# Data-Centric Systems and Applications

Intelligent data management is the backbone of all information processing and has hence been one of the core topics in computer science from its very start. This series is intended to offer an international platform for the timely publication of all topics relevant to the development of data-centric systems and applications. All books show a strong practical or application relevance as well as a thorough scientific basis. They are therefore of particular interest to both researchers and professionals wishing to acquire detailed knowledge about concepts of which they need to make intelligent use when designing advanced solutions for their own problems.

Special emphasis is laid upon:

- Scientifically solid and detailed explanations of practically relevant concepts and techniques
  (what does it do)
- Detailed explanations of the practical relevance and importance of concepts and techniques
  (why do we need it)
- Detailed explanation of gaps between theory and practice
  (why it does not work)

According to this focus of the series, submissions of advanced textbooks or books for advanced professional use are encouraged; these should preferably be authored books or monographs, but coherently edited, multi-author books are also envisaged (e.g. for emerging topics). On the other hand, overly technical topics (like physical data access, data compression etc.), latest research results that still need validation through the research community, or mostly product-related information for practitioners ("how to use Oracle 9i efficiently") are not encouraged.

Alejandro Vaisman • Esteban Zimányi

# Data Warehouse Systems

## Design and Implementation

Second Edition

Springer

Alejandro Vaisman
Instituto Tecnológico de Buenos Aires
Buenos Aires, Argentina

Esteban Zimányi
Université Libre de Bruxelles
Brussels, Belgium

ISSN 2197-9723 ISSN 2197-974X (electronic)
Data-Centric Systems and Applications
ISBN 978-3-662-65166-7 ISBN 978-3-662-65167-4 (eBook)
https://doi.org/10.1007/978-3-662-65167-4

*To Andrés and Manuel,*
*who bring me joy and*
*happiness day after day*
*A.V.*

*To Elena,*
*the star that shed light upon my path,*
*with all my love*
*E.Z.*

# Foreword to the Second Edition

Dear reader,

Assuming you are looking for a textbook on data warehousing and the analytical processing of data, I can assure you that you are certainly in the right spot. In fact, I could easily argue how panoramic and lucid the view from this spot is, and in the next few paragraphs, this is exactly what I am going to do.

Assembling a good book from the bits and pieces of writings, slides, and article commentaries that an author has in his folders, is no easy task. Even more, if the book is intended to serve as a textbook, it requires an extra dose of love and care for the students who are going to use it (and their instructors, too, in fact). The book you have at hand is the product of hard work and deep caring by our two esteemed colleagues, Alejandro Vaisman and Esteban Zimányi, who have invested a large amount of effort to produce a book that is (a) comprehensive, (b) up-to-date, (c) easy to follow, and, (d) useful and to-the-point. While the book is also addressing the researcher who, coming from a different background, wants to enter the area of data warehousing, as well as the newcomer to data processing, who might prefer to start the journey of working with data from the neat setup of data cubes, the book is perfectly suited as a textbook for advanced undergraduate and graduate courses in the area of data warehousing.

The book comprehensively covers all the fundamental modeling issues, and addresses also the practical aspects on querying and populating the warehouse. The usage of concrete examples, consistently revisited throughout the book, guide the student to understand the practical considerations, and a set of exercises help the instructor with the hands-on design of a course. For what it's worth, I have already used the first edition of the book for my graduate data warehouse course and will certainly switch to the new version in the years to come.

If you, dear reader, have already read the first edition of the book, you already know that the first part, covering the modeling fundamentals, and the second part, covering the practical usage of data warehousing are both

comprehensive and detailed. To the extent that the fundamentals have not changed (and are not really expected to change in the future), apart from a set of extensions spread throughout the first part of the book, the main improvements concern readability on the one hand, and the technological advances on the other. Specifically, the dedicated chapter 7 on practical data analysis with lots of examples over a specific example, as well as the new topics covering partitioning and parallel data processing in the physical management of the data warehouse provide an even more easy path to the novice reader into the areas of querying and managing the warehouse.

I would like, however, to take the opportunity and direct your attention to the really new features of this second edition, which are found in the last unit of the book, concerning advanced areas of data warehousing. This part goes beyond the traditional data warehousing modeling and implementation and is practically completely refreshed compared to the first edition of the book. The chapter on temporal and multiversion warehousing covers the problem of time encoding for evolving facts and the management of versions. The part on spatial warehouses has been significantly updated. There is a brand-new chapter on graph data processing, and its application to graph warehousing and graph OLAP. Last but extremely significant, the crown jewel of the book, a brand-new chapter on the management of Big Data and the usage of Hadoop, Spark and Kylin, as well as the coverage of distributed, in-memory, columnar, and Not-Only-SQL DBMS's in the context of analytical data processing. Recent advents like data processing in the cloud, polystores and data lakes are also covered in the chapter.

Based on all that, dear reader, I can only invite you to dive into the contents of the book, feeling certain that, once you have completed its reading (or maybe, targeted parts of it), you will join me in expressing our gratitude to Alejandro and Esteban, for providing such a comprehensive textbook for the field of data warehousing in the first place, and for keeping it up to date with the recent developments, in this, current, second edition.

Ioannina, Greece Panos Vassiliadis

# Foreword to the First Edition

Having worked with data warehouses for almost 20 years, I was both honored and excited when two veteran authors in the field asked me to write a foreword for their new book and sent me a PDF file with the current draft. Already the size of the PDF file gave me a first impression of a very comprehensive book, an impression that was heavily reinforced by reading the Table of Contents. After reading the entire book, I think it is quite simply the most comprehensive textbook about data warehousing on the market.

The book is very well suited for one or more data warehouse courses, ranging from the most basic to the most advanced. It has all the features that are necessary to make a good textbook. First, a running case study, based on the Northwind database known from Microsoft's tools, is used to illustrate all aspects using many detailed figures and examples. Second, key terms and concepts are highlighted in the text for better reading and understanding. Third, review questions are provided at the end of each chapter so students can quickly check their understanding. Fourth, the many detailed exercises for each chapter put the presented knowledge into action, yielding deep learning and taking students through all the steps needed to develop a data warehouse. Finally, the book shows how to implement data warehouses using leading industrial and open-source tools, concretely Microsoft's suite of data warehouse tools, giving students the essential hands-on experience that enables them to put the knowledge into practice.

For the complete database novice, there is even an introductory chapter on standard database concepts and design, making the book self-contained even for this group. It is quite impressive to cover all this material, usually the topic of an entire textbook, without making it a dense read. Next, the book provides a good introduction to basic multidimensional concepts, later moving on to advanced concepts such as summarizability. A complete overview of the data warehouse and online analytical processing (OLAP) "architecture stack" is given. For the conceptual modeling of the data warehouse, a concise and intuitive graphical notation is used, a full specification of which is given in

an appendix, along with a methodology for the modeling and the translation to (logical-level) relational schemas.

Later, the book provides a lot of useful knowledge about designing and querying data warehouses, including a detailed, yet easy to read, description of the de facto standard OLAP query language: MultiDimensional eXpressions (MDX). I certainly learned a thing or two about MDX in a short time. The chapter on extract-transform-load (ETL) takes a refreshingly different approach by using a graphical notation based on the Business Process Modeling Notation (BPMN), thus treating the ETL flow at a higher and more understandable level. Unlike most other data warehouse books, this book also provides comprehensive coverage on analytics, including data mining and reporting, and on how to implement these using industrial tools. The book even has a chapter on methodology issues such as requirements capture and the data warehouse development process, again something not covered by most data warehouse textbooks.

However, the one thing that really sets this book apart from its peers is the coverage of advanced data warehouse topics, such as spatial databases and data warehouses, spatiotemporal or mobility databases and data warehouses, and semantic web data warehouses. The book also provides a useful overview of novel "big data" technologies like Hadoop and novel database and data warehouse architectures like in-memory database systems, column store systems, and right-time data warehouses. These advanced topics are a distinguishing feature not found in other textbooks.

Finally, the book concludes by pointing to a number of exciting directions for future research in data warehousing, making it an interesting read even for seasoned data warehouse researchers.

A famous quote by IBM veteran Bruce Lindsay states that "relational databases are the foundation of Western civilization." Similarly, I would say that "data warehouses are the foundation of twenty-first-century enterprises." And this book is in turn an excellent foundation for building those data warehouses, from the simplest to the most complex.

Happy reading!

Aalborg, Denmark Torben Bach Pedersen

# 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 to have 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 held 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 day-to-day operations rather than for data analysis and decision making. As a consequence, new database technologies for these specific tasks 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 their evolution and discovering strategic information such as trends, correlations, and the like. Data warehousing is a well-established and mature technology used by organizations 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 aim at providing in-depth coverage of the main topics in the field, yet keeping a simple and understandable style.

Throughout the book, we cover 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 placed on querying using the de facto standard language MDX (MultiDimensional eXpressions) as well as the popular language DAX (Data Analysis eXpressions). Though there are many practical books covering these languages, academic books have largely ignored them. We also provide in-depth coverage of the extraction, transformation, and loading (ETL) processes. In addition, we study how key performance indicators (KPIs) and dashboards are built on top of data warehouses. An important topic that we also cover in this book is temporal and multiversion data warehouses, in which the evolution over time of the data and the schema of a data warehouse are taken into account. 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 mobility data warehouses, which allow the analysis of data produced by objects that change their position in space and time, like cars or pedestrians. Data warehousing and OLAP on graph databases and on the semantic web are also studied. Finally, big data technologies led to the concept of big data warehouses, which are also covered in this book.

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. In particular, the chapter on logical design includes a complete description of how to define an OLAP cube in Microsoft SQL Analysis Services using both the multidimensional and the tabular models. Similarly, the chapter on physical design illustrates how to optimize SQL Server and Analysis Services applications. Further, in the chapter on ETL we give a complete example of a process that loads the Northwind data warehouse, implemented using Integration Services. We also use Analysis Services for defining KPIs, and use Reporting Services to show how dashboards can be implemented. To illustrate spatial and spatiotemporal concepts we use the open-source database PostgreSQL, its spatial extension PostGIS, and its mobility extension MobilityDB. In this way, the reader can replicate most of the examples and queries

presented in the book. Finally, in the chapter on graph data warehouses we use Neo4j.

We also include review questions and exercises for all the chapters in order to help the reader verify that the concepts have been well understood. Support material for the book is available online at http://cs.ulb.ac.be/DWSDIbook2e/. This includes electronic versions of the figures, slides for each chapter, solutions to the exercises, and other pedagogic material that can be used by instructors using this book as a course text.

This second edition of the book updates several chapters with new results and technologies that have appeared since the publication of the first edition. In Chaps. 5, 6, and 7, the tabular model and DAX have been included. Chapter 15 covers big data warehouse technologies, which have considerably evolved since the first edition. Further, we have added new chapters covering temporal, multiversion, and graph data warehouses. Also, all application examples that make use of software tools have been updated to the latest versions of them. In addition to this new material, all chapters of the first edition have been revised and updated with the feedback obtained through seven years of teaching at undergraduate and graduate levels, and to professional teams in different industries.

## Organization of the Book and Teaching Paths

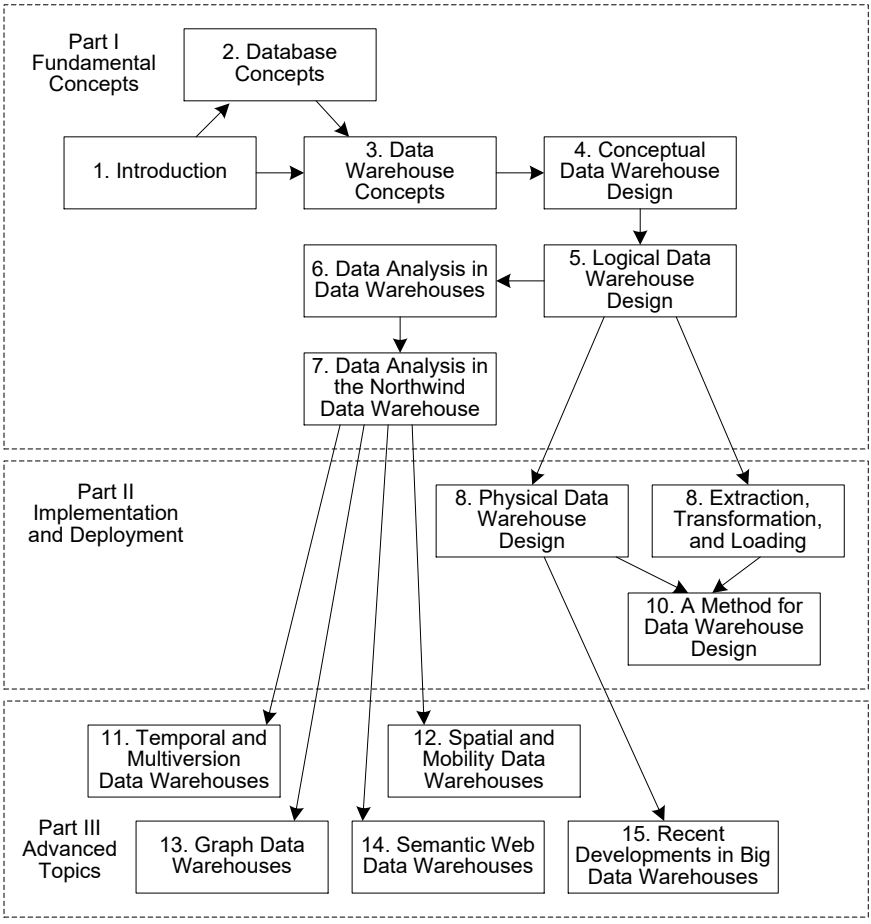
Part I of the book starts with Chap. 1, giving a historical overview of data warehousing and OLAP. Chapter 2 introduces the main concepts of relational databases needed in the remainder of the book. We also introduce the case study that we will use throughout the book, based on the well-known Northwind database. Data warehouses and the multidimensional model are introduced in Chap. 3, as well as the suite of tools provided by SQL Server. Chapter 4 deals with conceptual data warehouse design, while Chap. 5 is devoted to logical data warehouse design. Part I closes with Chaps. 6 and 7, which study SQL/OLAP, the extension of SQL with OLAP features, as well as MDX and DAX.

Part II covers data warehouse implementation issues. This part starts with Chap. 8, which tackles classical physical data warehouse design, focusing on indexing, view materialization, and database partitioning. Chapter 9 studies conceptual modeling and implementation of ETL processes. Finally, Chap. 10 provides a comprehensive method for data warehouse design.

Part III covers advanced data warehouse topics. This part starts with Chap. 11, which studies temporal and multiversion data warehouses, for both *data* and *schema* evolution of the data warehouse. Then, in Chap. 12, we study spatial data warehouses and their exploitation, denoted spatial OLAP (SOLAP), illustrating the problem with a spatial extension of the Northwind data warehouse denoted GeoNorthwind. We query this data warehouse

using PostGIS, PostgreSQL's spatial extension. The chapter also covers mobility data warehousing, using MobilityDB, a spatiotemporal extension of PostgreSQL. Chapters 13 and 14 address OLAP analysis over graph data represented, respectively, natively using property graphs in Neo4j and using RDF triples as advocated by the semantic web. Chapter 15 studies how novel techniques and technologies for distributed data storage and processing can be applied to the field of data warehousing. 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 to offer advanced graduate courses on data warehousing.




Relationships between the chapters of this book

# Acknowledgments

We would like to thank Innoviris, the Brussels Institute for Research and Innovation, which funded Alejandro Vaisman's work through the OSCB project; without its financial support, the first edition of this book would never have been possible. As mentioned above, some content of this book finds its roots in a previous book written by one of the authors in collaboration with Elzbieta Malinowski. We would like to thank her for all the work we did together in making the previous book a reality. This gave us the impetus to start this new book.

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 discuss with research colleagues from all around the world, and we exchanged viewpoints about the subject with them. The anonymous reviewers of these conferences and journals provided us with insightful comments and suggestions that contributed significantly to improve the work presented in this book. We would like to thank Zineb El Akkaoui, with whom we have explored the use of BPMN for ETL processes, and Judith Awiti, who continued this work. A very special thanks to Waqas Ahmed, a doctoral student of our laboratory, with whom we explored the issue of temporal and multiversion data warehouses. Waqas also suggested to include tabular modeling and DAX in the second edition of the book, and without his invaluable help, all the material related to the tabular model and DAX would have not been possible. A special thanks to Mahmoud Sakr, Arthur Lesuisse, Mohammed Bakli, and Maxime Schoemans, who worked with one of the authors in the development of MobilityDB, a spatiotemporal extension of PostgreSQL and PostGIS that was used for mobility data warehouses. This work follows that of Benoit Foé, Julien Lusiela, and Xianling Li, who explored this topic in the context of their master's thesis. Arthur Lesuisse also provided invaluable help in setting up all the computer infrastructure we needed, especially for spatializing the Northwind database. He also contributed in enhancing some of the figures of this book. Thanks also to Leticia Gómez from the Buenos Aires Technological Institute for her help on the im-

plementation of graph data warehouses and for her advice on the topic of big data technologies. Bart Kuijpers, from Hasselt University, also worked with us during our research on graph data warehousing and OLAP. We also want to thank Lorena Etcheverry, who contributed with comments, exercises, and solutions in Chap. 14.

Special thanks go to Panos Vassiliadis, professor at the University of Ioannina in Greece, who kindly agreed to write the foreword for this second edition. Finally, we would like to warmly thank Ralf Gerstner of Springer for his continued interest in this book. The enthusiastic welcome given to our book proposal for the first edition and the continuous encouragements to write the second edition gave us enormous impetus to pursue our project to its end.

Alejandro Vaisman
Buenos Aires, Argentina,
February 2022

Esteban Zimányi
Brussels, Belgium

# About the Authors

**Alejandro Vaisman** is a professor at the Instituto Tecnológico de Buenos Aires, where he also chairs the graduate program in data science. He has been a professor and chair of the master's program in data mining at the University of Buenos Aires (UBA) and professor at Universidad de la República in Uruguay. He received a BE degree in civil engineering, and a BCS degree and a doctorate in computer science from the UBA, under the supervision of Prof. Alberto Mendelzon, from the University of Toronto (UoT). He has been a postdoctoral fellow at UoT, and visiting researcher at UoT, Universidad Politécnica de Madrid, Universidad de Chile, University of Hasselt, and Université Libre de Bruxelles (ULB). His research interests are in the field of databases, business intelligence, and geographic information systems. He has authored and coauthored many scientific papers published at major conferences and in major journals.

**Esteban Zimányi** is a professor and a director of the Department of Computer and Decision Engineering (CoDE) of Université Libre de Bruxelles (ULB). He started his studies at the Universidad Autónoma de Centro América, Costa Rica, and received a BCS degree and a doctorate in computer science from ULB. His current research interests include spatiotemporal and mobility databases, data warehouses and business intelligence, geographic information systems, as well as semantic web. He has coauthored and coedited eight books and published many papers on these topics. He was editor-in-chief of the *Journal on Data Semantics* (JoDS) published by Springer from 2012 to 2020. He coordinated the Erasmus Mundus master's and doctorate programmes "Information Technologies for Business Intelligence" (IT4BI) and "Big Data Management and Analytics" (BDMA) as well as the Marie Skłodowska-Curie doctorate programme "Data Engineering for Data Science" (DEDS).

# Contents