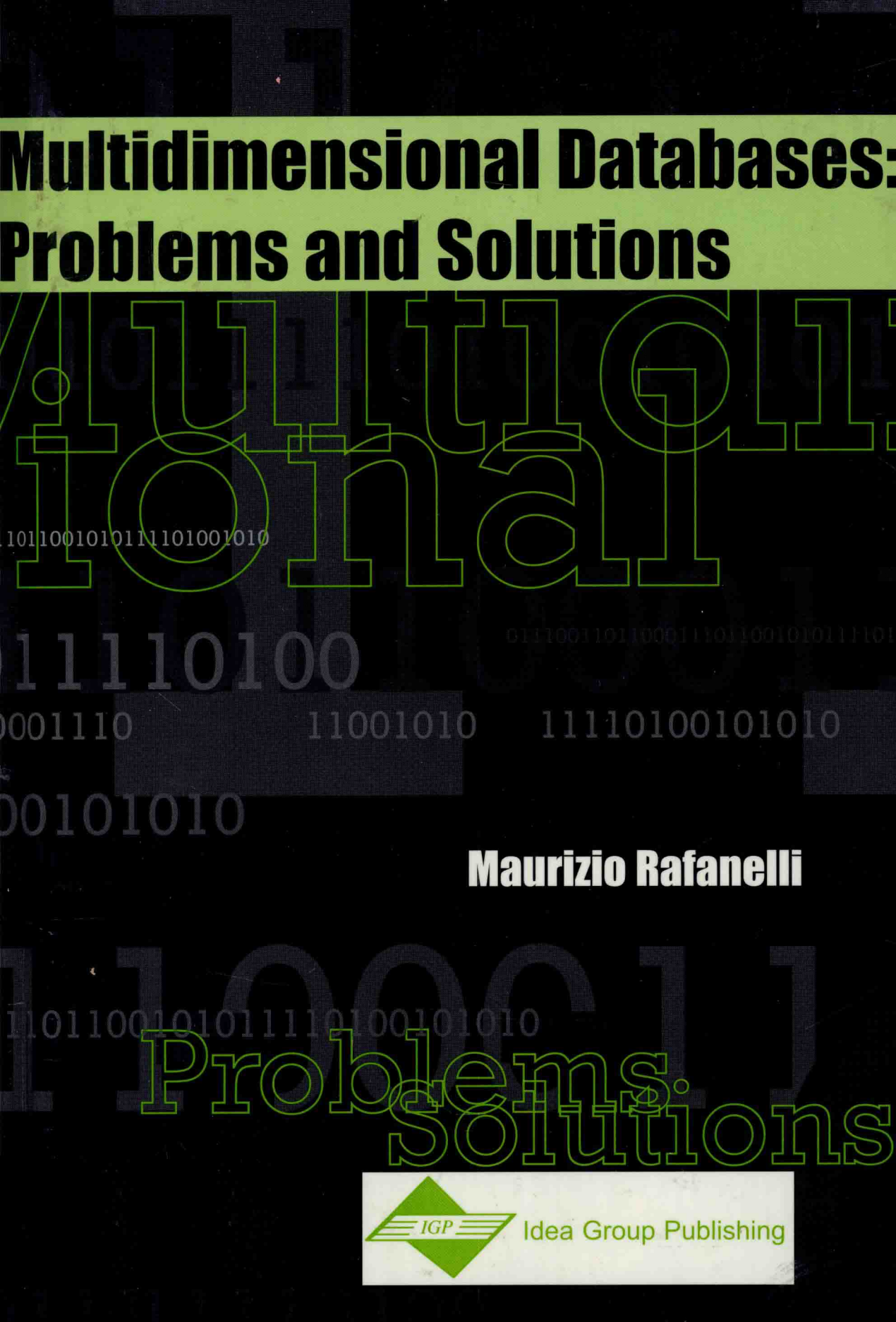
Multidimensional Databases: Problems and Solutions

Maurizio Rafanelli



IGP Idea Group Publishing

Multidimensional Databases: Problems and Solutions

Maurizio Rafanelli
Istituto di Analisi dei Sistemi ed Informatica – C.N.R., Italy

IDEA GROUP PUBLISHING
Hershey • London • Melbourne • Singapore • Beijing

Acquisition Editor:	Mehdi Khosrow-Pour
Senior Managing Editor:	Jan Travers
Managing Editor	Amanda Appicello
Development Editor:	Michele Rossi
Copy Editor:	Maria Boyer
Typesetter:	Amanda Lutz
Cover Design:	Weston Pritts
Printed at:	Integrated Book Technology

Published in the United States of America by
Idea Group Publishing (an imprint of Idea Group Inc.)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@idea-group.com
Web site: http://www.idea-group.com

and in the United Kingdom by
Idea Group Publishing (an imprint of Idea Group Inc.)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 3313
Web site: http://www.eurospan.co.uk

Library of Congress Cataloging-in-Publication Data

Rafanelli, M. (Maurizio)
Multidimensional databases : problems and solutions / Maurizio Rafanelli.
p. cm.
ISBN 1-59140-053-8 (hard cover) -- ISBN 1-59140-086-4 (ebook)
1. Multidimensional databases. I. Title.
QA76.9.D3 R219 2003
005.74--dc21

2002153246

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Foreword

Some months ago, I was invited by Idea Group Publishing to write a book on multidimensional databases because of my long experience in this area (more than 20 years). I accepted because I felt there was a need to create a point of reference to the most important issues in this particular field of research.

This presented me with the opportunity of writing a brief history of this field and of distinguishing between what is new in recent research and what is merely a renaming of old concepts. As many authors have rightly pointed out, a large number of concepts, operations, etc. had already been proposed in the '80s, appearing in papers, journals, conferences, workshops, and books. Important papers by Chin; Denning; Klug; Malvestuto; McCharty; Ghosh; Olken; Meral and Gultekin Ozsoyoglu; Rafanelli; Sato; Schlorer; Shoshani; Su; Tansel; and Wong document the history of this field.

Many of the most important and well-known researchers have contributed to this book, and each of them has written about their own specialist field. I hope, after many months of work, that this book will represent a milestone.

It is not only a review of past papers, but also of current research projects, and I hope it will encourage the search for new solutions to the many problems that are still open. There have been incredible advances in technology and ever-increasing demands from users in the most diverse applicative areas (finance, medicine, statistics, business, etc.). This presents a challenge for researchers to discover new methodological tools for investigating both old and new research areas in the database field and, in particular, that of multidimensional databases.

I am very proud to have had the cooperation of such distinguished scientists, who have given prestige and significance to this work, and I would like to thank them with all my heart for their valuable contribution.

Preface

The term "multidimensional data" generally refers to data in which a given *fact* is quantified by a set of *measures*, obtained by applying one more or less complex aggregative function (from count or sum to average or percent, and so on) to raw data. Such measures are characterized by a set of variables, called *dimensions*. In reality, a dimension often consists of a more complex structure than a simple variable, as we will see in the following chapter. Multidimensional data can be modeled by different representations, depending on the application field which uses them. For example, some years ago the term "multidimensional data" referred essentially to statistical data, that is, data whose use was (and is) basically for socio-economic analysis. The visual representation used most was (and is) the table (even if histograms, cakes, graphics, etc. are used too). Recently, the metaphor of the data cube, already proposed at the beginning of the '80s, was taken up again and used for new applications—such as On-Line Analytical Processing (OLAP)—which refer to business analysis.

Many of the problems concerning statistical multidimensional databases and many of the concepts defined in this context, especially if referring to models, operators, and algebras in general, have been up again and enlarged for new management system types and new applications, such as OLAP. Studies on this data type started at the beginning of the '80s. In this introduction we would like to give a brief history of the various topics in this research area, covering the period of the last 20 or more years.

At the beginning of the 1980s, a group of researchers began to look at some of the problems which arose when they considered data obtained by applying simple aggregation functions (count or sum) to row disaggregate data. This was the main reason that prompted some researchers to organize a workshop in California (1st SDBM (1981)) on the main topics of statistical databases (the term used, at that time, for multidimensional aggregate data). This name was coined because such data, organized in a database as flat files or multidimensional "tables," were mainly used to carry out statistical analysis or socio-economic type applica-

tions, such as census data on national production and consumption patterns, etc., as discussed in Shoshani (1982), Brown, Navathe, & Su (1983), and Shoshani & Wong (1985). However, they were also used for business applications, such as financial summary reports, sales forecasting, etc., as described in Wong (1984). Generally, these tables are represented by using two dimensions (for this reason they are also called "flat tables"), but, in reality, often a row and/or a column each consists of two or more dimensions.

As mentioned above, since 1981 a series of conferences have been organized: 1st SDBM (1981), 2nd SDBM (1983), 3rd SSDBM (1986), 4th SSDBM (1988), 5th SSDBM (1990), 6th SSDBM (1992), 7th SSDBM (1994), 8th SSDBM (1996), 9th SSDBM (1997), 10th SSDBM (1998), 11th SSDBM (1999), 12th SSDBM (2000), 13th SSDBM (2001), and 14th SSDBM (2002). Their purpose was to discuss issues of interest regarding statistical (and scientific) database management and to propose original solutions to the problems which arose in this area, both from the theoretical and from the application point of view. Recently, these conferences have also covered other new applications, in particular, on-line analytical processing and, in general, data warehousing. The papers presented at the conferences mentioned above cover practically all the topics of this area, from data models to new operators and relative query languages, from temporal summary table management to data privacy, from physical storage to metadata management, from graphic and visual interfaces to query optimization. Important issues on statistical databases are discussed in Shoshani (1982), Denning, Nichelson, Sande, & Shoshani (1983), Shoshani & Wong (1985), Rafanelli (1989), Rafanelli (1990), and in a book edited by Michalewicz (1990).

The first problem studied in the literature and regarding statistical multidimensional databases was that of privacy. At the end of the '70s, a few papers appeared in conferences and journals, for example, Yu & Chin (1977), Chin (1978), Denning D.E., Denning P.J., & Schwartz (1979). In them the authors began to study "how" to protect the individual privacy and, moreover, the privacy of companies, organizations, etc., which can be violated by clever manipulation of summary data. This manipulation could lead to the exact or approximate disclosure of confidential data or individuals. In the subsequent years other papers researched this topic further: Denning (1980), Schlorer (1980), Beck (1980), Schlorer (1981), Chin & Ozsoyoglu (1982), Denning (1983), Denning & Schlorer (1983), Chin (1986), Gusfield (1988), Adams & Wortmann (1989), Malvestuto, Moscarini, & Rafanelli (1991), and Michalewicz (1991). Recent proposals, such as Malvestuto & Moscarini. (1996b), Kao (1997), Chu (1997), Adam, Gangopadyay, & Holowczak. (1999), and Kleinberg, Papadimitriou, & Raghavan (2000), have been made to guarantee the privacy of individual records. These proposals include mechanisms of inference control, which can be said to be effective only if

sensitive summary data are neither explicitly nor implicitly released. There have also been proposals on how to answer complex queries (in the sense of structured data, such as a table). The privacy problem is widely discussed in Chapter 11 of this book.

During the '80s it was immediately evident that the above-mentioned (statistical) multidimensional tables had a more complex data structure than the conventional disaggregate data which were, in general, represented by relations (bi-dimensional tables in which the row, called "tuple," had the meaning of "instance of relationship" and the column represented an attribute of the same relation). From a given point of view, each attribute of the relation represented one of its "dimensions," even if in the literature the classic relations are considered as corresponding to 0-dimensional tables, all of whose attributes are essentially measure attributes, as affirmed in Gyssens & Lakshmanan (1997).

Instead, a statistical, multi-dimensional table was characterized by different elements. First of all, the concept of "summary attribute" was introduced; see, for example, the 1st SDBM (1981), the 2nd SDBM (1983), Shoshani (1982), Rafanelli & Ricci (1983), and Shoshani & Wong (1985). The summary attribute represents the result (summary) of aggregation operations performed on disaggregate data, as defined in Shoshani (1982), and in Bezenchek, Rafanelli, & Tininini (1996a), i.e., the "measure" of the aggregation is carried out. This point will be described in more detail in the following chapter. The previous attributes, called "category attributes" in order to distinguish them from the others, were enriched with a descriptive characteristic, inheriting and reinforcing the concept of table dimension.

The multidimensionality of the statistical tables was emphasized many years ago. For example, Rafanelli & Ricci (1983) wrote: "Category attributes represent a cross-product of an *n-dimensional space*...." Later, in Shoshani & Wong (1985), speaking on statistical databases, we find the following phrases: "In addition to statistical operators, such as sampling and aggregation, the access of the data is of a different nature. For example, it is quite common to access a region in *multidimensional space*, or to find materials with certain approximate properties, or cases that fit a statistical pattern. For such cases *multidimensional data structures* and search methods are desirable."

And in Fortunato, Rafanelli, Ricci, & Sebastio (1986), the authors affirm: "Every complex table can always be decomposed into different simple tables...The logical description of a simple table takes its *multidimensionality* into account." Again, Rafanelli (1991) wrote: "Multidimensionality is a dominant feature in SSDBs. Many of the characteristics and requirements of SSDBs can be traced to this feature...The difficulty of dealing with multidimensional spaces is further com-

pounded by the fact that each dimension can itself have a complex (usually hierarchical) structure."

Finally, in Rafanelli and Shoshani (1990a) and, subsequently, in Shoshani & Rafanelli (1991), we find the following phrases:" *Multi-dimensionality*; typically a multidimensional space defined by the category attributes, is associated with a single summary attribute (for example, the three-dimensional space defined by "state," "race," and "year" can be associated with "population." The implication is that a combination of values from "state," "race," and "year" (e.g., Alabama, Black, 1989) is necessary to characterize a single population value (e.g., 21,373)). *Classification hierarchies*—a classification relationship often exists for the categories. For example specific items: "fruit," "vegetable," "grain," etc. can be classified as "agricultural products"; "tables," "chairs," "sofas," etc. as "furniture." Moreover, "agricultural products," "furniture," etc. can themselves be classified as "products," thus forming a three-level classification hierarchy."

Recently studies have been made to discuss different problems regarding hierarchy. This topic will be discussed in more detail in Chapter 4. For example, the possibility of browsing along different levels of the same dimension, where each level represents the different granularity of the dimension instances, will be discussed. At the beginning, the possibility of describing the dimension of a table at a different level of aggregation (the above-mentioned granularity) and consequently the summary data described by this dimension, was seen as reclassifying a category attribute as another category attribute which represented the same more or less aggregate dimension, as written in Rafanelli & Ricci (1985). For example, the category attribute "months" of the dimension "time" could be *reclassified* as a new category attribute "quarter" by a function which groups the first three months into the first quarter, the second three months into the second quarter, and so on. Linked to the issue of the different dimensions which characterize and describe an aggregate data, there is the problem of the different classifications (see Malvestuto, Rafanelli, & Zuffada, 1988; Malvestuto & Zuffada, 1989) of the same dimension which appears, as a descriptive variable, in different databases. This topic will also be briefly discussed in Chapter 4.

Another issue which attracted the interest of scientific researchers from the beginning of the '80s was the conceptual data models for multidimensional databases and the operators to manipulate these data. The first graphical model, called SUBJECT, was proposed by Chan & Shoshani at the 7th VLDB Conference (Chan & Shoshani, 1981). Subsequently, different other graphical models were proposed: GRASS (see Rafanelli & Ricci, 1983; Rafanelli, 1987), SAM* (see Su, 1983), STORM (see Rafanelli & Shoshani, 1990a; Rafanelli & Shoshani, 1990b, 1990c; Shoshani & Rafanelli, 1991), ADAMO (see Bezenchek, Massari, & Rafanelli, 1994; Bezenchek, Rafanelli, & Tininini, 1996b; Tininini, Bezenchek,

& Rafanelli, 1996). This last model is briefly discussed in the following chapter. At the same time, models for summary data were proposed in Johnson (1981), Rafanelli & Ricci (1988), Chen, McNamee, & Melkanoff (1988), Sato (1991), Rafanelli & Ricci (1991), Malvestuto (1993), and Rafanelli & Ricci (1993).

More recently, a variety of formal multidimensional data models have been proposed, such as in Li & Wang (1996), Tininini, Bezenchek, & Rafanelli (1996), Agrawal, Gupta, & Sarawagi (1997), Datta & Thomas (1997), Gyssens & Lakshmanan (1997), Cabibbo & Torlone (1997), Lehner (1998), Vassiliadis (1998), Gingras & Lakshmanan (1998), Franconi & Sattler (1999), Pedersen & Jensen (1999), Nguyen, Tjoa, & Wagner (2000), and Pedersen, Jensen, & Dyreson (2001), both by academics and by industrial communities, even if there is not yet a broad consensus on a common terminology and formalism, as discussed in Chapter 3. At the same time, different proposals were presented regarding the algebras (operators with different characteristics from the classic relational operators, as discussed in Chapter 5, and the relative query languages with aggregate functions), for example, in Klug (1982), Fortunato, Rafanelli, Ricci, & Sebastio (1986), Ozsoyoglu, Ozsoyoglu, & Mata (1985), and Ozsoyoglu, Ozsoyoglu, & Matos (1987). Visual query languages were also proposed, such as in Wong & Kuo (1982), Rafanelli & Ricci (1990), and Meo Evoli, Rafanelli, & Ricci (1994). New elaborations or redefinitions of many of these operators have been recently proposed (see, for example, Gray, Bosworth, Layman, & Pirahesh, 1996; Gray et al., 1997; Gyssens & Lakshmanan, 1997; Cabibbo & Torlone, 1998; Lehner, 1998; Pedersen, Jensen, & Dyreson, 2001) with regard to OLAP applications, sometimes simply changing their name, other times extending the original operator, but, in particular, giving a rigorous formal definition of them. Similarities and differences between these two kinds of operators will be briefly discussed in Chapter 5.

Multidimensional querying is often based on the metaphor of the data cube and the concepts of facts, measures, and dimensions. It is often an exploratory process, performed by navigating along dimensions and measures, increasing/decreasing the level of detail, and focusing on specific subparts of the cube that appear "promising" for the required information. More recently, important results on query languages for bags (e.g., those in Albert, 1991; Libkin & Wong, 1993; Grumbach & Milo, 1996) have led to a more "natural" characterization of aggregate functions. Several approaches to the problem of querying multidimensional data have been based on extensions of the relational algebra and calculus and/or of SQL, the most common relational query language. In Chapter 9, techniques to retrieve multidimensional (aggregate) data and problems of evaluation, related to the efficient data retrieval and calculation (known as the problem of *rewriting a query using views*), are discussed.

Another problem which arose at the beginning of the '80s was the study of metadata. One of the first papers in which this issue was discussed and this important concept defined was McCarthy (1982). In it the author provides an excellent overview of metadata, including statistical metadata. As it is well known, metadata are data about data, i.e., systematic descriptive information about data content and organization. Metadata are necessary to specify information about multidimensional aggregate statistical data. They can provide the definition of logical models, as well as a more detailed documentation. Well-defined and differentiated metadata are necessary to allow software links between different logical and physical representations; between multidimensional aggregate statistical databases, application programs, and user interfaces; as well as between multiple distributed and heterogeneous systems (see 2nd SDBM, 1983; Rafanelli, 1991). Studies on metadata for multidimensional aggregate statistical data can be found in McCarthy (1982) and Wong (1984).

A problem which seriously impedes the integrated use of databases, and, in particular, the manipulation of multidimensional aggregate databases, is the use of different units of measure, for example, dollars or pesetas for currency, pounds or Kilograms for fruit production, and so on. There have been few studies on this issue. It was discussed in general in Karr & Loveman (1978) and Gehani (1982). Other authors, like Sparr (1981) and Rafanelli, Bezenchek, & Tininini (1996), studied this problem, focusing on statistical multidimensional aggregate databases.

Also the problem regarding the counting unit (i.e., if the counting unit is 106 and a summary value "10" actually represents "10 million") was briefly discussed in Rafanelli, Bezenchek, & Tininini (1996) and will be discussed in Chapter 1.

The management of the time in MDDB is a topic which was initially studied in Rafanelli (1990), and subsequently, in Yang & Widom (1998, 2000), and in Mendelzon & Vaisman (2000, 2001). One problem which arises is the changing of the semantics of a given term at two different moments in time. For instance, suppose we have two states, S1 and S2, that, at time t1, have a given boundary. If at time t2 state S1 gives state S2 part of its territory, then the boundary, the area, the population, etc. of the two states change, but their names remain the same. This means that we call objects, which are different in time, by the same name.

In recent years there has been growing interest in multidimensional database systems, essentially to perform on-line analysis of transaction-based business data, such as retail store transactions, and to consolidate and summarize data in order to put enterprise data into multidimensional perspectives. OLAP is considered a tool for developing business analysis and decision support applications. As explained in many papers, OLAP is an area of active commercial and research interest which developed in recent years, especially with the advances in hardware for on-line mass storage which have made the warehousing of large amounts

of data possible. One of the goals of OLAP tools is to provide fast answers to queries which aggregate the warehouse data. The term OLAP was coined in Codd, Codd, & Salley (1993), in order to characterize the dynamic aspects (with particular reference to performance issues) of such data. Subsequently there were other papers on OLAP applications, for example, Codd, Codd, & Salley (1993), Finkelstein (1995), and Priebe & Pernul (2000). Traditional data models, such as the ER model (Chen, 1976) and the relational model (Codd, 1970), do not provide good support for OLAP applications. As a result, new data models based on a multi-dimensional view of data have emerged. In all the chapters in this book, OLAP is referred to from different points of view: basic concepts, models, operators, time management, dynamic data cubes, materialized views, querying, incomplete information , etc. are some of the topics discussed.

Another issue to which some researchers have focused their attention on is the *interoperability* between multidimensional (aggregate) data and other data types, essentially geographic data. In fact, at present, neither models or algebras which include the possibility to represent, store, and manage complex data of this type exist. Consequently, it is necessary to study how to answer queries which have both multidimensional operators, and topological operators, i.e., queries of the type:

1. "Select the regions which border on the Tuscany region and in which the oil production was greater than 10,000 hectolitres in the years from 1990 to 1999," or
2. "Select the regions crossed by the Danube River and in which the average fruit production in a given year was greater than the previous year, in the period 1990-1999."

This problem is widely discussed in Chapter 13.

The aim of this book is to provide the first text on multidimensional databases and on the main topics of this area, discussing properties and peculiarities of such databases. Various well-known researchers, who are experts in one or more fields of multidimensional databases, were personally invited to write a chapter on their research area for this volume. As a result, this book presents the history, the current state, and future trends of each of the above-mentioned areas.

In particular, in Chapter 1, Maurizio Rafanelli presents the basic notions regarding multidimensional (aggregate) databases by referring to different definitions given for them in the literature. He discusses the differences which exist between the disaggregate, microdata and the aggregate, macrodata, and the importance, in this context, of the macrodata. He presents the process which makes it possible to obtain aggregate data (macrodata), from a very large set of raw data (microdata) (see Rafanelli, Bezenchek, & Tininini, 1996). He also introduces defi-

nitions of the different data structures proposed in the literature for multidimensional aggregate data (statistical object, table, cube, multidimensional aggregate data [MAD]). In this context he also introduces the concept of dimension and of the possible hierarchies which could be present in each dimension and which characterize the measure of the multidimensional aggregate data. Finally, he describes the different approaches used in the varying fields of application, such as the statistical environment and On-Line Analytical Processing (OLAP). Both of them use these data, but the approaches mentioned above depend on the different aspects emphasized in them. For example, most of the work they carry out is characterized by their change of use (from statistical and socio-economic type applications to the analysis of transaction-based business data). At the end of the chapter, the author explains the necessity to use a graph model to represent multidimensional aggregate data and discusses, in particular, the ADAMO model. Finally, he discusses tabular models, used both in SDB and OLAP, and gives a set of definitions regarding OLAP terminology.

Arie Shoshani, in Chapter 2, discusses the multidimensionality in statistical, OLAP, and scientific databases. He affirms that the term "multidimensional database" typically refers to a collection of objects, each represented as a point in a multidimensional space. Even data that is represented in a tabular form, such as relations, can be thought of as multidimensional data, if each row (tuple) is considered an object, and the columns (attributes) are considered the dimensions. The problem of viewing high-dimensional data to identify clusters, outliers, and various patterns has been the subject of several research projects. An important reason for viewing data in the multidimensional space is summarization. This need is most obvious in databases that represent statistical data or in databases used for decision support. These are referred to as "Statistical Databases" and "On-Line Analytical Processing" (OLAP), respectively. In the OLAP literature the multidimensional space is referred to as a "cube." Therefore, the Author affirms that another important aspect which links Statistical and OLAP databases is that each dimension can have a category hierarchy associated with it. Dimension hierarchies can become fairly complex depending on the type of dimension.

In Chapter 3, Riccardo Torlone presents the requirements that an ideal conceptual multidimensional data model should fulfill. Because it is widely accepted that traditional conceptual data models, like the entity-relationship model, are not well suited to describe the multidimensional and aggregative nature of OLAP applications, a variety of multidimensional data models have been recently proposed (both by the academic and industrial communities), but it should be said that a consensus on a common terminology and formalism has not emerged yet. Then, he presents a first version of the above-mentioned ideal conceptual multidimensional data model from a general point of view. Far from being complete, this

model aims to capture the core of the various multidimensional data models proposed in the scientific literature and the means adopted by OLAP systems to represent and manipulate data. The model relies on two main and agreed concepts: dimension and cube. A dimension represents a business perspective under which data analysis is to be performed and is organized in a hierarchy of levels. The levels of a dimension correspond to data domains at different granularity. A cube represents factual data on which the analysis is focused, and associated measures with coordinates, defined over a set of dimension levels. Therefore, a number of extensions of this model and a survey on various multidimensional models proposed in the literature are discussed.

In Chapter 4, Elaheh Pourabbas and Maurizio Rafanelli discuss the different types of hierarchies which can appear in a dimension of a multidimensional (aggregate) data structure. They study their behavior when the user browses along them or when he manipulates them by using operators, such as roll-up, drill-down, etc. Moreover, they classify four different basic types of hierarchies and three forms of data abstraction, and discuss their characteristics. These hierarchies divide a single dimension into different levels of aggregation. Depending on them, the authors discuss the characteristics of some OLAP operators which refer to hierarchies in order to maintain the data cube consistency. In particular, they discuss the important concept of a summarizable function and different types of mapping between two category attributes (which represent the instance of a dimension in a multidimensional data structure). Then, they propose a set of operators for changing the hierarchy structure. The issues discussed provide modeling flexibility during the schema design phase and correct data analysis. In particular, they deal more with multiple hierarchies, and introduce the multiplicity of a hierarchy as a semantic variant of a simple one.

In Chapter 5, Maurizio Rafanelli discusses the set of operators proposed by different authors in the literature. He draws distinctions between operators for statistical aggregate data (and between the relational approach and the multidimensional approach) and the operators for OLAP. A comparison between the two sets is also made, highlighting their similarities and differences. In particular, he discusses the operators for multidimensional aggregate data which extend relational algebra and relational calculus, then the operators for multidimensional aggregate data defined in a tabular environment, and finally the operators for OLAP applications. A comparison between the OLAP terminology and the multidimensional aggregate (statistical) database terminology is also made. Finally, the author resumes giving particular emphasis to the operators which form the skeleton of the basic algebra.

In Chapter 6, Alberto Mendelzon and Alejandro Vaisman show the need for a temporal approach to OLAP, giving a review of temporal database concepts

and work regarding time in multidimensional databases. They argue that, in the presence of dimension updates which trigger changes on the granularity of the facts, a temporal model supporting schema versioning is required. They show that existing data models are not suitable for these requirements. Therefore, they present a new temporal multidimensional model which makes it possible to keep track of the history of a multidimensional database, and introduce a temporal query language, called *TOLAP*, supporting this temporal model. A preliminary *TOLAP* implementation is described, and the results of experiments are commented on. At the end of the chapter, the authors suggest possible research directions.

In Chapter 7, Mirek Riedewald, Divyakant Agrawal, and Amr El Abbadi discuss several techniques for update-efficient aggregation on MOLAP data cubes. First they give an introduction and a background, focusing on data cubes and aggregate query types. Then they define an invertible aggregate operator and affirm that the existence of inverse operations enables the construction of elegant techniques for speeding up queries on MOLAP data cubes that do not require additional storage as opposed to materializing the original cube. Then very sparse data set problems are briefly discussed and the Progressive Data Cube (pCube) is proposed and described, in order to incorporate any hierarchical multidimensional index structure, e.g., R-tree, quad tree, etc. Following these are techniques that explicitly address the sparseness issue are presented. As the authors say, "the similarity between all approaches is that they try to find an appropriate balance between query, update, and storage cost. While earlier proposals mostly focused on query and storage aspects, large data sets with frequent updates created a need for more dynamic solutions."

In Chapter 9, Stefano Paraboschi, Giuseppe Sindoni, Elena Baralis and Ernest Teniente present the view selection problem, first showing how a materialized view can be a query optimization solution applicable to generic databases. A general approach can be developed as an extension of multi-query optimization techniques, where it is important to detect the query parts that are common to separate query plans, in order to compute them only once but reuse their results many times. View materializations have proven to be an important query optimization solution for multidimensional databases. The reasons for the success of view materialization in multidimensional databases are twofold: first, the extremely high ratio between queries and updates, typical of all data analysis systems, makes the task to maintain all the materialized views considerably easier; also, the limited variety of queries typically supported by multidimensional databases permits the design of sophisticated models that are able to represent well the contribution that a view materialization can offer to view computation.

In particular, this chapter presents a model where a star schema is enriched with functional dependencies that identify hierarchies among the dimensional at-

tributes. Queries, and potential view materializations, can be represented as nodes on a lattice. Ordering on lattice nodes means that a query can be computed using another query. Typical multidimensional database operations like roll-up and drill-down correspond nicely to typical lattice operations: meet and join. In the second part of the chapter the authors present the important characteristics of the solutions for materialized views that have been devised for commercial DBMS-based analysis tools. The chapter will base this analysis on the papers by research groups of DBMS vendors that have appeared in technical conferences and on the analysis of the technical documentation accompanying these systems.

In Chapter 9, Leonardo Tininini discusses techniques to retrieve multidimensional (aggregate) data proposed in the literature and which are based on the idea of determining the cube of interest and then navigating along the dimensions, increasing or decreasing the level of detail or selecting specific subparts of the cube. He gives a brief presentation of query languages based on an extension of the relational algebra and those based on a calculus (again an extension of the relational one), where queries are expressed in a more declarative way. He also presents visual languages, which usually rely on an underling algebra or calculus, and are based on a more interactive and iconic querying paradigm. In addition, he focuses on the problem of query evaluation, i.e., issues related to efficient data retrieval and calculation, possibly (often necessarily) using precomputed data, known in the literature as the problem of *rewriting a query using views*. He also presents some results on the equivalence and rewriting of aggregate queries that can be used to optimize the evaluation of queries on multidimensional data. Finally, he briefly outlines how specific techniques of indexing can significantly improve query evaluation in the multidimensional context.

In Chapter 10, Curtis E. Dyreson, Torben B. Pedersen, and Christian S. Jensen, after a brief introduction and a background of the topic, explain uncertainty in the data and in the aggregated data/metadata, and then discuss future directions of research. In particular, they present techniques for handling incompleteness in base data, derived data, and metadata. With regard to the incomplete measures, the authors outline the problems posed by three of the kinds of incomplete information that could appear in a measure attribute: unknown, imprecise, and probabilistic values. Then they discuss incompleteness in a grouping of attributes, focusing on problems and techniques for the grouping of exclusive and inclusive disjunctive values only. The study of incomplete information in a hierarchy follows. In it the incomplete derived data are considered and some additional strategies for improving the responsiveness of queries on the incomplete data are briefly discussed. Finally, incomplete information in metadata is discussed, with regard to non-covering, non-onto, and non-strict hierarchies.

In Chapter 11, Marina Moscarini and Francesco M. Malvestuto discuss the compromise of individual privacy through statistical queries on summary tables, reviewing some recent results on the inference problem which only refer to simple count and sum queries with data of real and non-negative real types. Attacks on confidentiality come from sensitive statistical queries whose answers allow a knowledgeable user to determine exactly or estimate accurately the value of a confidential field in some individual records. To guarantee privacy of individual records, a mechanism of inference control must be embodied in the (statistical) multidimensional aggregate database interface according to a security policy, which can be said to be effective only if sensitive summary statistics are neither explicitly nor implicitly released. The different control mechanisms proposed are analyzed as to their effectiveness and efficiency. In this chapter, the authors also review some recent results on the computational complexity of the inference problem for additive queries on a statistical database, with regard to simple count and sum queries, which will be referred to as simple "additive queries," with data of real and non-negative real types. They met with some computationally hard problems (on strong safety and strong p-safety), which are likely to have no efficient solutions, and conclude that there still remains much work to do to solve the problem of the security of statistical databases in a satisfactory way. Future and emerging research trends are listed to answer all the complexity questions raised by the inference problem discussed in this chapter, as well as the case of data of integer and non-negative integer types. Other lines of research will have to cover multidimensional tables and special classes of hypergraphs for which safety tests can be efficiently worked out.

In Chapter 12, Andrea Calí, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati revise the various approaches for dealing with schema and data integration in data warehousing. They explain why data integration is important in data warehousing, present a comprehensive framework for data integration in data warehousing, and then survey existing approaches both from an academic and from an industrial perspective. They affirm that a fundamental aspect in the design of a data warehouse system is the process of acquiring the raw data from a set of relevant information sources. They will call the component of a data warehouse system that deals with this process a "source integration system." The main goal of a source integration system is to deal with the transfer of data, from the set of sources constituting the application-oriented operational environment, to the data warehouse. Since sources are typically autonomous, distributed, and heterogeneous, this task has to deal with the problem of cleaning, reconciling, and integrating data coming from the sources. The design of a source integration system is a very complex task, which comprises several different issues. The authors high-