User Support System for Designing Decisional Database

Fatma Abdelhédi, Gilles Zurfluh
University of Toulouse1 Capitole – IRIT *(UMR 5505* - France)
{Fatma.Abdelhedi, Gilles.Zurfluh}@irit.fr

***Abstract*— The design of a multidimensional schema is usually performed by a specialist (computer scientist). According to data-driven, requirement-driven or hybrid-driven approaches, he determines the facts and axis of analysis. Such an approach assumes that the decision maker expresses, more or less formally, analysis needs and communicate them to the computer scientist. We propose multidimensional schema designing by the decision maker himself following a hybrid-driven approach. Through a process of assistance successively viewing intermediate schemas from sources, the decision maker gradually built his multidimensional schema. As what determined the measures studied, the analysis dimensions and hierarchies within dimensions. A software tool named SelfStar based on this principle has been developed and validated with decision makers.**

Keywords-Multidimensional model; design process; decisional Data-base; decision-makers' requirements; data-source

I. INTRODUCTION

Decisional Data Base (DDB) allows decision maker to analyze data structured according to a multidimensional schema (star or constellation). Multidimensional schema design has motivated numerous works which can be derived from data-driven, requirement-driven or hybrid-driven approaches [1][2][3]. These approaches, to perform as well as they are, all involve a designer who is only able to design a multidimensional schema from decision makers' requirements or data source schema. The development of such a schema is a cumbersome process because it requires decision makers to articulate their needs to a designer and that it captures well the need to translate them to a adapted schema form.

Decision makers do not generally mastered formalisms and IT (Information Technology) tools but are experts in their field work. The SelfStar project fits into this fact and aims to define the data warehouse process dedicated to casual users (decision makers). It propose an approach and a software tool allowing to the decision-maker to design himself his DDBs incrementally, according to his requirements, and this without recourse to an expert designer. This paper focuses on the constellation schema [4] design.

This paper focuses on the incremental design of a constellation schema from the relational schema of a data source. This work is part SelfStar project developed by our team and aimed to decision-makers to design their data-warehouses.

The paper is structured as following. In Section 2, we present briefly several approaches to design a multidimensional schema. The justification of SelfStar system is provided in Section 3. In Section 4, the input and the output of the system are successively defined. Section 5 is devoted to a presentation of multidimensional schema process. Section 6 describes the architecture of a case tool allowing to experiment these mechanisms.

II. RELATED WORK

Many studies have been devoted to the multidimensional schema approach. These approaches can be classified into 3 categories: data-driven, requirement-driven, and hybrid-driven approaches. Data-driven approach [1][5] uses database source to generate a set of candidate multidimensional schemas and presents the drawback of not taking into account decision makers' requirements, consequently, going towards a possible failure of expectations of decision maker. Requirement-driven approach [2][6] takes into the needs of decision makers, but the sources are ignored at first. After that, multidimensional schema has to be mapped on the data sources. So, the designer can discover that the requirements do not correspond to sources.

Hybrid-driven approaches [3][7] combines and integrate the benefits of these two previous approaches. In fact, this approach designs on the one hand candidate schemas from the data (data-driven approach) and on the other hand multidimensional schema from decision makers' requirements (requirement-driven approach). A designer must map these two types of schemas to obtain a consistent multidimensional schema. Romero and Abelló [3] propose an automatic method Multidimensional Design by Examples (MDBE) following a hybrid-driven approach. To generate multidimensional schemas, this method takes as input, on one hand, the decision makers' requirements expressed as SQL queries, and on the other hand, the relational data source.

Source analysis is provided by SQL queries and knowledge of relational source schema. Therefore, the multidimensional schema design requires an expert (a computer scientist) to formulate SQL [15] query and analyze data source. Pinet and Schneider [8] propose to generate a multidimensional schema from a conceptual schema using UML notations. This approach represents source classes with a directed acyclic graph. The user chooses a node from this graph to design a fact. All connected nodes to this chosen fact represent the potential dimensions of this fact. However, in our opinion, this representation of multidimensional schema is complex for the decision maker because of the number of generated nodes to represent the dimension hierarchies. However, interactions with the decision maker choices are limited to the facts.

To our knowledge, few works [7] try to allow users to participate in the process of multidimensional schema.

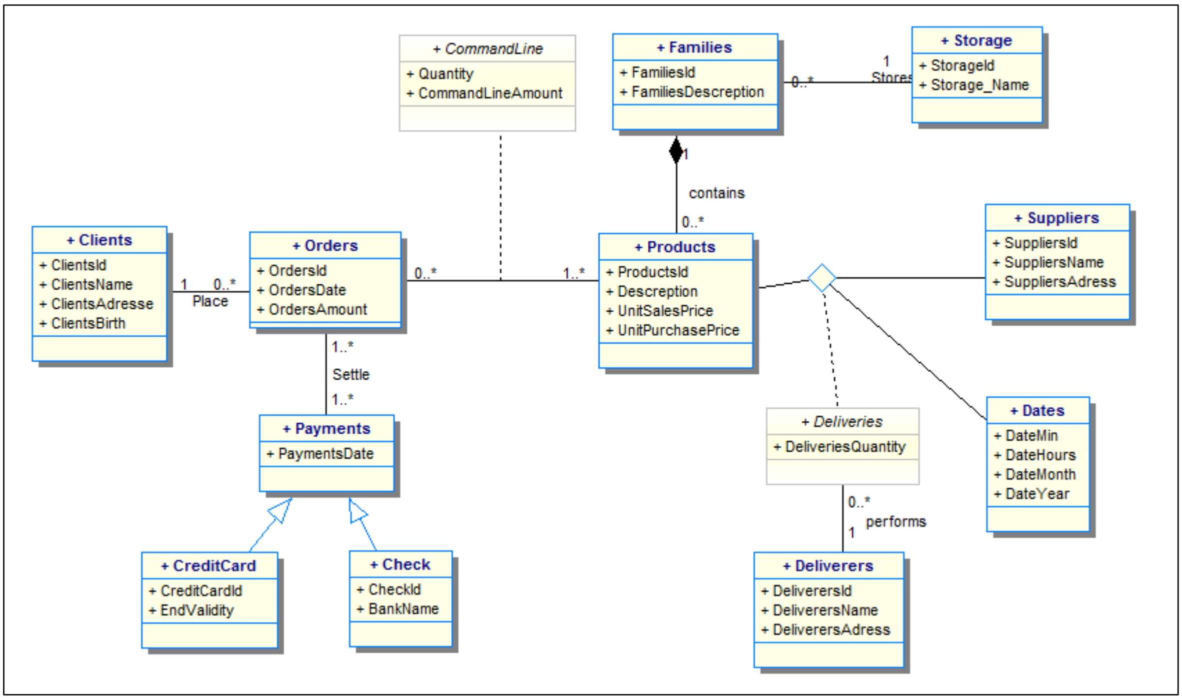

Figure 1. UML Class Diagram of products sales ans stock management

III. JUSTIFICATION OF SELFSTAR PROJECT

In the current approach, the decision makers should ask the specialists of the data management (administrators, computer scientist) each time they wish to get a new decisional database or to evolve the deprecated multidimensional schemas. This is related primarily to the complexity of:

- principles of designing of a decisional database
- ETL Processes ensuring the periodic load of the decisional database from there [9][10].

SelfStar project aims to propose a full approach and a software environment to allow to the decision maker to design a constellation schema. The design process of the schema is based on a hybrid approach: it starts from the database schema (class diagram CD) and the decision makers' requirements. This process is incremental: the decision maker integrates progressively the analysis needs in 3 successive intermediate schemas.

The advantage of such process is twofold:

- The decision maker becomes independent to express his decisional requirements,
- The process automatically controls the correspondence between analysis requirements and sources.

However, the decision maker, even if he knows his needs, is obviously confronted with a double complexity:

- Data source organization (Relational, Entity-Association or UML diagram),
- Process of the star or constellation schema design (fact(s), dimensions, hierarchy).

SelfStar project aims to propose formalized mechanisms to palliate this complexity.

IV. INPUT AND OUTPUT OF SELFSTAR

SelfStar system allows constellation schemas design from a data source and the decision-maker's requirements. In order to illustrate the input/output of the design process, we present successively an example of source schema, an expression of analysis needs and finally the decisional base schema resulting from the process.

A. Data source: Input of SelfStar

Data source is described by UML language. The choice of this formalism is justified by the semantic richness of the data model [2] and also by the correspondence between this language with the models entity-association and relational. Figure 2 gives the example of a classes diagram (CD) describing a management of the sales.

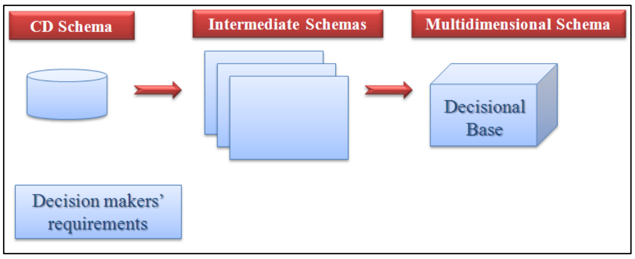

Figure 2. Our design process that allows a decision-maker to build a data-mart schema

The CD not being easily exploitable in the state, it is the subject thus of a reprocessing. This principle was already proposed in [11] for the Entity-Association model and in [7] for the UML CD. According to this principle, the conceptual schema is transformed into a simplified schema containing only object classes and binary links of type 1..N between these classes. The transformation of a DCL into an exploitable schema is carried out as follows:

- An object's classes or association's classes of the source becomes a class in the exploitable schema,
- A binary association links type 1..N is deferred in the state,
- A binary association links type M..N is transformed into a class bound by 2 links 1..N,
- A link of aggregation or composition is processed as an association links (these type of links are not meaningful in the multidimensional schema),
- A link of heritage disappears; the subclass of the source becomes a class and it will be on the same level as the super-class by inheriting its attributes and links (it preserves the semantics of the data).

The exploitable schema obtained is an oriented acyclic graph.

B. Requirements: Input of SelfStar

When a decision maker wants to analyze the data of a source, it can express his requirements:

- Either in the form of SQL request [12] ; but this proves to be difficult for decision maker except IT, especially when these requests make use of clauses Group By and Having;
- Or in the dashboards forms corresponding to the expected analysis results.

In SelfStar, the decision maker integrates his requirements by himself through a graphic interface and this without formalizing them beforehand.

C. Data-mart: output of SelfStar

From a data source and analysis needs, SelfStar system visualizes a decisional database on which the decision maker can apply its queries. The schema of Figure3 describe a decisional database allowing the decision maker to analyze the number of Orders (Orders_Nb measure) and the delivered quantity accumulated (Orders Amount measure) according to Products and Dates dimensions. Each dimension is associated to a set of organized parameters into hierarchies.

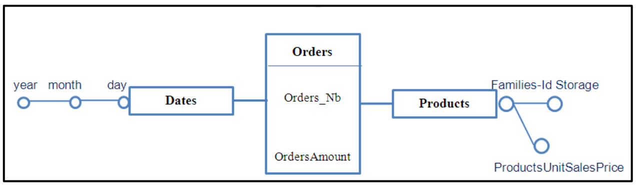

Figure 3. Data-mart schema

V. PERSONNALIZATION PROCESS

A. Multidimensional schema design

The proposed process in SelfStar (Figure 3) has a starting point the data source schema (CD) and the decision-makers' requirements (informal). It contains four successive steps in which the decision maker will interact with the system to integrate progressively his requirements. Each step produces a new schema more comprehensive than the previous one. The 4th schema corresponds to that of the decisional database, i.e. the desired result (Figure 3). The decisional schema design is performed incrementally.

First step consists to display a set of candidates facts in the intermediate schema number 1 (noted IS1). InIS1, the decision maker chooses the fact that he wants to analyze with the measures and their aggregations functions.

Second step generates automatically the intermediate schema number 2 (IS2); it proposes all possible associated dimensions with the selected fact. InIS2, the decision maker will be able to indicate dimensions according to which they wish to analyze the fact.

Third step generates an intermediate schema number3 (IS3) containing the constellation schema (fact(s) + dimensions) with all the possible hierarchies. In IS3, the decision maker will choose each responding hierarchies to their requirements.

Fourth step produces the multidimensional schema data-mart. At this level SelfStar will record the personalization metadata which later will allow the decision maker to elaborate the newest multidimensional schemas. All algorithms related to the different phases of the process are presented in [13].

B. Personnalization meta-data

Industrial experience of our team [14] showed that (1) a data source (sales or production BD) contains frequently from 30 to 60 objects classes and many links and (2) the analyzes carried by the same decision maker are usually very similar in terms of facts and dimensions.

According to (1), we consider that the decision maker isn't able to choose the fact directly from the conceptual source schema because this schema is considered complex for any person (except computer scientist). We have decided to show for the decision maker a simplified presentation of the source (noted IS1) that is extracted automatically from the source schema.

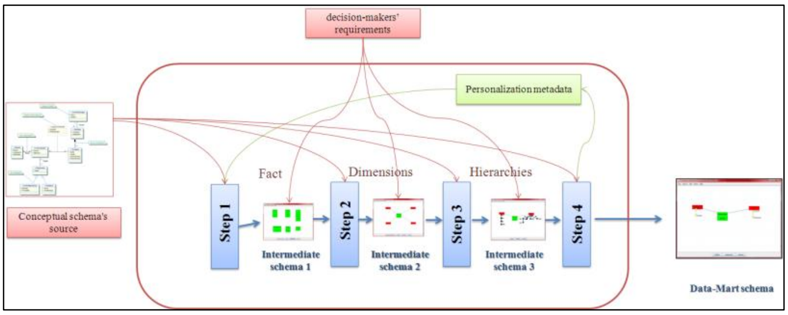

Figure 4. The transformation from data-source and decision makers' requirements to data-mart schema

The second point noted (2), led us to define a personalization mechanism on the IS1; this mechanism can extract, from the CD of source, the most significant classes for helping the decision maker.

The simplification of a source is based on the extraction of classes which must be analyzed by the decision maker (candidates facts). This extraction method is based on the personalization technical developed in our team and which inspired from the work of [12]. SelfStar record the personalization metadata whenever a decision maker develops a decisional schema. It is based on the scheduling of the candidate facts when the same decision maker will design a new multidimensional schema. Figure5 presents the algorithm that producing the personalization metadata.

```
Algorithm : PersoWeight
Input :  ES, CW, MS, U        -- Exploitable Schema, Classes Weight,
Multidimensional Schema and User
Output: CW                      --new weight values
begin
for i←1 to p do                 -- from each fact from MS
  s ← source($F_i$)
$P_s$← $P_s$+ Coef    -- increases the class source weight
for x in Dim($F_i$) do
    s ← source (x)
    Ps ← Ps +2
end for
end for
```

Figure 5. Generation personalization metadata

VI. CASE TOOL

A. Experimentation

SelfStar project was implanted in order to experiment all the proposed mechanisms. Figure 6 presents the software architecture. JAVA was used to develop this software. We used JAXP (Java API for XML Processing) that is a set of API including the SAX, DOM, XSLT (eXtensible Stylesheet Language Transformations) and XPath. The set of schemas used by SelfStar are with XML. We use JAXP tools to transform them. To view all XML schemas (intermediate schemas), we have developed a visualization module. It allows also the interaction between the system and the decision maker via the interface. This module uses the JGraph library.

Soon activation, the software demands to the decision maker identifier of relational database (source). CD source is transformed into an exploitable schema through the XSLT API (XSLT generator). The DOM generator creates then intermediate schema 1 (IS1) through by analyzing exploitable schema and using basic metadata (calculated weight). IS1 is displayed and the decision maker can choose directly on the screen fact(s) to analyze. This process is repeated to integrate choice of dimensions and attributes in the following intermediate schemas (IS2 and IS3).

Finally, the multidimensional schema will be displayed and personalization metadata are generated. They will subsequently be used to guide the decision maker in developing new multidimensional schemas. It should be noted that during the first analysis of the source, only the basic metadata are available. With each new analysis of this source, the metadata is enriched and will better define the profile of the decision maker.

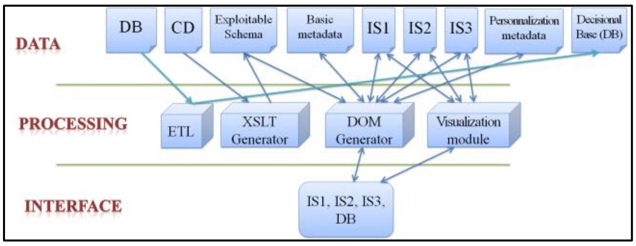

Figure 6. Software architecture of SelfStar

This paper does not deal ETL process. But, SelfStar has a module for loading data warehouse from sources. This module is automatically generated by SelfStar after the development of multidimensional schema.

B. *Validation*

To validate our approach and our tool, we used marks management application at University of Toulouse. The database source is composed of extracted data from the management education package (APOGEE [16] available in most french universities). We chose three decision makers of staff responsible for managing marks and awarding degrees.

The source is a relational database containing training descriptions, students and results. Figure 7 shows an extract from the schema describing the database source.

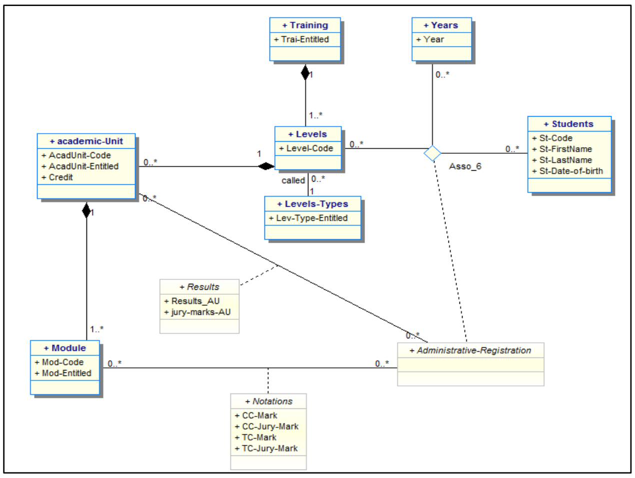

Figure 7. *CD extract of Database source*

We asked three employees of the University, having similar activities but independent create their own decisional databases using the database sources for analysis. It is noted that these decision makers use Business Object [17], occasionally.

(1) An official school: analyzes the students' absenteeism in the exams; observes the dates of issuance of marks of exams, etc.

(2) A training manager: defines the rate of exam success by students, analyzes the evolution of result per scholar year, etc.

(3) An academic manager of an academic year: examines student grades per subject per teacher (average, standard deviation, etc.), analyzes the results per subject, etc

After a brief training regarding the using of SelfStar, these three decision-makers showed certain affluence to quickly design new data-marts and manipulate them with Business-Object. The scheduling of the candidate facts was realized through the mechanism of personalization.

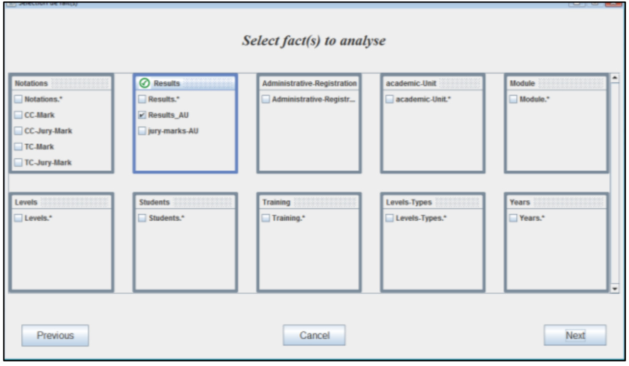

Figure 8. *Intermedite schema n°1*

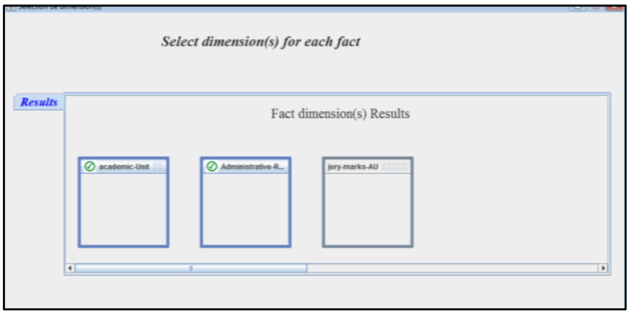

Figure 9. *Intermedite schema n°2*

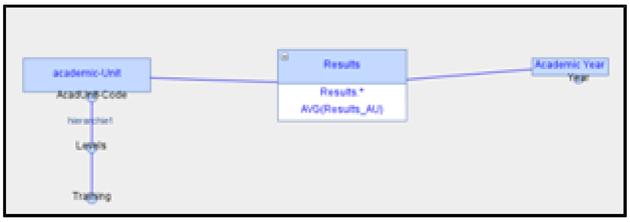

Figure 10. *Multidimensional schema*

VII. CONCLUSION AND FUTURE WORK

This paper provided an approach to design a multidimensional schema from the data sources schema to be analyzed; the decision makers' requirements are integrated progressively throughout the process. This approach is original in the measurement that it allows for the decision maker to design progressively his multidimensional schema without resorting to a database specialist (computer scientist). It is distinct clearly from the others data-driven, requirement-driven or hybrid-driven approaches in which the user is not directly involved.

REFERENCES

[1] F. Ravat and O. Teste, « Personalization and OLAP databases », New Trends in Data Warehousing and Data Analysis, vol. 3 , pp. 1–22, 2009.

[2] N. Prat, J. Akoka, and I. Comyn-Wattiau, « A UML-based data warehouse design method », Decision Support Systems, vol. 42, no 3, pp. 1449–1473, 2006.

[3] O. Romero and A. Abelló, « Automatic validation of requirements to support multidimensional design », Data & Knowledge Engineering, vol. 69, no 9, pp. 917–942, December, 2010.

[4] R. Kimball, « The data warehouse toolkit : Practical Techniques for Building Dimensional Data Warehouses », John Wiley and Sons, ISBN : 0-471-15337-0, 1996.

[5] D. L. Moody and M. A. R. Kortink, « From enterprise models to dimensional models: a methodology for data warehouse and data mart design », DMDW’00, Sweden, vol. 5, 2000.

[6] J. Trujillo, S. Lujan-Mora, and I. Y. Song, « Applying UML and XML for designing and interchanging information for data warehouses and OLAP applications », Journal of Database Management (JDM), vol. 15, no 1, pp. 41–72, 2004.

[7] P. Giorgini, S. Rizzi, and M. Garzetti, « Goal-oriented requirement analysis for data warehouse design », in Proceedings of the 8th ACM international workshop on Data warehousing and OLAP, 2005, pp. 47–56.

[8] F. Pinet and M. Schneider, « A unified object constraint model for designing and implementing multidimensional systems », Journal on Data Semantics XIII, pp. 37–71, 2009.

[9] P. Vassiliadis, « A survey of Extract–transform–Load technology », International Journal of Data Warehousing and Mining (IJDWM), vol. 5, no 3, pp. 1–27, 2009.

[10] F. Atigui, F. Ravat, O. Teste, and G. Zurfluh, « Using OCL for Automatically Producing Multidimensional Models and ETLProcesses », Data Warehousing and Knowledge Discovery, pp. 42–53, September 2012.

[11] I. Y. Song, R. Khare, Y. An, S. Lee, S. P. Kim, J. Kim, and Y. S. Moon, « SAMSTAR: An automatic tool for generating star schemas from an entity-relationship diagram », Conceptual Modeling-ER 2008, pp. 522–523, 2008.

[12] F. Ravat, O. Teste, R. Tournier, and G. Zurfluh, « Graphical querying of multidimensional databases », in Advances in Databases and Information Systems, pp. 298–313, 2007.

[13] F. Abdelhédi, G. Pujolle, O. Teste, and G. Zurfluh, « Computer-Aided Data-Mart Design », ICEIS, pp. 239-246, June - 2011.

[14] E. Annoni, F. Ravat, O. Teste, and G. Zurfluh, « Towards multidimensional requirement design », Data Warehousing and Knowledge Discovery, pp. 75–84, 2006.

[15] « Accueil oracle ». [Online]. Available: http://docs.oracle.com/cd/B19306_01/server.102/b14200/toc.htm. [Accessed: 12/25/2012].

[16] « Accueil Amue - Amue ». [Online]. Available: http://www.amue.fr/. [Accessed: 12/25/2012].

[17] « SAP France - SAP BusinessObjects - Solutions de Business Intelligence (BI) et de gestion de la performance ». [Online]. Available: http://www.sap.com/france/solutions/sapbusinessobjects/index.epx. [Accessed: 12/25/2012].