A General Model for Data Warehouses

Michel Schneider
Blaise Pascal University, France

INTRODUCTION

Basically, the schema of a data warehouse lies on two kinds of elements: facts and dimensions. Facts are used to memorize measures about situations or events. Dimensions are used to analyse these measures, particularly through aggregation operations (counting, summation, average, etc.). To fix the ideas let us consider the analysis of the sales in a shop according to the product type and to the month in the year. Each sale of a product is a fact. One can characterize it by a quantity. One can calculate an aggregation function on the quantities of several facts. For example, one can make the sum of quantities sold for the product type "mineral water" during January in 2001, 2002 and 2003. Product type is a criterion of the dimension Product. Month and Year are criteria of the dimension Time. A quantity is so connected both with a type of product and with a month of one year. This type of connection concerns the organization of facts with regard to dimensions. On the other hand a month is connected to one year. This type of connection concerns the organization of criteria within a dimension. The possibilities of fact analysis depend on these two forms of connection and on the schema of the warehouse. This schema is chosen by the designer in accordance with the users needs.

Determining the schema of a data warehouse cannot be achieved without adequate modelling of dimensions and facts. In this article we present a general model for dimensions and facts and their relationships. This model will facilitate greatly the choice of the schema and its manipulation by the users.

BACKGROUND

Concerning the modelling of dimensions, the objective is to find an organization which corresponds to the analysis operations and which provides strict control over the aggregation operations. In particular it is important to avoid double-counting or summation of non-additive data. Many studies have been devoted to this problem. Most recommend organizing the criteria (we said also members) of a given dimension into hierarchies with which the aggregation paths can be explicitly defined. In (Pourabbas, 1999), hierarchies are defined by means of a containment function. In (Lehner, 1998), the organization of a dimension results from the functional dependences which exist between its members, and a multi-dimensional normal form is defined. In (Hùsemann, 2000), the functional dependences are also used to design the dimensions and to relate facts to dimensions. In (Abello, 2001), relationships between levels in a hierarchy are apprehended through the Part-Whole semantics. In (Tsois, 2001), dimensions are organized around the notion of a dimension path which is a set of drilling relationships. The model is centered on a parent-child (one to many) relationship type. A drilling relationship describes how the members of a children level can be grouped into sets that correspond to members of the parent level. In (Vassiliadis, 2000), a dimension is viewed as a lattice and two functions "anc" and "desc" are used to perform the roll up and the drill down operations. Pedersen (1999) proposes an extended multidimensional data model which is also based on a lattice structure, and which provides non-strict hierarchies (i.e. too many relationships between the different levels in a dimension).

Modelling of facts and their relationships has not received so much attention. Facts are generally considered in a simple fashion which consists in relating a fact with the roots of the dimensions. However, there is a need for considering more sophisticated structures where the same set of dimensions are connected to different fact types and where several fact types are inter-connected. The model described in (Pedersen, 1999) permits some possibilities in this direction but is not able to represent all the situations.

Apart from these studies it is important to note various propositions (Agrawal, 1997; Datta, 1999; Gyssens, 1997; Nguyen, 2000) for cubic models where the primary objective is the definition of an algebra for multidimensional analysis. Other works must also be mentioned. In (Golfarelli, 1998), a solution is proposed to

derive multidimensional structures from E/R shemas. In (Hurtado, 2001) are established conditions for reasoning about summarizability in multidimensional structures.

MAIN THRUST

Our objective in this article is to propose a generic model based on our personal research work and which integrates existing models. This model can be used to apprehend the sharing of dimensions in various ways and to describe different relationships between fact types. Using this model, we will also define the notion of well-formed warehouse structures. Such structures have desirable properties for applications. We suggest a graph representation for such structures which can help the users in designing and requesting a data warehouse.

Modelling Facts

A fact is used to record measures or states concerning an event or a situation. Measures and states can be analysed through different criteria organized in dimensions.

A fact type is a structure

fact_name[(fact_key),
(list_of_reference_attributes), (list_of_fact_attributes)]

where

- fact_name is the name of the type;
- fact_key is a list of attribute names; the concatenation of these attributes identifies each instance of the type;
- list_of_reference_attributes is a list of attribute names; each attribute references a member in a dimension or another fact instance;
- list_of_fact_attributes is a list of attribute names; each attribute is a measure for the fact.

The set of referenced dimensions comprises the dimensions which are directly referenced through the list_of_reference_attributes, but also the dimensions which are indirectly referenced through other facts.

Each fact attribute can be analysed along each of the referenced dimensions. Analysis is achieved through the computing of aggregate functions on the values of this attribute.

As an example let us consider the following fact type for memorizing the sales in a set of stores.

Sales[(ticket_number, product_key), (time_key, product_key, store_key),
(price_per_unit, quantity)]

The key is (ticket_number, product_key). This means that there is an instance of Sales for each different product of a ticket. There are three dimension references: time_key, product_key, store_key. There are two fact attributes: price_per_unit, quantity. The fact attributes can be analysed through aggregate operations by using the three dimensions.

There may be no fact attribute; in this case a fact records the occurrence of an event or a situation. In such cases, analysis consists in counting occurrences satisfying a certain number of conditions.

For the needs of an application, it is possible to introduce different fact types sharing certain dimensions and having references between them.

Modelling Dimensions

The different criteria which are needed to conduct analysis along a dimension are introduced through members. A member is a specific attribute (or a group of attributes) taking its values on a well defined domain. For example, the dimension TIME can include members such as DAY, MONTH, YEAR, etc. Analysing a fact attribute A along a member M means that we are interested in computing aggregate functions on the values of A for any grouping defined by the values of M. In the article we will also use the notation Mij for the j-th member of i-th dimension.

Members of a dimension are generally organized

Figure 1. A typical hierarchy in a dimension