Basic Notions on Multidimensional Aggregate Data

Maurizio Rafanelli
Italian National Research Council, Italy

INTRODUCTION

The term *multidimensional aggregate data* (MAD; see Rafanelli, 2003) generally refers to data in which a given *fact* is quantified by a set of *measures* obtained applying one more or less complex *aggregative function* (count, sum, average, percent, etc.) to row data, measures that are characterized by a set of variables, called *dimensions*. MAD can be modeled by different representations, depending on the application field which uses them. For example, some years ago this term referred essentially to statistical data, that is, data whose use is essentially of socio-economic analysis. Recently, the metaphor of the data cube was taken up again and used for new applications, such as On-Line Analytical Processing (OLAP), which refer to aggregate and non aggregate data for business analysis.

BACKGROUND

Generally there are two broad classes of multidimensional (statistical) data: microdata and macrodata. The former refers to SDBs containing elementary or raw data, that is, records of individual entities or events. The latter refers to databases containing multidimensional aggregate data, often shown as statistical tables, that result from the application of aggregate functions on raw data. Microdata are generally stored as relations in a relational database, and when an aggregation function is applied to them, the result is a complex data, called *macrodata*, which consists of a descriptive part and a summary part. The latter is called *summary attribute* or measure, and is characterized by the descriptive part mentioned above called *metadata*. Its simplest definition is "data describing data."

In order to model aggregate data, we define these data from both a conceptual and a logical point of view. The main difference between them is that in the case of conceptual data, which we will call *multidimensional aggregate data* (MAD), we do not consider their physical storage, while in the case of logical data, which we will call the *multidimensional aggregate data structure* (MADS), we refer explicitly to their physical storage. It is useful to remember that the *multidimensionality* concept was introduced by Shoshani and Wong (1985) describing a MAD as a mapping from the domains of the category attributes (independent variable) to the (numerical) domains of the summary attributes (dependent variable). Each category attribute often represents a level of a hierarchy present in that dimension of that MAD and ranges over an n-dimensional space (the space of the n-tuples of category attribute instances), from which derives the concept of *multidimensionality*.

We give now some definitions useful in describing the aggregation process, that is, the process that allows one to obtain the multidimensional *aggregate* database from a relational *disaggregate* database.

BASIC NOTIONS

Let Θ be the database universe, that is, the set of all the relations that form the very large relational database in which raw data (microdata) are stored. Let R be the subset of Θ relative to all the relations used in the definition of the multidimensional aggregate (macro) database and which, therefore, refers to all the phenomena studied. Note that each phenomenon consists of one or more *facts* that are the physical objects stored in the database. Let $R_x = \{R_i\}_{i=1,\ldots,h}$ be the set of all the relations $R_1, R_2, \ldots, R_h$ (each of them with attributes different in number and names), which refer to the x-th phenomenon. Let $A^1_1, A^1_2, \ldots, A^1_{k_1}$ be the set of attributes of the relation R_1, where the superscript refers to the index which characterizes the considered relation, k_1, is the number of attributes of this relation (i.e., its cardinality), each of which has a definition domain $\Delta^i_1, \Delta^i_2, \ldots, \Delta^i_{k_1}$, and likewise for the other relations. To clarify how the subsets of R to be aggregated are characterized, let us analyze the concept of the *category attribute*. A category attribute is the result of an abstraction on one or more attributes of the microdata; analogously its *instances* are the result of an abstraction on the (numerical, Boolean, string, etc.) values actually associated with the single microdata.

Definition 1. Let R be the set of all the relations used in the definition of a multidimensional aggregate database, let Ω be the set of all the attributes which appear in R, let

$A_x \in \Omega$ be a generic attribute of this database, and let a_{xy} be one of its instances (with y = 1,…, k, where k is the cardinality of the definition domain of A_x). The logical predicate ($A_x = a_{xy}$), defined on the microdata of R, is called *base predicate*.

Definition 2. The *base set* of the base predicate ($A_x = a_{xy}$) is the subset of Θ consisting of all microdata which satisfy the base predicate. In the following such a subset will be denoted by $B_{A_x} = a_{xy}$.

Let A be the subset of all the attributes of Ω that will become descriptive (or category) attributes or measures of all the MADs that will form the multidimensional aggregate database at the end of the aggregation process. Then A is the set of all and only the attributes that describe all the facts that appear in the multidimensional aggregate database. Many of these attributes appear in different relations of R. Different attributes can contribute to form one *hierarchy*. Different hierarchies can belong to the same *dimension*, on the condition that pairs of hierarchies have at least one attribute in common. Note that parallel hierarchies, called *specialization hierarchies*, can exist. Moreover, other attributes, which do not appear in A, can complete the hierarchies mentioned above (on the condition that the relationship between them and the other attributes of the same hierarchy is defined). A^* is the set of these last attributes plus the attributes of A. We call these hierarchies *primitive hierarchies* because all the hierarchies that refer to one of them are included in it. Analogously, we call *primitive dimension* the dimension which includes all its primitive hierarchies.

Let H be the set of all the hierarchies (including the specialized hierarchies) defined in A^*. Let D be the set of all the dimensions defined in A^* (which can consist of different hierarchies). Note that the users often give the name of a dimension to descriptive variables of a MAD which are, in reality, levels of a hierarchy relative to this dimension. Let Δ be the set of all the definition domains (i.e., of all the instances) of the attributes of A, and let Δ* be the set of all the definition domains of the attributes of A^* which also include all the possible instances that each attribute can assume (therefore, also including the instances not present in the relations of Θ). We call *primitive domains* these definition domains. This means that all the attributes (and all the relative instances) which appear in the multidimensional aggregate database are part of A^* and Δ* respectively.

Category attributes are not the only metadata of multidimensional aggregate data: several other properties may provide a semantic description of the summary data. Among them we consider, in particular, the following:

- The *aggregation type*, which is the function type applied to microdata (e.g., count, sum, average, etc.) to obtain the macrodata (i.e., a MAD, see Rafanelli & Ricci, 1993), and which defines the *summary type* of the measure. This property must always be specified.
- The *data type*, which is the type of summary attribute (e.g., real, integer, non-negative real, non-negative integer).
- The *fact* F_j described by the multidimensional aggregate table considered (e.g., production, population, income, life-expectancy).
- Other properties may be missing, for example "data source" (which may be unknown), "unit of measure," and "unit of count," as defined in the following.

Let Γ be the set of the *functional dependencies* which are possibly present in the multidimensional aggregate database and which, therefore, exist among groups of attributes. Given a phenomenon x and given the set of relations $R_x \subset R$, we consider the subset of R_x formed only by the relations involved in the building of the fact F_j. We call this subset an *aggregation relation*, and denote it by R_j^x, where $R_j^x = \{R_{j,1}, \ldots, R_{j,s}\}^x$. Every fact F_j has its own descriptive space formed by s category attributes (where s is the cardinality of the j-th fact), which are a subset of all the attributes in the relations R_j^x. We denote the set of the above-mentioned s category attributes by $A_j^x = \{A_{j,ks}\}^x = \{A_{j1}, \ldots, A_{js}\}^x$. We call the relation β_j^x, formed by these attributes, a *base relation* of the fact F_j.

The measure values are the result of the aggregation process, i.e., of the application of the aggregation function to the base relation of the fact. The fact obtained by this aggregation process is called *base fact*, because its representation cannot even be disaggregated (i.e., only more aggregate views can be obtained). Each fact consists of a set of materialized views, obtained by applying different operators of aggregation (roll-up, group-by), or of reduction of the definition domains of its category attributes (dice). This set of materialized views defines the *lattice* of this fact. The source of this lattice is formed by the total of all the summary category instances of the base fact, and the sink formed by all the summary category instances at the lowest level of disaggregation.

Let $F = \{F_j\}$ be the set of all the *fact names* described by the multidimensional aggregate database. Let $S = \{S_j\}$ be the set of all the subjects described in the facts, in other words, the "what is" of the summary attributes (Cars, People, Fruit, Workers, Dollars, etc.). Let $R_j^x = \{R_{j,1}, \ldots, R_{j,s}\}^x$ be the subset of the relations in the microdatabase which are involved in the x-th fact. Let $A_j^x = \{A_{j,ks}\}^x = \{A_{j1},$