Merge of X-ETL and XCube towards a Standard Hybrid Method for Designing Data Warehouses

Nawfal El Moukhi*[1], Ikram El Azami[2]
Abdelaaziz Mouloudi[3]
MISC Laboratory, Faculty of Sciences
Ibn Tofail University, Kenitra, Morocco

Abdelali Elmounadi[4]
LASTIMI Laboratory
Mohammadia School of Engineers,
Mohammed V University, Rabat, Morocco

***Abstract*—There is no doubt that the hybrid approach is the best paradigm for designing effective multidimensional schemas. Its strength lies in its ability to combine the top-down and bottom-up approaches, thus exploiting the advantages of both approaches. In this paper, the authors try to identify and analyze the different hybrid methods developed for building data warehouses. The analysis revealed that the existing methods are too complicated and time consuming in the deployment phase. In order to solve this problem, the authors introduced a new hybrid method that is easy to use and saves a huge amount of deployment time. This new method consists of two main steps: the first data driven step allows an analysis of the source models by using the X-ETL method and gives rise to star models. The second requirements driven step performs a semantic analysis of the needs expressed in natural language by using the XCube Assist method. This analysis allows to improve the quality of star models generated by the X-ETL method without the intervention of a designer.**

Keywords*—Data warehouse design; hybrid method; relational model; multidimensional model; star model; X-ETL; semantic analysis; XCube assist

I. Introduction

There is no doubt that the data warehouse design is a crucial phase in any project of setting-up a decision support system. Therefore, it is very important for data warehouse designers to follow a well-founded and a well-consolidated design methodology which allows them to get a model that best meets the decision-maker's needs during the analysis and the data mining phases. Thus, several researchers have been interested in this question and a lot of research has been done on the design of data warehouse models. These works have led to the development of several design methods but none of them has been the subject of consensus to become a standard in the field. According to the adopted paradigm, these methods can be divided into two different categories: data-driven methods and others driven by requirements.

The data-driven methods, also known as supply-driven methods, aim to define multidimensional schemas by using model-driven engineering techniques on data sources. The approach followed by these methods offers the advantage of fully automating the design process but neglects the user's needs; contrary to requirements-driven methods that define multidimensional models on the basis of business goals and decision-maker's needs. However, these latter methods can lead to incompatibility between needs and data sources. These two approaches are contradictory and complementary at the same time since each of them has an advantage missed in the other. This complementarity has inspired researchers to define a new hybrid approach that combines these two classical approaches and includes all their advantages.

Today, the hybrid approach is the best paradigm for developing a unified and efficient design method that meets both user's needs and the issue of data availability. Thus, several studies have recently been conducted to develop hybrid methods for designing data warehouses. However, the methods developed so far remain too complicated and require a lot of time for deployment. In this sense, the authors introduced through this paper, a new hybrid method that is easy to use with a reasonable deployment time. The new hybrid method is based on a combination between the data-driven method X-ETL [1] and the requirements-driven method XCube Assist [2]. The paper is organized as follows: Section 2 presents the different hybrid methods developed to date as well as a critical analysis and comparison between these different methods. Sections 3 and 4 summarize the X-ETL methodology and the XCube Assist methodology and present the advantages and disadvantages of each of these methods. In Section 5, the authors present their new hybrid method and describe the entire semi-automatic process. Section 6 applies the new method on a case study from the literature.

II. Related Work

There are multitudes of methods for designing data warehouses. Among these methods, there are those that follow the hybrid approach.

- Cabibbo & Torlone, 1998 [3]: present one of the most frequently cited multidimensional design methods. It allows to generate a logical schema from ER (Entity-Relationship) diagrams and to produce multidimensional schemas in the form of relational databases or multidimensional tables. At first glance, this method appears to follow a data-driven paradigm since it allows for in-depth analysis of data sources. However, no formal rules for identifying multidimensional concepts from data sources are provided. Indeed, multidimensional concepts are identified from user requirements, hence its consideration as a hybrid method. This method consists of four essential steps: The first two steps aim to identify facts and dimensions and to restructure the

*Corresponding Authors

entity-association diagram. These two steps can be performed simultaneously to benefit from the feedback of each step. The authors even suggest that they should be carried out iteratively in order to refine the results obtained. After manually identifying the multidimensional concepts, each fact is represented as an entity and the dimensions missed in the schema that could be derived from external sources or associated metadata must be added. At this stage, it is also necessary to refine the levels of each dimension by means of the following transformations: (i) replacing many-to-many relationships, (ii) adding new concepts to represent new levels of interest, (iii) selecting a simple identifier for each entity level and (iv) deleting irrelevant concepts. Finally, two final steps aim to derive the multidimensional scheme. Some clues are provided to derive a multidimensional graph that will be directly mapped into the multidimensional schema. In general, this method remains informal like Kimball's. However, it established the foundations that were later used by the other methods;

- **Böehnlein & Ulbrich-vom Ende, 1999** [4]: present a hybrid method for deriving logical schemas from SER (Structured Entity Relationship) diagrams. According to the authors, SER is a better alternative for identifying multidimensional structures since it allows to visualize the existence dependencies between objects. This method has three main steps:

  - Pre-process: This step consists of transforming the ER diagram into a SER diagram.
  - Step 1: consists of identifying business measures from goals. The authors suggest looking for business events to discover interesting measures. Once these measures are identified, they are mapped to one or more objects in the SER diagram and will eventually generate facts.
  - Step 2: According to the authors, the hierarchical structure of SER diagrams is useful for identifying potential aggregation hierarchies. Thus, aggregation dimensions and hierarchies are identified by means of direct and transitive functional dependencies. However, the authors point out that the discovery of dimensions is a creative task that must be complemented by a good knowledge of the application field.
  - Step 3: In this step, each fact table is created using all the primary keys of its analysis dimensions as a compound primary key and denormalizing or normalizing the aggregation hierarchies. Finally, a star or snowflake schema is derived;
- **Bonifati et al., 2001** [5]: is the first method to introduce a formal hybrid approach and generating at the end a multidimensional schema. It is a semi-automatic method that consists of three main steps:

  - The first demand-driven step: consists of collecting user needs through interviews and by using the Goal/Question/Metrics paradigm (GQM). GQM consists of a set of forms and guidelines developed in four phases: (i) a first phase aims to formulate the goals in abstract terms, (ii) a second phase to identify the goals through interviews, (iii) a phase of integration and reduction of the number of goals identified by grouping those with similarities and finally, (iv) a more detailed analysis and description of each goal. Then, the authors present guidelines for deriving a logical multidimensional schema from requirements.
  - The second supply-driven step: conducts a comprehensive analysis of the ER diagrams of the data sources and produces graphs that can eventually give rise to star schemas. These graphs are created as follows:
    + Potential fact entities are marked according to the number of its additive attributes. Each identified fact is taken as the central node of a graph.
    + The dimensions are identified by means of many-to-one and one-to-one relationships from the central node. In addition, many-to-many relationships are transformed into one-to-many relationships.

  An algorithm is presented at the end to derive the snowflake schemas of each graph. These schemas are then transformed into star schemas by flattening the dimension hierarchies (denormalized dimensions).

  - The third step aims to integrate and conciliate the two paradigms and generate a feasible solution that best reflects the user's needs. It allows you to map demand-driven schemas to supply-driven schemas in three main steps:
    + Terminology analysis: Before integration, both demand-driven and supply-driven schemas must be converted into a common terminology language.
    + Schema matching: supply-driven schemas are compared with demand-driven schemas. A match occurs if both schemas have the same fact. Some metrics on the number of measures and dimensions are calculated.
    + Ranking and selection: supply-driven schemas are ranked according to the metrics calculated in the previous step and presented to the user;
- **Giorgini, Rizzi & Garzetti, 2008** [6]: present a hybrid method that consists first of all in gathering multidimensional requirements and then in mapping them on the data sources during a conciliation process. The method can also be considered purely demand-driven if the user does not wish to consider the data sources. According to the authors, it is important to design the organization setting in which the data warehouse will operate (organization modeling) and to capture the functional and non-functional requirements of the data warehouse (decisional modeling). If the method follows a hybrid paradigm, the next step would be to match the requirements with the ER diagrams or the relational diagrams describing the operational sources. This mapping phase consists of three steps:

  - Requirements mapping: the facts, dimensions and measures identified during the requirements analysis are mapped on the data sources. Depending on the type of data sources, the authors present a set of tips for

mapping each concept. For example, facts are mapped on entities or n-ary associations in ER diagrams and on relations in relational diagrams.
- Hierarchy construction: For each identified fact, data sources are analyzed in order to search for functional dependencies based on the algorithm discussed in [7].
- Refinement: this step aims to reorganize the fact shcema in order to better meet the user's needs. During this process, it is possible to distinguish between available concepts (mapped from requirements), unavailable concepts (requested in the requirements but cannot be mapped to the data sources) and what is available and not necessary. The authors propose to use this information to reorganize the dimensions (grafting and pruning of aggregation hierarchies) and/or to try to find new directions for analysis;

- **Mazón, Trujillo & Lechtenbörger, 2007** [8]: present a semi-automatic method that first allows to obtain a conceptual schema from the users' needs, then to verify and correct this schema in comparison to relational data sources using QVT (Query / View / Transformation). In the first demand-driven phase, users have to state their requirements by means of business goals. These goals will be used to derive information requirements and both must be modeled by an adaptation of the i* framework. At the end of this phase, a multidimensional schema must be derived from this formalization and must be modeled by using a UML extension (UML profile) proposed by the authors. For the second phase of verification, the authors propose five QVT relations based on multidimensional normal forms (MNF) to align the conceptual schema derived from the requirements with the relational schema of the data sources. These relations are presented as follows:

  - 1MNF (a): a functional dependency in the conceptual schema must have a corresponding functional dependency in the relational schema.
  - 1MNF (b): Functional dependencies among dimension levels contained in the source databases must be represented as aggregation relationships in the conceptual schema. Therefore, they complement the conceptual schema with additional aggregation hierarchies contained in the sources.
  - 1MNF (c): additive measures derived from other measures must be identified in the conceptual schema.
  - 1MNF (d): measures must be assigned to facts so that the atomic levels of the fact form a key. In other words, the authors require to place the measure in a fact while preserving the appropriate granularity of the data.
  - 2MNF and 3MNF: these two forms require the use of concept specializations when structural NULL values in data sources do not guarantee the completeness of the information.

Recently, the authors have improved the demonstration phase in [9] [10] [11] by proposing two new approaches to detecte facts and multidimensional metadata in order to further exploit data source schemas;

- **Romero & Abelló, 2010** [12]: propose a hybrid method for deriving multidimensional conceptual schemas from the needs expressed in SQL queries. it is a fully automatic method that was introduced for the first time in a previous paper [13]. Unlike the methods mentioned above, it performs both phases (data-driven and demand-driven) simultaneously and at the same time. In this manner, each paradigm benefits from the feedback of the other. This method allows to produce constellation schemas from requirements expressed as SQL queries and the logical schema of relational data sources even when the latter are denormalized. The construction of the multidimensional schema is done via two different steps:

  - The first step extracts the multidimensional knowledge contained in the SQL query (i.e. the multidimensional role played by each concept in the query as well as the conceptual relationships between concepts), that is properly stored in a graph. At this point, the logical schema of the data sources will play a crucial role in inferring the conceptual relationships among the concepts.
  - The second step validates each multidimensional graph according to a set of constraints defined by the authors. These constraints must be respected in order to place the data in a multidimensional space and produce a data cube without summarizability problems. If the validation process fails, the method ends because the requested data could not be analyzed from a multidimensional point of view. Otherwise, a multidimensional schema is directly derived from the multidimensional graph.

Unlike data-driven methods, this method focuses on data that interests the end user by considering his or her needs expressed in SQL queries. At the same time, and unlike requirement-driven methods, it is able to offer new interesting multidimensional knowledge that is ignored by the user by analyzing only concepts that are closely related to the user's needs. Finally, solid and significant multidimensional schemas are proposed at the end of the validation process;

The analysis of these different methods led to distinguish two types of hybrid design:

- Fully hybrid methods: these are methods that follow a fully hybrid process. The process steps can only be interpreted in the whole process which follows a hybrid approach;
- Compound hybrid methods: These methods consist of a data-driven phase and a demand-driven phase, and each of these two phases results in a multidimensional model. In other words, it is the fusion between a data-driven and a demand-driven method that gives these methods a hybrid character. This type can also be divided into two subtypes, namely:

- Sequential hybrid methods: in this subtype, the two methods are executed sequentially. The first method results in a multidimensional model that is an input to the second method. In this way, the second method uses the results of the first one to produce a more complete

and comprehensive multidimensional model which consider both data sources and user requirements.

- Parallel hybrid methods: allow both methods to be executed independently and at the same time. Each of the two methods results in a multidimensional model and it is only at the end of the process that these two models are merged to obtain a final result.

In addition to this identified typology, and in order to better analyze these different methods, the authors used the following criteria: operating mode (guide - semi-automatic - automatic) / input data / tool / cost / consumed time. Table I lists the methods already mentioned and analyses them using these criteria.

According to the comparative table, five hybrid methods between six are composed. However, and despite their diversity, none of these composite methods offers a reasonable time of realization as that offered by the Romero & Abello method. In addition, two of them (the Cabibbo & Torlone and Böehnlein & Ulbrich-vom Ende methods) are presented only as a simple implementation guide and the other three (Bonifati et al. / Giorgini, Rizzi & Garzetti / Mazón, Trujillo & Lechtenbörger) are semi-automatic unlike the Romero and Abello method which is completely automatic. It should also be noted that in these composite methods, only the method of Bonifati et al. which is parallel while the other four are sequential. This choice of the research community is largely due to a major disadvantage of the parallel approach. This disadvantage is the low exploitation of the results of the data-driven phase and the demand-driven phase since the two phases are executed independently.

TABLE. I. COMPARATIVE TABLE OF THE DIFFERENT HYBRID METHODS

Hybrid methods	**Type**	**Operating mode**	**Input data**	**Tool**	**Time consumption**
Cabibbo & Torlone, 1998	Compound hybrid: sequential	Guide	ER	No	Very high
Böehnlein & Ulbrich-vom Ende, 1999	Compound hybrid: sequential	Guide	SER	No	Very high
Bonifati et al., 2001	Compound hybrid: parallel	Semi-automatic	ER	No	Very high
Giorgini, Rizzi & Garzetti, 2008	Compound hybrid: sequential	Semi-automatic	ER Relational	Yes	Very high
Mazón, Trujillo & Lechtenbörger, 2007	Compound hybrid: sequential	Semi-automatic	Relational	Yes	Very high
Romero & Abelló, 2010a	Fully hybrid	automatic	Relational	No	Average

It is clear that the Romero and Abello method is the best hybrid method for designing multidimensional schemas. However, this method has a major disadvantage since its use requires either an intervention by the developer who is supposed to understand the users' needs and formalize them in the form of requests, or that users of the BI system have some knowledge of databases and query languages and in the latter case the method cannot be used by all users.

In order to overcome these gaps, the authors tried through this paper to present a new hybrid method of the sequential compound type that allows to offer relevant results in a reasonable time and to collect and analyze requirements expressed in a natural language that can be used by all users and without the intervention of any developer or designer. To carry out this new method, the authors used their previous works [1], [2], [14]–[17] by combining the data-driven method X-ETL with the requirements-driven method XCube Assist.

III. METHODOLOGY

A. X-ETL Method

The X-ETL method is a data-driven method that transforms a relational model into a multidimensional model in a completely automatic way. It is based on 3 main components.

- Relational meta-model: to which the input source model must be conform. This metamodel is composed of three main elements: a database containing tables and these tables contain columns;
- Multidimensional meta-model: to which the output model must be conform. It is a meta-model describing the elements and relationships in a star model. Thus, it is composed of a single fact table containing one or more measures and related at least to 2 or more dimensions;
- Transformation engine.

The transformation engine is the core of the X-ETL method since it is the engine that transforms the relational model into a multidimensional model. This engine is based on a set of rules for detecting facts and dimensions tables. The most important of these rules states that the relationship between a fact table and a dimension table is many-to-one and can never be one-to-many or many-to-many. Thus, the engine first detects the fact tables by calculating the number of foreign keys in each table and retaining those containing the highest number. Then, the program uses the cardinalities to detect the dimensions that are directly related to the fact table. Finally, the program identifies the dimensions indirectly related to the fact table by using the principle of the transitivity of cardinalities. A star model validated by the multidimensional meta-model is generated at the end of the process. Fig. 1 represents the Framework of the X-ETL method:

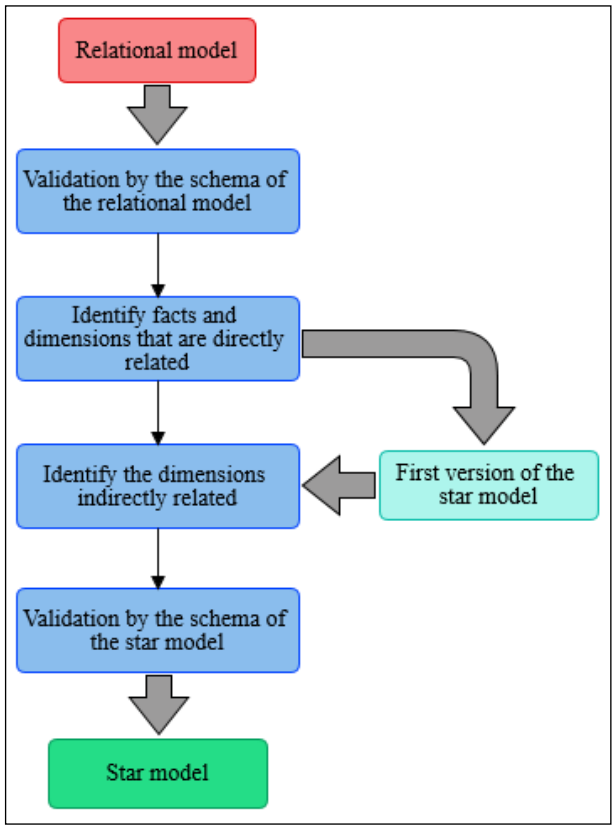

Fig. 1. Framework of X-ETL Method.

In order to evaluate the X-ETL method, the authors have used two relational models from the literature with their multidimensional reference models. The first example represents a sales activity and the second an activity in the field of medicine. The choice of two different domains allowed to test the validity of the method for different fields of application. The comparison between the results obtained with the X-ETL method and the multidimensional reference models revealed a similarity of the fact tables and the majority of the dimensional tables with some differences in the measures and the attributes hierarchy. According to this comparison, and since the appropriate choice of fact table and dimensions guarantees the availability of the necessary data for a possible aggregation of the data and modification of the attributes hierarchy, it can be said that the X-ETL method allows to get satisfactory results.

However, and like any data-driven method, the X-ETL method has some disadvantages. The first of these disadvantages lies in the nature of the method, which is based only on an input model to perform all the next steps and generate a multidimensional model at the end of the process. Therefore, the success of the whole process and the accuracy of the resulting multidimensional models depend largely on the quality of the input model. If the latter one is poorly designed, there is a high risk that the program will retain the wrong fact tables and thus generate a multidimensional model that makes no sense. The second disadvantage is the risk of retaining unnecessary dimensions since the program considers only the constraint of modeling to identify these dimensions. Therefore, if the program offers the advantage of generating multidimensional models with all possible dimensions, there is a risk that these models may be overloaded and generate an over-information of the user of the decision-making system.

The disadvantages of the X-ETL method are mainly due to the approach followed, which is limited to data analysis and excludes the users' and decision-makers' needs. Thus, and in order to overcome these disadvantages and obtain more effective results, it is necessary to integrate a demand-driven method that will allow the multidimensional models generated to be further refined.

B. XCube Assist Method

It is a semi-automatic method that allows to generate multidimensional models from the users' needs. This method is mainly based on the semantic analysis of decision questions expressed in natural language and a comparison with the search history and the source model. As shown in Fig. 2, the history file structure contains a set of reference questions that have been retained by the system with answers in the form of pre-defined metrics and multidimensional models.

As a result, semantic comparison with search history helps to guide the user in his decision-making search by offering similar reference questions, metrics and corresponding multidimensional models. If no correspondence has been detected or the proposed metrics do not answer the user's question, the system retains the new question or metric and moves on to a semantic comparison between the user's search and the table and field names of the source relational model. This semantic comparison helps to detect semantic relationships between terms searched by the user and terms in the components of the relational model. So, the table that has a semantic relationship, especially of the synonymy type, with the strategic axis defined by the user is retained as a fact table and the tables that have a relationship with the analysis dimensions are retained as dimensions. At the end of the process, the XCube Assist program generates a star model for the user and adds it to the reference models for possible use in future research.

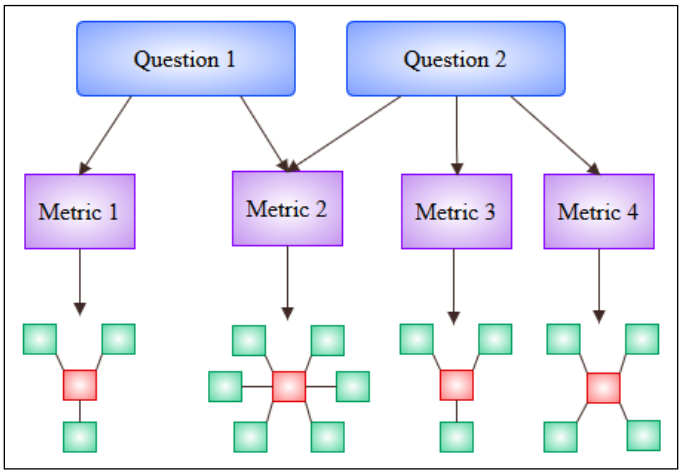

Fig. 2. The Structure of the History Files.

The test of the XCube Assist method on an example from the literature and the comparison of the results obtained with the multidimensional model provided, revealed that the program always retains the right fact table even when the user expresses his need with expressions different from the table and fields names of the relational model, and this is due to the semantic analysis performed previously. The most important thing is that the user must know exactly what he is looking for and identify his need regardless of the jargon and the language used to express it. The comparison also showed that the method does not generate a complete multidimensional model but only the part that meets the user's needs. However, the recursivity of the method allows, not only to feed the data warehouse with new multidimensional models, but also to improve the quality of the multidimensional models already generated and thus to ensure that these models evolve according to the new needs expressed.

While the XCube Assist method has the advantage of verifying the availability of data in the source model, it has the disadvantage of not considering the constraints of the cardinalities of the relational model. Consequently, the resulting multidimensional model may contain dimensions that cannot be related to the fact table because of a relational constraint and their retention requires deep transformations in the source model. It is true that the definition of this type of dimension can be useful for designers in order to carry out these transformations. However, the non-checking of the constraint of the cardinalities results in retaining this type of dimension in the same way as the other dimensions and without any particular indication that allows designers to distinguish them from the other dimensions.

Therefore, the completeness of the models generated by the method depends on the complementarity of the needs expressed and the assimilation by users of all the multidimensional possibilities that data sources can offer. In addition, users may sometimes have latent needs that need to be stimulated to express them.

According to this analysis, it is clear that the XCube Assist method can lead to more complete and efficient results if it is combined with a data-driven method that will integrate data constraints into the process while exposing to the user all the decision-making potential hidden in the data sources.

C. The Hybrid Method HX-ETL

The new hybrid method consists of combining the X-ETL method with the XCube Assist method. As shown in Fig. 3, it is sequential method composed of two successive steps:

- A first data-driven step: which consists of executing the X-ETL method on a relational source model to get star models. This method is only based on modeling constraints to retain the components of the star schema.
- A second demand-driven step: this step is based on the XCube Assist method to identify users' needs and conciliate them with the results of the previous step in order to generate valid multidimensional models that meet users' needs.

The conciliation stage is not limited to a simple mapping of needs on the results of the X-ETL method. Rather, it allows an exchange of information between the second step and the results of the first step. This interactivity in the process not only allows the verification of the modeling constraints of the requested multidimensional model but also to guide users in their research and stimulate their latent needs by proposing new dimensions. Indeed, the semantic analysis of the user-defined need and its comparison with each fact table of the star models give rise to a percentage of correspondence. The fact table that has the highest match rate while exceeding the 50% threshold is the one that best meets the user's need. This fact table is therefore retained in the final multidimensional schema and the dimensions to which it is related in the star schema will be proposed to the user even when these analytical dimensions have not been expressed in his need. In this way, the system will help the user to enrich his multidimensional model with new dimensions by exposing to him all the dimensional potential offered by the data. Fig. 4 shows the BPMN diagram that illustrates the interactivity between the two stages of the HX-ETL method.

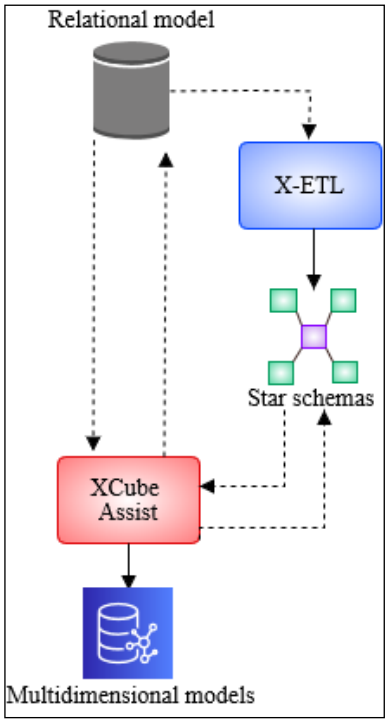

Fig. 3. General Schema of the New Hybrid Method.

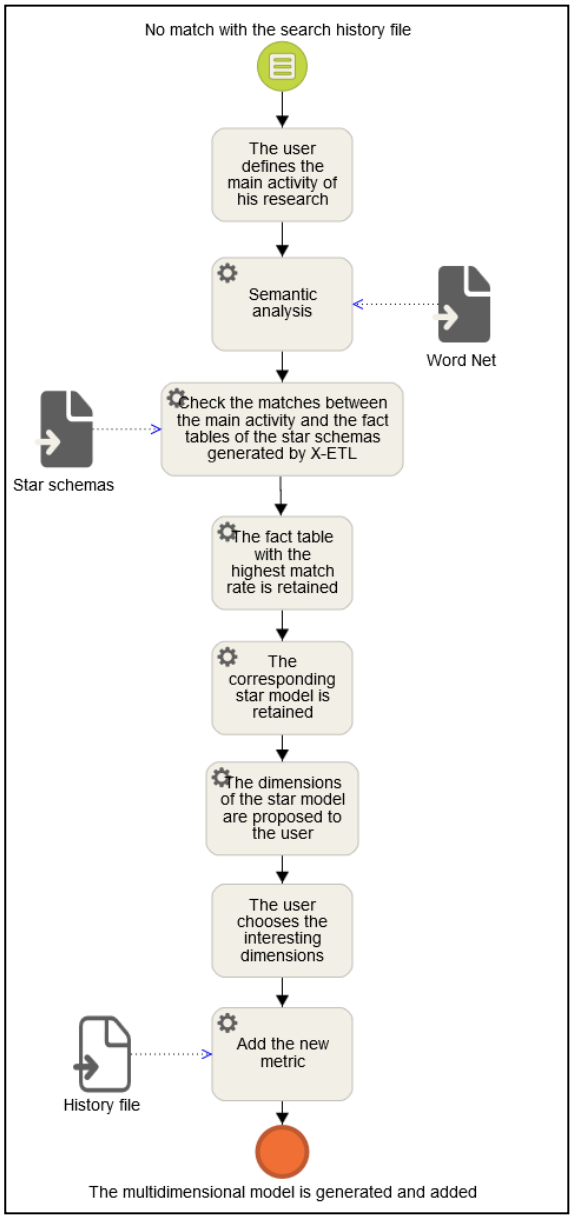

Fig. 4. BPMN Schema of the First Semantic Comparison in the HX-ETL Method

The algorithm below represents the process of the first semantic comparison between the defined need and the results of the X-ETL method.

1 **BEGIN**
2 Retain the main activity of the research defined by the user
3 Perform Semantic analysis basing on the WordNet library
4 Check matches between the main activity and the fact tables of the star schema generated by X-ETL
5 Retain the table with the highest match rate as fact table
6 Retain the corresponding star model
7 Propose the dimensions of the star model to the user
8 Retain the interesting dimensions chosen by the user
9 Add the new metric to the history file
10 Generate and add the multidimensional model
11 **END**

Since the star models do not represent all the data of the source model, a second comparison between the expressed need and the tables of this model is performed. This comparison identifies the fact table and dimensions in the source model in case the program does not find a match in the results of the X-ETL method. This phase is of crucial importance as it allows designers to be informed of multidimensional components that are requested by users but that have not been retained by the X-ETL method because of a modelling constraint. In this way, designers will then be able to introduce modifications and readjustments in the source relational model to make it more appropriate for the decision-making system and more flexible for multidimensional modeling. Fig. 5 shows the BPMN schema of this second semantic comparison.

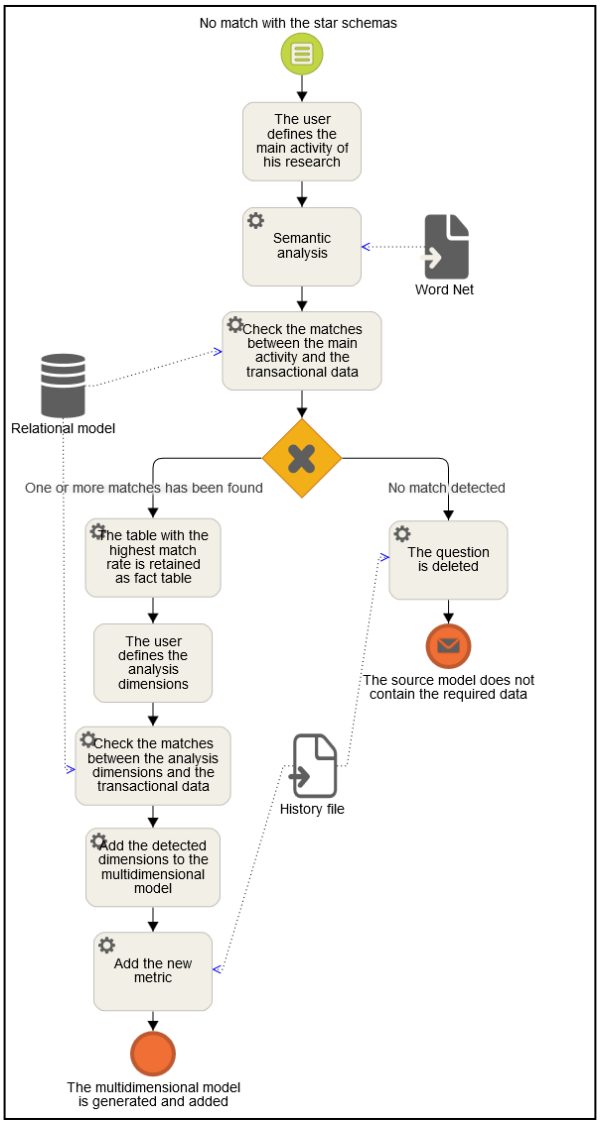

Fig. 5. BPMN Schema of the Second Semantic Comparison in the HX-ETL Method.

Below is the algorithm that describes the process of the second semantic comparison in the new hybrid method.

```
BEGIN
    Retain the main activity of the research defined by the user
    Perform Semantic analysis basing on the WordNet library
    Check matches between the main activity and the transactional Data basing on the relational model
    IF one or more matches has been found THEN
        Retain the table with the highest match rate as fact table
        Define the analysis dimensions by the user
        Check the matches between the analysis dimensions and the transactional data
        Add the detected dimensions to the multidimensional model
        Add the new metric to the history file
        Generate and add the multidimensional model
    ELSE
        Delete the question from the history file
        Return message explaining that the source model does not contain the required data
    ENDIF
END
```

Fig. 6 represents the process of the data-driven step as well as the requirements-driven step and the conciliation phase between these two steps.

The multidimensional model generated at the end of the process contains 3 types of dimensions:

- Dimensions retained from the star schema: they are the dimensions that are retained during the first semantic comparison. These dimensions are valid since they are requested by users and at the same time meet modeling constraints. These dimensions are labelled in the final multidimensional model by the expression < dimension type = valid >;
- Dimensions retained from the transactional model: These dimensions are retained at the end of the second semantic comparison between the user's need and the transactional model. These dimensions do not meet the modeling constraints and are therefore labelled by the expression < dimension type = invalid >;
- Non-existent dimensions: This third type represents the dimensions that are requested by the users but that do not exist anywhere. They are marked in the final result by the expression < dimension type = nonexistent >.

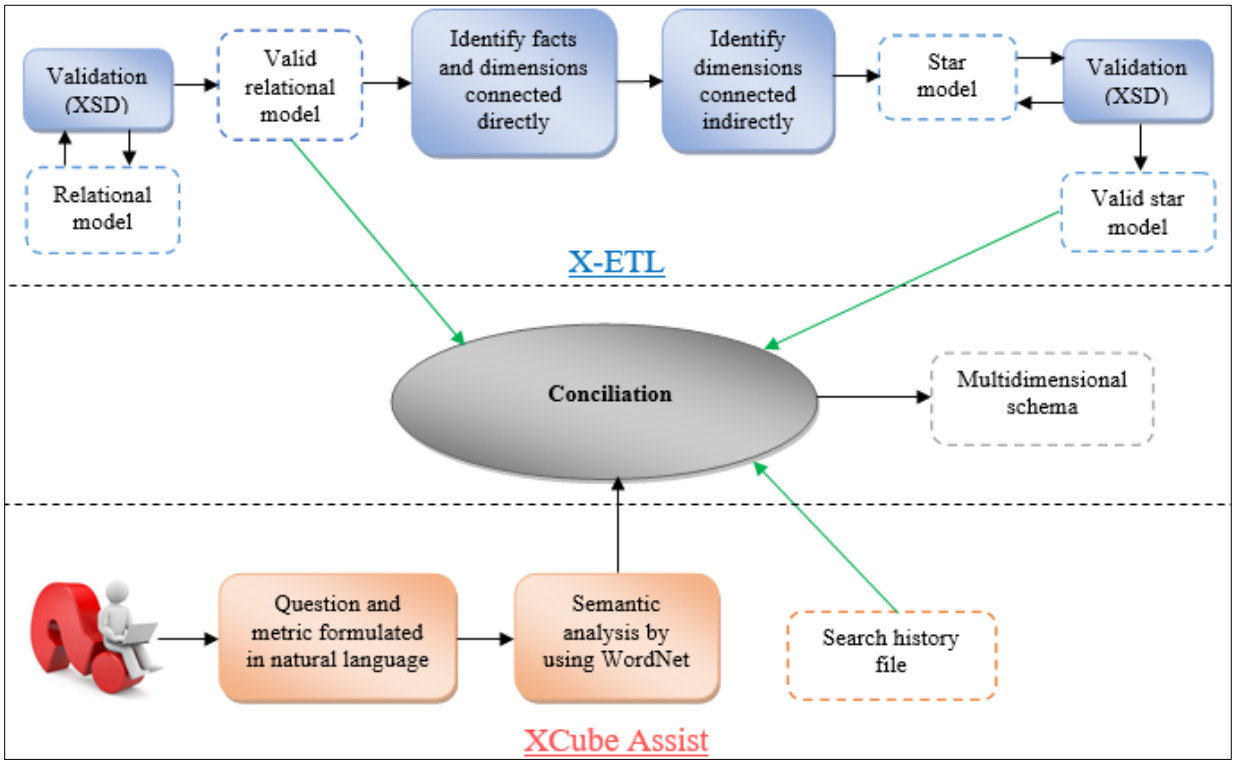

Fig. 6. Detailed Schema of the New Hybrid Method.

IV. Empirical Case Study

In order to test and evaluate the HX-ETL method, the authors used the "sales" example shown in Fig. 7 and previously used in the X-ETL method and the XCube Assist method. The use of the same example will help to verify and show the improvements that will be made by the hybrid method compared to the other two methods.

The first step of the hybrid method will give rise to the results of the X-ETL method. They are the three star models shown in the Table II.

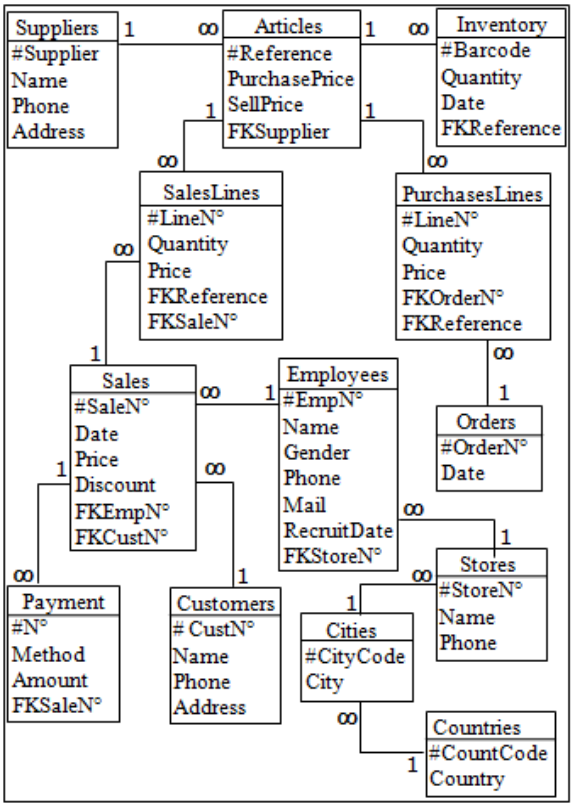

Fig. 7. Relational Model of the Sales Example.

TABLE. II. The Star Models Generated by the First Step of the Hybrid Method

	Fact tables	Dimension tables
1st star model	SalesLines	Articles – Customers – Suppliers – Sales – Stores – Cities – Countries – Employees – Time.
2nd star model	Sales	Employees – Customers – Stores – Cities – Countries – Time.
3rd star model	PurchasesLines	Orders – Articles – Suppliers – Time.

In the second requirement-driven step, assuming that the user asks the following question: **Are our products sold enough?** Once the question is submitted, the system will perform a semantic analysis to look for possible matches with the reference questions stored in the history file. If a match is found it will be proposed to the user. Assuming now that the system does not detect any correspondence and retains the new question. In this case, the user will have to determine the metric that will be able to answer his question by first defining the strategic axis of his research. Assuming, for example, that the user defines the term "Sales" as a strategic axis. The program will perform a semantic analysis of this term and make a comparison with the fact tables of the star models generated at the end of the first step. In this example, the system will retain the "Sales" fact table since it has the highest match rate. Then, the system will extract the corresponding star model (2nd model) and propose all its dimensions to the user: Employees, Customers, Stores, Cities, Countries and Time. Assuming, for example, that the user selects all the proposed dimensions and requests two more dimensions: Products and Categories. In this case, the program will move on to the second semantic comparison with the transactional model, which will result in a single correspondence between Products and Articles. This last table will be proposed to the user who will have the choice to select it or not. At the end of the process, the multidimensional model shown in Fig. 8 will be generated as an XML file.

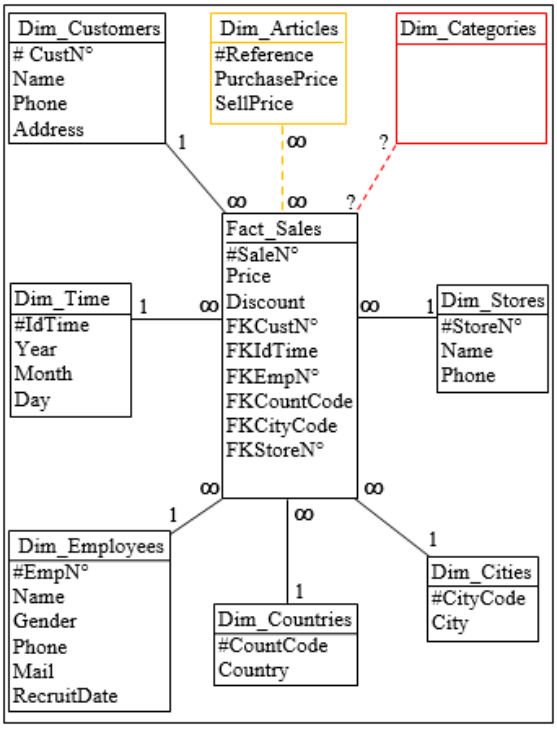

Fig. 8. Multidimensional Model Generated by the HX-ETL Method.

V. Analysis and Criticism

Compared to the other hybrid methods cited in the second section, and according to the results obtained in the empirical case study, it is possible to state that the new method is the first formal method that allows both to obtain satisfactory results in a reasonable time while being based on needs expressed in natural language and data sources. However, the new method has some limitations:

- The method does not allow to generate hierarchies in dimensions;
- The measures are always selected manually in the fact table (X-ETL);
- The method does not allow to automatically merge two or more tables to obtain a single dimension. In the multidimensional model in Fig. 8, the tables "Dim_Stores", "Dim_Cities" and "Dim_Countries" can be merged into a single table: "Dim_Place".

Therefore, the research perspectives will be to introduce improvements in the HX-ETL method to overcome these limitations while automating the method process as much as possible.

VI. Conclusion

In general, the hybrid approach considers both the analysis needs and the data for the construction of the schema. Nowadays, this approach is the one that is the subject of more investigation. The general idea is to build candidate schemas from the data (bottom-up approach) and compare them with the schemas defined according to users' requirements (top-down approach). Thus, the constructed schema is a response to real analysis needs and it is also possible to implement it with data sources. However, there are several design logics in this approach: there are fully hybrid methods, sequential compound hybrid methods and parallel compound hybrid methods.

The authors tried through this paper to present a new hybrid method of the sequential compound type by combining their data-driven method called X-ETL with the requirements-driven method XCube Assist. This combination enabled to collect and analyze needs expressed in natural language while obtaining relevant results in a reasonable time. The new method also allows to demonstrate the shortcomings and deficiencies of the source model for eventual improvements by identifying in the final results the dimensions that cannot be modelled and the dimensions that are solicited and non-existent.

References

[1] N. El Moukhi, I. El Azami, A. Mouloudi, and A. Elmounadi, “X-ETL: A Data-Driven Approach for Designing Star Schemas,” Int. J. Recent Contrib. Eng. Sci. IT, vol. 7, no. 1, pp. 4–21, Mar. 2019.

[2] N. El Moukhi, I. El Azami, A. Mouloudi, and A. ElMounadi, “Requirements-based approach for multidimensional design,” Procedia Comput. Sci., vol. 148, pp. 333–342, 2019.

[3] L. Cabibbo and R. Torlone, “A Logical Approach to Multidimensional Databases,” in Proceedings of the 6th International Conference on Extending Database Technology: Advances in Database Technology, Berlin, Heidelberg, 1998, pp. 183–197.

[4] M. Boehnlein and A. Ulbrich-vom Ende, “Deriving initial data warehouse structures from the conceptual data models of the underlying operational information systems,” in Proceedings of the 2nd ACM international workshop on Data warehousing and OLAP-DOLAP ’99, Kansas City, Missouri, United States, 1999, pp. 15–21.

[5] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, and S. Paraboschi, “Designing data marts for data warehouses,” ACM Trans. Softw. Eng. Methodol., vol. 10, no. 4, pp. 452–483, Oct. 2001.

[6] P. Giorgini, S. Rizzi, and M. Garzetti, “GRAnD: A goal-oriented approach to requirement analysis in data warehouses,” Decis. Support Syst., vol. 45, no. 1, pp. 4–21, Apr. 2008.

[7] M. Golfarelli and S. Rizzi, Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill Education, 2009.

[8] J.-N. Mazón, J. Trujillo, and J. Lechtenbörger, “Reconciling requirement-driven data warehouses with data sources via multidimensional normal forms,” Data Knowl. Eng., vol. 63, no. 3, pp. 725–751, Dec. 2007.

[9] J. Pardillo, J.-N. Mazón, and J. Trujillo, “Model-Driven Metadata for OLAP Cubes from the Conceptual Modelling of Data Warehouses,” in Data Warehousing and Knowledge Discovery, Berlin, Heidelberg, 2008, pp. 13–22.

[10] J.-N. Mazón and J. Trujillo, “A hybrid model driven development framework for the multidimensional modeling of data warehouses!,” ACM SIGMOD Rec., vol. 38, no. 2, p. 12, Oct. 2009.

[11] A. Carmè, J.-N. Mazón, and S. Rizzi, “A Model-Driven Heuristic Approach for Detecting Multidimensional Facts in Relational Data Sources,” in Data Warehousing and Knowledge Discovery, Berlin, Heidelberg, 2010, pp. 13–24.

[12] O. Romero and A. Abelló, “Automatic validation of requirements to support multidimensional design,” Data Knowl. Eng., vol. 69, no. 9, pp. 917–942, Sep. 2010.

[13] O. Romero and A. Abelló, “Multidimensional Design by Examples,” in Data Warehousing and Knowledge Discovery, Berlin, Heidelberg, 2006, pp. 85–94.

[14] N. El Moukhi, I. El Azami, A. Mouloudi, and A. El Mounadi, “Requirements-driven modeling for decision-making systems,” in 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, 2018, pp. 1–7.

[15] N. El Moukhi, I. El Azami, A. Mouloudi, and A. El Mounadi, “Towards a new automatic data warehouse design method,” Electron. J. Inf. Technol., no. 11, pp. 1–22, 2018.

[16] N. El Moukhi, I. El Azami, and A. Mouloudi, “Towards a new method for designing multidimensional models,” Int. J. Bus. Inf. Syst., vol. 28, no. 1, pp. 18–41, 2018.

[17] N. El Moukhi, I. El Azami, and A. Mouloudi, “X-ETL: A new method for designing multidimensional models,” in 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), Rabat, 2017, pp. 1–6.