Information and Software Technology 54 (2012) 360–379

Contents lists available at SciVerse ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Hybrid methodology for data warehouse conceptual design by UML schemas

Francesco Di Tria *, Ezio Lefons, Filippo Tangorra

Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, Via Orabona 4, 70125 Bari, Italy

ARTICLE INFO

Article history:
Received 11 November 2010
Received in revised form 29 September 2011
Accepted 17 November 2011
Available online 6 December 2011

Keywords:
Data warehouse
Conceptual design
Requirement analysis
*i** Framework
UML multidimensional model
Data modelling

ABSTRACT

Context: Data warehouse conceptual design is based on the metaphor of the cube, which can be derived from either requirement-driven or data-driven methodologies. Each methodology has its own advantages. The first allows designers to obtain a conceptual schema very close to the user needs but it may be not supported by the effective data availability. On the contrary, the second ensures a perfect traceability and consistence with the data sources—in fact, it guarantees the presence of data to be used in analytical processing—but does not preserve from missing business user needs. To face this issue, the necessity emerged in the last years to define hybrid methodologies for conceptual design.
Objective: The objective of the paper is to use a hybrid methodology based on different multidimensional models in order to gather all advantages of each of them.
Method: The proposed methodology integrates the requirement-driven strategy with the data-driven one, in that order, possibly performing alterations of functional dependencies on UML multidimensional schemas reconciled with data sources.
Results: As case study, we illustrate how our methodology can be applied to the university environment. Furthermore, we evaluate quantitatively the benefits of this methodology by comparing it with some popular and conventional methodologies.
Conclusion: In conclusion, we highlight how the hybrid methodology improves the conceptual schema quality. Finally, we outline our present work devoted to introduce automatic design techniques in the methodology on the basis of the logical programming.

1. Introduction

The data warehouse conceptual design is the most crucial step to correctly represent the domain of interest and it is the milestone on which the different viewpoints of decision makers and Informatics must agree [1]. Therefore, it is very important for data warehouse designers to follow a consolidated and robust conceptual design methodology, as the development of a data warehouse is a very expensive process, even today that there exist several software tools covering all the steps of the data warehouse lifecycle and offering pre-packaged solutions.

The *requirement-driven* and the *data-driven* approaches are antithetical to each other [1] and designers are forced to choose one of these. In fact, a requirement-driven approach, also known as *demand-driven* or *goal-oriented* methodology, aims to define multidimensional schemas using business goals resulting from the needs of the decision makers. The data sources are considered later, when the Extraction, Transformation, and Loading (ETL) phase must be designed. In this step, the multidimensional concepts (such as facts, dimensions, and measures) have to be mapped on the data sources in order to program the feeding plan of the data warehouse. At this point, it may happen that the designer discovers that the needed data are not available. Conversely, a data-driven approach, also known as *supply-driven* methodology, aims to define multidimensional schemas through a *made-by-expert* reengineering process of the data sources, minimizing the participation of end users and, consequently, going towards a possible failure of their expectations.

Nowadays, several efforts to define a unified strategy that integrates the benefits of these two main approaches have led to the definition of hybrid methodologies, classified in pure hybrid methodologies and integration-derived hybrid methodologies. The former group includes all the methodologies that perform the design process considering simultaneously the data sources and the business goals [2,3]. The latter comprises methodologies that combine and integrate a data-driven approach with a requirement-driven one, which, in turn, can be divided into sequential hybrid and parallel hybrid methodologies. In sequential hybrid methodologies, the two stages are executed according to a prefixed order, and the output of the first stage is used as input of the second stage [4–6]. In parallel hybrid methodologies, the two stages are executed independently and, at the end, the comparison and integration of the schemas coming from the different stages are performed [7].

* Corresponding author.
E-mail addresses: francescoditria@di.uniba.it (F. Di Tria), lefons@di.uniba.it (E. Lefons), tangorra@di.uniba.it (F. Tangorra).

0950-5849/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2011.11.004

Abbreviations

ETL	Extraction, Transformation, and Loading	CRUI	Conference of Italian University Rectors
DFM	Dimensional Fact Model [11]	CWM	Common Warehouse Metamodel [27]
QVT	Query-View-Transformation [26]	MOF	Meta-Object Facility [28]
MIUR	Italian Ministry of Education, University and Research		

The survey of the current methodologies in [8] addresses the necessity of hybrid approaches in multidimensional modelling, defines a unified terminology for multidimensional concepts, and also introduces a set of useful comparison criteria to evaluate methodologies on the basis of their capabilities to adequately represent such multidimensional concepts. An important comparison criterion is the identification of the inputs to be provided to a methodology. As an example, methodologies can work on conceptual or logical schemas, whereas a conceptual schema can be expressed according to the Entity/Relationship (E/R) or the Unified Modeling Language (UML) formalism and a logical schema can be either a relational schema or an XML schema.

Other important issues in this research topic are the actual lack of tools supporting automatic multidimensional modelling [8] and the fewness of CASE tools to support automatic design [9]. While methodologies based on a requirement-driven approach suffer from drawbacks related to the comprehension of user needs expressed according to a natural language, in methodologies based on a data-driven approach the adoption of automatic techniques is encouraged by well-structured data sources, and, additionally, by the presence of functional dependencies [10].

The contribution of the paper is the definition of a sequential hybrid methodology that takes into account both the advantages of data modelling, as in the Dimensional Fact Model (DFM) [11], and the strong formalization of user requirements. Such a formalization is represented by UML multidimensional schemas [12] obtained from the i^* framework [13]. These UML multidimensional schemas, due to their high level of standardization and formal representation of multidimensional concepts, can be effectively used for the automation of the design processes. In the paper, we show how a large part of the steps performed in our design of multidimensional schemas can be done automatically.

The paper is organized as follows. In Section 2, the related work about current hybrid methodologies is presented. In Section 3, the drawbacks related to the data-driven and requirement-driven methodologies are summarized, along with their multidimensional models. In Section 4, we present our methodology to integrate requirement-driven and data-driven approaches and, then, techniques to automatize such integration. In Section 5, we present a real case study to illustrate the proposed methodology from the practical viewpoint. Then, in Section 6, we quantitatively compare the different methodologies, highlighting the benefits introduced by our approach. Finally, we present some concluding remarks and the directions of our future work.

2. Related work

This section discusses the features of some current hybrid methodologies, for which we provide a description according to the classification schema previously presented: pure hybrid methodologies ([2,3]) and integration-derived methodologies that are, in turn, sub-classified as sequential ([4–6,14]) and parallel ([7]).

2.1. Pure hybrid methodology

The pure hybrid methodology is based on the idea that the user requirements can be entirely derived by defining a preliminary workload that contains all the analytical queries the end users intend to execute in order to extract information from the data warehouse ([2,3]). On that assumption, the authors propose an algorithm able to automatically create a graph (whose nodes are the tables of the data sources and edges are the joins between tables) that aims to identify whether each table has to be considered a fact table or a dimensional table. They claim that labelling all nodes correctly generates a valid multidimensional schema. The labels are assigned by examining the role played by tables and attributes in the SQL queries included in the preliminary workload. As an example, a table whose primary key appears in a group-by clause is labelled as dimensional level. However, the authors do not consider that queries coming from business goals may have neither syntactic nor semantic regard to the data sources. In the paper, there is no mention about how to face possible syntactical or semantic incompatibilities. Moreover, the algorithm simply assigns a tag to each table of the data source but no new schema is produced. Nevertheless, this methodology represents a good starting point for further refinements, as it can help designers to quickly and automatically identify facts and dimensions in data sources in the early stages of the design process.

2.2. Sequential hybrid methodology

As concerns sequential hybrid methods, in [4], the authors present a framework to be used for the conceptual design of data warehouses. Such a framework starts from the analysis of the business goals defined by decision makers. Using these goals, a schema representing information requirements is first produced. Then, an initial conceptual schema is suitably derived by discovering facts and dimensions from the information requirements. At this point, in order to take also the data sources into account, the conceptual schema is reconciled with the logical schemas of the data sources by applying multidimensional normal forms [15]. This strategy was already present in [16], where, initially, it was not considered as a hybrid model. In fact, reconciling a requirement-derived schema with data sources does not suffice by itself to define a hybrid methodology. Reconciling means only verifying whether an initial conceptual schema agrees with data sources, whereas decision makers would want to obtain some kind of information that is not effectively available because of the lack of data. On the contrary, data-driven methodologies allow the designers to manually modify, through a reengineering process of the data sources, the functional dependencies in a logical schema, by deleting unnecessary relationships and introducing useful ones (for example, adding dimensions to cubes and creating hierarchies for aggregation paths). Furthermore, designers can add attributes derived from the raw data (as computed measures).

In [6], Schneider defines a sequential hybrid methodology that adopts a graph-based model. This allows the designer to easily build a graphical schema, the so-called data warehouse graph, which can be mapped to relational or object models. So, the designer can produce a multidimensional schema that best fits user needs and, then, verify its compatibility with the data source schemas. The author claims that automatic techniques can be used to check the compatibility and that semantic incompatibilities arising during the matching can be solved using an ontology.