A STATISTICAL DATA MODEL

François BRY

Introduction

In order to avoid the particularities of one data base, it is useful to have a data-description formalism at our disposal. Such a formalism is a data model (Gardarin (1982), Tsichritzis (1980)). We present here the principle of a statistical data model we have conceived. It give the object types and their composition rules, in order to describe a wide class of statistical data bases. It has been designed ot only to be a tool to build a data base management system (DBMS), but also to establish an essential common language between managers of an application and data processing engineers.

Though the most recent models, the relational and entity-relationship models, tend to satisfy the data independence from storage device (physical independence) and the data independence from applications (logical independence), they have been conceived to describe data bases which are very different from statistical ones. Furthermore, these 'classical' models do not provide concepts of statistical groupings.

Particularities of statistical data bases

Particularities of statistical data bases are listed below as: data movement characteristics, logical data structure characteristics and physical data structure characteristics.

Data movement characteristics:

- There is no updating of statistical records: after being validated, a statistical file is never modified.

- Large sets of new data are periodically inserted in a statistical data base. A statistical data base may increase in size of 10% at a time.

- The very most part of queries needs to read a file in its entirety.

It must be pointed that the 'classical' DBMS, and the underlying data models, were designed to make updating, inserting small sets of records, easier. Usually, they optimize the reading-time of small sets of data to the prejudice of reading-time of large ones.

Logical data structure characteristics:

- The set of values of some data can be known 'a priori' (for example, a demographic inquiry in a country lead to a set of places of birth which can be known without listing the whole file. Moreover, it is meaningful to know places where nobody is born).

- For statistical treatment, some data have to be described by 'hierarchical nomenclature'. A nomenclature is a part of the meaning of data as well as a way to express queries.

Note that 'classical' data models do not allow to treat 'a priori sets'.

Physical data structure characteristics:

- Most of the time, statistical files have fixed length records.

- Statistical files are very often of small density: a large part of data have zero value.

These two points can lead to appropriate file organisation allowing data-compacting. We note that 'classical' DDBMS are supported by file organisation allowing variable record length.

The model

The statistical data model we present here tends to take into account these characteristics. It structures the basic data by two object types: nomenclature and data-matrix. Three types are used to express the basic data: criterion, summable-data and non-summable-data.

A nomenclature is a balanced tree, whose levels are used to define 'a priori sets' and to express groupings of basic data. Although nomenclature can be see as 'meta-data', new nomenclatures can be constructed and old ones can be modified. This allows us to take into account the specific role of time in almost all statistical sources: the structuration of data is evolutive and thus accepts distorsions in time.

In order to aim at the construction of macro-data (or aggregated data) criteria are described by nomenclatures. Summable data and non-summable data allow us to distinguish between macro-data that can be built by successive aggregations and those data that must be built with the initial data.

Instead of relation, the statistical model structure the data in data-matrix with an undefined number of dimensions and several levels of groupings. Dimensions are expressed by criteria and

levels by nomenclatures. At each data-matrix is attached a set of summable and non-summable data. The elements of a data-matrix are sets of numerical values, which are the values of the set of summable and non-summable data.

For this model any query result is a data-matrix structured upon the same data model. This allows the statistician to build his macro-data himself at the aggregation level required, by successive uses of the system. This statistical data model is founded upon the assumption that expressing new data-matrices in order to calculate from those stored in the data base is a good way to query a statistical data base.

The structuration of data in data-matrix allows us to support a statistical DBMS by a file organisation of constant length record. So it is possible to see a data-matrix as a 'classical' statistical file, and there is no need of extracting data from the base to use a statistical package or already existent statistical programs.

This statistical model allows four abstraction levels (Smith (1977)): two generalisations to form the data-matrix, another one not explained here and an aggregation to form some families of data-matrix. Like the relationnal model, it allows [k:n] associations (Chen (1976)).

This model has been designed in order to build a management system for a very large data base: its size is about 10**8 bytes. One of our main purpose was to provide statisticians with an easy-to-use system. We are now trying to validate this data model by constructing a statistical DBMS upon it.

(see Bry (1984) for a detailed exposure and complete references)

References

Bry F. (1984),
Un modele de Donnees Statistique, rapport de recherche, I.R.T., to appear in september '84 (in french).

Chen P.P.S. (1976),
The Entity-Relationship Model - Toward a Uniform View of Data, ACM Transactions on Database Systems, Vol 1, No 1, march '76.

Gardarin G. (1982),
Bases de Donnees, Paris.

Tsichritzis D.C. and Lochovsky F.H. (1980),
Data Models, London.

Smith J.M. and Smith D.C.P. (1977),
Database Abstraction: Aggregation and Generalization, ACM Transactions on Database Systems, Vol 2, No 2, june '77.
