Data-Centric Systems and Applications

Elzbieta Malinowski · Esteban Zimányi

Advanced Data Warehouse Design

From Conventional to Spatial and Temporal Applications

With 152 Figures and 10 Tables

Springer

Elzbieta Malinowski
Universidad de Costa Rica
School of Computer &
Information Science
San Pedro, San José
Costa Rica
emalinow@cariari.ucr.ac.cr

Esteban Zimányi
Université Libre Bruxelles
Dept. of Computer &
Decision Engineering (CoDE)
Avenue F.D. Roosevelt 50
1050 Bruxelles
Belgium
ezimanyi@ulb.ac.be

2nd corrected printing 2009

ISBN: 978-3-540-74404-7 e-ISBN: 978-3-540-74405-4

Library of Congress Control Number: 2007941791

ACM Computing Classification: H.2, H.3, H.4, J.1, J.2

Cover design: KünkelLopka, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To Yamil,
my one and only love and husband,
who throughout 25 years has been
a patient friend, an untiring supporter,
and a wise adviser
E.M.

To Fabienne and Elena,
with all my love and tenderness,
for all the joy and the beautiful
moments that I share with them
E.Z.

Foreword

When I was asked to write the foreword of this book, I was both honored and enthusiastic about contributing to the first book covering the topic of spatial data warehousing. In spite of over ten years of scientific literature regarding spatial data warehousing, spatial data mining, spatial online analytical processing, and spatial datacubes, Malinowski and Zimányi are the first to invest the energy necessary to provide a global view of spatial data warehousing in a coherent manner. This book introduces the reader to the basic concepts in an academic style that remains easy to read. Although one might think initially that there is nothing special about spatial data warehouses, this book shows the opposite by introducing fundamental extensions to the development of spatial data warehouses. On the one hand, there are extensions required for the design and implementation phases, and on the other hand there are extensions that improve the efficiency of these phases. This book offers a broad coverage of these extensions while at the same time guiding the reader to more detailed papers, thanks to a representative bibliography giving a sample of the main papers in the field. In addition, this book contains a glossary that is very useful for the reader since the datacube paradigm (also called analytical or multidimensional database paradigm) is very different from the traditional database paradigm (also called transactional or operational database paradigm).

Of particular interest in this book is the coverage of spatial datacubes, of their structure, and of implementation details in a data warehouse perspective. In particular, the different types of spatial hierarchies, spatial measures, and spatial data implementations are of interest. As a technology of choice to exploit data warehouses, OLAP technology is discussed with examples from Microsoft SQL Server Analysis Services 2005 and from Oracle 10g OLAP Option. Spatial OLAP (SOLAP) is also mentioned for spatial data warehouses.

In addition to defining spatial and temporal components of datacubes, the book proposes a method that is based on an extended entity-relationship approach, the spatial and temporal extensions of the MADS model, and a waterfall-based design approach. Readers already using the entity-relationship

approach will immediately benefit from the proposed concepts. Nevertheless, all readers should benefit from reading this book, since the underlying fundamental concepts can be transposed to UML or other proprietary modeling languages. Considering past contributions of Malinowski and Zimányi in the field, this book provides an up-to-date solution that is coherent with their previous work. This fact alone is worth the interest of the reader since it contributes to the building and dissemination of knowledge in our field.

Typical of design method books where authors suggest their own recipe, this book proposes one approach covering both theoretical concepts and design phases in an extensive and coherent manner. This is a welcome contribution despite the fact that different approaches have always existed in system design and will always exist, explicitly or implicitly. This is necessary for method improvement and for different application contexts. Habitually, these methods converge towards the issues to tackle (the "what" and "why") while they typically diverge about the ways to solve these issues (the "how" and "when"). Consequently, one benefit of this book is to provide the readers with the necessary background to explore other elements of solutions beyond those cited. Typically, these solutions are not as well structured as this book, since knowledge of a research team is fragmented over different papers written at different times. Personally, I with my research team must confess to doing exactly such piecemeal dissemination regarding the Perceptory alternative in spite of years of practical usage in over 40 countries thanks to UML, geospatial standards, and adaptability to designers' own agile or waterfall methods. In this regard, the readers will be very grateful to Malinowski and Zimányi for having grouped under one cover, in a coherent and homogeneous manner, their views about conventional data warehouses, spatial data warehouses, and temporal data warehouses.

Finally, the arrival of the first book on a topic (other than a collaborative collection of chapters) can always be considered as a milestone which indicates that a field has reached a new level of maturity. As such, this book is a significant milestone on the SOLAP portal timeline (http://www.spatialbi.com) and recognizes Malinowski and Zimányi's contributions to the field.

Yvan Bédard
Creator of the acronym SOLAP and the first SOLAP commercial technology
Holder of the Canada NSERC Industrial Research Chair in Geospatial Databases for Decision Support
Department of Geomatics Sciences & Centre for Research in Geomatics
Université Laval
Québec City, Canada

Preface

Data warehouses are databases of a specific kind that periodically collect information about the activities being performed by an organization. This information is then accumulated over a period of time for the purpose of analyzing how it evolves and, thus, for discovering strategic information such as trends, correlations, and the like. Data warehouses are increasingly being used by many organizations in many sectors to improve their operations and to better achieve their objectives. For example, a data warehouse application can be used in an organization to analyze customers' behavior. By understanding its customers, the organization is then able to better address their specific needs and expectations.

This book advocates a *conceptual* approach to designing data warehouse applications. For many decades, databases have been designed starting from a conceptual-design phase. This phase purposely leaves out any implementation considerations in order to capture as precisely as possible the requirements of the application from the users' perspective. When this has been done, a logical-design phase aims at translating these requirements according to a particular implementation paradigm, such as the relational or the object-relational paradigm. In the final stage, specific considerations concerning a particular implementation platform are taken into account in order to build an operational application. This separation of concerns offers many advantages and has contributed significantly to the widespread use of database applications in almost every aspect of our everyday life.

The situation is different for data warehouses, which are typically designed starting at a logical level, followed then by a physical-design phase. This state of affairs precludes an accurate description of users' analytical requirements, in particular because these requirements are driven by the technical limitations imposed by current data warehouse tools. There has recently been increased interest in the research community in devising conceptual models for data warehouse applications. Nevertheless, this area is still at an early stage and there is not yet a consensus about the specific characteristics that such a conceptual model should have.

This book presents the MultiDim model, a conceptual model for data warehouse design. We used as a starting point the classical entity-relationship model, thus taking into account the experience gained in more than four decades of applying conceptual modeling to traditional databases. The MultiDim model was designed in order to cope with the specific requirements of data warehouses and analytical applications. Therefore, it adopts a multidimensional view of information: *measures*, which are indicators that allow one to evaluate specific activities of an organization, are analyzed using various *dimensions* or perspectives. In addition, measures under analysis can be expressed at various levels of detail with the help of *hierarchies*, which define multiple abstraction levels for the dimensions.

The book covers the design of conventional data warehouses and also addresses two innovative domains that have recently been introduced to extend the capabilities of data warehouse systems, namely, the management of spatial and temporal (or time-varying) information.

Spatial information allows us to describe how objects are located in space: persons live in specific locations and go to other locations for their work and for their leisure activities. These locations belong to particular subdivisions of the Earth's surface, such as counties, states, countries, and continents. Spatial information has been successfully used for many years in a wide range of application domains, such as environmental applications, cadastral and land use management, utilities management, and transportation and logistics, to name but a few. However, current data warehouse systems do not provide support for managing spatial information and, thus, location information must be represented using traditional alphanumeric data types. The extension of data warehouses in order to support spatial information opens up a new spectrum of possibilities, since this allows the inherent semantics of spatial information to be taken into account. For example, this could allow one to monitor the evolution of urban green areas and their impact on biodiversity, and the evolution of the use of public transportation systems with respect to the use of private cars. In addition, displaying the results of analysis using cartographic representations enables the discovery of information that would be impossible to obtain otherwise. We define in this book a conceptual model for designing spatial data warehouses. This proposal has its roots in the work that has been performed in the fields of spatial databases and geographic information systems.

On the other hand, temporal information captures the evolution in time of real-world phenomena: persons change their homes or their jobs, which induces a modification of their behavior, new products and technologies appear while others are discontinued, etc. While current data warehouses allow one to keep track of the evolution of measures under analysis, they are not able to capture the evolution of the dimensions used for analyzing those measures. The need to support the evolution in time of all elements composing a data warehouse has been acknowledged for many years. For example, if reorganizations of sales districts are performed periodically to adapt to changing market conditions,

it is necessary to keep track of these changes in order to analyze their impact on sales. However, the solutions provided so far for dealing with these issues are not satisfactory. We propose in this book a conceptual model for designing temporal data warehouses. Our proposal is based on work performed in the domain of temporal databases.

While this book proposes a conceptual approach to designing data warehouses, we also discuss how the conceptual specifications can be translated into logical and physical specifications for implementation with current data warehouse systems. We have chosen Microsoft Analysis Services 2005 and Oracle 10g as representative platforms for the implementation of our conceptual model. For conventional and temporal data warehouses, we have envisaged implementation on both platforms. However, for spatial data warehouses, we have used only Oracle 10g as the target platform, since a spatial extension of the database management system is required. By providing effective implementations of our conceptual model, we aim at promoting the use of the conceptual approach in data warehouse design.

As experience has shown, designing data warehouse applications is a complex and costly process. Therefore, providing methodological support for this task is of paramount importance. This book provides a method that incorporates conceptual design into the data warehouse design process. We cover two alternative, and complementary, approaches to dealing with this problem. The first one, called *analysis-driven design*, starts from the analysis requirements of decision-making users. The second approach, called *source-driven design*, starts by analyzing the information in the sources providing data to the warehouse. These two approaches can be combined into an iterative development process in order to ensure the correspondance of the analysis requirements with the available information. Obviously, the method must also take into account the specificities of the spatial or temporal information. However, since data warehouse design is a relatively recent domain, our proposal constitutes only a preliminary step in this respect; more research must be done as well as an evaluation of our method in a variety of data warehouse design projects.

This book is targeted at practitioners, graduate students, and researchers interested in the design issues of data warehouse applications. Practitioners and domain experts from industry with various backgrounds can benefit from the material covered in this book. It can be used by database experts who wish to approach the field of data warehouse design but have little knowledge about it. The book can also be used by experienced data warehouse designers who wish to enlarge the analysis possibilities of their applications by including spatial or temporal information. Furthermore, experts in spatial databases or in geographic information systems could profit from the data warehouse vision when they are building innovative spatial analytical applications. The book can also be used for teaching graduate or advanced undergraduate students, since it provides a clear and a concise presentation of the major concepts and results in the new fields of conceptual data warehouse design and of spatial and temporal data warehouses. Finally, researchers can find an introduction

to the state of the art on the design of conventional, spatial, and temporal data warehouses; many references are provided to deepen their knowledge of these topics.

The book should allow readers to acquire both a broad perspective and an intimate knowledge of the field of data warehouse design. Visual notations and examples are intensively used to illustrate the use of the various constructs. The book is purposely written in an informal style that should make it readable without requiring a specific background in computer science, mathematics, or logic. Nevertheless, we have taken special care to make all definitions and arguments as precise as possible, and to provide the interested reader with formal definitions of all concepts.

Support material for the book has been made available online at the address http://cs.ulb.ac.be/ADWDbook/. This includes electronic versions of the figures and pedagogic material that can be used by instructors using this book as a course text.

We would like to thank the Cooperation Department of the Université Libre de Bruxelles, which funded the sojourn of Elzbieta Malinowski in Brussels; without its financial support, this book would never have been possible. Parts of the material included in this book have been previously presented in conferences or published in journals. At these conferences, we had the opportunity to have discussions with research colleagues from all around the world, and we exchanged various views about the subject with them. The anonymous reviewers of these conferences and journals provided us with insightful comments and suggestions that contributed significantly to improving the work presented in this book. We are also grateful to Serge Boucher, Boris Verhaegen, and Frédéric Servais, researchers in our department, who helped us to enhance considerably the figures of this book. Our special thanks to Yvan Bédard, professor at the Laval University, who kindly agreed to write the foreword for this book, even when it was only in draft form. Finally, we would like to warmly thank Ralf Gerstner from Springer for his continued interest in this book. The warm welcome given to our book proposal and his enthusiasm and encouragement throughout its writing helped significantly in giving us impetus to pursue our project to its end.

Elzbieta Malinowski
Esteban Zimányi
October 2007

Contents

1 Introduction 1
1.1 Overview 2
1.1.1 Conventional Data Warehouses 2
1.1.2 Spatial Databases and Spatial Data Warehouses 4
1.1.3 Temporal Databases and Temporal Data Warehouses . . 5
1.1.4 Conceptual Modeling for Databases and Data Warehouses 6
1.1.5 A Method for Data Warehouse Design 7
1.2 Motivation for the Book 8
1.3 Objective of the Book and its Contributions to Research 11
1.3.1 Conventional Data Warehouses 12
1.3.2 Spatial Data Warehouses 13
1.3.3 Temporal Data Warehouses 13
1.4 Organization of the Book 14
Review Questions 16

2 Introduction to Databases and Data Warehouses 17
2.1 Database Concepts 18
2.2 The Entity-Relationship Model 19
2.3 Logical Database Design 23
2.3.1 The Relational Model 23
2.3.2 The Object-Relational Model 32
2.4 Physical Database Design 37
2.5 Data Warehouses 40
2.6 The Multidimensional Model 43
2.6.1 Hierarchies 44
2.6.2 Measure Aggregation 46
2.6.3 OLAP Operations 47
2.7 Logical Data Warehouse Design 49
2.8 Physical Data Warehouse Design 51
2.9 Data Warehouse Architecture 55

2.9.1 Back-End Tier . . . 56
2.9.2 Data Warehouse Tier . . . 57
2.9.3 OLAP Tier . . . 58
2.9.4 Front-End Tier . . . 58
2.9.5 Variations of the Architecture . . . 59
2.10 Analysis Services 2005 . . . 59
2.10.1 Defining an Analysis Services Database . . . 60
2.10.2 Data Sources . . . 61
2.10.3 Data Source Views . . . 61
2.10.4 Dimensions . . . 62
2.10.5 Cubes . . . 64
2.11 Oracle 10g with the OLAP Option . . . 66
2.11.1 Multidimensional Model . . . 67
2.11.2 Multidimensional Database Design . . . 68
2.11.3 Data Source Management . . . 69
2.11.4 Dimensions . . . 70
2.11.5 Cubes . . . 71
2.12 Conclusion . . . 73
Review Questions . . . 73

3 Conventional Data Warehouses . . . 77
3.1 MultiDim: A Conceptual Multidimensional Model . . . 78
3.2 Data Warehouse Hierarchies . . . 81
3.2.1 Simple Hierarchies . . . 83
3.2.2 Nonstrict Hierarchies . . . 89
3.2.3 Alternative Hierarchies . . . 95
3.2.4 Parallel Hierarchies . . . 95
3.3 Advanced Modeling Aspects . . . 99
3.3.1 Modeling of Complex Hierarchies . . . 99
3.3.2 Role-Playing Dimensions . . . 101
3.3.3 Fact Dimensions . . . 103
3.3.4 Multivalued Dimensions . . . 103
3.4 Metamodel of the MultiDim Model . . . 108
3.5 Mapping to the Relational and Object-Relational Models . . . 109
3.5.1 Rationale . . . 109
3.5.2 Mapping Rules . . . 110
3.6 Logical Representation of Hierarchies . . . 114
3.6.1 Simple Hierarchies . . . 114
3.6.2 Nonstrict Hierarchies . . . 122
3.6.3 Alternative Hierarchies . . . 125
3.6.4 Parallel Hierarchies . . . 125
3.7 Implementing Hierarchies . . . 126
3.7.1 Hierarchies in Analysis Services 2005 . . . 126
3.7.2 Hierarchies in Oracle OLAP 10g . . . 128
3.8 Related Work . . . 130

3.9 Summary . . . 132
Review Questions . . . 134

4 Spatial Data Warehouses . . . 137
4.1 Spatial Databases: General Concepts . . . 138
4.1.1 Spatial Objects . . . 138
4.1.2 Spatial Data Types . . . 138
4.1.3 Reference Systems . . . 140
4.1.4 Topological Relationships . . . 140
4.1.5 Conceptual Models for Spatial Data . . . 142
4.1.6 Implementation Models for Spatial Data . . . 142
4.1.7 Models for Storing Collections of Spatial Objects . . . 143
4.1.8 Architecture of Spatial Systems . . . 144
4.2 Spatial Extension of the MultiDim Model . . . 145
4.3 Spatial Levels . . . 147
4.4 Spatial Hierarchies . . . 147
4.4.1 Hierarchy Classification . . . 147
4.4.2 Topological Relationships Between Spatial Levels . . . 153
4.5 Spatial Fact Relationships . . . 156
4.6 Spatiality and Measures . . . 157
4.6.1 Spatial Measures . . . 157
4.6.2 Conventional Measures Resulting from Spatial Operations . . . 160
4.7 Metamodel of the Spatially Extended MultiDim Model . . . 161
4.8 Rationale of the Logical-Level Representation . . . 163
4.8.1 Using the Object-Relational Model . . . 163
4.8.2 Using Spatial Extensions of DBMSs . . . 164
4.8.3 Preserving Semantics . . . 165
4.9 Object-Relational Representation of Spatial Data Warehouses . 166
4.9.1 Spatial Levels . . . 166
4.9.2 Spatial Attributes . . . 168
4.9.3 Spatial Hierarchies . . . 169
4.9.4 Spatial Fact Relationships . . . 174
4.9.5 Measures . . . 176
4.10 Summary of the Mapping Rules . . . 178
4.11 Related Work . . . 179
4.12 Summary . . . 182
Review Questions . . . 183

5 Temporal Data Warehouses . . . 185
5.1 Slowly Changing Dimensions . . . 186
5.2 Temporal Databases: General Concepts . . . 189
5.2.1 Temporality Types . . . 189
5.2.2 Temporal Data Types . . . 190
5.2.3 Synchronization Relationships . . . 191

5.2.4 Conceptual and Logical Models for Temporal Databases 193
5.3 Temporal Extension of the MultiDim Model 194
5.3.1 Temporality Types 194
5.3.2 Overview of the Model 196
5.4 Temporal Support for Levels 199
5.5 Temporal Hierarchies 200
5.5.1 Nontemporal Relationships Between Temporal Levels . . 200
5.5.2 Temporal Relationships Between Nontemporal Levels . . 202
5.5.3 Temporal Relationships Between Temporal Levels 202
5.5.4 Instant and Lifespan Cardinalities 203
5.6 Temporal Fact Relationships 205
5.7 Temporal Measures 206
5.7.1 Temporal Support for Measures 206
5.7.2 Measure Aggregation for Temporal Relationships 211
5.8 Managing Different Temporal Granularities 211
5.8.1 Conversion Between Granularities 212
5.8.2 Different Granularities in Measures and Dimensions 212
5.8.3 Different Granularities in the Source Systems and in the Data Warehouse 214
5.9 Metamodel of the Temporally Extended MultiDim Model 215
5.10 Rationale of the Logical-Level Representation 217
5.11 Logical Representation of Temporal Data Warehouses 218
5.11.1 Temporality Types 218
5.11.2 Levels with Temporal Support 220
5.11.3 Parent-Child Relationships 224
5.11.4 Fact Relationships and Temporal Measures 230
5.12 Summary of the Mapping Rules 232
5.13 Implementation Considerations 233
5.13.1 Integrity Constraints 234
5.13.2 Measure Aggregation 238
5.14 Related Work 241
5.14.1 Types of Temporal Support 242
5.14.2 Conceptual Models for Temporal Data Warehouses 242
5.14.3 Logical Representation 244
5.14.4 Temporal Granularity 246
5.15 Summary 246
Review Questions 248

6 Designing Conventional Data Warehouses 251
6.1 Current Approaches to Data Warehouse Design 252
6.1.1 Data Mart and Data Warehouse Design 252
6.1.2 Design Phases 254
6.1.3 Requirements Specification for Data Warehouse Design . 254
6.2 A Method for Data Warehouse Design 256
6.3 A University Case Study 257

6.4 Requirements Specification . . . 259
6.4.1 Analysis-Driven Approach . . . 259
6.4.2 Source-Driven Approach . . . 267
6.4.3 Analysis/Source-Driven Approach . . . 271
6.5 Conceptual Design . . . 271
6.5.1 Analysis-Driven Approach . . . 272
6.5.2 Source-Driven Approach . . . 281
6.5.3 Analysis/Source-Driven Approach . . . 284
6.6 Characterization of the Various Approaches . . . 286
6.6.1 Analysis-Driven Approach . . . 286
6.6.2 Source-Driven Approach . . . 288
6.6.3 Analysis/Source-Driven Approach . . . 289
6.7 Logical Design . . . 289
6.7.1 Logical Representation of Data Warehouse Schemas . . . 289
6.7.2 Defining ETL Processes . . . 293
6.8 Physical Design . . . 294
6.8.1 Data Warehouse Schema Implementation . . . 294
6.8.2 Implementation of ETL Processes . . . 300
6.9 Method Summary . . . 301
6.9.1 Analysis-Driven Approach . . . 302
6.9.2 Source-Driven Approach . . . 302
6.9.3 Analysis/Source-Driven Approach . . . 303
6.10 Related Work . . . 304
6.10.1 Overall Methods . . . 306
6.10.2 Requirements Specification . . . 307
6.11 Summary . . . 311
Review Questions . . . 312

7 Designing Spatial and Temporal Data Warehouses . . . 315
7.1 Current Approaches to the Design of Spatial and Temporal Databases . . . 316
7.2 A Risk Management Case Study . . . 316
7.3 A Method for Spatial-Data-Warehouse Design . . . 318
7.3.1 Requirements Specification and Conceptual Design . . . 318
7.3.2 Logical and Physical Design . . . 329
7.4 Revisiting the University Case Study . . . 332
7.5 A Method for Temporal-Data-Warehouse Design . . . 333
7.5.1 Requirements Specification and Conceptual Design . . . 334
7.5.2 Logical and Physical Design . . . 341
7.6 Method Summary . . . 345
7.6.1 Analysis-Driven Approach . . . 345
7.6.2 Source-Driven Approach . . . 346
7.6.3 Analysis/Source-Driven Approach . . . 347
7.7 Related Work . . . 349
7.8 Summary . . . 350

Review Questions . . . 351

8 Conclusions and Future Work . . . 353
8.1 Conclusions . . . 353
8.2 Future Work . . . 356
8.2.1 Conventional Data Warehouses . . . 356
8.2.2 Spatial Data Warehouses . . . 357
8.2.3 Temporal Data Warehouses . . . 359
8.2.4 Spatiotemporal Data Warehouses . . . 360
8.2.5 Design Methods . . . 361
Review Questions . . . 362

A Formalization of the MultiDim Model . . . 363
A.1 Notation . . . 363
A.2 Predefined Data Types . . . 363
A.3 Metavariables . . . 364
A.4 Abstract Syntax . . . 365
A.5 Examples Using the Abstract Syntax . . . 367
A.5.1 Conventional Data Warehouse . . . 367
A.5.2 Spatial Data Warehouse . . . 369
A.5.3 Temporal Data Warehouse . . . 372
A.6 Semantics . . . 374
A.6.1 Semantics of the Predefined Data Types . . . 375
A.6.2 The Space Model . . . 375
A.6.3 The Time Model . . . 379
A.6.4 Semantic Domains . . . 380
A.6.5 Auxiliary Functions . . . 380
A.6.6 Semantic Functions . . . 383

B Graphical Notation . . . 391
B.1 Entity-Relationship Model . . . 391
B.2 Relational and Object-Relational Models . . . 393
B.3 Conventional Data Warehouses . . . 394
B.4 Spatial Data Warehouses . . . 396
B.5 Temporal Data Warehouses . . . 397

References . . . 399

Glossary . . . 419

Index . . . 433

Abstract

Decision support systems are interactive, computer-based information systems that provide data and analysis tools in order to assist managers at various levels of an organization in the process of decision making. Data warehouses have been developed and deployed as an integral part of decision support systems.

A data warehouse is a database that allows the storage of high volumes of historical data required for analytical purposes. This data is extracted from operational databases, transformed into a coherent whole, and loaded into a data warehouse during an extraction-transformation-loading (ETL) process.

Data in data warehouses can be dynamically manipulated using online analytical processing (OLAP) systems. Data warehouse and OLAP systems rely on a multidimensional model that includes measures, dimensions, and hierarchies. Measures are usually numeric values that are used for quantitative evaluation of aspects of an organization. Dimensions provide various analysis perspectives, while hierarchies allow measures to be analyzed at various levels of detail.

Currently, both designers and users find it difficult to specify the multidimensional elements required for analysis. One reason for this is the lack of conceptual models for data warehouse design, which would allow one to express data requirements on an abstract level without considering implementation details. Another problem is that many kinds of complex hierarchies that arise in real-world situations are not addressed by current data warehouse and OLAP systems.

In order to help designers to build conceptual models for decision support systems and to help users to better understand the data to be analyzed, we propose in this book the MultiDim model, a conceptual model that can be used to represent multidimensional data for data warehouse and OLAP applications. Our model is based mainly on the existing constructs of the entity-relationship (ER) model, such as entity types, relationship types, and attributes with their usual semantics, which allow us to represent the concepts of dimensions, hierarchies, and measures. The model also includes a conceptual

classification of various kinds of hierarchies that exist in real-world situations and proposes graphical notations for them.

Users of data warehouse and OLAP systems increasingly require the inclusion of spatial data. The advantage of using spatial data in the analysis process is widely recognized, since it reveals patterns that are difficult to discover otherwise. However, although data warehouses typically include a spatial or location dimension, this dimension is usually represented in an alphanumeric format. Furthermore, there is still no systematic study that analyzes the inclusion and management of hierarchies and measures that are represented using spatial data.

With the aim of satisfying the growing requirements of decision-making users, we have extended the MultiDim model by allowing the inclusion of spatial data in the various elements composing the multidimensional model. The novelty of our contribution lies in the fact that a multidimensional model is seldom used for representing spatial data. To succeed in our aim, we applied the research achievements in the field of spatial databases to the specific features of a multidimensional model. The spatial extension of a multidimensional model raises several issues, to which we refer in this book, such as the influence of the various topological relationships between spatial levels in a hierarchy on the procedures required for measure aggregation, the aggregation of spatial measures, and the inclusion of spatial measures without the presence of spatial dimensions.

One of the essential characteristics of multidimensional models is the presence of a time dimension that allows one to keep track of changes in measures. However, the time dimension cannot be used for recording changes in other dimensions. This is a serious restriction of current multidimensional models, since in many cases users need to analyze how measures may be influenced by changes in dimension data. As a consequence, specific applications must be developed to cope with these changes. Further, there is still a lack of a comprehensive analysis to determine how the concepts developed for providing temporal support in conventional databases might be applied to data warehouses.

In order to allow users to keep track of temporal changes in all elements of a multidimensional model, we have introduced a temporal extension of the MultiDim model. This extension is based on research done in the area of temporal databases, which have been successfully used for modeling time-varying information for several decades. We propose the inclusion of various temporality types, such as valid time and transaction time, which are obtained from the source systems, in addition to the loading time generated in a data warehouse. We use this temporal support to provide a conceptual representation of time-varying dimensions, hierarchies, and measures. We also refer to specific constraints that should be imposed on time-varying hierarchies and to the problem when the time granularity in the source systems and the data warehouse differ.

The design of data warehouses is not an easy task. It requires one to consider all phases, from requirements specification to the final implementation, including the ETL process. Data warehouse design should also take into account the fact that the inclusion of data items in a data warehouse depends on both the users' needs and the availability of data in the source systems. However, currently, designers must rely on their experience, owing to the lack of a methodological framework that considers these aspects.

In order to assist developers during the data warehouse design process, we propose a method for the design of conventional, spatial, and temporal data warehouses. We refer to various phases, namely, requirements specification, conceptual design, logical design, and physical design. We include three different approaches to requirements specification depending on whether the users, the operational data sources, or both are the driving force in the process of requirements gathering. We show how each approach leads to the creation of a conceptual multidimensional model. We also present the logical- and physical-design phases that refer to data warehouse structures and the ETL process.

To ensure the correctness of the proposed conceptual model, we formally define the model providing its syntax and semantics. Throughout the book, we illustrate the concepts using real-world examples with the aim of demonstrating the usability of our conceptual model. Furthermore, we show how the conceptual specifications can be implemented in relational and object-relational databases. We do this using two representative data warehouse platforms, Microsoft Analysis Services 2005 and Oracle 10g with the OLAP and Spatial extensions.