Society for Science and Education
United Kingdom
TMLAI Transactions on Machine Learning and Artificial Intelligence
Volume 5 No 4, 2017
ISSN 2054-7390

Automating the Conceptual Modeling of Data Warehouse in Information System ERP Type

Ayoub Elotmani, Pr. Halima Bouden
Laboratoire d'Informatique, Recherche Opérationnelle et Statistique Appliquée
Université Abdelmalek Essaadi.
Tetouan, Maroc.
Elotmaniayoub@gmail.com, bouden.halima@gmail.com

ABSTRACT

An information system type ERP (enterprise resources planning) aims to control the information flow and ensure the integrity of unique repository. However, the strategic level of this organization is looking to note a knowledge for vocation to decide and forecast. This feasible with the introduction of the business intelligence, which allowing storing historical data and organized in storage space: data warehouse (DWH), reduced to data mart (DM) that are an extract of the DWH adopted to a specific use, it is modelled for OLAP (on line analytical processing) analyses.

This article is part of a critical study, about the three Data Warehouse design approaches, in order to make the best approach adaptable to an ERP database. On the order hand, to study the proposals of the automation of the phases of the design of the warehouse.

Keyword: ERP, data warehouse, OLAP, MDA, reducing, Genetic Algorithm, Multiple Corresponding Analysis.

1 Introduction

The information system type ERP collect all operational data of the company in a unique relational database in the logical sense. The need for use of these data in the decision-making process, business intelligence that are dedicated to the management of the company, are based on relational sources to build multidimensional form patterns, easily accessible by decisions-makers, and which adapts well to the OLAP analyses. They are generally group them into two storage spaces: data warehouses (DWH), otherwise reduced to data mart (DM) for a particular use.

Small and medium-sized companies do not allow the high cost of a warehouse. As a result, they build their DM directly on their transactional basis OLTP. The present work is concerned with this alternative. It offers a benchmark of the automatics steps to aid the construction of DM's star schema. To develop the best approach to construction of the patterns of data mart dedicated specifically to the relational source of information system ERP type. In order to make proposals for improvement and evolution.

This article is organized as follows: in section 2, we briefly presenting the Enterprise resource planning ERP, and the data warehouse. In addition, the description of the three approaches of design of the data warehouse. A comparative study of these three approaches is presented in section 3. Section 4, is devoted

DOI: 10.14738/tmlai.54.3330
Publication Date: 15th August, 2017
URL: http://dx.doi.org/10.14738/tmlai.54.3330

to the detailed presentation of the process of developing a multidimensional schema that meets the limits of the comparative study.

2 From Relational to Multidimensional

2.1 ERP

ERP or "Enterprise Resource Planning" software can be defined as an application computer customizable, modular and integrated, which aims to integrate and optimize the process of management of the company by providing a single repository and relying on standard management rules. [15].

In General, an ERP system contains the following general characteristics:

- An integrated system that has the feature to integrate several functions of the company [3]. It extends to the whole of the business processes, the MRP concept developed in the 1970s for industrial management.
- A modular tool composed of a set of application modules (usually signed by the same publisher) that shares a single, logical database. Integrating these modules around a single repository of data is likely to ensure workflow collaboration.
- A "configurable" system allows combining standardization and adaptation to the company [3]. He opposed the proprietary software developed for the specific needs of the company. It allows quick adaptation to the operating rules offering tools of development or customization of application add-ons [8].

Therefore, the ERP is software integrated with the characteristic to cover several or all functions of the enterprise, each managed by a function module, the integration of these modules around one unique repository of data.

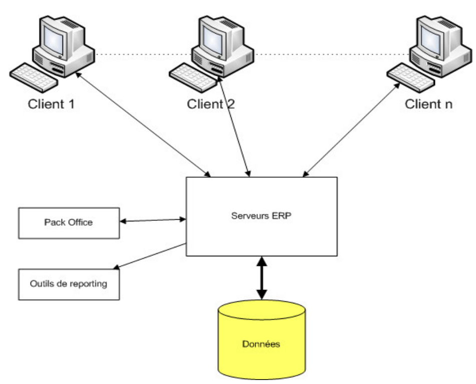

Figure 1: Third Architecture of an Enterprise Resource Planning

2.2 Data WareHouse

The classic definition of a warehouse given:

"A data warehouse is a collection of subject-oriented data, integrated, non-volatile, histories and organized for the support of a decision-making process."

We detail these features:

- Subject-Oriented: data warehouses are organized by topic. For example, a chain of food stores organizes data from its inventory to sales carried out by product and store, during a period.
- Integrated: The data from different sources must be integrated before their storage in the data warehouse. Integration (for example, put in correspondence of formats), to have a consistency of information
- Non-volatile: the difference of the operational data, data warehouse are permanent and cannot be changed. The refreshment of the warehouse is to load new data, without changing or losing existing ones.
- Time-Variant: the evolution of the data is essential for the decision-making process, which, for example, uses prediction techniques based on past trends to predict future developments.

2.3 State of the art and related work

This work is part of the construction of multidimensional data warehouses. The construction of a DWH made from the data of an organization.

The decision-making information system literature proposes three classical approaches to data warehouse design: (i) top-down approach, (ii) bottom-up approach, and (iii) mixed approach.

2.3.1 Top-down approach

The top-down approach is used when economic problems are well known. This approach provides a synergy between business topics and provides a unique version of the truth.

The top-down approach is consistent with Bill Inmon's view that the data warehouse must respond of requirements of all organization’s users, not just a certain group.

This type of approach adopted in [7] where the author is based on the identification of the facts and the attributes of the dimensions to elaborate the set of star schemas. Of course, this design method produces a solution that responds well to analytical needs Expressed by future users.

Then it allows constructing multidimensional schemas from the analysis of the needs of the decision makers, the problems of conformity between the OLAP needs (On-line Analytical Processing) and the source of the ERP are manipulated when loading the data in the warehouse.

This method is based on three essential phases:

(i) Data collection, (ii) needs specification, and (iii) formalization of needs

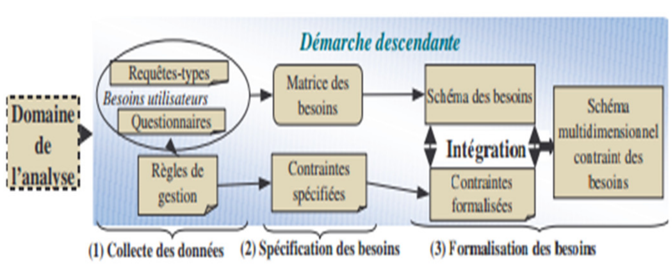

Figure 2: schematic of the top-down approach

Data collection

The data collection phase consists of determining the initial requirements by defining the relevant OLAP queries through questionnaires in order to better position the need.

Requirements specification

This phase allows analysing and organizing data in queries, and translating the easy to integrate in the multidimensional schema constraints management rules.

Formalisation of needs

After collecting the data and specifies the needs of users, the remaining phase is to formalize these multidimensional schema needs. In addition to the definition of the multidimensional facts of this representation must meet the decision-making needs sets.

This construction goes through four steps:

Definition of the facts

Identical measures analysed through parameters collection automatically.

Definition of the dimensions

Each setting characterized by the same measures is automatically grouped into a dimension. Enriched by the addition of attributes low.

Definition of hierarchies

The Organization of the parameters grouped in each dimension in hierarchies.

- Multidimensional schema definition.

According to needs, collection dimensions are automatically assigned to the facts that fit.

2.3.2 The bottom-up approach:

The bottom-up approach is based on the information contained in the data source. It is a flexible method that allows the organization to go further with lower costs, by creating independent datamarts from the data source schema (E / A) relationships. In this case, it is necessary to have a designer of the multidimensional schema with a dual competence computer and trades. For it is the designer of the multidimensional schema that must elaborate the different interests of the organization by analyzing the relational schema.

This approach is consistent with Ralph Kimball's view that the data warehouse should be easily understood by users and provide the correct answers as soon as possible.

The construction of the data warehouse schemes is based on the model of the existing relations between entities this type of approach is adopted by [6] and [9]. This approach considers that the relevant information is in the data source.

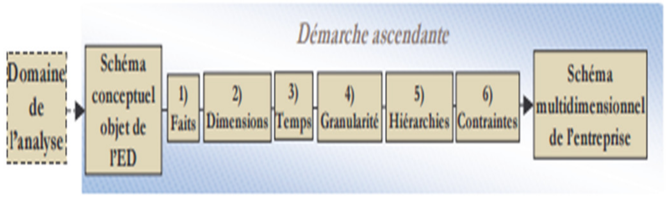

Figure 3: schematic of the bottom-up approach

2.3.3 The Mixed Design Approach:

This type of approach has been adopted in [10] and [2], consists in combining these two design approaches. The source-guided design phase automatically generates candidate DWH schemas. These schemes consist in producing potential facts based on the entities and associations of the source model. This approach has also been taken up by [18], which generate ideal schemes by fusion of analytical needs and then validates these ideal schemes obtained by projection on a power source.

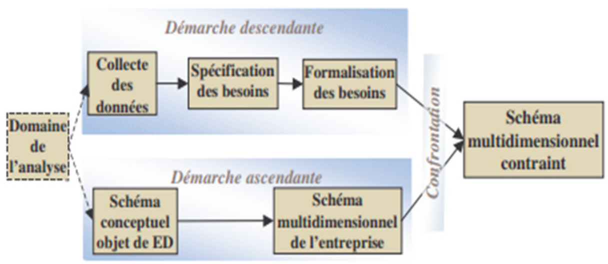

Figure 4: schematic of the mixed approach

Thus, the top-down approach builds candidate schemes that respond well to queries expressed by end users. On the other hand, the bottom-up approach is based on the existing relationships between the data of the relational source and follow a more structured approach to design the company's decision database. On the other hand, the mixed approach is the combination of the two approaches, which consists in constructing multidimensional schemas based on the requests of the decision-makers; on the other hand, it is based on the relational schema source.

The focus of this study is criticism of the top-down approach based on OLAP needs, and the bottom-up, mixed approach to DM design, which is based on the source (ERP data). For drawing the best approach, in order to make improvements.

3 Comparative study of DM design approaches

Reporting and analysing data are the main mission in the process of designing the data warehouse, in fact taking into consideration the source of data is necessary, however the decision makers will be confused while designing the data warehouse so it is such a priority and necessity to show that if neglecting the source . The data will be unstructured in terms of relationships between its characters. Hence, that will lead to the fact that is meaningless for the decision makers to contribute in such cases till the data is structured.

In designing the data warehouse guided by the needs of decision makers, the source data did not taken into consideration. Is right, only the needs of decision makers are necessary to build the data warehouse scheme. So the multidimensional schema design may become incorrect if the existing relationships between the data are not clearly identified to the designer. On the other, hand unnecessary for decision makers if the problems of the correspondence with the source relational scheme experienced a total divergence.

The authors of [12] propose a design method based on UML modeling. At the conceptual level, the need of decision makers is translated into an UML class diagram format in order to obtain a multidimensional scheme. This approach had drawbacks because user requirements alone are taken into consideration, and the unavailability of the source data influences the implementation of the conceptual schema produced.

On the other hand, the design guided by source exploits the data of the source whatever the level of relevance. Therefore, this case the perimeter of analysis is sometimes wide which requires time and it is difficult to manage the needs of analysts that can be evolved over time or from decision maker to the others. Therefore, designers must always consult the source relational schema that can be uncontrolled.

To remedy this limitation, the authors of [6] propose to the decision-makers to specify the relevant source data. The disadvantage of this proposal is that designers should consult source databases know that they do not always master them.

This approach shares the relational scheme to build facts for independent data mart. However, there may be problems when integrating datamarts into a coherent data warehouse. We can't find the nodes to link datamarts, to construct a coherent constellation schema of a data warehouse

This approach stems from the needs of the company developed by decision-makers, while the top-down approach aims at integration and consistency of data at the enterprise-wide level.

Another proposition in [11], the authors propose a multidimensional schema model constructed from an UML conceptual schema that represents the source. The designer chooses a node of this model to define a fact. All the facts selected are linked with potential dimensions of this fact. However, in a management information system ERP type the relational schemes are wide. Therefore, in our opinion, it is complex for decision makers to represent dimension hierarchies because of the number of nodes generated.

Despite the importance of the fact-finding step from the company's data model, none of these approaches presented a precise method for doing so. While it is always necessary to review the need of end users. In order to remedy this, the mixed approach proposes combining the two approaches [5] and [16]. Indeed, this approach builds the candidate schemas from the data (bottom - up approach) that is to say the determination of these axes of analysis from the relational sources. Beyond this the multidimensional schemas are based on analytical needs (top-down approach). The designer must compare these two types of schemas to obtain a multidimensional result consistent and responding to the needs of decision makers.

[16] Emphasize on automatizing a suitable solution that lies on the Multidimensional Design by Examples (MDBE), which follows a mixed approach. To generate multidimensional schemas, this method takes as input on the one hand the needs of the decision makers expressed in the form of SQL queries and on the other hand the source of relational data. The querying of the sources is ensured by SQL queries and a knowledge of the relational schema of the source. Therefore, building the multidimensional schema requires an expert to formulate SQL queries and query the data sources.

In the shadow of our knowledge, little work involves directly the decision-maker in the process of developing a multidimensional scheme. Note that the article of [11] proposes to the decision-maker to reduce the star schema generated by the system in order to match the scheme to its analysis needs. However, interactions with the decision-maker remain limited.

4 Proposal

In order to respond on the exiting limits approaches,

We consider that the ascending step is relayed separately but in parallel with the downward step, that is to say, the designer can, in the ascending phase, consult the information collected during the downward step. This is proposed in [1] the authors propose an approach (i) mixed, (ii) unified and (iii) automatic for the design of multidimensional DWH. It is a 'mixed' approach to modeling since the schema of the warehouse must be built from the needs of decision-makers and validated against the data sources.

This approach is called 'automatic' because it proposes a set of rules allowing the automatic generation of the logical and physical diagrams from the conceptual schema. In this sense, Model-Driven Architecture (MDA) offers a set of automatic transformation models. In MDA's advocates the development of three types of models: i) the model of the requirements (Computation Independent Model: CIM). (ii) The Analysis and design model (Platform Independent Model: PIM). (iii) The concrete design or code model (Platform Specific Model: PSM)

The MDA's principles must build a model of the requirements (CIM) that automatically turns this in a sequence of models in order to get a physical model adapted to the chosen platform.

In this proposal the first step is not treated, the two following steps allow to build the independent model (PIM) platforms. Which is divided into two different levels of PIM. The first level represents the conceptual model (PIM1) and allows describing the multidimensional structure of the warehouse [13] and specifying the associated ETL treatments. The second level (PIM2) presents different succession of logical models produced automatically from the (PIM1). Finally, the different logical models are translated automatically into a set of platforms of physical implementation models (PSMs).

4.1 Model PIM

In this proposal, the first step is not treated. The next two steps allow building platform independent models (PIMs). That is divided into two different levels of PIM. The first level represents the conceptual model (PIM1) and allows to describe the multidimensional structure of the warehouse [13] and to specify the associated ETL treatment. The second level (PIM2) presents different succession of logical models produced automatically from (PIM1). Finally, the different logical models are translated automatically into a set of physical implantation platform (PSM) models.

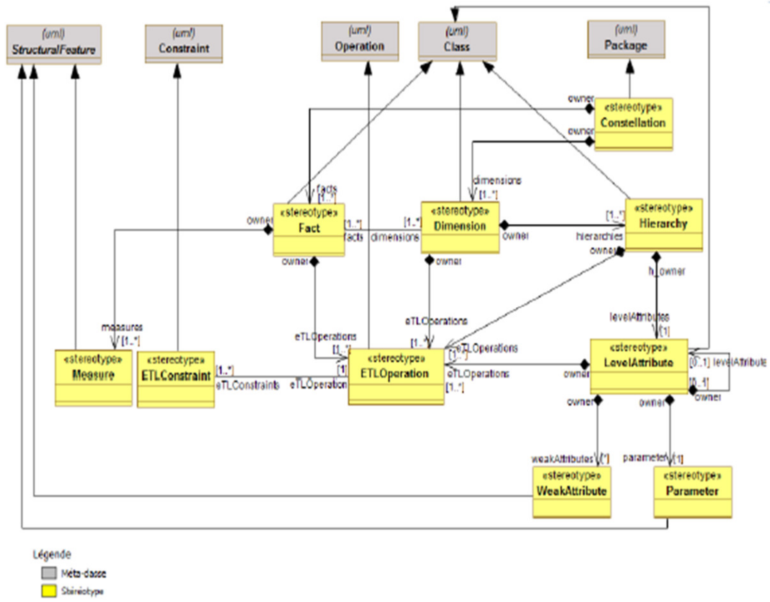

Figure 5 : UML Profile for DWH (DWP)

The UML profile for the DWH explicit in figure 5, composed by all the components of the data warehouse.

In the context of DWHs, static properties are the measurements for the facts, parameters, and weak attributes of the component hierarchies dimensions. The operations concern particularly the warehouse loading processes modelled by the stereotype "ETLOperation" (extension of the meta-class "Operation"). We assume that each fact, dimension, hierarchy, and attribute level must have at least one ETL operation. These operations are described by a list of constraints defined by the stereotype "ETLConstraint". All stereotypes defined in this way have public visibility with the exception of operations and ETL constraints ("ETLOperation" and "ETLConstraint"). The latter must be invisible to the decision-maker. Explain by [1].

The conceptual model (PIM1) is used to represent data that respond to the needs of decision makers in a multidimensional format. This representation is based on the static structure of this model. Therefore, we focus on concepts of fact, dimension, measurement, hierarchy, parameter and weak attribute.

On the other hand, the dynamic aspect of the conceptual model presents the loading operations that concern each multidimensional element. These processes are described using UML processes to validate the multidimensional elements in relation to the different operational sources. After the conceptual model definition (PIM1), the set of rules is defined for the automatic generation of the chosen logical models (PIM2).

4.2 The PSM model

Generally, DWH design stops at logical level. The benefit of this proposal covers the whole cycle of development from formalization of need to physical level. Each constructed logic model is transformed into a logical model through the application of a set of rules of passage to the materialized view.

The physical models (PSM) described by the materialized views are produced by the assembly of the conceptual (PIM1) and logic (PIM2) models. It is a multi-model transformation: one or more input models are transformed into one or more output model (s). Indeed, each dimension of the conceptual model is translated into an implanted physical dimension. The second input model is the logic model. At this level, each table is transformed into a materialized view.

After the automatic implementation of data warehouse, we always leave a margin where the number of the dimensions are many, for this we thought to look for a method to reducing the dimensions of data warehouse after the conceptual modeling that we adopted as a proposal. According to our studies, we find beforehand a proposal to reducing the multidimensionality of the Olap cubes with the genetic algorithm and the multiple correspondence analysis.

In this context, the authors of [17] propose an approach for reducing the data warehouse's dimensions by identifying a subset of dimensions that may contain the most amount data of the original data warehouse.

This approach uses Genetic Algorithms (GAs) that have proven their efficiency in solving optimization problems, to identify the subset of dimensions to keep. Also used Multiple Correspondence Analysis (MCA) to evaluate different combinations of dimensions proposed by this GAs. It is based on the elimination of certain dimensions of the Olap cubes by merging the groups of candidate facts created in order to be regrouped and replaced by a unique fact.

MCA is a multidimensional analysis method of an array containing qualitative data; it is considered a qualitative data matrix. Where a set of individuals in the lines is described by a set of attributes in the columns. A special data structures for the ACM analysis method including tow table:

- Table of Condensed Coding (TCC): It is a table with n rows and p columns. At the intersection of the row (variable) and the column (individual).
- Complete Disjunctive Table (CDT): This table is built from the TCC and has n rows includes the modalities of variables associated with the m columns.

The MCA offers the possibility of comparing two rows or columns by calculating the distance between them. Such a technique makes it possible to calculate the degree of similarity between these rows or columns. This distance calculation can be done using the distance square to compare two points of individuals (two line profiles) or a column arrangement (two vertical profiles).

In this approach, the goal is to reduce the number of dimensions of a data warehouse. It is a matter of selecting the dimensions contain the most amount information for decision makers from the initial dimensions. In this case, the use of the GAs approach is interesting.

Each individual designates the dimensions of the warehouse, which also represents a possible solution to the problem. Will be represented by a chromosome containing the same number of genes as data warehouse dimensions. The function's adaptation used to evaluate the quality of each individual in the different successive populations corresponds to the calculation of the distances between the individual's terms of the MCA. We can deduce this calculation by the modality of equation to calculate the distance between two terms points. Which represents the degree of resemblance between two individuals, for each modality. After the sum of all the corresponding calculated values of the modalities of all dimensions, we can evaluate the new distance decreases between individuals.

The following algorithm proposed by [17] for this purpose:

```
for i from 1 to 10 do
for j from 1 to 6 do
TCC[i][j] ← rondom(1,0)
end for
end for
for i from 1 to (N=10) do
j=1
for k from 1 to length(dim1[]) do
if (ind[i][j] = dim1[k])
then TDC[i][k] ← 1
end if
end for
j[m=1 to 5]=j[m-1] + length (dim[p=2 to 6][])
for k[m=1 to 5]=1 tolength (dim[p=2 to 6] []) do
if (ind[i][j] = dim[p=2 to 6][k(m=1 to 5)]
then TDC[i][k(m=1 to 5)] ← 1
end if
end for
end for
for i from 1 to 10 do
for j from 1 to 19 do
if (TDC[i][j] 1)
then TDC[i][j] ← 0
end if
end for
end for
for k from 1 to 6 do
for j from 1 to 19 do
Occ← 0
for i from 1 to 10 do
som← 0
if TDC[i][j] = 1
then occ← occ+1
end if
end for
A= occ/N
B=(TDC[i][j]-xref)²
X=B/A
som← som+X
end for
dist← som/p
end for
{ind1(i+1), ind2(i+1), ind3(i+1), ind4(i+1)} ← min {ind1(i), ind2(i)…, ind10(i)}
{ind5(i+1), ind6(i+1)} ← crossover ( ind1(i+1), ind2(i+1), pc )
{ind7(i+1), ind8(i+1)} ← crossover ( ind3(i+1), ind4(i+1), pc )
{ind9(i+1)} ← mutation ( ind1(i+1), pm )
{ind10(i+1)} ← mutation ( ind3(i+1), pm )
```

Figure 6: The reduction corresponding algorithm

We can obtain in the example treated by [17] that the reduction of the dimensionality of the OLAP cube using the GA and the MCA is followed by a reduction on the lines of the data warehouse. We also notice the lines that are similar. Become doable to eliminate and replace them with a single line and summarize the measurements of each in the new line. In FIG. X, the algorithm corresponding to the process of identifying the duplicate lines and the replacement by the correspondent.

```
while i<=n do
for (j,k) from 1 to p' do
tab[k] ←CDTm[i][j]
end for
i++
while (j,k)<= p' do
if tab[k]==CDTm[i][j] then (j,k)++
else i++
end if
m=1
measure=0
if k==p' then
l[m]=i
measure= measure+measure[i]
m++
end if
```

Figure 7: Corresponding algorithm to the identification of duplicated lines.

5 Discussion

Our study focuses on the automatic construction of multidimensional data warehouse schemes to bypass problems of decision making in ERP-type management information systems. The solution proposed by [1] is adaptable for our case because it responds to the need to design and build a physical data warehouse scheme automatically - this solution is based on the principle of model-driven architecture - for to produce historically, aggregated, business-oriented information, described in a unique and consistent way, mainly based on the basic data of the ERP. Thus, to remedy this problem, the problem of correspondence between schemas built based on user requirements with the relational schema of the ERP, which is sometimes wide. This proposal develops a mixed approach that combines the two top-down and bottom-up approaches.

In the next article, we will discuss a problematic which will improves this work. For processing large amounts of data. These include handling large non-relational databases, data warehouse support. That allows easy access to a Cloud environment. The results obtained can be used to circumvent the problem of decision-making regarding the introduction of ERP in cloud environments.

6 Evolution and Conclusion:

This document presents an ERP as an unavoidable tool that integrates all the company's processes and facilitates the flow of data. The ERP contains many important data for decision-makers. The disadvantage is that an ERP generates only reports of particular problems. For this, we have thought of defining a solution that is based on relational data, in order to build a data warehouse automatically. We take into account the needs defined in the end users.

We have studied the three bottom-up, top-down and mixed-design approaches of the data warehouse, and discuss the boundaries of each. In order to draw the mixed approach as a suitable approach to our case. With the solution presented in [1] whose authors propose a model-driven approach, for the design and automatic construction of a data warehouse.

The future work is to propose a solution for processing large amounts of data. These include manipulating large non-relational databases, data warehouse support. That allows facility to access to a Cloud environment. That goes with the new revolution in ERP implementation in cloud environment.

REFERENCE

[1] Atigui F., Ravat F., Teste O., Zurfluh G., Démarche dirigée par les modèles pour la conception d'entrepôts de données multidimensionnelles. 26éme Journées des Bases de données Avancées BDA, Toulouse, 2010.

[2] Bonifati A., Cattaneo F., Ceri S., Fuggetta A., et Paraboschi S. « Designing Data Marts for Data Warehouse », ACM Transactions on Software Engineering and Methodology, vol. 10(4), p. 452-483. Octobre 2001

[3] Davenport, Thomas H. « Putting the enterprise into the enterprise system ». Harvard business review 76 (4). 1998

[4] Feki J., Hachaichi Y. Du relationnel au multidimensionnel : Conception de magasins de données. Sfax, janvier 2007.

[5] Giorgini P., Rizzi S. et Garzetti M., Goal-oriented requirement analysis for data warehouse design. ACM International Workshop on Data Warehousing and OLAP (DOLAP'05), p. 47-56. Bremen, Novembre 4-5, 2005.

[6] HÜSEMANN B., LECHTENBÖRGER J. et VOSSEN G., Conceptual data warehouse modeling. 2nd international Workshop on Design and Management of Data Warehouse (DMDW'00), p. 6.1 – 6.11, Vol. 28, Suède, June, 2000.

[7] KIMBALL R. The Data Warehouse Toolkit : Practical Techniques for Building Dimensional Data Warehouses. New York, 374 p, 1996

[8] Lequeux. 1999. Manager avec les ERP. Editions d'Organisation.

[9] Pang C., Taylor, K., Zhang, X. et Cameron, M. Generating Multidimensional Schemata from Relational Aggregation Queries. pp. 584–589, 2004.

[10] Phipps, C., Davis K., Automating data warehouse conceptual schema design and evaluation. 4th International Workshop on Design and Management of Data Warehouse (DMDW'02), p. 23-32, Canada. May 272002.

[11] Pinet. F, et Schneider. M., A Unified Object Constraint Model for Designing and Implementing Multidimensional Systems. Journal on Data Semantics XIII, pp. 37-71, 2009

[12] Prat N., J. Akoka, I. Comyn-Wattiau. "A UML-based data warehouse design method", Decision Support System, Vol.42(3), p. 1449-1473, 2006.

[13] Ravat F. Modèles et outils pour la conception et la manipulation de systèmes d'aide à la décision. Université des Sciences Sociales, Toulouse, décembre 2007.

[14] Ravat F., Teste O., Tournier R., G. Zurfluh. « Querying Multidimensional Data bases ». 11th East-European Conferance on Advances in Databases and Information Systems (ADBIS'07) p.298-313, Varna, septembre 2007.

[15] Reix, Robert, Bernard Fallery, Michel Kalika, et Frantz Rowe. Systèmes d'information et management des organisations. 6e édition. Paris. 2011

[16] Romero. O, et Abelló. A. 2010. Automatic validation of requirements to support multidimensional design. Data & Knowledge Engineering. Pp. 917-942, Septembre 2010.

[17] Semeh B., NAOUALI S., Reducing the multidimensionality of OLAP cubes with Genetic Algorithms and Multiple Correspondence Analysis. The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015). Pp 452 – 459. 2015

[18] Soussi A., Feki, J., Gargouri. F. Approche semi-automatisée de conception de schémas multidimensionnels valides, Journée Entrepôts de Données et l'Analyse en ligne (EDA'05), Revue des Nouvelles Technologies de l'Information, RNTI-B-1, Cépadues, P71-90. Juin 2005.