Using Perspective Schemata to Model the ETL Process

Valéria M. Pequeno, João Carlos G. M. Pires

Abstract—Data Warehouses (DWs) are repositories which contain the unified history of an enterprise for decision support. The data must be extracted from information sources, transformed and integrated into a staging area, then loaded (ETL) into the DW, using ETL tools. These tools focus on data movement, where the models are only used as a means to this aim. Under a conceptual viewpoint, the authors want to innovate the ETL process in two ways: 1) to make clear compatibility between models in a declarative fashion, using correspondence assertions and 2) to identify the instances of different sources that represent the same entity in the real-world.

This paper presents the overview of the proposed framework to model the ETL process, which is based on the use of a reference model and perspective schemata. This approach provides the designer with a better understanding of the semantic associated with the ETL process.

Keywords—conceptual data model, correspondence assertions, data warehouse, data integration, ETL process, object relational database.

I. INTRODUCTION

NOWADAYS enterprises have an increasing need to take decisions as fast and accurate as possible. The competition is very high and who holds reliable and strategic information in the right moment has the advantage. One way to achieve such a goal is by making use of data warehousing, which provides processes and technologies to build integrated data repositories called Data Warehouses (DWs). The data in those repositories represents the unified history of an enterprise at a suitable level of detail to be useful for analysis [1]. The data warehouse will act as a source for a wide range of applications such as OLAP (On-Line Analytical Processing), data mining and other advanced analysis techniques. A simplified architecture of a DW system is presented in Fig. 1.

![Simplified data warehouse architecture](image)

In architecture presented in Fig. 1, the data must be extracted from different information sources, transformed and integrated to be loaded (ETL) into the DW. DW data is then delivered to Data Marts (DM), probably with some more changes. DMs are subsets of DW data designed to serve specific demands of a particular group of users. Moreover, either a DW or a DM should be created and accessed through metadata that provides detailed documentation for data in the DW system, such as applied transformations and origin of data.

The ETL process is the major stage in any DW system. It deals with the complexity inherent in the multiplicity of autonomous and heterogeneous data sources, such as where the data is and which data represents the same concept. Although important, the ETL tasks are made in an ad hoc fashion, most of them by different automatic (or semi-automatic) tools while others are manual. Some of these tools were developed in the academic world (e.g. ARKTOS tool and IBIS tool), and others in the market (e.g. Sunopsis and BO) – see [2] for a survey. However, they either do not deal with conceptual modelling or they use ad hoc formalism.

Nowadays, these problems are the hardest to deal with because DWs are growing rapidly in two ways. Firstly, DW increases in data volume from 20 to 100 terabytes or even pedabytes [3]. Secondly, DW expands data models complexity (both in sources and in DW), which implies the rise of the difficulty of managing and understanding these models [4], [5]. Although there are specialized tools with graphical interface to do the mapping between the source information and the DW system, they are mostly procedural. It means that the knowledge of procedures and policies are hidden in diverse codes, which are totally dependent on experts and technicians. These tools focus strongly on data movement, as the models are only used as a means to this aim.

In an ETL process for building a DW system, it is not concerned just with mapping between schemata, but also with an effective data integration that expresses a unified view of the enterprise. In this context, it is crucial to have a conceptual reference model [6], [3], [7]. A Reference Model (RM) is an abstract framework that provides a common semantic that can be used to guide the development of other models and help with data consistency [3].

There are, mainly, two approaches to build a DW, the top-down and the bottom-up. In the top-down approach [1] a DW is first built followed by DMs. On the other hand, the bottom-up approach [8] starts with the DMs to get the DW. Nevertheless, in spite of the seeming differences, “both approaches collect data from source systems into a single data store (named data warehouse in the top-down approach and staging area in the bottom-up one), from which DMs are populated [9].”
In addition, both strategies need to keep historical information of component entities and classification entities, as well as keep information about events that occur in the business (stored in transaction entities [10]). Both approaches, earlier or later, store this information on a multidimensional form; moreover, in the top-down approach, the data integration is done using a normalized model with extensions to keep historical information. The ETL process is present in both approaches, although the execution of it differs between the two strategies. In this research, both approaches are equally applicable.

In order to consider the growing complexity of source and DW data models, it is proposed to take a declarative approach, which is based on making explicit the relationship between data sources and data warehouse taking into account the Reference Model, independently of the ETL process involved. Furthermore, the ETL process itself can use this information.

II. OVERVIEW

The present proposal offers a way to express the existing data models (source, DW, DM, RM) and the relationship between them. The approach is based on:

1) Schema language (L_s), which is used to describe the actual data models (source, DW, DM, RM). The formal framework focuses on an object-relational paradigm, which includes definitions adopted by the main concepts of object and relational models as they are widely accepted in literature – cf. [11], [12].

2) Perspective schema language (L_Ps) is used to describe perspective schemata. A perspective schema is a special kind of schema that describes a data model (part or whole) (target schema) in terms of other data models (base schemata). L_Ps mainly extends L_s with two components: Correspondence Assertions (CAs) and Matching Functions (MFs). Correspondence Assertions formally specify the relationship between schema components. Matching functions indicate when two data entities represent the same instance of the real world. L_Ps includes data transformations, such as names conversion and data type conversions. When the target schema is described in the scope of a perspective schema, instead of just referring an existing schema, the perspective schema is called view schema.

Fig. 2 illustrates the basic components of the proposed architecture and their relationships. The schemata RM, DW, DM, S1,...,Sn, S’1, S’2 represent, respectively, the reference model, the data warehouse, a data mart, the source schemata S1,...,Sn, the view schemata S’1 (a viewpoint of schema S1), and S’2 (an integrated viewpoint of schemata S2 and S1). The relationship between the RM and the other schemata is shown through the perspective schemata P_{S’1,RM},..., P_{S’n,RM,P_{RM},DW} (denoted by the solid arrows). Once the perspective schemata P_{S’1,RM},..., P_{S’n,RM,P_{RM},DW} is declared, a new perspective schema (P_{S’1,s’2,...,s’n,DW}) between the DW schema and the source schemata (designated by a dotted arrow) can be deduced.

This inferred perspective schema shows the direct relationship between the DW and its source information, and can be used to automatically materialize the ETL process. All schemata (including the perspective ones) are stored in a metadata repository.

The remainder of this paper is laid out as follows. Sections III and IV illustrate, through examples, the proposal discussed in this paper: Section III shows the languages to describe schemata and perspective schemata, and Section IV presents the process of inference. Section V briefly reviews related work on conceptual models for DWs and for ETL. The paper ends with Section VI, which points out the new features of the approach presented here and in ongoing or planned future work on this topic.

III. SCHEMA LANGUAGES

This Section introduces the languages used to describe the schemata and perspective schemata.

In the remainder of the paper, consider a simple sales scenario comprising a data source S1, a reference model RM, and a data warehouse DW. The schemata are shown in Fig. 3. S1 and RM include information about sales of products, being that in itemunitprice, the values are stored in dollars while in saleitemunitprice, they are stored in euros. DW contains historical information about customers and products, as well as summarized information regarding sales. The relation DW.saleitem stores the amount (prodsold_amount) and the quantity total (prodsold_qty) of each product (prodid, item, date) sold per customer (custid) and per sale date (datesk). The historical data is kept using three properties: startdate, enddate, and current_flag. Startdate and enddate indicate the historical range of when each tuple was current. Startdate stores the date when the tuple was inserted into the relation. Enddate stores the date when the tuple is no longer current, or the null value in case it is still current. Current_flag store true or false indicating if the tuple is the current one or not. The properties prodid, custid, and datesk are surrogate keys automatically generated by the system. Though, it is not shown in Fig. 3, all relations have keys and some of them have foreign keys.
A. Basic Schema Language

The schema language \( L_S \) is used to describe all actual schemata (source, DW, RM, and DM) of a DW system. It is general, flexible and expressive enough to allow the projection of traditional and non-traditional applications, and so enable the manipulation of complex data – cf. [4], [13], [5].

This language includes definitions adopted by the main concepts of object and relational models as they are widely accepted in literature. Thus, \( L_S \) includes the notions of type, typed value, object, tuple, property (in the relational literature, it is named attribute [14]), class, relation, signature method, method, key, foreign key, schema, state of a schema, and path expression. All these concepts are formally defined in [15] and are not explained here. However, they are close to the normal definition. Instead of the formal definition, we present an example to clarify the notation of \( L_S \):

**Example 1:** The relation PRODUCT of the schema \( S_1 \) presented in Fig. 3 has the following declaration in \( L_S \):

\[
\text{PRODUCT (prod_id}_{S1}, \text{prod_name}_{S1})
\]

B. Perspective Schema Language

The language \( L_{PS} \) is used to define perspective schemata. A perspective schema describes a data model, part or whole (target schema), in terms of other data models (base schemata). The language \( L_{PS} \) includes new concepts beyond those in language \( L_S \):

- to express the subset of the target schema that will be necessary in the perspective schema;
- to determine when two objects/tuples are distinct representations of the same object in the real-world (named the instance matching problem);
- to establish the semantic correspondence between schemata’s components.

The target schema may have much more information than it is required to represent in a perspective schema, namely when the target schema is the Reference Model. Hence, it is required to clearly indicate which elements of the target schema are in the scope of the perspective schema. This is done in \( L_{PS} \) using ‘require’ declarations. For instance, consider the perspective schema \( P_{S1:RM} \) between the schemata \( RM \) (the target schema) and \( S_1 \) (the base schema) both as presented in Fig. 3. For this perspective schema, the following relations from \( RM \) are needed:

- \( \text{require(PRODUCT, } \{ \text{pid}_{RM}, \text{pname}_{RM} \} \) \)
- \( \text{require(SALEITEM, } \{ \text{sid}_{RM}, \text{pid}_{RM}, \text{quantity}_{RM}, \text{uniprice}_{RM} \} \) \)
- \( \text{require(SALE, } \{ \text{sid}_{RM}, \text{sale_date}_{RM}, \text{cid}_{RM} \} \) \)
- \( \text{require(CUSTOMER, } \{ \text{cid}_{RM}, \text{name}_{RM}, \text{caddress}_{RM} \} \) \)

Note that, for instance, the properties: \( ptype \) from \( RM \).PRODUCT, \( \text{posted_date}_{RM} \) and \( \text{discount}_{RM} \) from \( RM \).SALE, and \( \text{region}_{RM} \) and \( \text{cphone}_{RM} \) from \( RM \).CUSTOMER are not shown as being required.

1) Matching Functions: From a conceptual viewpoint, it is essential to provide a way to identify instances of different schemata that represent the same entity in the real-world. The proposal here is to use matching functions, which can include various techniques for matching instances, including some of those used in data cleaning, such as lookup tables, user-defined functions, heuristics and past matching. These functions, as occur in [16], define a 1:1 correspondence between the objects/tuples in families of corresponding classes/relations.

In particular, this work is based on the following matching function signature:

\[
\text{match : } ((S_1[R_1], \tau_1) \times (S_2[R_2], \tau_2)) \rightarrow \text{Boolean},
\]

being \( S_1 \) schema names, \( R_i \) class/relation names, and \( \tau_i \) the data type of the instances of \( R_i \), for \( i \in \{1, 2\} \). When both arguments are instanced, \( \text{match} \) verifies whether two instances
are semantically equivalent or not. If only one argument is
instanced, e.g. $S_1, R_1$, then it obtains the semantically equi-
alent $S_2, R_2$ instance of the given $S_1, R_1$ instance, returning
true when it is possible, and false when nothing is found or
when there is more than one instance to match.

In some scenarios one-to-many correspondence between
instances are common, e.g. when historical data is stored in
the DW. In this case, a variant of match should be used, which
has the following form:

\[
\text{match} : ((S_1 [R_1], \tau_1) \times (S_2 [R_2 (\text{predicate})], \tau_2)) \rightarrow \text{Boolean}.
\] (2)

**predicate** is a boolean condition that determines the context
in which the instance matching must be applied in $S_1, R_2$.

Some examples of matching functions signatures involving
schemata of Fig. 3 are presented in Fig. 4. The implementation
of the matching functions shall be externally provided, since
their code is very close to the application domain.

2) **Correspondence Assertions**: The semantic correspondence between schemata’s components is declared
in this proposal through the Correspondence Assertions (CAs), which are used to formally assert the correspondence
between schema components in a declarative fashion. CAs are
classified in four groups: Property Correspondence Assertion (PCA), Extension Correspondence Assertion (ECA), Summation Correspondence Assertion (SCA), and Aggregation Correspondence Assertion (ACA). Examples of CAs are shown in Fig. 5 and explained in that Section.

Property CAs relate properties of a target schema to the properties of base schemata. They allow for dealing with
several kinds of semantic heterogeneity such as: naming conflict (for instance synonyms and homonyms properties),
data representation conflict (that occurs when similar contents are represented by different data types), and encoding conflict (that occurs when similar contents are represented by different formats of data or unit of measures). Furthermore, it can declare: i) Boolean conditions where the property’s value depends on the validation of one (or several) condition(s); ii) calculations where the value of a property is the result of a calculation involving two or more properties of a same instance.

For example, the PCAs $\psi_1$ and $\psi_3$ (see Fig. 5) deal with,
respectively, naming conflict and encoding conflict. $\psi_1$ links the property $\text{id}_{RM}$ of $RM.PRODUCT$ to the property $\text{prod_id}_{S_1}$ of $S_1.PRODUCT$. $\psi_3$ assigns $\text{unitprice}_{RM}$ of $S_{RM}.ITEM$ to $\text{unitprice}_{S_1}$ of $S_1$ using the function $\text{dollarToEuro}$ to convert currencies from dollars (stored in $\text{unitprice}_{S_1}$) to euros (stored in $\text{unitprice}_{RM}$).

The Extension CAs are used to describe which objects/tuples of a base schema should have a corresponding
semantically equivalent object/tuple in the target schema. For instance, the relation $RM.PRODUCT$ is linked to relation
$S_1.PRODUCT$ through the ECA $\psi_4$ presented in Fig. 5. $\psi_4$ determines that $RM.PRODUCT$ and $S_1.PRODUCT$ are equivalent, i.e., for each tuple of $PRODUCT$ of the schema $S_1$ there is one semantically equivalent tuple in $PRODUCT$ of the schema $RM$, and vice-versa.

There are five different kinds of ECAs: equivalence, selection, difference, union, and intersection. The ECA of union
is not close to union of sets, rather it indicates a relation similar
to the natural outer-join of the usual relational models.

In order to exemplify a SCA of normalization, consider the source schema $S_2$ (not presented in

\[
\text{match} : (S_1 [R_1], \tau_1) \times (RM [PRODUCT], \tau_2) \rightarrow \text{Boolean}
\]

\[
\text{match} : (RM [PRODUCT], \tau_2) \times (DW [PRODUCT (current_flag = true)], \tau_3) \rightarrow \text{Boolean}
\]

**Fig. 4. Examples of matching function signatures.**
### Property Correspondence Assertions (PCAs):

| 
| \( \psi_1: P_{RM}^{PRODUCT} \rightarrow S_1^{PRODUCT} \) |
| 
| \( \psi_2: P_{RM}^{PRODUCT} \rightarrow S_2^{PRODUCT} \) |
| 
| \( \psi_3: P_{RM}^{ITEM} \rightarrow S_1^{ITEM} \) |
| 
| \( \psi_4: P_{RM}^{ITEM} \rightarrow S_2^{ITEM} \) |

### Extension Correspondence Assertions (ECAs):

| 
| \( \psi_4: \) Count all male students whose grades are above 75 |
| 
| \( \psi_5: \) Count all male students whose grades are above 75 |

### Summation Correspondence Assertions (SCAs):

| 
| \( \psi_6: P_{RM}^{ITEM} \rightarrow S_1^{ITEM} \) |
| 
| \( \psi_7: P_{RM}^{ITEM} \rightarrow S_2^{ITEM} \) |

### Aggregation Correspondence Assertions (ACA):

| 
| \( \psi_8: P_{RM}^{PRODUCT} \rightarrow S_1^{PRODUCT} \) |
| 
| \( \psi_9: P_{RM}^{PRODUCT} \rightarrow S_2^{PRODUCT} \) |

Fig. 5. Examples of correspondence assertion.

---

**IV. INFERRING NEW PERSPECTIVE SCHEMATA**

After the perspective schemata has been defined, the next step is to infer a new perspective schema, based on all perspective schemata prior defined (see Fig. 6). This inferred perspective will connect the DW directly to its sources and can be used to automatically materialize the ETL process. For example, consider the perspective schemata \( P_{RM}^{DW} \) and \( P_{S1,RM}^{DW} \), which directly connects the DW to the source \( S_1 \). This schemata will have the same classes, relations, keys, and foreign keys as defined in original \( P_{RM}^{DW} \); and new CAs and new match function signatures will be created.

---

![Diagram](Image) Fig. 6. Sketch of the creation of an inferred perspective schema.
function signatures. Based on these rules, the system will generate all possible CAs to the new perspective as well as all match function signatures.

A pseudo-code with the essence of the process to generate a new perspective is shown in Fig. 8. In Fig. 8 \( P_7 \) is a perspective schema from the reference model to the data warehouse; \( P_7, 1 \neq j \neq n \), are perspective schemata from source schemata to the reference model; and \( P_7 \) is the inferred perspective schema from source schemata to the data warehouse. All elements of the perspective schemata are grouped in lists: \( classList, relationList, keyList, caList, \) and \( mfList \). The three first lists hold require declarations of, respectively, classes, relations, and keys and foreign keys. \( caList \) contains correspondence assertion declarations, and \( mfList \) has match function signatures.

The inference mechanism can be illustrated more clearly through the running example. For each CA, originally defined to \( P_{RM,DW} \) (relating elements of the DW to elements of the RM) will exist a new CA in \( P_{S_i,DW} \). Fig. 5 only presents some CAs of \( P_{S_i,RM} \) and of \( P_{RM,DW} \). Based on these CAs, two new CAs can be inferred: \( \psi_3 \) and \( \psi_5 \) shown in Fig. 7(a). The process to deduce \( \psi_3 \) is presented in Fig. 7(b).

For each match function signature originally defined to \( P_{RM,DW} \) will exists a new match function signature in \( P_{S_i,DW} \). Fig. 4 only presents one match function signature of \( P_{RM,DW} \). Based on this match function signature and on the ECAs presented in Fig. 5, a new match function signature can be created, shown in Fig. 9.

The inference mechanism has been developed as part of a proof-of-concept prototype using a Prolog language. Beside this module, the prototype consist of more five modules, such as the schema manager, and the ISCO translator. The schema manager module is employed by the designer to manage the schemata (in language \( L_S \)) as well as the perspective schemata (in language \( L_{PS} \)). The ISCO translator performs the mapping between schemata written in \( L_S \) or \( L_{PS} \) languages to schemata defined in a language programming called Information Systems Construction language (ISCO) [18]. ISCO is based on a contextual constraint logic programming that allows the construction of information systems. It can define (object) relational schemata, represent data, and transparently access data from various heterogeneous sources in a uniform way, like a mediator system [19]. Thus, it is possible to access data from information sources using the perspective schema in ISCO. Furthermore, once the perspective schema from source schemata to the global schema has been inferred, as well as the new match functions have been implemented, it can be translated to ISCO language and so the data of the global schema can be queried.

Another advantage of the proposed work is that a particular kind of schema evolution in a DW is transparent (when changes occur in the source schemata or new ones are added), since it is enough to automatically generate a new perspective schema from the source schemata to the data warehouse one.

V. RELATED WORK

This Section addresses related work in the fields of conceptual modelling for data warehousing and ETL.

Available literature quotes that the conceptual models for DW have focused on technical aspects such as multidimensional data models (e.g. [20], [21], [4], [22], [23], [24]) as well as the materialized view definition and maintenance (e.g. [25]). In particular, the most conceptual multidimensional models are extensions to the Entity-Relationship model (e.g. [26], [27], [28], [29]) or extensions to UML (e.g. [30], [31], [32]). There are only some works involving conceptual models based on non-multidimensional aspects [33], [25], [6].

There are various approaches in existence [34], [6], [35] for dealing with the ETL activities in a conceptual setting. So far, the authors of this paper are not aware of any research that precisely deals with mappings (structural and instance) between the sources and the DW, and with the problem of semantic heterogeneity in a whole conceptual level.
Reference in [6] presented a framework adopted in the Data Warehouse Quality project. Similar to this study, their proposal includes a reference model (cited as “enterprise model”) designed using an Enriched Entity-Relationship (EER) model. However, unlike the authors’ research, all their schemata, including the DW schema, were formed by relational structures, which were defined as views over the reference model. Their proposal provided the user with various levels of abstraction: conceptual, logical, and physical. In their conceptual level, they introduce the notion of intermodel assertions that precisely capture the structure of an EER schema or allow for the specification of the relationship between diverse schemata. However, any transformation (e.g., restructuring of schema and values) or mapping of instances is deferred for the logical level. In addition, they did not deal with complex data, integrity constraints, and path expressions, as this research does.

Reference in [34] proposed a conceptual model for dealing with inter-relationships of attributes and concepts and with the ETL activities in the early stages of a DW project. Their research included a rich graphical notation and dealt with several kinds of transformation presented in the usual ETL process, such as surrogate key transformation, checks for null values, primary key violations, aggregate values, and data conversion. However, they did not mention the matching of instances, neither did their work have any mechanisms to verify if the conceptual model was legal or not.

Reference in [35] focuses on an ontology-based approach to determine the mapping between attributes from the source schemata and the DW schema, as well as to identify the ETL transformations required for correctly moving data from source information to the DW. Their ontology, based on a common vocabulary as well as a set of data annotations (both provided by the designer), allows formal and explicit description of the semantic of the sources and the DW schemata. However, their strategy requires a deep knowledge of all schemata involved in the DW system, in what is usually not an usual task. It is dispensable in the proposal presented by the authors of the current paper as each schema (source or DW) needs to be related only to the reference model one. Additionally, in [35] there is nothing about the matching of instances and their ETL operations being a subset of transformations treated by us.

VI. CONCLUSIONS AND FUTURE WORK

This paper proposes a declarative approach to make explicit the relationship between data sources and the Data Warehouse (DW), not only at a structural level, but also at an instance level, independently of the ETL process involved. The approach considers the use of a Reference Model (RM) and can be divided in three steps:

1) to describe all source schemata, the DW schema, a Data Mart (DM) schema, and the (RM) schema, using the schema language $L_S$.

2) to describe perspective schemata to relate the schemata declared in step 1, using the perspective schema language $L_{P,S}$.

3) to infer a new perspective schema from source schemata to the DW schema based on all perspective schemata defined in the step 2.

It has been shown how the actual (DW, DM, RM, sources) and perspective schemata can be described in the proposed language and how new perspective can be deduced. This process was illustrated with some examples. An advantage to this approach is that it provides designers/users with a better understanding of the semantic associated with the ETL process. Moreover, the designers can describe the DW without concerns about where the sources are or how they are stored. It is possible because all schemata in the DW system are related to the RM through the perspective schemata, and based on the latter, the direct relationship between the DW and its sources is automatically generated. Moreover, the schema evolution, when the source schemata changes or new ones are added, becomes transparent to the DW, due to use of the reference model and of the inference process.

A prototype Prolog-based has been developed to allow the description of schemata and perspective schemata in the proposed language as well as to infer new perspective schemata based on other ones. The matching functions can be implemented using Prolog itself or external functions. In addition, the prototype include translators from the proposed language to the ISCO one. ISCO [18] allows access to heterogeneous data sources and to perform arbitrary computations. Thus, user-queries can be done, in a transparent way, to access the information sources, like occurs in mediator systems [19].

A mediator strategy in a data warehouse context is not appropriate. It is due to particular nature of the DW, which uses huge data volume and require that user-queries can be answer very quickly and efficiently. Thus, for future work, investigations will be made into how the perspective schemata can be used to automate the materialization of the ETL process. Another important direction for future work is the development of a graphical user-friendly interface to declare the schemata in the proposed language, and thus, to hide some syntax details.

REFERENCES


