# INCLUSION OF TIME-VARYING MEASURES IN TEMPORAL DATA WAREHOUSES

Elzbieta Malinowski* and E. Zimányi
*Department of Informatics & Networks,Université Libre de Bruxelles*
*50 av.F.D.Roosevelt, 1050 Brussels, Belgium*
*emalinow@ulb.ac.be, ezimanyi@ulb.ac.be*

Keywords: Temporal data warehouses, time-varying measures, data warehouse design

Abstract: Data Warehouses (DWs) integrate data from different source systems that may have temporal support. However, current DWs only allow to track changes for measures indicating the time when a specific measure value is valid. In this way, applications such as fraud detection cannot be easily implemented since they require to know the time when changes in source systems have occurred. In this work, based on the research related to Temporal Databases, we propose the inclusion of time-varying measures changing the current role of the time dimension. First, we refer to different temporal types that are allowed in our model. Then, we study different scenarios that show the usefulness of inclusion of different temporal types. Further, since measures can be aggregated before being inserted into DWs, we discuss the issues related to different time granularities between source systems and DWs and to measure aggregations.

## 1 INTRODUCTION

Data warehouses (DWs) store and provide access to large volumes of historical data supporting the decision-making process. The structure of DWs is based on a multidimensional view of data usually represented as a *star schema*, consisting of fact and dimension tables. A *fact table* contains numeric data called *measures*. *Dimensions* are used for exploring the measures from different analysis perspectives.

Current multidimensional models include an omnipresent time dimension that serves as a time-varying indicator for measures, e.g., total sales in March 2005; however, this dimension cannot be used for representing the time when changes in other dimensions have occurred, e.g., when a product has changed its ingredients. Therefore, usual multidimensional models are not symmetric in representing changes for measures and dimensions.

On the other hand, Temporal Databases (TDBs) allow to represent and manage time-varying information. Two different temporal types[2] are considered: *valid time* (VT) and *transaction time* (TT) that allow to know, respectively, when the data is true in the modeled reality and current in the database. If both temporal types are used, they define *bitemporal time* (BT). These temporal types are used for representing *events*, i.e., something that happens at a particular time point, or *states*, i.e., something that has extent over time. For the former *an instant* is used; it is represented as a non-decomposable time unit called *granule* with the size called *granularity*. A state is represented by an *interval* or *period* indicating the time between two instants.

Temporal Data Warehouses (TDWs) join the research achievements of TDBs and DWs in order to manage time-varying multidimensional data. TDWs raise several issues, e.g., consistent temporal aggregations, storage methods, etc. However, very little attention has been drawn to conceptual modeling for TDWs and to the analysis of which temporal support should be included in TDWs considering that TDBs and DWs are semantically different.

Firstly, DW data is integrated from existing source systems whereas TDB data is inserted by users. Secondly, DW data is neither modified nor deleted[3] while TDB data can be changed by users directly. Finally,

[*] The work of E. Malinowski was funded by a scholarship of the Cooperation Department of the Université Libre de Bruxelles.

[2] Usually called time dimensions; however, we use the term “dimension” in the multidimensional context.

[3] We ignore modifications due to errors during data loading and deletion for purging DW data.

DWs are designed according to users' analysis needs based on the multidimensional model where measures and dimensions play different roles. TDB design is concerned with transactional applications where all data is handled in a similar manner.

In this paper, we introduce temporal extensions for the MultiDimER model (Malinowski and Zimányi, 2005). Due to space limitations, we only refer to measures[4]. Section 2 briefly recalls the main features of the MultiDimER model and Section 3 describes temporal types allowed in the model. Further, since source systems and DWs may have different time granularities[5], e.g., source data is introduced on a daily basis yet DW data is aggregated by month, we consider two different situations: when measures are not aggregated before integration into a TDW and when these aggregations are realized. We refer to the former in Section 4 presenting several scenarios, for which different temporal types are required. Section 5 considers the latter and refers to the mapping between different time granularities and to aggregation of measures. Finally, Section 6 surveys works related to TDWs and Section 7 gives the conclusions.

## 2 OVERVIEW OF THE MULTIDIMER MODEL

In the *MultiDimER* model (Malinowski and Zimányi, 2005) a *schema* is defined as a finite set of dimensions and fact relationships. A *dimension* is an abstract concept for grouping data that shares a common semantic meaning. It represents either a level, or one or more hierarchies. Levels correspond to entity types (Figure 1 a). Every instance of a level is called a *member*.

A *hierarchy* contains several related levels (Figure 1 b). It express different structures according to the criteria used for analysis (Figure 1 c), e.g., geographical location. *Cardinalities* (Figure 1 d) indicate the minimum and the maximum numbers of members in one level that can be related to a member in another level. Given two consecutive levels of a hierarchy, the higher level is called *parent* and the lower level is called *child*. A level of a hierarchy that does not have a child level is called *leaf*.

Levels contain one or several *key attributes* (represented in bold and italic in Figure 1) and may also have other *descriptive attributes*. A key attribute of a parent level defines how child members are grouped. A key attribute in a leaf level or in a level forming a dimension without hierarchy indicates the granularity of measures in the associated fact relationship.

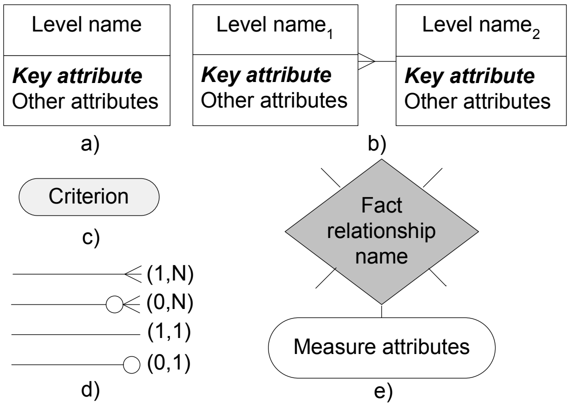

Figure 1: Notations for multidimensional model: a) one-level dimension, b) hierarchy, c) analysis criterion, d) cardinalities, and e) fact relationship.

A *fact relationship* (Figure 1 e) represents an $n$-ary relationship between leaf levels. It may contain attributes commonly called *measures*.

## 3 TEMPORAL TYPES IN TDWs

Current DWs do not offer different temporal types, thus users may have difficulties in expressing their needs for some kinds of applications, e.g., for fraud detection. In the temporal extension of the MultiDimER model we allow to include VT, TT, or bitemporal time (BT) coming from source systems and additionally, the time when data is loaded into a TDW.

The inclusion of VT for representing when the data is valid in the modeled reality is important for TDW applications since it allows to aggregate measures correctly (Eder et al., 2002).

Regarding TT, three different approaches exist: (1) ignoring TT (Body et al., 2003; Mendelzon and Vaisman, 2003), (2) transforming TT from source systems to represent VT (Martín and Abelló, 2003), or (3) considering TT generated in a TDW in the same way as TT is used in TDBs (Martín and Abelló, 2003; Koncilia, 2003), i.e., allowing to know when data was inserted, modified, or deleted from DWs. However, using the first approach traceability applications, e.g., for fraud detection, cannot be implemented. The second approach is semantically incorrect because data may be included in databases after their period of validity has expired, e.g., client's previous address. In the third approach, since TDW data is neither modified nor deleted, TT generated in a TDW represents indeed the time when data was loaded into a TDW. This time is called in our model *data warehouse load-*

[4] In (Malinowski and Zimányi, 2006) time-varying dimensions have been introduced.

[5] We consider the granularity as a time precision in which measure values are recorded.

*ing time* (DWLT).

Further, in the modeling process, application requirements determine the type of temporal support needed in each element of a TDW (attributes, levels, measures, etc.). Obviously that depends on whether or not the different data sources of the TDW provide temporal support. For example, *snapshot* systems (Jarke et al., 2003), which in order to find the changes require to compare data with the previous versions, do not offer any temporal support except VT that may be included as a user-defined attribute. On the other hand, *logged* systems (Jarke et al., 2003), which register all actions in the log files, contain TT; they also may include VT similar to snapshot systems.

# 4 TEMPORAL SUPPORT FOR NON-AGGREGATED MEASURES

In this section we refer to the case when time granularities attached to measures in source systems and in a TDW are the same, i.e., measures are not aggregated. Considering that temporal support in TDWs depends on both the availability of temporal types in source systems and the kind of required analysis, we present next examples that refer to these two aspects.

**Case 1. Sources: non-temporal, TDWs: DWLT** In real-world situations, many sources can be non-temporal or temporal support is implemented in an ad-hoc manner that can be both inefficient and difficult to automate. Nevertheless, decision-making users may require the history of how source data has evolved (Yang and Widom, 1998). Thus, the measure values can be timestamped when loaded to the TDW. An example is given in in Figure 2 representing the schema for analysis of the history of Product inventory considering different suppliers and warehouses.

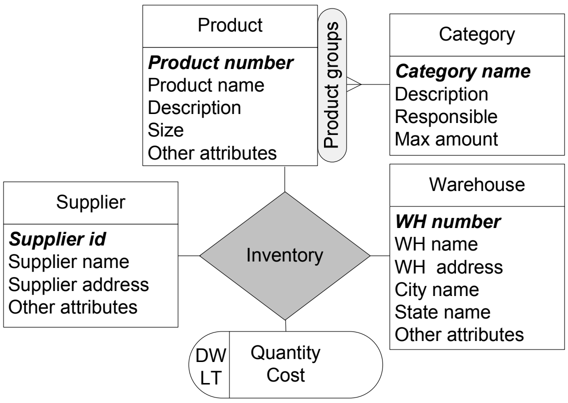

Figure 2: Inclusion of DWLT for measures.

The important question is whether it is necessary to have the time dimension in the model after including temporal types for measures. If the time dimension has only the attributes that contain a granule, this dimension is not required anymore. The additional information, e.g., if it is the week day, the last day of the month, can be obtained applying time manipulation functions. However, in some TDW applications this calculation can be very time-consuming or some data cannot be acquired at all, e.g., occurred events[6]. Thus, whether this dimension will be included depends on users' requirements and the DBMS capabilities.

**Case 2. Sources and TDWs: VT** This case occurs when source systems can offer VT, which is also required in a TDW. Figure 3 gives an example of an event model used for the analysis of banking transactions. Different types of queries can be formulated for this model. For example, analysis of clients' behavior considering the maximum or minimum withdraw, the total number of transactions during lunch hours, etc. This can help, for example, to avoid cancellation of an account or to promote some new services.

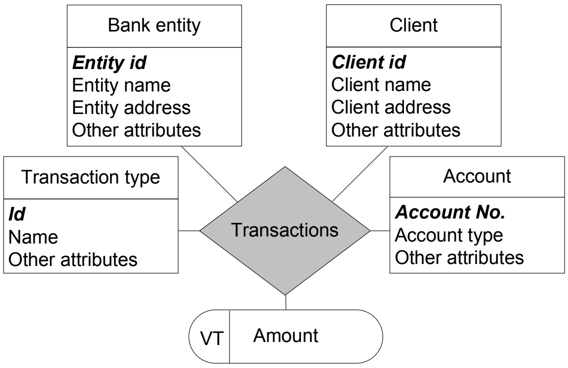

Figure 3: Inclusion of VT for measures.

**Case 3. Sources: TT, TDWs: VT** In this case, users require to know either the time when an event occurred in reality or a period of validity for data representing state. However, source systems can only offer the time when data was modified in a source system, i.e., TT. Thus, the analysis if TT can be used for approximating VT should be made. For example, if a measure represents clients' account balance, VT for this measure can be calculated considering transaction times of two consecutive operations.

Nevertheless, TT cannot always be used for calculating VT, since some data can be inserted in source systems (registering TT) when they are not valid in the modeled reality, e.g., employee's previous salary. In many applications, only the user knows VT.

[6] In Costa Rica when an event such as earthquake occurs, sales of water bottles and canned food increases.

**Case 4. Sources: VT, TDWs: VT and DWLT** In the previous two cases, we include VT in a TDW, which is the most common practice. However, the addition of DWLT can give the information since when the data has been available for the decision-making process helping to better understand decisions made in the past and to adjust loading frequencies.

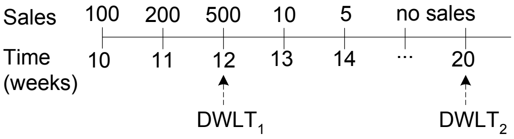

Figure 4: Example of having VT and DWLT.

For example, based on a growing tendency of product sales during weeks 10, 11 and 12 (Figure 4), it was decided to buy more products. However, only in the next DW load, occurred eight weeks later, a sudden decrease of sales has been revealed. Thus, an additional analysis can be performed to understand the causes of these changes in sales behavior. Further, the decision of more frequent loads may be taken.

**Case 5. Sources: TT, TDWs: TT (DWLT, VT)** DW data can be needed for traceability applications (e.g., for fraud detection) where changes to data and time when they have occurred should be available. That is possible if source systems have TT.

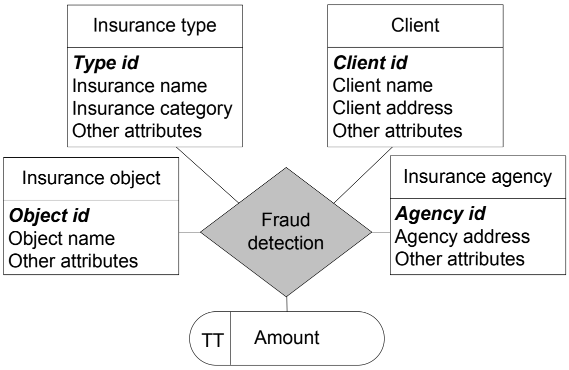

Figure 5: Example of a TDW for insurance company with TT for representing time of measure changes.

An example given in Figure 5 is used for an insurance company having as an analysis focus the amount of insurance payments. Since investigators suspect an internal fraud by modification of the amount of insurance paid to clients, the detailed information is required indicating when changes in measure values have been introduced. Further, the inclusion of DWLT would give the additional information since when data has been available for the investigation process while the inclusion of VT would allow to know when the payment was received by client. In many real systems, the combination of both, TT and VT, i.e., BT will be included.

**Case 6. Sources: BT, TDWs: BT and DWLT** TDW data should offer a timely consistent representation of information (Bruckner and Tjoa, 2002). Since some delay may occur between the time when the data is valid in the reality, when it is known in the sources, and when it is stored in the DW, it is sometimes necessary to include VT, TT and DWLT. Figure 6 shows an example of the usefulness of having these three temporal types. This example is based on the conceptual model for managing temporal consistency in active DWs (Bruckner and Tjoa, 2002).

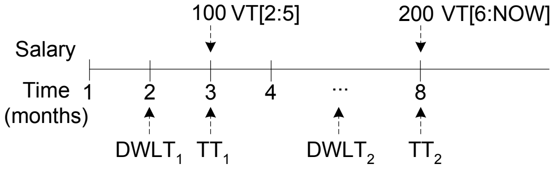

Figure 6: Example of having VT, TT, and DWLT.

In this example, a salary 100 with VT from the second to fifth months was stored at the third month ($TT_1$) in a source system. Afterwards, at the eighth month ($TT_2$) a new salary was inserted with value 200 and VT from the sixth month until NOW. When data was loaded into TDW at $DWLT_1$, the value of the salary was unknown. In the next loading $DWLT_2$ the value 100 was stored in the TDW. However, depending on which instant of time users want to analyze different values can be retrieved[7].

# 5 TEMPORAL SUPPORT FOR AGGREGATED MEASURES

In this section, we will analyze how to match different time granularities between source and TDW systems and how to aggregate measures to which these time granules are attached. We also refer to temporal types that can be used for aggregated measures in TDWs.

## 5.1 Different time granularities between source systems and TDWs

Since TDW measures can be aggregated with regard to time before loading, an adequate mapping between multiple time granularities of a source system and a TDW should be considered. Two mappings are distinguished: *regular* and *irregular* (Dyrsen, 1994). In

[7] For more details and analysis, readers can refer to (Bruckner and Tjoa, 2002).

the former, some conversion constant exists, so if one granule is represented by an integer it can be converted to another by a simple multiply or divide strategy, e.g., minutes and hours or days and weeks.

In irregular mappings, granules cannot be converted by a simple multiply or divide, e.g., month and days, since each month is formed by a different number of days. Thus, the mapping between different granules must be specified explicitly.

Further, some mappings between different granularities are not allowed (Bettini et al., 2000; Dyrsen, 1994), e.g., between weeks and months since a week can belong to two months. Nevertheless, this situation can be found in DW applications, e.g., the analysis of employees' salaries for each month having some employees with a salary received on weekly basis. We call the mapping of such granularities *forced*.

## 5.2 Aggregation of measures with VT

After considering the mapping between different time granularities, the aggregation of measure values must be realized taking into account (1) applied functions, e.g., sum, average and (2) type of measures, e.g., flow or stock (Lenz and Shoshani, 1997).

However, in some cases the procedures for measure aggregations could be complex. A simplifying example is given in Figure 7 where the time granularity in sources (month) requires a regular mapping to DW granularity (quarter). This example includes different cases: (1) the same salary is paid during several months overlapping different quarters (salary 20 and 40), (2) during a quarter different amounts of salary can be paid (quarter 2), and (3) during several months of a quarter an employee does not receive a salary (quarter 3). The required measure is average salary per quarter. For the first quarter, the average value is calculated easily. The second quarter, simple average does not work, thus the weighted mean value may be given instead. However, for the third quarter, a user should indicate how the value must be specified. In the example, we opt for giving an undefined value.

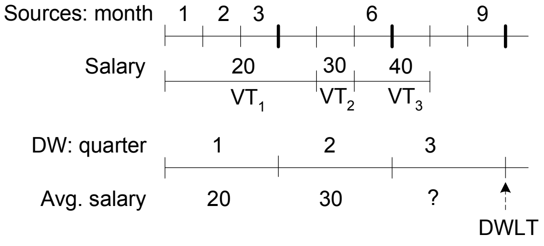

Figure 7: Example of coercion function for salary.

Real situations could be more complicated demanding clear specifications of *coercion functions* (Merlo et al., 1999) or *semantic assumptions* (Bettini et al., 2000). The idea of coercion functions is not new in the TDB research community, e.g., (Merlo et al., 1999) use them for calculating values attached to timestamps of different granularities between subtypes and supertypes or (Bettini et al., 2000) for proposing a new framework for TDBs.

It should be noted that coercion functions are always required for the forced mapping, since a finer time granule can map to more than one coarser time granule, e.g., a week to two months. Therefore, measure values to which a finer granule is attached must be distributed. For example, suppose that a salary is paid on weekly basis and this measure is stored into a TDW with a granule month. If the week belongs to two months, e.g., January and February, a user may specify that the percentage of salary that is assigned for a month is obtained from the percentage of the week contained in the month (e.g., 2 days from 7).

## 5.3 Temporal types for aggregated measures in TDWs

For aggregated measures, if source systems are nontemporal, only DWLT can be included; if TT forms part of source systems, this time will not be included in a TDW. The purpose of having TT is to analyze changes occurred to individual data, and TT for aggregated data is meaningless.

On the other hand, VT may exist in source systems for every individual measure. If measure values are aggregated regarding time, VT must be adjusted to the corresponding TDW granule. For example, the VT of the aggregated measure of salary equal 20 in Figure 7 is equal 1 (quarter 1), even though VT for this salary in a source system overlaps also quarter 2.

# 6 RELATED WORK

Most works related to TDWs include VT, e.g., (Body et al., 2003; Eder et al., 2002; Mendelzon and Vaisman, 2003). The inclusion of TT is a less common practice and in Section 3 we already discussed existing approaches. Only (Bruckner and Tjoa, 2002) discuss the inclusion of VT, TT, and DWLT for active data warehouses. However, unlike our approach, they limit the usefulness of these temporal types for only active DWs and do not offer a conceptual model that includes these types.

Very few conceptual models for TDWs have been proposed, e.g., (Body et al., 2003; Eder et al., 2002; Mendelzon and Vaisman, 2003). These models formally describe the temporal support for multidimensional models. However, they are mainly concerned about the temporal querying of data, correct aggregations, or evolutions of the multidimensional structure. None of them refer to the features discussed in this work, i.e., the inclusion of different temporal types for measures and the problem of different time granularities between source systems and TDWs.

There are many works in TDBs related to transformations from finer to coarser (or vice versa) granularities. For example, (Dyrsen, 1994) defines mappings between different granularities as explained in Section 5.1 while (Bettini et al., 2000) and (Merlo et al., 1999) refer to the problem of conversion of different time granularities and of handling data attached to these granules[8]. On the other hand, multiple time granularities for measures and dimensions are implicitly considered in (Eder et al., 2002). They mainly focus on correct measure distributions between different temporal versions of dimension members.

Even though the aspect of managing data with multiple time granularities is widely investigated in TDBs, this is still an open research in TDWs.

# 7 CONCLUSIONS

TDWs extend DWs allowing to represent time-varying multidimensional data. This extension is based on the research achievements of TDBs and should consider the semantic differences between TDBs and DWs.

Based on a conceptual multidimensional model called MultiDimER, we offer a temporal extension for levels, hierarchies, and measures, ensuring that all TDW elements are treated symmetrically. In this paper, we referred to time-varying measures.

First, we proposed the inclusion in TDWs of different temporal types. Afterwards, we referred to two different situations when the time granularity for representing TDW measures is either the same or coarser than the one in source systems. For the former, we presented several cases justifying the inclusion of TT, VT, or BT from source systems and of DWLT generated in a TDW. For the latter, we referred to existing proposals in TDBs that can be used in TDWs for transformations of different time granularities and for adequate handling of aggregations for measures. Further, we presented different temporal types that may be included for aggregated data, i.e., VT and DWLT.

The inclusion of temporal types in conceptual models allows to consider temporal semantics as an integral part of TDWs. Further, it allows to expand the analysis spectrum for decision-making users.

[8] More detailed references can be found, for example in (Bettini et al., 2000).

# REFERENCES

Bettini, C., Jajodia, S., and Wang, X. (2000). *Time Granularities in Databases, Data Mining, and Temporal Reasoning*. Springer.

Body, M., Miquel, M., Bédard, Y., and Tchounikine, A. (2003). Handling evolution in multidimensional structures. In *Proc. of the 19th Int. Conf. on Data Engineering*, pages 581–592.

Bruckner, R. and Tjoa, A. (2002). Capturing delays and valid times in data warehouses – towards timely consistent analyses. *Journal of Intelligent Information Systems*, 19(2):169–190.

Dyrsen, C. (1994). *Valid-Time Indeterminacy*. PhD thesis, University of Arizona.

Eder, J., Koncilia, C., and Morzy, T. (2002). The COMET metamodel for temporal data warehouses. In *Proc. of the 14th Int. Conf. on Advanced Information Systems Engineering*, pages 83–99.

Jarke, M., Lenzerini, M., Y.Vassiluiou, and Vassiliadis, P., editors (2003). *Fundamentals of Data Warehouse*. Springer.

Koncilia, C. (2003). A bi-temporal data warehouse model. In *Proc. of Short Papers of the 15th Int. Conf. on Advanced Information Systems Engineering*, pages 77–80.

Lenz, H. and Shoshani, A. (1997). Summarizability in OLAP and statistical databases. In *Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management*, pages 132–143.

Malinowski, E. and Zimányi, E. (2005). Hierarchies in a multidimensional model: from conceptual modeling to logical representation. Accepted for publication in Data & Knowledge Engineering.

Malinowski, E. and Zimányi, E. (2006). A conceptual solution for representing time in data warehouse dimensions. In *Proc. of the 3rd Asia-Pacific Conf. on Conceptual Modelling*. Accepted.

Martín, C. and Abelló, A. (2003). A temporal study of data sources to load a corporate data warehouse. In *Proc. of the 5th Int. Conf. on Data Warehousing and Knowledge Discovery*, pages 109–118.

Mendelzon, A. and Vaisman, A. (2003). Time in multidimensional databases. In Rafanelli, M., editor, *Multidimensional Databases: Problems and Solutions*, pages 166–199. Idea Group Publishing.

Merlo, I., Bertino, E., Ferrari, E., and Guerrini, G. (1999). A temporal object-oriented data model with multiple granularities. In *6th Int. Workshop on Temporal Representation and Reasoning*, pages 73–81.

Yang, J. and Widom, J. (1998). Mantaining temporal views over non-temporal information source for data warehousing. In *Proc. of the 6th Int. Conf. on Extending Database Technology*, pages 389–403.