Preface

Since the late 1970s, relational database technology has been adopted by most organizations to store their essential data. However, nowadays, the needs of these organizations are not the same as they used to be. On the one hand, increasing market dynamics and competitiveness led to the need of having the right information at the right time. Managers need to be properly informed in order to take appropriate decisions to keep up with business successfully. On the other hand, data possessed by organizations are usually scattered among different systems, each one devised for a particular kind of business activity. Further, these systems may also be distributed geographically in different branches of the organization.

Traditional database systems are not well suited for these new requirements, since they were devised to support the day-to-day operations rather than for data analysis and decision making. As a consequence, new database technologies for these specific tasks have emerged in the 1990s, namely, data warehousing and online analytical processing (OLAP), which involve architectures, algorithms, tools, and techniques for bringing together data from heterogeneous information sources into a single repository suited for analysis. In this repository, called a data warehouse, data are accumulated over a period of time for the purpose of analyzing its evolution and discovering strategic information such as trends, correlations, and the like. Data warehousing is nowadays a well-established and mature technology used by organizations in many sectors to improve their operations and better achieve their objectives.

Objective of the Book

This book is aimed at consolidating and transferring to the community the experience of many years of teaching and research in the field of databases and data warehouses conducted by the authors, individually as well as jointly.

However, this is not a compilation of the authors' past publications. On the contrary, the book aims at being a main textbook for undergraduate and graduate computer science courses on data warehousing and OLAP. As such, it is written in a pedagogical rather than research style to make the work of the instructor easier and to help the student understand the concepts being delivered. Researchers and practitioners who are interested in an introduction to the area of data warehousing will also find in the book a useful reference. In summary, we aimed at providing an in-depth coverage of the main topics in the field, yet keeping a simple and understandable style.

We describe next the main features that make this book different from other academic ones in the field. Throughout the book, we follow a methodology that covers all the phases of the data warehousing process, from requirements specification to implementation. Regarding data warehouse design, we make a clear distinction between the three abstraction levels of the American National Standards Institute (ANSI) database architecture, that is, conceptual, logical, and physical, unlike the usual approaches, which do not distinguish clearly between the conceptual and logical levels. A strong emphasis is given to querying using the de facto standard MDX (MultiDimensional eXpressions). Though there are many practical books covering this language, academic books have largely ignored it. We also provide an in-depth coverage of the extraction, transformation, and loading (ETL) processes. Unlike other books in the field, we devote a whole chapter to study how data mining techniques can be used to exploit the data warehouse. In addition, we study how key performance indicators (KPIs) and dashboards are built on top of data warehouses. Although there are many textbooks on spatial databases, this is not the case with spatial data warehouses, which we study in this book, together with trajectory data warehouses, which allow the analysis of data produced by objects that change their position in space and time, like cars or pedestrians. We also address several issues that we believe are likely to be relevant in the near future, like new database architectures such as column-store and in-memory databases, as well as data warehousing and OLAP on the semantic web.

A key characteristic that distinguishes this book from other textbooks is that we illustrate how the concepts introduced can be implemented using existing tools. Specifically, throughout the book we develop a case study based on the well-known Northwind database using representative tools of different kinds. As an example of a commercial implementation, we used the tools provided with Microsoft SQL Server, namely, Analysis Services, Integration Services, and Reporting Services. As an example of an open-source implementation, we used the Pentaho Business Analytics suite of products, which includes Pentaho Analysis Services, an OLAP engine commonly known as Mondrian, and Pentaho Data Integration, an ETL tool commonly known as Kettle. In particular, the chapter on logical design includes a complete description of how to define an OLAP cube in both Analysis Services and Mondrian. Similarly, the chapter on physical design

illustrates how to optimize SQL Server, Analysis Services, and Mondrian applications. Further, in the chapter on ETL we give a complete example of a process that loads the Northwind data warehouse, implemented using both Integration Services and Kettle. In the chapter on data analytics, we used Analysis Services for data mining and for defining key performance indicators, and we used Reporting Services to show how dashboards can be implemented. Finally, to illustrate spatial and spatiotemporal concepts, we used the GeoMondrian OLAP tool over the open-source database PostgreSQL and its spatial extension PostGIS. In this way, the reader can replicate most of the examples and queries presented in the book.

We have also included review questions and exercises for all the chapters in order to help the reader verify that the concepts in each chapter have been well understood. We strongly believe that being formal and precise in the presentation of the topics, implementing them on operational tools, and checking the acquired knowledge against an extensive list of questions and exercises provides a comprehensive learning path for the student.

In addition to the above, support material for the book has been made available online at the address http://cs.ulb.ac.be/DWSDIbook/. This includes electronic versions of the figures, slides for each chapter, solutions to the proposed exercises, and other pedagogic material that can be used by instructors using this book as a course text.

This book builds up from the book *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications* coauthored by one of the authors of the present work in collaboration with Elzbieta Malinowski and published by Springer in 2007. We would like to emphasize that the present book is not a new edition of the previous one but a completely new book with a different objective: While the previous book focused solely on data warehouse design, the present book provides a comprehensive coverage of the overall data warehouse process, from requirements specification to implementation and exploitation. Although approximatively 15% of the previous book was used as a starting point of the present one, this reused material has been adapted to cope with the new objectives of the book.

Organization of the Book and Teaching Paths

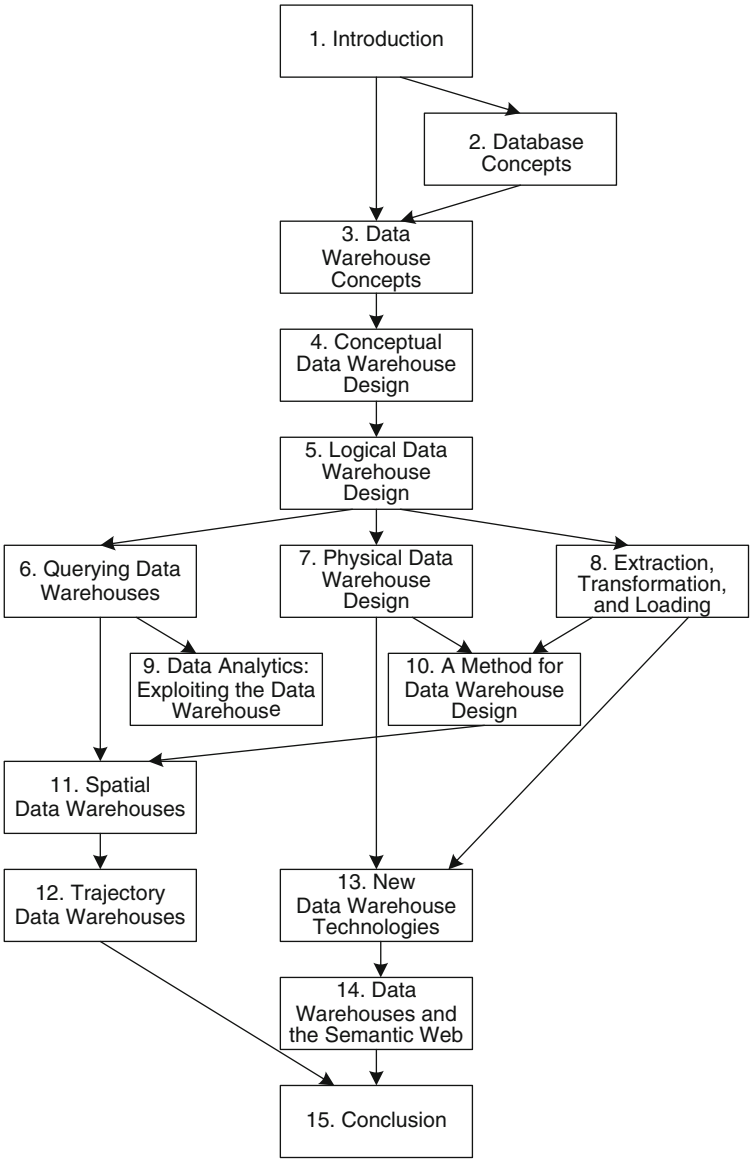
Part I of the book starts with Chap. 1 and provides a historical overview of data warehousing and OLAP. Chapter 2 introduces the main concepts of relational databases needed in the remainder of the book. We also introduce the case study that we will use throughout the book, which is based on the well-known Northwind database. Data warehouses and the multidimensional model are introduced in Chap. 3, as well as the tools provided by SQL Server and the Pentaho Business Analytics suite. Chapter 4 deals with conceptual data warehouse design, while Chap. 5 is devoted to logical data warehouse

design. Part I closes with Chap. 6, which studies MDX and SQL/OLAP, the extension of SQL with OLAP features.

Part II covers data warehouse implementation and exploitation issues. This part starts with Chap. 7, which tackles physical data warehouse design, focusing on indexing, view materialization, and database partitioning. Chapter 8 studies conceptual modeling and implementation of ETL processes. Chapter 9 studies data analytics as a way of exploiting the data warehouse for decision making. Chapter 10 closes Part II, providing a comprehensive method for data warehouse design.

Part III covers advanced data warehouse topics. This part starts with Chap. 11, which studies spatial data warehouses and their exploitation, denoted spatial OLAP (SOLAP). This is illustrated with an extension of the Northwind data warehouse with spatial data, denoted GeoNorthwind, and we query this data warehouse with a spatial extension of the MDX language. Chapter 12 covers trajectory data warehousing. Like in Chap. 11, we illustrate the problem by extending the Northwind data warehouse with trajectory data and show how this data warehouse can be queried extending SQL with spatiotemporal data types. Chapter 13 studies how novel techniques (like the MapReduce programming model) and technologies (like column-store and in-memory databases) can be applied in the field of data warehousing to allow large amounts of data to be processed. Chapter 14 addresses OLAP analysis over semantic web data. Finally, Chap. 15 concludes the book, pointing out what we believe will be the main challenges for data warehousing in the future. Appendix A summarizes the notations used in this book.

The figure below illustrates the overall structure of the book and the interdependencies between the chapters described above. Readers may refer to this figure to tailor their use of this book to their own particular interests. The dependency graph in the figure suggests many of the possible combinations that can be devised in order to offer advanced graduate courses on data warehousing. We can see that there is a path from Chaps. 1 to 6, covering a basic course. In addition, according to the course needs and the coverage depth given to each of the topics, this basic course can be naturally extended to include any combination of physical design, ETL process, data analytics, or even spatial databases. That means the book organization gives the lecturer enough flexibility to combine topics after the basic concepts have been delivered. For example, advanced courses can include a quick overview of Chaps. 1–5, and then they can be customized in many different ways. For example, if the lecturer wants to focus on querying, she could deliver the paths starting in Chap. 6. If she wants to focus on physical issues, she can follow the paths starting in Chaps. 7 and 8.



Relationships between the chapters of this book

http://www.springer.com/978-3-642-54654-9

Data Warehouse Systems
Design and Implementation
Vaisman, A.; Zimányi, E.
2014, XVI, 625 p. 133 illus., Hardcover
ISBN: 978-3-642-54654-9