A UML Profile for Modeling Data Warehouse Usage *

Veronika Stefanov and Beate List

Women's Postgraduate College for Internet Technologies
Institute of Software Technology and Interactive Systems
Vienna University of Technology
{stefanov, list}@wit.tuwien.ac.at

Abstract. Data Warehouse (DWH) systems represent a single source of information for analyzing the status, the development and the results of an organization.
Today's DWH systems provide many different services to different kinds of users. People involved in designing and managing DWH systems need to see the big picture of how the DWH is being used, to have an overview of the current situation, and to be able to visualize future scenarios. Currently, there is a lack of such general models in Data Warehousing.
We introduce the UML Profile for Modeling DWH Usage for modeling the different kinds of DWH usage on a conceptual level. It uses features of UML intended for the purpose of creating abstract, general models. The profile distinguishes four perspective of usage, and allows to model details of the users. The UML Profile is applied to examples illustrating some of the application scenarios.

1 Introduction

A Data Warehouse (DWH) system is more than just a big database. Defined as "a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process" [1], DWH systems represent a single source of information for analyzing the status, the development and the results of an organization [2]. Analysts and decision makers take measures such as the number of transactions per customer or the increase of sales during a promotion and use them to recognize trends or warning signs and to decide on future investments.

Today's DWH systems provide many different services to different kinds of users: Users retrieve summaries and reports relevant to them, or analyze data with specialized visualization tools. The system may send them messages via e-mail or sms, or provide a quick overview visualization in a dashboard or an intranet portal. Users need access to data at different times, some need it occasionally, others more often, suddenly urgently, or regularly and predictably.

* This research has been funded by the Austrian Federal Ministry for Education, Science, and Culture, and the European Social Fund (ESF) under grant 31.963/46-VII/9/2002.

People involved in managing, designing or evolving today's DWH systems need to see the big picture of all these different ways the DWH is being used, in order to have an overview of the current situation, and to be able to visualize future scenarios. Overview diagrams are needed to facilitate communication with users and decision makers.

Surprisingly, today there are no existing models to describe the different aspects of DWH usage on a conceptual level. There is a lack of general models that provide a broader view over several aspects, eventhough there exist many detailed models of sub-areas. We identify a need for a model that shows on the conceptual level:

1. Who are the users and how are they grouped together?
2. Which part of the DWH system do they use? How do they use it?
3. How intensely are which parts of the DWH being used by which users?
4. When do users need to use which part, and how time critical is it?
5. How important is it?

To fill this gap, we use the UML extension mechanism to specify the *UML Profile for Modeling DWH Usage*. Our profile uses some of the lesser known features of UML, intended for the purpose of creating preliminary models with a "less precise but more general representation" [3]. We have grouped the features of the profile into four perspectives, which focus on the most common application scenarios of DWH usage modeling (Section 3). The *UML Profile for Modeling DWH Usage* (Section 4) offers the following contributions:

- It allows to model who uses the DWH, to group the users, and to model their organizational affiliation, skill level, and an approximate number of instances for each user role.
- Modelers can show how often users use something, and how time critical and how important a certain usage is, as well as active or passive usage types.
- The model allows the analysis of the implications of changing scenarios (e.g. adding a component, increasing numbers of users) on various levels of detail.
- It can be used to identify critical patterns (many important accesses, rapid growth) and to identify parts of the DWH that are not used or not used very often or importantly.
- DWH usage models can be used to support the design of user access controls or personalized user interfaces.
- In general, the models make the overall structure of DWH usage visible on the conceptual level, thus replacing the custom of creating *ad hoc* diagrams and drawings for the communication with users and decision makers.

DWH usage models are intended to provide an overview without aiming at a design process. Compared to requirements analysis in Data Warehousing, our approach to DWH usage is broader, and not necessarily focused on a future system to be built. The UML Profile allows to model the users in detail and does not explicitly include (design) goals of any kind. In MDA [4] terms, our approach is located in the CIM (Computation Independent) area, where models are not necessarily intended to be transformed into code.

2 Modeling Data Warehouses: Background

Our approach applies UML to the Data Warehousing domain. It is aimed at encompassing all the different ways that users may use a DWH. Our goal is to provide an overview over all aspects of DWH usage, not only focussing on the data model. Nevertheless, due to the special characteristics of DWH data, it is necessary to take the data model especially into account.

DWH applications involve complex queries on large amounts of data, which are difficult to manage for human analysts. In Data Warehousing, data is often organized according to the multidimensional paradigm, which allows data access in a way that comes more natural to human analysts. The data is located in n-dimensional space, with the dimensions representing the different ways the data can be viewed and sorted (e.g. according to time, store, customer, product, etc.).

A multidimensional model, also called star schema or fact schema, is basically a relational model in the shape of a star (see Fig. 1 for an example). At the center of the star there is the *fact* table. It contains data on the subject of analysis (e.g. sales, transactions, repairs, admissions, expenses, etc.). The attributes of the fact table (e.g. cost, revenue, amount, duration, etc.) are called *measures* or *fact attributes*. The spokes/points of the star represent the *dimensions* according to which the data will be analyzed (sorted/aggregated by data, by store). The dimensions can be organized in hierarchies that are useful for aggregating data (e.g. store, city, region, country). Stars can share dimensions, thus creating a web of interconnected schemas that makes drill-across operations possible.

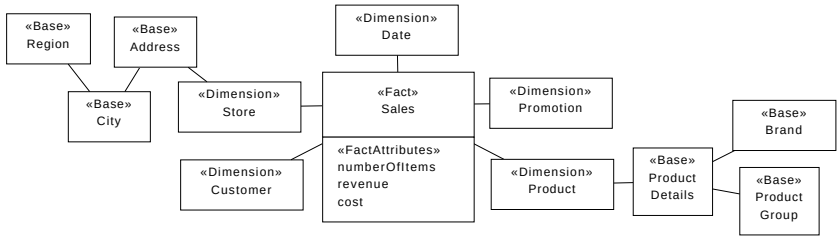


Fig. 1. A simple multidimensional model modeled in UML notation with stereotypes from [5]. Aggregation levels are only shown for the Product and the Store dimensions.

There are many approaches to modeling the multidimensional data structures of data warehouses (see [6–8] for comparisons), some of which are object-oriented models or based on the Unified Modeling Language (UML) [8, 5, 9].

For modeling multidimensional data, we choose to use the UML Profile of Luján-Mora et al. as described in [5]. This Profile allows to model not only the core features of multidimensional models (facts, measures, and dimensions), but also many advanced features such as degenerate dimensions or nonstrict and complete dimensional hierarchies, and also provides three levels of detail.

3 Data Warehouse Usage: Perspectives and Application Scenarios

In order to provide models of DWH usage that are useful to different application scenarios, we need to define our notion of usage. Our goal is to achieve a broad view of usage, while maintaining concise models.

Usage occurs between different kinds of *users* (i.e. roles of users, groups of users, external users) which use different *parts of the DWH system* (data marts, facts, overview dashboards) in different ways (only passively, very often, more resticted), as illustrated in Figure 2. For greater clarity, we have divided the general notion of DWH usage into four perspectives:

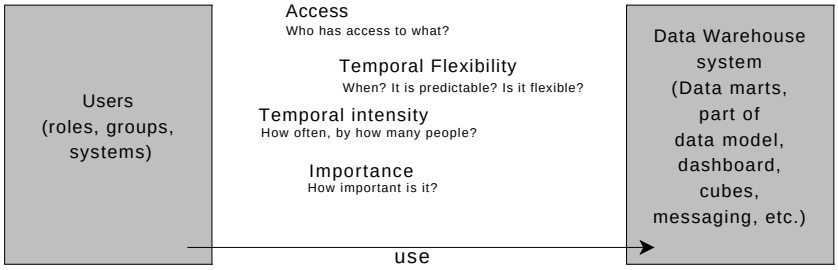


Fig. 2. Perspectives to consider when describing the usage of a data warehouse

1. **Access control:** Who is allowed to use what?
2. **Temporal intensity:** How often do they use it?
3. **Temporal flexibility:** Do they have to use it at a certain time, or can it wait? Is it predictable when they need to use it?
4. **Importance:** How important is this usage?

We have identified a number of application scenarios of DWH usage models. They vary with the target user group and the perspectives to be modeled, and offer modeling solutions for typical every-day requirements in DWH management, maintenance and (re-)design.

To gain an overview of the current system, *DWH engineers* and *architects* as well as *managers* can employ usage models containing details from the *access control perspective* and an approximate number of instances for the user roles or groups. This answers general questions such as "who is using this?", "how many people would complain if we remove this?", etc.

A more detailed model using the *access types* from the access perspective can serve as input for specifying *access restriction* policies, and/or for setting *predefined views* and queries in data access tools.

Temporal intensity and flexibility considerations can be used (a) during the *planing phase* of a DWH design project, user requirements have to be matched to the available resources. DWH Usage models offer a way to capture a general overview of both aspects. Designers can then procede from the usage models to more detailed models later on.

Additionally, if (b) changes become necessary to an existing DWH, usage models can help to *identify critical patterns*. For instance, if due to mergers or reorganizations the number of users in a certain area rapidly increases or decreases, the intensity and flexibility perspectives can provide an overview of the implications.

As in any real-life setting, often not all that is desirable can be achieved. With the help of usage models with elements from the *importance perspective*, managers can decide how to resolve *resource conflicts*.

4 UML Profile for Modeling Data Warehouse Usage

This section introduces our UML Profile for Modeling DWH Usage. We use the extension mechanism of UML and import elements from a well-known UML Profile of the Data Warehousing domain, in order to achieve a conceptually sound model, with (a) tool support and (b) well-known notation elements as additional advantages gained by choosing to extend UML.

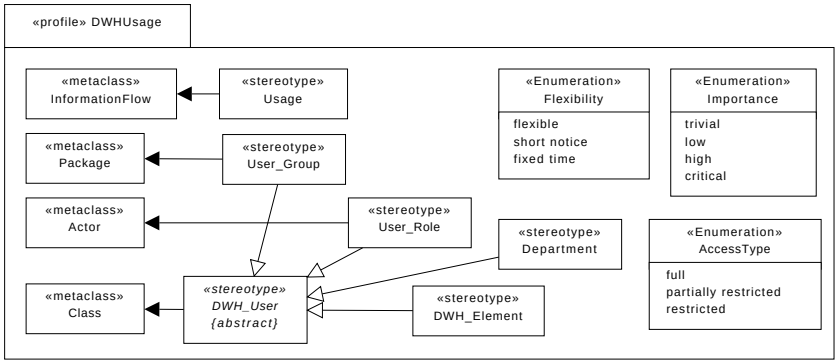


Fig. 3. The UML Profile for Modeling Data Warehouse Usage: Package contents

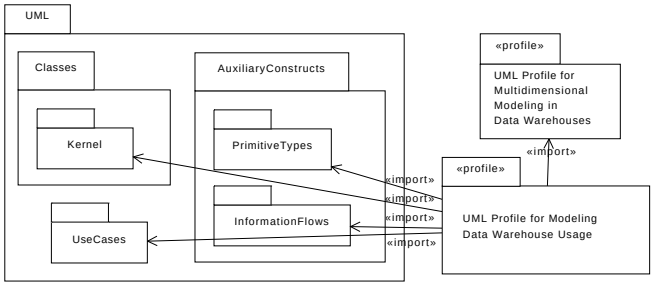
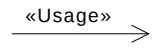


Fig. 4. The UML Profile for Modeling Data Warehouse Usage: Imports

Figure 3 gives an overview of our UML Profile and its stereotypes and supporting enumerations, and also shows which classes are used as base classes of the stereotypes. Figure 4 shows that for modeling multi-dimensional data models, we import the Profile of Luján-Mora et al. [5], and also additionally some packages of the UML metamodel (for the convenience of not having to use fully qualified names). Table 1 describes the characteristics of the stereotypes.

Name Base class Specializations Description Tag Definition	***DWH User*** Class User Role, User Group, and Department An entity using the DWH. Abstract. **numberOfInstances** Type: Integer, Multiplicity: 1 Description: The number of instances of this role. Visualized in the icon e.g. as a number in the "head" of an actor symbol. **skillLevel** Type: SkillLevel, Multiplicity: 1 Description: The skill level of the user, i.e. whether able to write queries	
Name Base class Generalization Description	**User Role** Actor DWH User A role that users/actors take when they access a DWH. One physical person (or external software system) may have several roles, and there may be several instances of one role.	
Name Base class Generalization Description	**User Group** Package DWH User Group of similar roles (orthogonal to Department)	
Name Generalization Description	**Department** DWH User Organizational department (orthogonal to User Group)	
Name Generalization Description	**DWH Element** DWH User An Element of the DWH system that users can access, and that can access other elements, e.g. a dashboard, a portal, etc.	
Name Base class Description Tag Definition	**Usage** InformationFlow A *Usage* indicates an "information channel" [3] between the DWH and its users. See text below for details. **accessType** Type: AccessType, Multiplicity: 1 Description: Indicates whether the access is (partially) restricted **temporalFlexibility** Type: Flexibility, Multiplicity: n, Default value: full Description: Indicates whether the usage is flexible in terms of time. **temporalIntensity** Type: String, Multiplicity: 1 Description: A textual description of the intensity of usage, e.g. number of instances per time interval, as scalar or as interval, probability range, etc. **importance** Type: Importance, Multiplicity: 1 Description: Indicates the level of importance attached to this usage.	
Name Stereotype Values	**SkillLevel** Enumeration {basic, intermediate, expert}	
Name Stereotype Values	**AccessType** Enumeration {full, partially restricted, restricted}	

Name	**Flexibility**
Stereotype	Enumeration
Values	{flexible, short notice, fixed time}
Name	**Importance**
Stereotype	Enumeration
Values	{trivial, low, high, critical}
Additionally, all elements imported from the UML Profile for multidimensional modeling in data warehouses [5].	

Table 1: The UML Profile for Modeling Data Warehouse Usage

Usage is defined as an InformationFlow, which is a type of directed relationship that specifies that information items circulate from sources to targets[1]. Information flows are defined in UML as a very general concept to be used in "*preliminary models, before having taken detailed modeling decisions on types or structures. One other purpose of information items and information flows is to abstract complex models by a less precise but more general representation of the information exchanged between entities of a system.*" [3] This makes information flows very suitable for our purpose, which is to provide models that capture overview of the general structure of DWH usage.

The direction of the Usage arrow indicates whether the users actively initiate the access to the DWH or whether they wait to receive messages from the system, i.e. *push* or *pull* mode. A user analyzing OLAP data would be pull, whereas an e-mail alert is push.

5 Examples

In this section we illustrate the use of the UML Profile for Modeling Data Warehouse Usage with a number of examples. The examples each focus on a subset of the features of the profile and together provide an overview over the perspectives described in Section 3.

5.1 Users accessing Data Warehouse data

Figure 5 focuses on the questions "who needs which data?" and "who should be allowed to see what?". Diagrams of this type can be used in discussions with (future) users and in a later stage of the DWH design process may serve as rough input for specifying access restriction controls. Diagrams like this make it possible to identify preliminary groups of users, based on their data needs.

For each hospital admission it is recorded who was admitted (Patient dimension), what was the primary Diagnosis, which bed the patient was given (Placement) and which Insurance will cover the expenses. Health care professionals (nurses, doctors, therapists) need to access data on the patient, the diagnosis and the placement (the latter resticted to the ward they are working at), whereas

[1] Sources and targets of an information flow may be: Actor, Node, UseCase, Artifact, Class, Component, Port, Property, Interface, Package, and InstanceSpecification [3].

the administration is interested in overall figures of how many patients were admitted where, but should not access patient details or diagnoses. For the billing clerks of the accounting department, all data for charging the hospital bills to the insurance companies has to be accessible. Finally, if the data is made available to medical researchers (e.g. research on the seasonal occurrence and duration of certain medical conditions), only aggregated patient data and diagnoses are relevant.

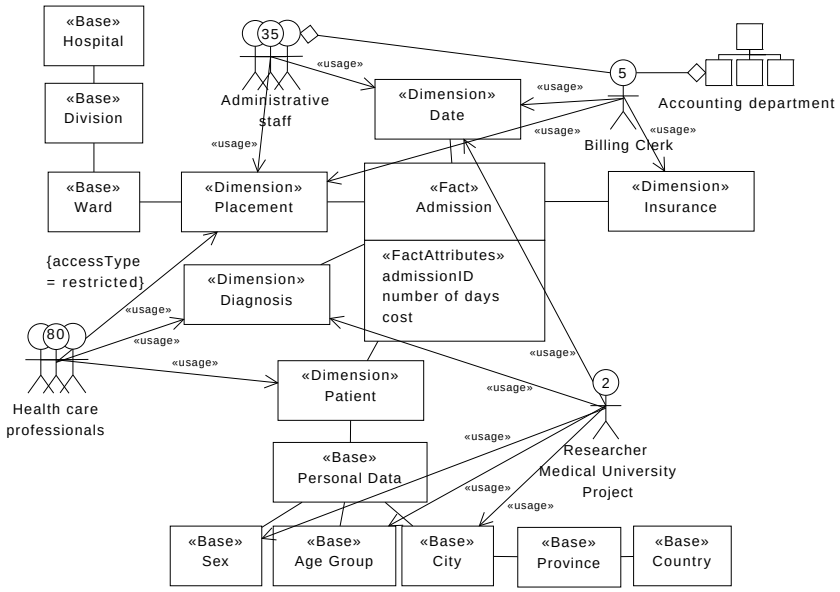


Fig. 5. Hospital admissions: Different roles and groups

5.2 Temporal aspects of DWH Usage, evaluating scenarios

The example shown in Figure 5.2 illustrates the concepts of Intensity and Flexibility as described in Section 3.

Consider a Sales fact, the typical example of Data Warehousing (see Section 2), which contains data on items sold, to be sorted, aggregated and analyzed by time, product group, store, etc. In this example we take the Sales fact as a whole (a "StarPackage") and focus on the different users from various parts of the enterprise who all want to access this data.

Branch managers want to analyze the sales of their branch regularly once a week, and more or less predictably at the same time. The marketing department on the other hand will want access to sales data occasionally, but not necessarily at a given time or urgently. Product managers need to access the data with a varying intensity: If their new product is launched for example, they will watch the sales closely, but not at other times.

A new component that provides sales history data to sales agents is added to the system. Via this component, 150 individuals will want to access the sales data. The diagram provides an overview of the situation and supports discussions about this design decision.

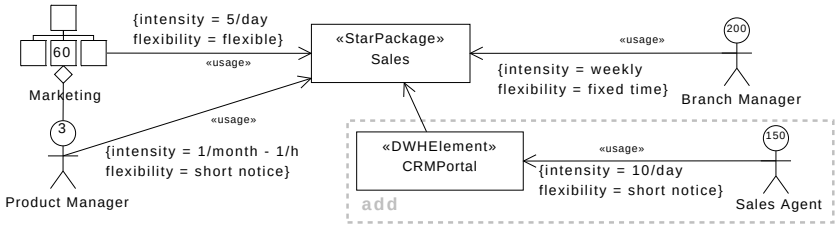


Fig. 6. Sales Example: Temporal aspects of DWH usage: Intensity and Flexibility

5.3 Importance in DWH Usage

Assessing the relative importance of features is crucial for design decisions. For our purpose, we restrict it to four levels, (critical, high, low, and trivial) and subsume economical as well as "political" importance (see [10] for examples of political issues in Data Warehousing) under one item, as shown in Figure 5.3. In this example the use of the attribute skillLevel and passive DWH usage (via an e-mail alert service) are also shown.

Sales and Marketing people need to access sales data for their everyday work, which is of high importance. Top managers occasionally browsing sales data should also be treated as important, which is an example of "political" and not so much economical importance. Aggregated sales data is also fed into the Intranet portal, but this considered a "nice to have" feature of trivial importance.

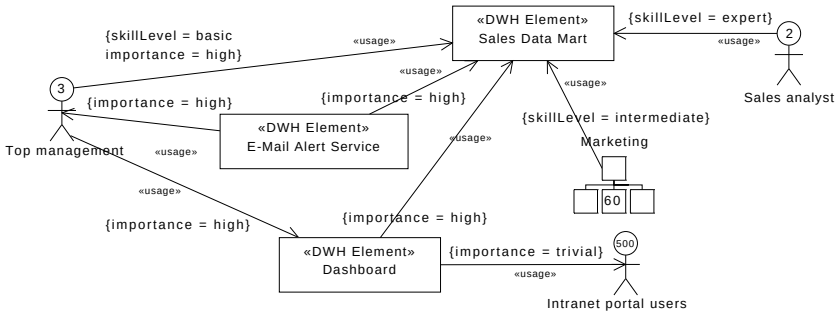


Fig. 7. Importance, passive usage (push instead of pull), and user skills

6 Related Work

Our approach to modeling DWH usage on the conceptual level as presented in this paper touches many different areas, with access control, temporal intensity, temporal flexibility and importance, active or passive usage, details of the users such as their skill level, number of instance or affiliation. To the best of our knowledge, there is no comparable work with the same focus. Nevertheless, regarding individual aspects, we can discuss our relationship to previous work.

Modeling the users who access different parts of data warehouse systems is also an issue for security and access control in Data Warehousing.In [11], the authors present a UML Profile for secure DWHs that includes user profiles and user roles contained in hierarchies. The users are granted privileges to access parts of a multidimensional data model. [12] define a different kind of authorization model integrated with MD modeling, also based on user roles. [13] or [14] are other example of approaches to modeling user access to parts of multidimensional models and/or OLAP operations.

As these approaches have a different aim in modeling users (i.e. grouping users with similar privileges for access control and security measures), they do not include the organizational affiliation of users, their skill level, or the importance, intensity or temporal flexibility of their access to the DWH.

Our approach can also be compared to Requirements Engineering. Explicitly modeling requirements is quite young in the Data Warehousing research field, although there are some very good examples available.

[15] describe an approach to model DWH requirements with UML use cases. [16] argue that DWH requirement analysis should fundamentally be based on goals, and present a model for goals and decisions. [17] distinguish supply- and demand-driven DWH design methods and describe a data warehouse requirement analysis based on goals that is suitable also for mixed situations, and [18] introduce an approach that is able to directly build a conceptual and logical data model from a requirements model.

Our UML profile for modeling DWH usage is not intended as a means to create data models. DWH usage models can be used to support the requirements analysis phase of a DWH project, but provide a more general outlook on the way users access a data warehouse.

7 Conclusion

Today's DWH systems provide many different services to different kinds of users. In order to have a big picture of the current situation and to visualize future scenarios, people involved in designing and managing today's DWH systems need an overview of all these different ways the DWH is being used.

In this paper, we have introduced the UML Profile for Modeling DWH Usage for modeling the different kinds of DWH usage on a conceptual level. It distinguishes four perspective of usage (access control, temporal intensity, temporal flexibility and importance) as well as active or passive usage, and allows

to model details of the users such as their skill level, number of instances, functional grouping or organizational affiliation. We base "usage" on UML information flows, which are intended for a more general representation of information exchanges, and import the elements of an existing profile for modeling the special multidimensional data models of DWHs.

References

1. Inmon, W.H., Hackathorn, R.D.: Using the data warehouse. Wiley-QED Publishing, Somerset, NJ, USA (1994)
2. Kimball, R., Reeves, L., Ross, M., Thornthwaite, W.: The Data Warehouse Lifecycle Toolkit. John Wiley & Sons, Inc. (1998)
3. Object Management Group, Inc.: UML 2.0 Superstructure. http://www.omg.org/cgi-bin/apps/doc?formal/05-07-04.pdf (2005)
4. Object Management Group, Inc.: Model Driven Architecture (MDA). http://www.omg.org/cgi-bin/doc?formal/03-06-01 (2004)
5. Luján-Mora, S., Trujillo, J., Song, I.Y.: A UML profile for multidimensional modeling in data warehouses. Data Knowl. Eng. **59**(3) (2006) 725–769
6. Vassiliadis, P., Sellis, T.K.: A Survey of Logical Models for OLAP Databases. SIGMOD Record **28**(4) (1999) 64–69
7. Blaschka, M., Sapia, C., Höfling, G., Dinter, B.: Finding Your Way through Multidimensional Data Models. In: DEXA '98. (1998) 198–203
8. Abelló, A., Samos, J., Saltor, F.: YAM^2 (Yet Another Multidimensional Model): An Extension of UML. In: IDEAS '02, IEEE Computer Society (2002) 172–181
9. Nguyen, T.B., Tjoa, A.M., Wagner, R.: An Object Oriented Multidimensional Data Model for OLAP. In: Web-Age Information Management (WAIM) 2000), Springer-Verlag (2000) 69–82
10. Demarest, M.: The politics of data warehousing. http://www.noumenal.com/marc/dwpoly.html (1997)
11. Ferández-Medina, E., Trujillo, J., Villarroel, R., Piattini, M.: Developing secure data warehouses with a uml extension. Inf. Syst. **32**(6) (2007) 826–856
12. Priebe, T., Pernul, G.: A pragmatic approach to conceptual modeling of olap security. In: ER '01: Proceedings of the 20th International Conference on Conceptual Modeling, London, UK, Springer-Verlag (2001) 311–324
13. Kirkgöze, R., Katic, N., Stolba, M., Tjoa, A.M.: A security concept for olap. In: DEXA '97: Proceedings of the 8th International Workshop on Database and Expert Systems Applications, Washington, DC, USA, IEEE Computer Society (1997) 0619
14. Wang, L., Jajodia, S., Wijesekera, D.: Securing olap data cubes against privacy breaches. IEEE Symposium on Security and Privacy **2004** (2004) 161
15. Bruckner, R., List, B., Schiefer, J.: Developing requirements for data warehouse systems with use cases. In: Proceedings 7th Americas Conference on Information Systems. (2001) 329–335
16. Prakash, N., Gosain, A.: Requirements driven data warehouse development. In: CAiSE Short Paper Proceedings. (2003)
17. Giorgini, P., Rizzi, S., Garzetti, M.: Goal-oriented requirement analysis for data warehouse design. In: Proceedings DOLAP '05, ACM Press (2005) 47–56
18. Mazon, J.N., Trujillo, J., Serrano, M., Piattini, M.: Designing data warehouses: From business requirement analysis to multidimensional modeling. In: Proc. 1st Int. Workshop on Requirements Eng. for Business Need and IT Alignment. (2005)