Semi-automatic Data Warehouse Design methodologies: a survey

Wafa Tebourski[1],Wahiba Ben Abdessalem Karâa[2] and Henda Ben Ghezala[3]

[1] Computer Science Department, High Institute of Management, University of Tunis, Bouchoucha, Bardo-Tunis, Tunisie

[2] Computer Science Department, High Institute of Management, University of Tunis, Bouchoucha, Bardo-Tunis, Tunisie

[3] Computer Science Department, National School of Computer Sciences, University of Manouba Manouba, Manouba, Tunisie

Abstract

Data warehouses are used in making strategic decisions. A data warehouse is a collection of integrated, historised data originating from heterogeneous sources which gives rise to business bases (data store). Several approaches proposed semi-automatic building data warehouses. In this paper, we present an overview of those works dedicated to the design of data warehouses and a comparison of these different approaches.

Keywords: *Data warehouses, multidimensional modeling, Bottom-up approach, Top-down approach, mixed approach.*

1. Introduction

Faced to the large amounts of data and the remarkable diversification of their sources, a scientific and economic interest to explore the reservoirs of knowledge is established. Hence, we use a process of decision support where users seek models of information interpretation, hidden knowledge extraction and potentially useful information from the available data to improve product and services quality for the company's strategic differentiation.

Systems for Decision Support (DSS) are flexible and interactive information systems that help decision makers to extract useful information, to identify and solve problems and make decisions [22]. The DSS processes the information from different sources in one place, the information are consistent and familiar to the user. The DSS combines and standardize databases, allowing analysis and decision making.

Among the decision support systems, data warehouse systems are possibly the most used in the world. Traditional information model systems are not able to analyze complex data on a large number of areas such as complex calculations, aggregations ... etc. Thus, multidimensional modeling was proposed. The multidimensional model aims at presenting the data in a structured and intuitive way to solve the transactional models' difficulties and meet the decision makers' needs. The multidimensional model is based on two fundamental concepts: fact and dimension.

Many researchers have focused on the design of data warehouse schemas. This design is a complex task. Several works have been proposed. Our purpose is to present a comparative study between these different approaches, based on several criteria.

This article is organized as follows: In Section 2, we present the different design approaches of data warehouses. In Section 3, we present the different research works related to multidimensional modeling at the conceptual and logical level. In Section 4, we propose a comparative study between the different models.

2. Data warehouses design approaches

2.1 Sources based approaches (ascending/Bottom-up)

Approaches directed by sources perform the extraction of data from heterogeneous sources. These approaches integrate the data into a storage space accessible by all decision makers. The design of data warehouse is based on a detailed analysis of data models, generally the entity-relationship model (E/R). Such approaches facilitate the ETL (Extraction-Transformation-Load) processes since each entity and relation in the source model will be presented by multidimensional concepts.

Various studies have been conducted in this context such as [6], [14], [16].

2.1.1 Golfarelli et al's approach

The authors [14] suggest a formal model: Dimensional Fact Model (DFM) which is a multidimensional graphical model, clearly differencing concepts such as facts, dimensions, measures and hierarchies. This model is presented as a set of tree structured patterns of events.

In this approach, the derivation of a pattern is performed using a two-steps process:

(i) The first step consists in establishing facts.

(ii) The second step ensures the construction of an tree attribute; it removes irrelevant attributes from the tree, identify and analyze related dimensions, measures and hierarchies for each fact.

2.1.2 The approach of Hüsemann et al.

This approach [6] is used for star logical modeling. Such study is conducting using a four-steps process:
The analysis and specification requirements: the experts select the relevant attributes of the model E / A source and specify their use (measure of fact or dimension identifier). Additional requirements are added using complex derived measures;

(i) The conceptual modeling: at this stage, a transformation of the semi-formal specification requirements of the multidimensional conceptual schema is accomplished;

(ii) The logical modeling: This step converts conceptual schemas into logical patterns respecting the logic model referred (usually relational or multidimensional) via transformation rules;

(iii) The Physical modeling: This step performs a physical implementation of logic diagrams.

2.1.3 The approach of Romero et al.

The goal of this approach [16] is to identify the multidimensional concepts from domain ontology. This approach is based on four criteria allowing exploration of the multidimensional concepts. These latter's are: (i) the multidimensional model, (ii) the constraint of multidimensional space arrangement, (iii) the integrity of constraint base, and (iv) of the additivity constraint.

2.2 Requirements based approaches (descendants / Top-down)

This type of approaches presents the phases of requirement specification and derivation of conceptual schemas. These approaches attempt to reduce the risk of failure of the decisional information system.
In this context several researches have been directed towards the design of data warehouses based on requirements such as [12], [13], [17] and [21]

2.2.1 Kimball's approach

This approach [21] is a requirements based approach aiming to derive a logical design of the data warehouse. This is an informal method, which introduces a detailed multidimensional concept to give rise to multidimensional schemas guide.
The proposed method has two axes:
The bus architecture: aims to identify all the data marts which the designer seeks to build. Data marts are defined as a pragmatic set of related facts. The next step consists in classifying the different dimensions of each data mart. Hence, an ad hoc matrix is constructed to capture the multidimensional requirements and point out associations between different data marts.
The cycle of multidimensional life is driven by a five steps process: (i) project planning, (ii) business requirements definition, (iii) choice of technology, (iv) data modeling, and (v) specification and development of an application.

2.2.2 The approach of Cabbibo and Torlone

The requirement based approach [13] presents the design methodology of the most cited. This

approach allows the generating of a logic diagram of ER (Entity Relationship or n-ary) diagrams. In addition, it can produce multidimensional schemas in terms of relational databases or multidimensional matrix. This is an informal method performs depth analysis of data sources, but does not provide formal rules.
However, this method introduces the basic foundations that will be used later in the literature, put in place the foundations used by the rest of methodologies. The proposed method consists of four steps:

(i) The first and the second stage allow restructuring facts and dimensions identification as well as the ER diagram.
(ii) The third and the fourth steps provide the multidimensional diagram.

2.2.3 The approach of Mazôn et al.

This proposal is requirement-based approach. The authors [12] integrates the business objectives of the company in the specification of the requirements using i* technology. This approach is based on three steps:

(i) Definition of business goals: consists in specifying the main objectives of a company. These goals can be classified into three levels of abstraction: strategic, decision-making and informational.

(ii) Modeling requirements using i* technology: This step identifies the data warehouse users, the business goals of the organization and the relationship between these two components.

(iii) Transformation of the obtained i* model into a multidimensional model: using heuristics [23].

2.2.4 The approach of Giorgini et al.

This approach [17] begins with the specification requirements phase and carried out using two organizational model and decision-making model. After that, a construction step is performed to provide the conceptual model which is derived from the relational model in the decision-making perspective and subsequently refined using the hierarchies of the organizational model diagram.

2.3 Mixed Approaches

This type of approach incorporates both bottom-up and top-down approaches in an attempt to take advantage of their benefits. Some researches have focused on mixed approaches, such as the [3], [4] and [18].

2.3.1 The approach of Bonifati et al.

Bonifati et al proposed a semi-automatic approach [3] based on both of requirements and sources, called as mixed approach. This method consists in three phases:

(i) Bottom-up analysis: This step examines the E/R model of the data source to construct star schemas candidates using patterns. We note that the bottom-up analysis can generate a large number of candidate patterns. An algorithm transforms each N-M association in n one to many through dealing with E/R model as a graph.
(ii) The top-down analysis: This step collects the analyzed requirements that will be refined and aggregated in a tabular report of abstraction. This step outputs the star schemas ideals.
(iii) Integration: This step makes the perfect match for each pattern from the top-down analysis with all the candidates produced by the bottom-up analysis to meet the requirements of decision diagrams.

2.3.2 The approach of Nabli et al.

The approach proposed by [4] is a mixed approach of using the automated design of data marts and data warehouse starting from semi-structured OLAP requirements expressed in a tabular form.
This method contains three steps: (i) The acquisition OLAP requirements, assisted through ontology, (ii) the generation of patterns of data marts and (iii) the generation of the warehouse schema.

2.3.3 The approach of Giorgini et al.

Giorgini et al introduced a hybrid approach consisting in three phases: (i) requirements analysis, (ii) matching requirements with sources and (iii) refinement [18]:

(i) The requirement analysis phase: this step generates a decision model and organizational model.
(ii) Matching requirements with sources: in this step, the decision model is mapped to a data source E/R through jointing on organizational model.
(iii) Refinement: The multidimensional model is enriched through the construction of hierarchies and their refinement.

3 Multidimensional modeling

Multidimensional modeling has two aspects: (i) conceptual aspect aims to make realistic modeling, (ii) the logical aspect presenting this reality.

3.1 Conceptual modeling

Several works dedicated to conceptual modeling of data warehouses and data marts are proposed. Three models are established in this context: (i) models based on the extension of the model Entity/ Relationships, (ii) models based on the UML extension, and (iii) personalized models.

3.1.1 Extension of the Model E/R

For modeling data warehouses, several extensions of the model E/R (entity-relationship) have been proposed. We discuss, such as, Starer [15], ME / R [7] and ERA [9]. The fact is a new modeled concept in E/R representing real data having the same properties. Indeed, the entity concepts present the different hierarchical levels of dimensions. However, the relations modelize associations between entities or entering facts and entities.

3.1.2 UML extension

Several methods for modeling data warehouses have been proposed in the literature, these methods are based on particular object paradigm, using UML: [1], [2], [23], [26]. In [23], [24], the authors propose a multidimensional object-oriented model (GOLD) for integrating UML in the multidimensional modeling.
They used a set of value-typed stereotypes. The fact is modeled using a class stereotyped "FACT" containing measures (atomic or calculated attributes). The dimension is represented using a class stereotyped "DIMENDION", hierarchies are represented by classes with the stereotype "BASE". The relations between two levels of a dimension in a hierarchy are modeled using the stereotype "ROLLUP UP."

3.1.3 Ad hoc or Personalized model

The Ad hoc models represent another class of modeling methods in the literature [5], [6], [8], [13], [14], [20] .The Dimensional Fact Model (DFM) proposed by [14] and extended by Rizzi [25], consists on a set of facts diagrams. The studies works on this category of approaches offer a range of constructors based on the following concepts: facts, measure, dimension, hierarchy, descriptive attribute (low attribute), multiple arcs, shared hierarchies and parameters.

These different families of proposed conceptual models are based on several paradigms (entity-relationship, object ...etc). These models mainly focus on the multidimensional data modeling.

3.2 Logic modeling

At the logic level several presentations, are feasible of multidimensional modeling. The relational schema, suitable for decision-making, is presented in a star schema shaped, in snowflake or constellation:

a. Star Schema: it is a multidimensional representation of data given data [21] describing the fact placed in the center dimensions surrounding it. Every fact is a table called fact table, consisting of a set of attributes representing activity measures and foreign keys that reference the dimensions. Each dimension corresponds to a table, called dimension table containing attributes (strong or weak) and primary key to ensure the joints with the fact table. In this context, we are talking about a star logic model.

b. Snowflake schema: it consists in building a separate table for each hierarchical level of a dimension. In this model, the joints are numerous; the fact table includes a foreign key at the hierarchy level of each dimension.

c. Constellation schema: it is a collection of star schemas that divide the common dimensions.

4 Comparative study

The comparison between these three approaches seems essential, we rely on seven criteria classified into four categories such as: (i) the inputs occur on different kinds of data sources, (ii) the outputs articulating the goals which is the data warehouses, or the data marts creation, (iii) design focuses on logical or conceptual, formal and informal modeling of data representation, (iv) methods engineering requirements summarizing the specification requirements.

Figure 1 shows a comparison between the different approaches of data warehouses design based on these criteria.

From the level of design perspectives [6], [14] and [16] use logical schema. However, [12], [13], [17], [18] and [21] introduce conceptual schema.

The majority of these approaches aim to create the data warehouse as the work [6], [12], [13], [14], [16], [17] and [18]. However, the rest leads to create data marts such work [4] and [21].

All those approaches are formal methods except the approach of [13] and [21] where an informal modeling is introduced.

The design of data warehouse is based on several data sources namely the relational schema such as the works of [6], [12], [13], [14], [17], [18] and [21].Other methods [4] and [16] use ontology as data sources.

Given the complexity of the requirements specification, some research have developed their own techniques such as i * technique proposed by Mazon et al [12], the TROPOS method Giorgini [17], [18] and the GQM method (Goals / Question / Metric) [3].

Approaches / Criteria	*Approaches directed by sources*			*Approaches directed by requirements*				*Mixed approaches*		
	Golfarelli et al.[14]	Hüsemann et al.[6]	Romero et al.[16]	Kimball et al[21]	Cabibbo et al.[13]	Mazón et al.[12]	Giorgini et al.[17]	Bonifati et al.[3]	Abelló et al.[4]	Giorgini et al.[18]
level Conceptual schema	-	+	+	-	-	+	+	-	+	+
level Logic schema	+	-	-	+	+	-	-	+	-	-
Goals Data warehouse	+	+	+	-	+	+	+	+	-	+
Goals Data marts	-	-	-	+	-	-	-	-	+	-
Modeling Formal	+	+	+	-	-	+	+	+	+	+
Modeling Informal	-	-	-	+	+	-	-	-	-	-
Type of data sources ER diagram	-	-	-	-	+	-	-	-	-	-
Type of data sources Ontology	-	-	+	-	-	-	-	-	+	-
Type of data sources Relational schema	+	+	-	+	-	+	+	+	-	+
Conceptual representation Others	-	-	-	Adopting a	Adopting a	-	DFM	-	-	-
Conceptual representation UML Extension	-	-	+	-		+	-	-	+	+
Conceptual representation Personalized	+	+	-	-		-	-	+	-	-
Method used for Requirements Specification	-	-	-	-	-	Technique i*	TROPOS	GQM (Goals /Question/	-	TROPOS

Fig. 1: comparison between the different approaches to data warehouse design

At the sight of this comparative analysis, we can infer that led sources based approaches are useful if the schema of the data source is simple and available. In this category, they suffer generally from requirements engineering patterns. In contrast, the requirements based approaches, focus on the requirements specification which are frequently variable and limitedly expressed.
Thus, the design of data warehouses cannot be exclusively based on data sources or requirements. Indeed, we find that both ascending and descending approaches are complementary and can be mixed together for better results, being the subject of the third approach called hybrid approaches.

4 Conclusion

In this paper, we have presented various research approaches data warehouses design. These studies were classified according to three trends: sources based approaches; requirements based approaches and mixed approaches. Our comparative study discusses various works of data warehouses design.
In the future, we propose to study the problem of modeling data warehouses and we introduce a new method.

References

[1] A.Abello, J.Samos, and F.Saltor. YAM2 (Yet another multidimensional model),"An extension of UML". In Proceedings of the International Database Engineering & Applications Symposium. Edmonton, Canada. 2002, July 17-19, (pp. 172-181).
[2] A.Abello, J.Samos, and F.Saltor. YAM2,"A multidimensional conceptual model extending UML". Information System,2006, (pp.541-567).
[3] A.Bonifati, F.Cattaneo, S.Ceri, A.Fuggetta, and S.Paraboschi,"Designing data marts for data warehouses". ACM Trans. Softw.Eng. Methodol, 2001, (pp.452–483).
[4] A.Nabli, J.Feki, F.Gargouri, "Automatic Construction of Multidimensional Schema from OLAP Requirements", Arab International Conference on Computer Systems and Applications (AICCSA'05), Reference proceedings 0-7803-7983-7/03 © 2005 IEEE, Cairo, Egypt, January 2005.
[5] A.Tsois, N.Karayannidis, and T.Sellis, "Conceptual Data Modeling for OLAP". In Proc. DMDW. Interlaken, Switzerland, 2001.
[6] B.Hüsemann, J.Lechtenbörger, G.Vossen, "Conceptual Data Warehouse Design". Design and Management of Data Warehouses, Sweden, 2000.
[7] C.Sapia, M.Blaschka, G.Hofling, and B.Dinter, "Extending the E/R model for the multidimensional paradigm". Lecture Notes in Computer Science, 1552, 1999,(pp. 105-116).
[8] D.L.Moody, and MA.R.Kortink, "From Enterprise Models to Dimensional Models: a Methodology for Data Warehouse and Data Mart Design". In Design and Management of Data Warehouses, 2000, (pp. 5).
[9] E.Franconi, U.Sattler, "A DataWarehouse Conceptual DataModel for Multidimensional Aggregation". In Proceedings of the Intl.Workshop on Design andManagement of DataWarehouses (DMDW1999), Technical University of Aachen (RWTH), 1999, (pp. 13–1–13–10).
[10] F.Liu, C.Yu, and W.Meng, "Personalized Web Search For Improving Retrieval Effectiveness", IEEE Transactions on Knowledge and Data Engineering, vol. 16, n1,2004, (pp. 28-40).
[11] J.P.Mc.Gowan, "A multiple model approach to personalized information access", Master Thesis in Computer Science, Faculty of science, University College Dublin, 2003.
[12] J.Mazón, J.Trujillo, M.Serrano , and M.Piattini, "Designing data warehouses: from business requirement analysis to multidimensional modeling", in: K. Cox, E. Dubois, Y. Pigneur, S.J. Bleistein, J. Verner, A.M. Davis, R. Wieringa (Eds.), REBNITA Requirements Engineering for Business Needs and IT Alignment, University of New South Wales Press, 2005.
[13] L.Cabibbo and R.Torlone, "A Logical Approach to Multidimensional Databases". In VIth International Conference on Extending Database Technology (EDBT 98), Valencia, Spain, volume 1377 of LNCS, Springer, 1998,(pp. 183–197).
[14] M.Golfarelli, D.Maio, and S.Rizzi, "The dimensional fact model: conceptual model for data warehouses", international Journal of Cooperative Information Systems 7, 1998.
[15] N.Tryfona, F.Busborg, and J.Christiansen, "A conceptual model for data warehouse design". In Proc. of ACM Second International Workshop on Data Warehousing and OLAP (DOAP'99), Kansas City, Missouri, USA, November,1999, (pp. 3–8).
[12] O.Romero, and A.Abelló, "Automating Multidimensional Design from Ontologies". DOLAP'07, Lisboa, Portugal, November 9, 2007.
[17] P.Giorgini, S.Rizzi, and M.Garzetti, "Goal-oriented Requirement Analysis for Data Warehouse Design". In Proc. of 8th Int. Workshop on Data Warehousing and OLAP, ACM Press, 2005, (pp. 47–56).
[18] P.Giorgini, S.Rizzi, and M.Garzetti, "A Goal-Oriented Approach to Requirement Analysis in Data Warehouses". In Decision Support Systems (DSS) jounal, Elsevier, (pp. 4-21), Vol 45 Issue 1, 2008.

[19] P.L.Tchienehom, " Modèle générique de profils pour la personnalisation de l'accès à l'information", In INFORSID, 2005,(pp. 269–284),.
[20] P.Vassiliadis, A.Simitsis, and S.Skiadopoulos, "Conceptual modeling for etl processes", In Theodoratos, 2002,(pp. 14–21).
[21] R.Kimball," The Data Warehouse Toolkit", John Wiley and Sons, Inc., New York,1996.
[22] S.Alter,"Decision Support System": Current Practices and Continuing Challenges. Massachusetts: Addison-Wesley Publishing Co., 1980, (pp.316).
[23] S.Luján-Mora, J.Trujillo, I.Y.Song, "Extending the UML for multidimensional modeling". In Proceedings of the International Conference on the Unified Modeling Language. Dresden, Germany, 2002,(pp. 290-304).

[24] S.Luján -Mora, J.Trujillo, and I.Y.Song, "A UML Profile for Multidimensional Modeling in Data Warehouses". Data and Knowledge Engineering, 2006, (pp.725–769).

[25] S.Rizzi, "Conceptual Modeling Solutions for the Data Warehouse. Data Warehouses and OLAP: Concepts, Architectures and Solutions", edited by R. Wrembel and C. Koncilia, 2007.

[26] T.B.Nguyen, A.M.Tjoa, and R.Wagner, "An object-oriented multidimensional data model for OLAP". In Proceedings of the International Conference on Web-Age Information Management Shanghai, China, 2000, (pp. 69-82).