

V. Geetha*, R.Manjupargavi and N.Subhalakshmi

Department of Computer Science, S.T.E.T. Women's College, Sundarakkottai, Mannargudi - 614 016, Thiruvarur (DT), Tamilnadu, India.

ttps://doi.org/10.56343/STET.116.012.001.002
http://stetjournals.com

Abstract

This paper deals with effective OLAP relationship builder for the software product development using the model Business Intelligence (BI) components. To capture the entity process that always stores the data in the form of open and closed datasets. However, the most interesting On Line Analytical Processing (OLAP) queries can not only be answered on internal data alone, but also external data must also be discovered (most often on the Web), acquired, integrated and (analytically) queried resulting in a new type of OLAP, exploratory OLAP. Here, Semantic Web(SW) technologies come to the rescue, as they allow semantics (ranging from very simple to very complex) to be specified for web-available resources. SW technologies do not only support capturing the "passive" semantics, but also support active inference and reasoning on the data. Finally, all the findings are discussed and a number of directions for future research are outlined, including SW support for intelligent Multidimensional(MD) querying, using SW technologies for providing context to data warehouses, and scalability issues. The main objective of this conclusion is that SW technologies are very relevant for the future of BI and OLAP, but that a number of new developments are needed to reach full potential.

Key words: Business Intelligence, Data Warehousing, OLAP, ETL, Semantic Web.

Received : May 2017 Revised and Accepted : August 2018

INTRODUCTION

Business Intelligence (BI) is aimed at gathering, transforming and summarizing available data from existing sources to generate analytical information suitable for decision-making tasks. The most widely used approach to BI has been the combination of Data Warehousing (DW), On Line Analytical Processing (OLAP) technologies and the Multidimensional (MD) data model (Rizzi *et al*., 2008). DW / OLAP technologies have been successfully applied for analytical purposes, but always in a well controlled "closed-world" scenario, where the set of data sources is rather static and well structured data is periodically loaded in batch mode applying heavy cleansing transformations. However, the eruption of XML and other richer semi-structured formats like RDF has opened up much more heterogeneous and open scenarios than those of such traditional in-house DW applications. The opportunity and importance of using unstructured and semi-structured data (either textual or not) in the decision making process in (Inmon *et al*., 2008). Nowadays, Web 2.0 sites and Linked Open Data initiatives are becoming sources of huge amounts of valuable semi-structured data. Currently no one questions the need of adding all this information to the traditional analysis of corporate processes. A significant amount of information and thus, knowledge that can be found in "unconventional" data sources like Web portals, social media, unstructured or less-structured data stores like product reviews, customer complaints, e-mails and so on.

Enterprises have started to look into such rich information sources to increase their profits and improve their products and services. As an example, populating a business report that shows the effect of a product campaign in a specific time period may require combining information from historical, structured data like product sales and customer data, residing in a Data Warehousing (DW), with sentiments extracted from Big Data (e.g., tweets) relating to products promoted by the respective campaign (Simitsis *et al*., 2012 ; Ghazal *et al*., 2013).

Thus, companies want to explore all these new data opportunities are that include them in their OLAP analyses, leading to a new type of OLAP: Exploratory OLAP. The main difference between Exploratory OLAP and the Traditional OLAP is naturally in the exploration of issue: new data sources, new ways of structuring data, new ways of putting data together and new ways of querying data. Whereas Traditional OLAP is Performed in a "closed-world" scenario based only on internal data, an essential part of Exploratory OLAP, to discover, acquire, integrate and analytically query new external data.

*Corresponding Author :
email: *kkmannaig@gmail.com*

P - ISSN 0973 - 9157
E - ISSN 2393 - 9249

The Semantic Web (SW) has been conceived as a means to build semantic spaces over Web published contents so that Web information can be effectively retrieved and processed by both humans and machines for a great variety of tasks.

A recent article introduced the concept of fusion cubes to mean cubes that, based on a core of internal multidimensional data, gradually merge with external data, in order to support self-service BI (Abell *et al*.,2013). The article provides a motivating example, which captures the essence of exploratory OLAP and shows as to why SW technologies are needed in this scenario. The example concerns a group of concerned citizens (watch Dogs) that want to monitor if the fishing catches being landed in the various EU countries respect the overall limits set up by the EU marine protection schemes and also as to how they are related to marine protection areas. The watch dogs want to analyze the data by Time, Location, and Species, where each of these three dimensions should be organized into a hierarchy of levels, e.g., Day-Week Month- Year, Port-Province-Country-Region and Subspecies- Species-Family. To do this, they must integrate statistical catch data (in a flat tabular format) with geographical data about marine protection areas (from public database, in SW format), fish population data (from various research databases, in a multitude of formats ranging from comma separated files to SW data), and finally with ontology data describing geo and species hierarchies (in SW formats). Reasoning capabilities are needed to perform the complex integration and resolve conflicts, e.g., contradicting catches data or species classifications. Interestingly, SW technologies are powerful enough to both model all these different types of data and provide the needed reasoning capabilities on top

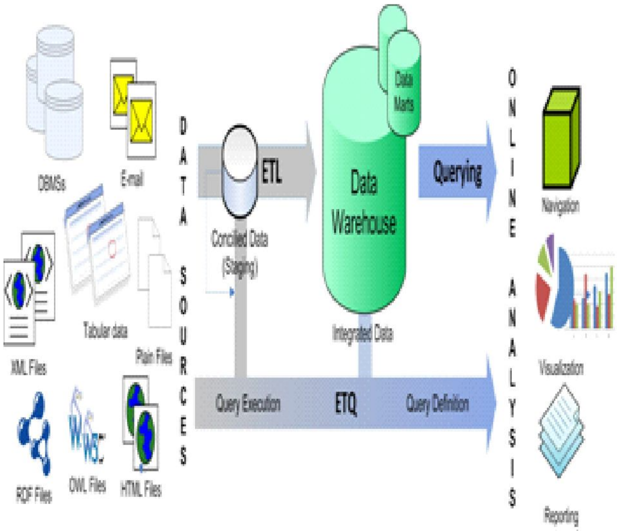

Fig.1. Show the DW / OLAP Elements and Dataflow

The main contributions of this paper can be summarized as follows:

- Propose a set of five novel criteria to categorize DW / OLAP systems,
- Analyze as to how these criteria affect the need for Semantics and the feasibility of the design and Data provisioning processes,
- Analyze as to how semantic-aware reasoning techniques can aid,
- Survey and categorize existing DW / OLAP work according to the five criteria and the three reasoning criteria and
- Identify research trends in this area.

METHODOLOGIES

Nowadays, a new trend of OLAP work has emerged, which applies SW technologies to mainly address data integration issues and the automation of data processing. The purpose of this paper is to categorize the main requirements of these new OLAP approaches, as well as to show as to how SW technologies can help to fulfill the new requirements.

As there are many papers proposing a large variety of system features, in this section we present a methodology that guides this survey and produces a clear picture of this intricate area.

We first present the characteristics of Traditional OLAP use cases to frame the area of interested. Then, five criteria related to the different relevant aspects of DW / OLAP systems are defined. By means of these criteria, in the rest of the paper, current approaches are categorized. Furthermore, the five criteria define a space that allows us to locate Exploratory OLAP use cases and to distinguish them from Traditional OLAP use cases. In addition, we use another three criteria related to expressiveness, reasoning and complexity and to characterize the existing work with regard to SW technologies.

The structure of OLAP systems

OLAP technology is aimed at gathering, transforming and summarizing available data from existing sources to generate analytical information suitable for decision-making tasks. Traditionally, OLAP has been associated with data warehouses (DW), following the three layered structure shown in Figure 1, namely:

- The data sources layer, which consists of all the potential data of any nature (e.g., relational, object oriented, semi-structured, and textual) that can help to fulfill the analysis goals,

- The integration layer, which transforms and cleanses the data gathered from the sources, as well as stores them in an appropriate format for the subsequent analysis (i.e., the DW), and

- The analysis layer, which contains a number of tools for extracting information and knowledge from the integrated data and presenting it to the analysts (i.e., OLAP cubes, charts, reports, etc).

As it is clear from this description, the integration model of Traditional OLAP systems (DW/OLAP) is based on a global schema (i.e., the DW schema), which is seen as a view over the underlying data source schemas (which is usually known as Global as View or GaV for short). In this integration model, query answering is simple. The external data sources are (implicitly) assumed to be known in advance as are the user needs guiding the design of the global schema. This works well when the sources and requirements are indeed known in advance, but encounters problems when this does not occur. For those cases, more flexible integration models are needed. In particular, the integration of external data schemas in terms of a global schema (often in the form of global domain ontology) has been studied from (Levy, 1998). From the global schema, local schemas can be derived; i.e., the local schemas are seen as (more specialized) views of the unified general global schema. The resulting integration model (usually known as Local as View or LaV for short) is thus highly extensible, at the expense of considerably more complicated query answering. Therefore, in this integration model the reasoning power of SW technologies is especially needed.

Materialization

Starting from the figure.1, we firstly find Materialization. This criterion concerns the level of materialization of the integrated data. In Traditional DWs, all the integrated data is fully materialized (i.e., Full) often including a data as called data staging area for performing transformations and cleansing. At the other extreme, Virtual DWs extract data from sources at query time, integrate them on the fly, return the result to the user and then throw away the integrated data. Notice that the ETQ processes described in the previous section fall in this category. A compromise, where some data is materialized, while other data, typically data with many changes, are extracted at query time, is sometimes used (Dayal *et al*., 2009). Closer to the Virtual DW, the Result Keeping approach first extracts data on-demand from sources and computes the result on the fly (e.g., for displaying in a dashboard), but then stores/keeps the results to allow repeated requests for the same result to be delivered quickly (Pederson *et al*., 2004). Complex ETL flows may actually have subparts each residing in different categories (i.e., Partial). For example, it is common to have an "on-line" flow that performs fast, but less thorough, on the fly integration in main memory for immediate use, while a parallel "off-line" flow performs more thorough integration for historical use and stores all data persistently (Vassiliadis and Simitsis, 2009). Here, SW technologies can be used to describe the data and the results, as well as the steps in between.

Transformations

Proceeding clockwise, the next one is Transformations. This criterion concerns the level of transformations applied to the source data during the integration process. In Traditional DWs, it is common to apply many Complex and significant transformations, e.g., creating versions of data, significant cleansing, computing holistic aggregates, etc. At the other end of the spectrum, some use cases demand only Lightweight transformations that can be done quickly on the fly (even for streaming data), e.g., moving averages, simple and approximate aggregations, renaming/ removing columns, etc.

Freshness

The next criterion is Freshness, which concerns how often the data integration process is performed (i.e., how often the DW is refreshed). Traditional DWs were refreshed periodically (e.g., daily, in batch mode). A variation of this is Micro-batches where the refreshment is run often (e.g., every 15 or 30 minutes), on the smaller batch of data accumulated in that period. Other DWs (e.g., the Virtual DWs mentioned above) refresh the data on demand, when requested by users. More recently, there has been a trend to refresh the DW even more frequently (e.g., with propagation delays of at most a minute or so).

Structuredness

The next criterion is Structuredness which concerns which types of data are found in the data sources or, more specifically, how Structured the least structured type of source data is structured. In Traditional OLAP cases, all sources consist of structured data, typically relational tables or in a few cases structured spreadsheets.

Extensibility

The next and last criterion is Extensibility. This criterion concerns how Dynamic the set of data sources can be, i.e., how easily new data sources could be brought into the system. In Traditional DW/OLAP, the same (mostly internal) Static data sources are used over and over, and new sources are only brought in at new major DW releases (i.e., at most a few times per year). Recently, there has been a trend to include new data sources, often from external data suppliers, into an existing DW more often in order to answer new questions, making the source set Evolving.

SW technologies for OLAP systems

SW technology is aimed at providing the necessary representation languages and tools to express semantic-based metadata. This focuses on semantics which is very useful for Exploratory OLAP systems, where the vast amount of unstructured or semi-structured sources demand new semantic-aware solutions that enable machine process able data integration.

SW technologies can aid the development of Exploratory OLAP systems in two aspects: on the one hand, ontologies serve the purpose of formally conceptualizing both the domain of interest and the business concepts. On the other hand, by means of semantic annotation, different data sources can be mapped to ontology concepts, resulting in a homogeneous conceptual space where we capture the meaning of the integrated elements.

Most ontology languages, such as the Web Ontology Language (OWL; the W3C recommendation), have strong foundations in logics and differ from other semantic-aware technologies (like diagrammatic languages such as UML or ER) in that they are machine process able and support reasoning. Thus, we can describe concepts and relationships but also infer implicit knowledge from that explicitly stated. Two main families of logic-based languages currently underlie most of the research done in this direction: Description Logics (DL) and Data log-related logics. Both approaches seek the same objective, but from different points of view. While DL focus on representing knowledge (and thus, the schema is expected to evolve), Data log is more focused on capturing the instances (and in this sense, closer to the database field).

SW technologies categorization criteria

Although logic-based languages are very appealing for their semantic-awareness and reasoning features, while reasoning is computationally hard. For this reason, most of the research done in this direction is focused on balancing the language expressiveness and the reasoning services provided according to each scenario. This trade-off is traditionally captured in terms of three criteria (Fig.3) Reasoning capabilities provided, Language expressiveness, and Computation complexity. Without loss of generality, in the remainder of this paper we focus on how research on OLAP makes use of logic-based ontology languages and the trade-off offered with reference to these criteria.

Reasoning

Starting clockwise from the top, the Reasoning criterion concerns the inference algorithms needed. We mainly talk about the use of Standard reasoning services (such as subsumption), non-Standard inferences (such as schema matching, transitive closures, temporal reasoning (Middle Fart and Pederson, 2011) and no use of reasoning (i.e., None). We say a reasoning service is Standard if it is supported by most reasoners. The typical inferences provided by DL reasoners are concept satisfiability, subsumption and query answering (Berlana *et al*., 2011). Concept satisfiability and subsumption sit at the terminological level, whereas query answering also deals with instances. Relevantly, very few DL languages (e.g., DL-Lite in (Nebot and Berlanga,2012) and the OWL2 QL profile, based on DL-Lite) properly support query answering which means that, in practice, query answering is prohibitively costly for large data sets, such as those in OLAP scenarios. Thus, most DL languages are typically used at the terminological level.

Computation

The next criterion is Computation. For this axis we do not mean classic theoretical computational complexity, but instead the feasibility of computing certain reasoning tasks under certain assumptions (i.e., in a given scenario). An expensive inference (e.g., computing the transitive closure of all properties in an ontology) computed once may indeed be more feasible than a relatively less complex reasoning task (e.g., computing subsumption in OWL DL ontologism) conducted relatively often (e.g., over a very large ontology and triggered by a certain event in the application GUI).

Comparison

In previous sections, we have introduced the four main stages of a DW / OLAP system (discovery, acquisition, integration and querying) and later, we have introduced a set of criteria to categorize current approaches. As we explained before, our main focus is to investigate how SW technologies can aid throughout these stages.

In practice, current approaches are traversal to these four conceptual stages and thus, they cannot be classified according to them. As in classical software design approaches, current solutions either focus on the MD schema design of OLAP systems (i.e., at the schema level), or on integration and provisioning of data (i.e., at the data / instances). For each of these two categories (schema vs. data), we have identified representative papers, and subsequently divided them into two subcategories (for a total of four categories of papers), depending on whether SW technologies are applied to satisfy the requirements of Traditional OLAP systems (here denoted Semantic-aware OLAP systems), or to support (to some extent) the new set of requirements of Exploratory systems. The papers were selected based on our experience, depending on how well they exemplify the categories.

For each of the relevant papers identified, we show its position in each of the five DW criteria, and also the value in the three SW ones. Horizontally, the table is divided into four parts corresponding to semantic-aware MD design, multidimensional query definition, semantic aware ETL processes, and ETQ processes. When one paper deals with issues, MD design and data provisioning, it appears twice in the table and is analyzed from both perspectives (which may result in apparently contradictory classifications, caused by the different viewpoints of the analysis).

Data Schema Design

MD design is a well-known paradigm in the area of DW and databases in general, always related to OLAP tools. It was popularized by Ralph Kimball at the logical level in (Pederson *et al*., 2004). Multidimensionality is based on the fact-dimension dichotomy. This paradigm aims at analyzing the fact (or subject of analysis) instances, from different analysis dimensions (i.e., points of view). Several measures (i.e., metrics) are available for each fact instance in order to gain insight. Furthermore, the MD model also provides foundations to study / analyze the available measures at various aggregation levels determined by the hierarchical structure of the dimensions. Indeed, aggregation is one of the main characteristics of the MD model, setting foundations for the well known roll-up and drill-down operators.

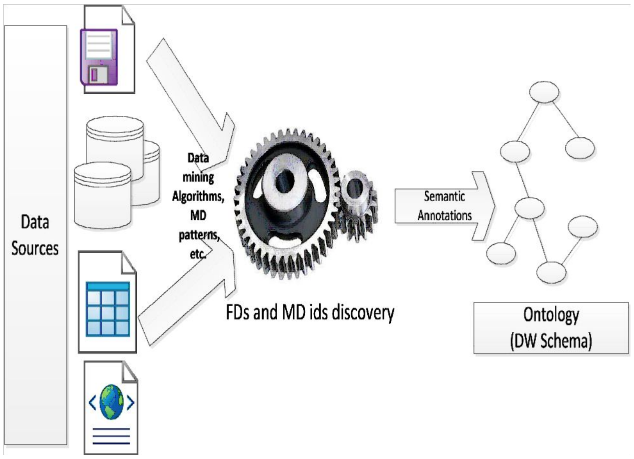

Fig.2. Show the Anthologies and Semantics Annotations

Fig.2 Anthologies for semantic Annotations: Ontologism as a passive actor to overcome Heterogeneities.

To automate MD design, classical approaches (e.g., Pederson *et al*., 2004) focus on the organization of data and assume relational (or homogeneous) and well structured sources and therefore, they are hardly effective (or feasible) in heterogeneous scenarios with disparate sources. Indeed, the more automatable they are, the more tied to a specific formalism or language (typically relational sources). Consequently, they do not tackle the integration of different data models.

Data Provisioning

Another challenge towards Exploratory OLAP is to shift the focus from DW-centric, Traditional ETL flows to broader data flows consisting of complex analytic operations, involving a plurality of different data types and sources, spanning multiple execution engines, and running under different freshness requirements and at different paces, ranging from slow or frequent batches to micro-batches or real-time processing. As with Traditional and Exploratory OLAP, where the latter requires a solution that captures, transforms, and presents fresh data in order to answer changing questions, we also see the need for Exploratory ETL processes. As in Figure 1, the processes are named these processes Extract, Transform, and Query (ETQ) in order to differentiate them from traditional Telling. ETQs should be able to gather data, apply computations, and produce dynamic reports or populate dashboards directly from –potentially evolving– user requirements. In addition, Figure.3 described the ETQ processes which deviate from Traditional processing in that they may affect various stages of the design, for example they may be used to populate DW constructs or to answer a business query by fetching data directly from the sources like on demand ETL. (Dayal *et al*., 2009).

With the widespread adoption of new technologies in the Web, such as XML and other richer semi structured formats like RDF, important and useful information is being captured in a large variety of data sources.

CHALLENGES

In this section, we summarize our findings and identify a list of challenges that require a fresh look in the future.

We divide our discussion between the two areas of interest in this survey, namely schema design and data provisioning, but we also comment on whether SW technologies are ready to fully support the needs of next generation OLAP systems.

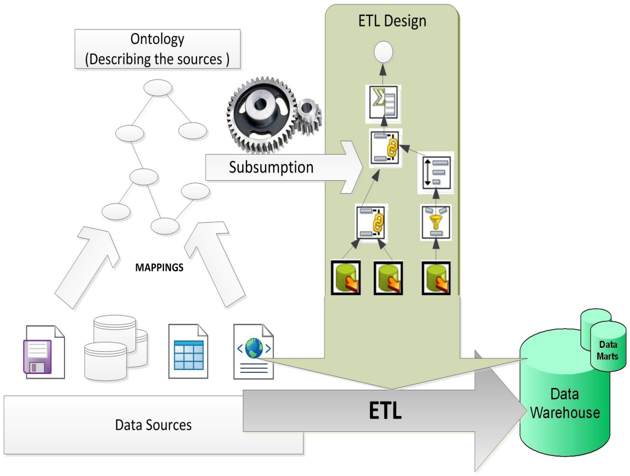

Fig.3. Show the ETL Design and Process

CONCLUSION

As enterprises tend to use broader information pools (e.g., social media, sensor data, documents) for enabling better decision making and strategic planning, traditional DW and OLAP technologies need to be adjusted and extended appropriately. SW technologies can help by enabling understanding and integrating the source data and its semantics better and thus, may assist in building semantic bridges across multiple information silos.

REFERENCES

Abell, A., Darmont, J., Etcheverry, L., Golfarelli, M., Mason, J.N., Naumann, F., Pedersen, T. B., Rizzi, S., Trujillo, J., Vassiliadis, P. and Vossen, G. 2013. Fusion cubes: Towards self-service business intelligence, *Int. J. on Data Warehousing and Mining,* 9 : 66-88. https://doi.org/10.4018/jdwm.2013040104

Berlanga, R., Romero, O., Simitsis, A., Nebot, V., Pedersen, T. B., Abell, A. and Aramburu, M.J. 2011. Semantic web technologies for business intelligence. In : *Business Intelligence Applications and the Web: Models, Systems and Technologies.* IGI Global, 310–339. https://doi.org/10.4018/978-1-61350-038-5.ch014

Dayal, U., Castellanos, M., Simitsis, A. and Wilkinson, K. 2009. Data integration flows for business intelligence, in 12th Int. Conf. on Extending Database Technology (EDBT), ser. ACM Int. Conf. Proceeding Series, 360 : 1–11. https://doi.org/10.1145/1516360.1516362

Ghazal, A., Rabl, T., Hu, M., Raab, F., Poess, M., Crolotte, A. and Jacobsen, H.A. 2013. Big Bench: towards an industry standard benchmark for big data analytics. In : *Proc. of ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'13)*, 1197–1208. https://doi.org/10.1145/2463676.2463712

Kampgen, B. and Harth, A. 2011. Transforming statistical linked data for use in OLAP systems. In : *7th Int. Conf. on Semantic Systems (I-SEMANTICS), ser. ACM Int. Conf. Proceeding,* 33–40. https://doi.org/10.1145/2063518.2063523

Levy, A.Y. 1998. The information manifold approach to data integration, *IEEE Intelligent Systems*, 13 : 12–16. https://doi.org/10.1109/5254.722342

Middelfart, M. and Pedersen, T.B. 2011. The meta-morphing model used in targit bi suite. In : ER Workshops, 364–370. https://doi.org/10.1007/978-3-642-24574-9_52

Nebot, V. and Berlanga, R. 2012. Building data warehouses with semantic web data, *Decision Support Systems,* 52 : 853–868. https://doi.org/10.1016/j.dss.2011.11.009

Pedersen, D., Pedersen, J. and Pedersen, T.B., 2004. Integrating XML data in the TARGIT OLAP system. In Proc : *20th Int. Conf. on Data Engineering (ICDE),* 778–781.

Rizzi, S., Abell, A., Lechtenb¨orger, J. and Trujillo, J. 2006. Research in data warehouse modeling and design: dead or alive? In Proc *: 9th Int. Workshop on Data Warehousing and OLAP (DOLAP),* 3–10. https://doi.org/10.1145/1183512.1183515

Simitsis, A., Wilkinson, K., Castellanos, M. and Dayal, U. 2012. Optimizing analytic data flows for multiple execution engines, In : *SIGMOD Conf.,* 829–840. https://doi.org/10.1145/2213836.2213963

Vassiliadis, P. and Simitsis, A. 2009. Near real time etl, in New Trends. In : *Data Warehousing and Data Analysis,* 1–31. https://doi.org/10.1007/978-0-387-87431-9_2