International Journal of Computer and Communication Technology

Volume 4 | Issue 2 Article 13

April 2013

Extending Uml for Multidimensional Modeling in Data Warehouse

Bakul Dhawan
University School of Information Technology, bakuldhawan@gmail.com

Anjana Gosain
University School of Information Technology, anjana_gosain@yahoo.com

Available online at www.interscience.in

Extending Uml for Multidimensional Modeling in Data Warehouse

Bakul Dhawan & Anjana Gosain
University School of Information Technology
E-mail: bakuldhawan@gmail.com, anjana_gosain@yahoo.com

Abstract **-** Multidimensional modeling is the foundation of data warehouses, MD databases, and On-Line Analytical Processing (OLAP) applications. Nowadays Dimensional modeling and object-orientation are becoming growing interest areas. In the past few years; there have been many proposals, for representing the MD properties at the conceptual level. However, none of them has been accepted as a standard for conceptual MD modeling. In this paper, we present an extension of the Unified Modeling Language (UML) using a UML profile for multidimensional databases. This profile is composed of a set of stereotypes, constraints and tagged values. We have extended the uml for representing the main multidimensional properties at the conceptual level such as the many-to-many relationships between facts and dimensions, degenerate dimensions, multiple and alternative path classification hierarchies, and non-strict and complete hierarchies and aggregate fact table.

Keywords - *Data warehouse,uml, multidimensional modeling, object-orientation.*

I. INTRODUCTION

Data warehouses are the cornerstone of data-driven decision support systems (DSS). They rely on a multidimensional model, providing users with a business-oriented view of data. Using On-Line Analytical Processing (OLAP) tools, decision makers may then navigate through and analyze multidimensional data. [16].The practitioners and researchers believe that the development of these systems should be based on a conceptual multidimensional model that allows designers to easily structure information into facts and dimensions A *Fact* which contain interesting measures of a business process (product sale, amount). A *Dimension* is used for analyzing the measures of a business process (customer, time, state). In this paper we have we have made use of the extensible mechanisms in uml that includes stereotypes, tagged values and constraints. All these three are together called as a uml profile. We have extended the uml for representing the main multidimensional properties at the conceptual level such as the many-to-many relationships between facts and dimensions, degenerate dimensions, multiple and alternative path classification hierarchies, and non-strict and complete hierarchies and aggregate fact table. This paper has been divided into following sections. In section 2 previous work performed by various authors has been discussed. In section 3 Multidimensional modeling and its properties and hierarchies have been discussed, In section 4 UML profile and its extensibility mechanisms like stereotypes, constraints and tagged values are being discussed. In section 5, we have discussed how uml can be extended for multidimensional modeling; section 5 represents the comparison between the various models proposed by authors.

We have based our scheme in UML for the following main reasons: [2], [3]

(i) UML is a well known standard modeling language known by most database designers; thereby designers can avoid learning a new notation.

(ii) Considers an information system's structural and dynamic properties at the conceptual level more naturally than do other approaches such as the Entity-Relationship model.

(iii) UML provides powerful mechanisms—such as the Object Constraint Language and the Object Query Language.

II. PREVIOUS WORKS

In [2] authors have discussed the various extensibility mechanisms like stereotypes, constraints and tagged values to elegantly represent main MD properties at the conceptual level. They have made use of the Object Constraint Language (OCL) to specify the constraints attached to the defined stereotypes, thereby avoiding an arbitrary use of these stereotypes.

In [3] authors have presented UML-based data warehouse design method that spans conceptual, logical and physical design. Starting from user requirements, the conceptual phase leads to a UML model. They have enriched uml with concepts relevant to multidimensional systems. The logical phase maps the enriched UML model into a multidimensional schema, independently of any implementation tool.

In [5] authors have summarized the UML Extensibility Mechanisms and introduced the main MD concepts such as fact, dimension, and hierarchy level . they have proposed the new UML extension (stereotypes, tagged values, and constraints) for MD modelling and represented MD extension in Rational Rose.

In [6] an Object Oriented multidimensional data model (OOMD, renamed GOLD) has been introduced. They have discussed the certain key issues in multidimensional modeling, such as derived measures, derived dimension attributes and the additivity on fact attributes along dimensions.

In [7] the object oriented OLAP framework using UML models is presented. It presents the concepts and techniques that allow the users to exploit simultaneously the features of OLAP and object systems. The dimension and measures are modeled in terms of class diagrams by using UML.

In [12] the authors have represented a novel approach to integrating data mining models into multidimensional models in order to accomplish the conceptual design of DWs with Association Rules (AR). They have provided a UML profile that allows specifying Association Rules mining models for DW at conceptual level in a clear and expressive way.

In [11] the authors have described summarizability issues in multidimensional modeling. Importantly. Every multidimensional model to be implemented must ensure summarizability because, otherwise, its violation can lead to incorrect results.

II. MULTIDIMENSIONAL MODELING

Multidimensional modeling structures the information into facts and dimensions. A fact contains measures or fact attributes and a dimension is used for analyzing. It contains dimension attributes and hierarchies. For example product sale is a fact and product price and quantity are its measures and time is a dimension which forms the following hierarchy year-month-week.

Multidimensional modeling properties -

Facts -

Many-to-many relationships with particular dimensions-many-to-one relationships exist between the fact and every particular dimension, and thus facts are usually considered to have many-to-many relationships between any of two dimensions [2]**.***Derived measures* derived attribute is represented by placing / next to a measure in the fact class. Derivation rules are specified in the brackets. *Additivity*-A measure is additive along a dimension if the SUM operator can be used to aggregate attribute values along all hierarchies defined on that dimension.[2]

Dimensions -

Multiple and alternative path classification hierarchies -Regarding dimensions, the classification hierarchies defined on certain dimension attributes are fundamental because the subsequent data analysis will be addressed by these classification hierarchies. A dimension attribute may also be aggregated (related) to more than one hierarchy, and therefore, multiple classification hierarchies and alternative path hierarchies are also relevant. [2]

Multiple classification hierarchy-
Product-type-family-group, Product-brand

Alternative path classification hierarchy -

Store-city-province-state, Store-sales-area-state

Non strict classification hierarchy and Complete classification hierarchy - Nevertheless, classification hierarchies are not so straightforward in most cases. The concepts of strictness and completeness are important, not only for conceptual purposes, but also for further design steps of MD modeling. [2] ‘‘Strictness’’ means that an object of a lower level of a hierarchy belongs to only one of a higher level, [2], [3], [6] e.g. a city is related to only one state. In ‘‘Completeness’’ all the members belong to one higher-class object and that object consists of those members only. [2],[3],[6] For example, we can say that the classification hierarchy between the state and city levels is ‘‘complete” only when a state is formed by all the cities recorded and all the cities that form the state are recorded.

Categorization of dimensions - OLAP scenarios sometimes become very large as the number of dimensions increases significantly, and therefore, this fact may lead to extremely sparse dimensions and data cubes. In this way, there are attributes that are normally valid for all elements within a dimension while others are only valid for a subset of elements (also known as the categorization of dimensions) [2].

IV. UML Profile

In order to model systems with certain specific needs, UML can be extended in two different ways: (i) the lightweight and (ii) the heavyweight way. The former implies a UML extension by means of a profile providing the stereotypes, tagged values and constraints needed in order to specify the peculiarities of the modeled system. In the latter, a completely new modeling language is proposed by extending the Meta Object Facility (MOF) [17], the modeling language from which UML is defined [12].

4.1 UML Extensibility Mechanism

The UML Extensibility Mechanism package is the sub package from the UML metamodel that specifies how specific UML model elements are customized and extended with new semantics by using stereotypes, tagged values, and constraints. A coherent set of such extensions, defined for specific purposes, constitutes a UML profile. For example, the UML 1.4 [10] includes a standard profile for modeling software development processes and another one for business modeling [5].*A profile* is a stereotyped package that contains model elements that have been customized for a specific domain or purpose by extending the metamodel using stereotypes, tagged definitions, and constraints [10].A profile consists of stereotypes, tagged values and constraints.

Stereotypes- A stereotype extends the vocabulary of the UML, allowing you to create new kinds of building blocks that are derived from existing ones but that are specific to your problem [11], [13].Graphically, a stereotype is rendered as a name enclosed by guillemots and placed above the name of another element (for example <<name>>) alternatively, you can render the stereotyped element by using a new icon associated with that stereotype.

A *tagged value* extends the properties of a UML building block, allowing you to create new information in that element's specification [11], [13]. Example: In the release team of a project that is responsible for assembling, testing and deploying releases, you might want to keep track of the version number and test results for each major subsystem.

A *constraint* extends the semantics of a UML building block, allowing you to add new rules or modify existing ones [11], [13]. Graphically, a constraint is rendered as a string enclosed by brackets and placed near the associated element(s) or connected to that element(s) by dependency relationships.

V. UML Extension Considering Aggregation

In this section we have represented how UML can be extended for multidimensional modeling. The MD modeling properties are represented by means of a UML class diagram in which the information is separated into facts and dimensions. Dimensions and facts are represented by *dimension classes* and *fact classes* respectively. Then, fact class is represented as composite class which is in a shared aggregation relationship with n dimension classes. In figure1. We have represented the following hierarchies which have also been explained in [14]. We have added a way to represent aggregate fact table.

An *aggregate fact table* Sales summary is being represented by using a derivation relationship between the basic fact table and aggregate fact table. This table would be having all the attributes same as the base fact table except the key. Aggregate fact table contains the precalculated summaries derived from the most granular fact table [10].Another fact table named credit_sales is also shown in the figure. Both the fact tables, product_sales and credit_sales have been derived from the data warehouse. Which is represented by a stereotype <<*derivation*>>.This can also be considered as a way to drill through. *Drill through* is drilling to the lower level of granularity, as stored in the source data warehouse repository. [10] we have also presented a way to separate the large and changing dimensions by using <<*composed of*>> as shown in the figure, the customer dimension is composed of personal information such as name, phone no. and other information such as purchase order, credit rating and income level

As discussed and represented by juan Trujillo in [14] , *Many-to-many relationships* between facts and particular dimensions can be considered by indicating the 1...* *cardinality* on the dimension class role. For example, in Figure 1, we can see that the fact class Sales has a many-to-many relationship with the dimension class Product and a one-to-many relationship with all other dimensions. By default, all measures in the fact class are considered *additive*. For nonadditive measures, additive rules are defined as constraints and are also placed in somewhere around the fact class. [5].Furthermore, as authors have discussed in [2, 3, 5] *derived measures* can be considered by placing a (constraint /) and their derivation rules are placed between braces in somewhere around the fact class, as shown in the Figure. With respect to dimensions, every classification hierarchy level of a dimension is specified by a class (called a *base class*). An association between the classes specifies the relationships between two levels of a classification hierarchy. As discussed in [2, 3, and 5] every classification hierarchy level must have an *identifying attribute (constraint {OID})* and a *descriptor attributes (constraint {D}).*The concepts of *strictness and non-strictness* are defined by the *multiplicity* 1 and 1...*In addition, defining the *{completeness} constraint* in the target associated class role addresses the completeness of a classification hierarchy. This approach considers all classification hierarchies non-complete by default [5].The *categorization of dimensions*, used to model additional features for an entity's subtypes, is considered by means of generalization-specialization relationships. However, only the dimension class can belong to both a classification and a specialization hierarchy at the same time e.g. in the figure it can be seen that customer and salesperson dimension can be generalized into a person dimension.

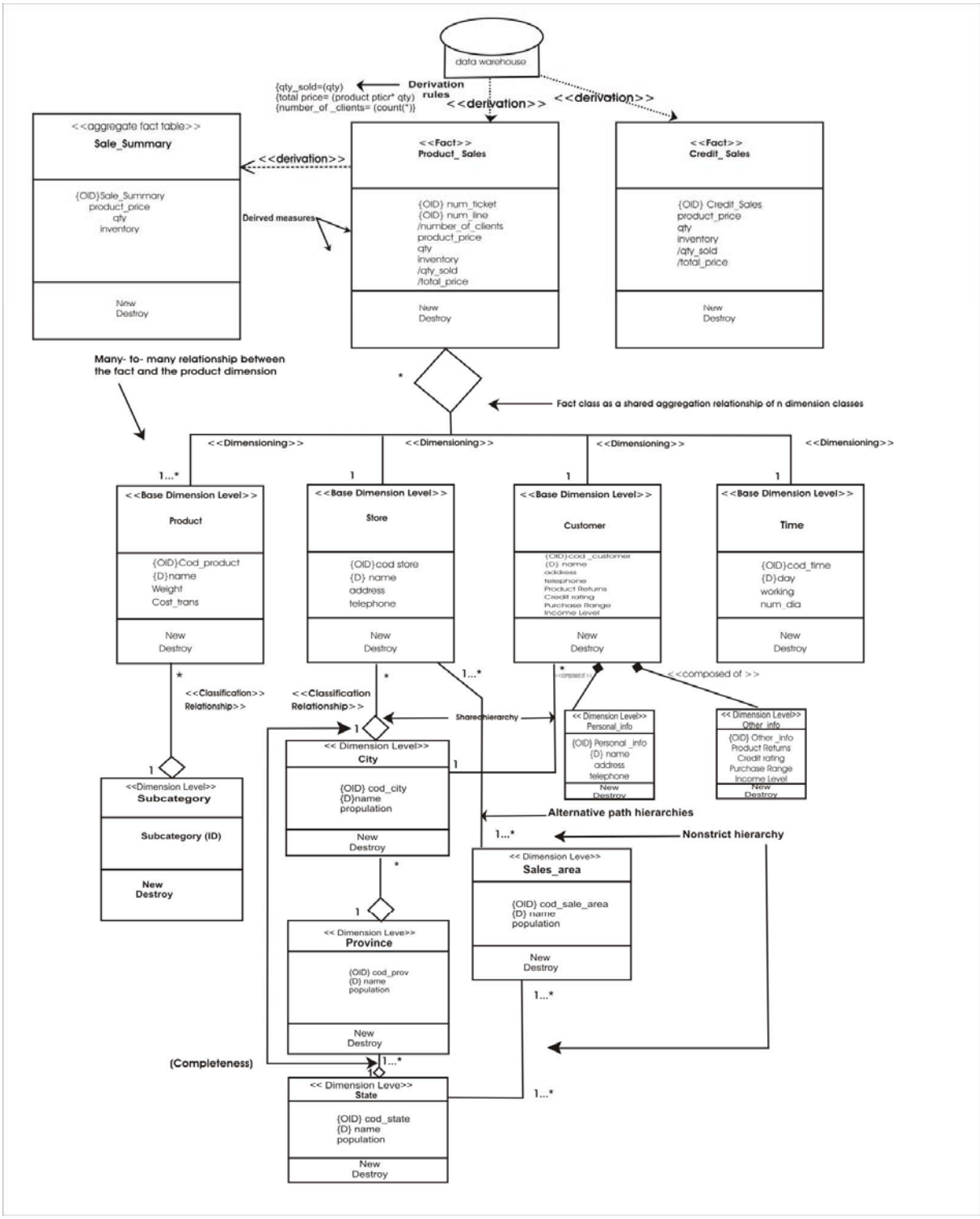

Fig 1: Extending Uml For Multidimensional Modeling

Using Aggregation

VI. COMPARATIVE STUDY

Previously many papers have discussed about multidimensional properties. Table 1 provides the comparison of these properties discussed by various authors.

Starting with the facts, only [8] has talked about multistars, rest [2] and [3] discuss them partially. Many-to many relationships and atomic measures is being discussed in all the papers. A derived measure is being supported by only [8], [3] and[6] and additivity is being discussed in all the papers except [5].

With reference to the dimensions, all the authors have discussed about Categorization of dimensions, nonstrict classification hierarchies and Multiple and alternative path classification hierarchies except [16] which doesn't represent Multiple and alternative path classification hierarchies. Complete classification hierarchies is not being discussed by [8] and [16].drill down is not discussed by [3] and [5] and roll up is not being discussed by [5] and [3].Drill across is only discussed in [8] and none of the papers have discussed about Drill through, Aggregate fact table, Separating large and changing dimensions.

Table 1. Comparision of other uml based multidimensional models

Multidimensional Modeling Properties	**Model**					
	[5]	**[8]**	**[6]**	**[16]**	**[2]**	**[3]**
FACTS						
Multistars	NO	YES	NO	NO	YES	YES
Many-to-many relationships with particular dimensions	YES	YES	YES	YES	YES	YES
Atomic measures	YES	YES	YES	YES	YES	YES
Derived measures	NO	YES	YES	NO	NO	YES
Additivity	NO	YES	YES	YES	YES	YES
DIMENSIONS						
Categorization of dimensions	YES	YES	YES	YES	YES	YES
Multiple and alternative path classification hierarchies	YES	YES	YES	NO	YES	YES
Nonstrict classification hierarchies	YES	YES	YES	YES	YES	YES
Complete classification hierarchies	YES	NO	YES	NO	YES	YES
Drill down	NO	YES	YES	YES	YES	NO
Roll up	NO	YES	YES	YES	YES	YES
Drill across	NO	YES	NO	NO	NO	NO
Drill through	NO	NO	NO	NO	NO	NO
Aggregate fact table	NO	NO	NO	NO	NO	NO
Separating large and changing dimensions	NO	NO	NO	NO	NO	NO

REFRENCES

[1]. G.Booch,J.Rumbaugh,and I. Jacobson : The Unified Modeling Language user Guide.

[2]. Sergio Luja´n-Mora, Juan Trujillo , Il-Yeol Song, : A UML Profile For Multidimensional Modelling In Data Warehouses. In: Data & Knowledge Engineering 59 (2006) 725–769.

[3]. Nicolas Prat a, Jacky Akoka b, Isabelle Comyn-Wattiau, : A UML-Based Data Warehouse Design Method. In: Decision Support Systems 42 (2006) 1449–1473.

[4]. Sueli de Fatima Poppi Borba,Aran Bey Tcholakian Morales : Extending the UML for Dimensional Models in Object-Oriented Database. In: Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05) 1529-4188/05 2005 IEEE

[5]. Sergio Luján-Mora, Juan Trujillo: extending Uml For Multidimensional Modelling. In: springer-verlag berlin Heidelberg 2002.

[6]. Juan Trujillo: The GOLD model: An Object Oriented multidimensional data model for multidimensional databases. In: Research Group of Logic Programming and Information Systems Dept. of Financial EconomicsUniversity of Alicante. E-03071. Alicante. Spain.

[7]. Madhu Bhan,D.E.Geetha, T.V.Suresh Kumar, K.Rajanikanth: Modeling of Object Oriented OLAP. In: M. S. Ramaiah Institute of Technology Bangalore – 560 054, India.

[8]. Alberto Abell´o,Jos´e Samos,F`elix Saltor,Dept. de Llenguatges i Sistemes Inform`atics: YAM2 (Yet Another Multidimensional Model): An extension of UML".

[9]. Robert Wrembel Poznań University of Technology, Poland Christian Koncilia Panoratio GmbH, Germany:Data Warehouses and OLAP: Concepts, Architectures and Solutions" .

[10]. Paulraj Ponniah, Data Warehousing Fundamentals: A Comprehensive Guide For It Professionals.

[11]. Jose-Norberto Mazón, Jens Lechtenbörger b, Juan Trujillo: A survey on summarizability issues in multidimensional modelling. In: Data & Knowledge Engineering 68 (2009) 1452–1469)

[12]. Jose Zubcoff a, Juan Trujillo b: A UML 2.0 Profile to Design Association Rule Mining Models in the Multidimensional Conceptual Modeling of Data Warehouses. In: Data & Knowledge Engineering 63 (2007) 44–62

[13]. Jasmine Farhad: The UML Extension Mechanisms. In: Dept of Computer Science,University College London,13-Dec-02

[14]. Juan Trujillo,Manuel Palomar,Jaime Gomez: Designing Data Warehouses with OO Conceptual Models. In: 2001 IEEE

[15]. Esperanza Marcos, Bel´en Vela, Jos´e Mar´ıa Cavero,:A Methodological Approach For Object-Relational Database Design Using UML. In: Informatik Forsch. Entw. (2004) 18: 152–164

[16]. Nicolas Prat, , Isabelle Comyn-Wattiau, Jacky Akok: Representation Of Aggregation Knowledge In Olap Systems. In: 18th European Conference on Information Systems

[17]. S. Luja´n-Mora, J. Trujillo, I. Song: Multidimensional Modeling with UML Package Diagrams .In: Proceedings of the 21st International Conference on Conceptual Modeling (ER'02), Tampere, Finland, October 2002, Lecture Notes in Computer Science, ol. 2503, Springer-Verlag, 2002, pp. 199–213.