DATABASE SYSTEMS

Introduction to Databases and Data Warehouses

Nenad Jukić
Loyola University Chicago

Susan Vrbsky
The University of Alabama

Svetlozar Nestorov
Loyola University Chicago

Prospect
Press

This book was previously published by: Pearson Education, Inc.

ORDERING INFORMATION:
Upon publication of the Prospect Press Edition,
students may order directly from these retailers:

Redshelf.com – eTextbooks and paperbacks

VitalSource.com – eTextbooks

Campus bookstores may order ebooks and paperbacks from Redshelf.com by contacting PO@Redshelf.com.

eTextbook:
ISBN: 978-1-943153-18-3

Paperback:
ISBN: 978-1-943153-19-0

For more information, contact Beth.Golub@ProspectPressVT.com

In memory of my mother, Ruža. To my father, Drago, and to Linda and Lee Albritton.
To my wife, Maria, and our kids, Maja, Niko, and Boris.

––Nenad

To Austin, who teaches me new things every day. To my students, who continually remind me of the joy of learning. To Brian, who recognizes my need for both.

––Susan

To my parents, Atanaska and Evtim.

––Svetlozar

BRIEF CONTENTS

Appendices

PREFACE

Database Systems: Introduction to Databases and Data Warehouses is an introductory yet comprehensive database textbook intended for use in undergraduate and graduate information systems database courses. Even though it is likely that students taking a course based on this book will have already been exposed to an introductory information systems course, the book itself can also be used with no prerequisite course taken. The book contains all necessary introductions, followed by a detailed coverage of database topics. Its goal is to provide a significant level of database expertise to its readers and users.

COVERAGE OF OPERATIONAL AND ANALYTICAL DATABASES

The current job market in fields such as information systems, business data analysis, and decision support requires both operational and analytical database systems competence. *Database Systems: Introduction to Databases and Data Warehouses* covers operational and analytical database systems in a comprehensive fashion, providing a theoretical foundation and meaningful hands-on experiences. Students in database courses based on this textbook will learn how to design and use operational *and* analytical databases and will be prepared to apply their knowledge in today's business environments, which require a solid understanding of both of these types of databases.

Both operational and analytical databases are now mainstream information systems topics. A responsible educational approach is to teach both of these topics in a meaningful way, even when the curriculum allows for only one database-related class. In our opinion and experience, it is pedagogically straightforward to teach a course based on this book. We have been teaching database courses (both in semester and quarter versions at both the graduate and undergraduate levels) based on the material presented in this book for several years. In every instance, we were able to cover all of the intended topics. Our courses based on this book have received overwhelmingly positive evaluation scores. In addition, we have received numerous written testimonies from former and current undergraduate and graduate students and their hiring companies (both in internship and full-time job scenarios) reporting the high level of preparedness of our students for jobs that require database competence.

FEATURES

Our coverage of fundamental topics related to the design and use of operational and analytical databases is divided into 10 chapters and 10 appendices.

Chapters 1 through 6 focus on operational database topics and include coverage of standard database issues, including *database requirements*, *ER modeling*, *relational modeling*, *database constraints, update anomalies*, *normalization*, *SQL*, *the database front-end*, and *data quality*.

Chapters 7 through 9 focus on analytical database topics and include coverage of data warehouse and data mart-related topics, including *data warehousing concepts*, *dimensional modeling* (*star schemas*), *data warehouse/data mart modeling approaches*, *the extraction/ transformation/load (ETL) process*, *online analytical processing (OLAP)/business intelligence (BI) functionalities*, and *the data warehouse/data mart front-end.*

Chapter 10 presents a higher-level (less-detailed) overview of *database administration*.

Appendices (**A, B, C, D, E, F, G, H, I,** and **J**) contain brief overviews of additional database and database-related topics, including *enhanced ER modeling (EER)*, *further normal forms (beyond 3NF)*, *enterprise resource planning (ERP)*, *data governance and master data management*, *object-oriented databases*, *distributed databases, parallel databases, cloud computing* , *data mining*, *XML*, *NoSQL databases*, and *Big Data*.

Suggestions on how to efficiently cover the presented topics in both semester and quarter versions of courses are included in the instructor's resources accompanying this book.

The book's Web site includes access to the free Web-based data modeling suite **ERDPlus** (*erdplus.com*), designed and developed in conjunction with the book. Students and instructors can use this data modeling suite, specifically designed for use in academic settings, to create ER diagrams, relational schemas, and dimensional models (star schemas). We encourage instructors and students to give this provided data modeling suite a try, and experience its simplicity, ease of use, portability, and fit for academic use. Of course, instructors and students are also welcome to use other tools for the purpose of creating ER diagrams, relational schemas, and dimensional models. Coverage of the included exercises is possible either by using the provided data modeling suite ERDPlus or by using other modeling tools and means for creating these diagrams (such as Visio, ERwin, ER/Studio, MS Word, MS Excel, MS PowerPoint, free drawing, etc.).

SUPPLEMENTS

Database Systems: Introduction to Databases and Data Warehouses is accompanied by a comprehensive set of supplementary materials for instructors and students. The supplements include:

A companion Web site (*dbtextbook.com*), which includes:

- Link to ERDPlus, a data modeling software suite (developed in conjunction with the book)
- SQL scripts and data sets
- Instructions for free access to DBMS and OLAP/BI software
- E-mail access to the authors

For Instructors

The Instructor Resources will be available at the Publisher's site
www.prospectpressvt.com/titles/jukic/instructor-resources/ and include the following:

- PowerPoint slides (versions for *quick coverage, typical coverage*, and *complete coverage*)
- An Instructor's Manual that includes:
  - Solutions for the end-of-chapter questions, exercises, and mini cases
  - Additional exercises (with solutions)
  - Sample syllabi and topic coverage plans
  - Project ideas
- A Test Item File that includes a comprehensive set of test questions in multiple choice, true/false, and essay format, each with a page reference and difficulty level.

PEDAGOGICAL APPROACH

Database Systems: Introduction to Databases and Data Warehouses combines clear descriptions of theoretical concepts, easy-to-understand examples, and comprehensive hands-on components, wherever appropriate and useful. For every skill that students acquire, multiple hands-on exercises and mini cases are given at the end of corresponding chapters.

Most of the chapters end with sections whose titles contain the phrase *"A Note About . . .".* These sections represent topics that can be studied as additional optional readings or be covered in the same fashion as the other content (depending on the level of class and the time available).

More about the pedagogy of each chapter is given in the following outline.

OUTLINE

Chapter 1: Introduction

MAIN TOPICS The introductory chapter gives a quick overview of basic database-related terms, concepts, and components, such as data and information, database management system (DBMS), database system development steps, and operational versus analytical databases.

PEDAGOGY The introductory topics are covered using short descriptions and brief introductory examples. This chapter is designed to be covered in a quick and to-the-point fashion, setting the stage for the following chapters.

Chapter 2: Database Requirements and ER Modeling

MAIN TOPICS This chapter provides comprehensive coverage of entity-relationship (ER) modeling as a conceptual method for formalizing user database requirements. A variation of Chen ER notation is used throughout the chapter, but other notations and conceptual data modeling methods are mentioned as well. The purpose of ER modeling as a method for collecting and visualizing requirements is stressed. The chapter covers ER constructs—entities (including weak entities); attributes (regular, unique, composite, multivalued, derived); and 1:1, 1:N, and M:N relationships (binary and unary).

ADDITIONAL NOTES Notes at the end of this chapter discuss several additional issues related to ER modeling (M:N relationships with multiple instances between the same entities, associative entities, and ternary and higher-degree relationships).

PEDAGOGY This chapter is example-driven and presents comprehensive examples of the collection of requirements and the creation of ER models. Exercises, mini cases, and free software (ERDPlus—*ER diagram feature*) reinforce the introduced concepts. This chapter is designed to be covered in depth by first describing concepts relating to ER modeling and requirements visualization to the students, and then reinforcing the concepts through multiple hands-on exercises.

Chapter 3: Relational Database Modeling

MAIN TOPICS This chapter provides comprehensive coverage of the relational database model, including relational concepts, relational schema, integrity constraints, and user-defined constraints. It covers the process of mapping ER diagrams (entities, attributes, binary and unary 1:1, 1:N, and M:N relationships) into relational schemas.

ADDITIONAL NOTES Notes at the end of this chapter discuss several additional issues related to relational database modeling (mapping associative entities, mapping ternary relationships, designer-created primary keys and autonumber option, and the necessity of both ER and relational modeling).

PEDAGOGY This chapter is example-driven and presents comprehensive examples of the mapping of ER constructs and the creation of relational schemas. Exercises, mini cases, and software (ERDPlus—*relational schema feature*) reinforce the concepts that have been introduced. This chapter is designed to be covered in depth by first relating the relational database modeling concepts to the students, and then reinforcing the concepts through multiple hand-on exercises.

Chapter 4: Update Operations, Update Anomalies, and Normalization

MAIN TOPICS This chapter describes update operations (insert, delete, and modify) and provides coverage of normalization and update anomalies (as justification for normalization). It introduces and discusses the concept of functional dependencies. It covers first normal form (1NF), second normal form (2NF), and third normal form (3NF) (other normal forms are covered in Appendix B).

ADDITIONAL NOTES Notes at the end of this chapter discuss several additional issues related to normalization (normalization exceptions, denormalization, normalization versus ER modeling, adding tables for streamlining database content).

PEDAGOGY This chapter is example-driven and presents comprehensive examples of update operations, update anomalies, and the normalization process. It provides exercises to reinforce the introduced concepts. This chapter is designed to be covered in depth by first relating the update and normalization concepts to students and then reinforcing the concepts through multiple hand-on exercises.

Chapter 5: SQL

MAIN TOPICS This chapter provides comprehensive coverage of SQL (Structured Query Language). It covers SQL statements for creating, updating, and querying relational databases. Coverage of commands for retrieval of data includes the SELECT statement (with multiple conditions using AND, OR, and NOT operators), aggregate functions (SUM, COUNT, AVG, MIN, MAX), GROUP BY, ORDER BY, HAVING, nested queries, UNION and INTERSECT operators, IN, EXISTS, various joins, and an overview of other SQL statements and functionalities.

ADDITIONAL NOTES Notes at the end of this chapter discuss several additional issues related to SQL (inappropriate use of observed values in SQL, SQL standard, and SQL syntax differences in various popular RDBMS packages).

PEDAGOGY This chapter is example-driven and presents comprehensive examples of concepts that have been introduced. It is designed to explain the process of building, populating, and querying a relational database using SQL statements. It contains examples of SQL commands doing this, executed in their natural consecutive order. The companion Web site (*dbtextbook. com*) presents scripts containing all SQL statements from this chapter for six popular DBMS packages (Oracle, MySQL, Microsoft SQL Server, PostgreSQL, Teradata, and IBM DB2). Instructors can use these scripts to copy, paste, and execute SQL statements directly in an RDBMS (of their choice) during the lectures based on this chapter. By doing so, instructors can *introduce students to SQL commands* and, at the same time, *demonstrate created, populated, and queried databases* Data sets, exercises, and mini cases reinforce the concepts that have been introduced. In addition, the companion Web site (*dbtextbook.com*) provides instructions on how to obtain free, unlimited access to state-of- the-art relational DBMS software. This chapter is designed to be covered in depth by first relating the SQL concepts to the students and then reinforcing the concepts through multiple hands-on exercises.

Chapter 6: Database Implementation and Use

MAIN TOPICS This chapter includes coverage of data quality issues—accuracy, completeness, consistency, uniqueness, timeliness, and conformity of data. These topics are covered in the context of data stored within database systems. This chapter also covers the design and use of database front-end interfaces (database forms, reports, and applications), referential integrity options (delete and update options: cascade, restrict, set-tonull, and set-to-default), indexes, and the implementation of user-defined constraints.

ADDITIONAL NOTES An additional note at the end of this chapter discusses assertions and triggers.

PEDAGOGY This chapter is designed to provide quick but meaningful coverage of the most fundamental database implementation issues and those database use issues not covered in Chapter 5. It presents examples of concepts that have been introduced. Hands-on exercises reinforce the concepts that have been introduced.

Chapter 7: Basic Data Warehousing Concepts

MAIN TOPICS This chapter defines the terms data warehouse and data mart and introduces basic data warehouse components and concepts (source systems, ETL-extraction/ transformation/load, integrated analytical data repository, subject-oriented databases, and OLAP/BI front end). It also gives an overview of data warehouse system development steps.

PEDAGOGY The introductory data warehousing topics are covered using short descriptions and brief introductory examples. This chapter is designed to be covered in a quick and to-the-point fashion, setting the stage for the following two chapters.

Chapter 8: Data Warehouse and Data Mart Modeling

MAIN TOPICS*:* This chapter covers dimensional modeling, a conceptual and logical data design technique used for designing analytical databases (e.g., data warehouses or data marts). It describes concepts such as fact and dimension tables, star schema, snowflake schema, constellations, and slowly changing dimensions. It also covers ER modeling as a technique for modeling analytical databases (as opposed to ER modeling as a technique for modeling operational databases, covered in Chapter 2). This chapter also gives an overview of different development approaches to data warehousing projects: data warehouse bus architecture (a.k.a. the Kimball approach), including a discussion of conformed dimensions, normalized data warehouse (a.k.a. the Inmon approach), and independent data marts.

ADDITIONAL NOTES*:* An additional note at the end of this chapter compares dimensional modeling and ER modeling as data warehouse/data mart design techniques.

PEDAGOGY: This chapter is example-driven and presents comprehensive examples of dimensional models (star schemas) based on single and multiple sources, detailed and aggregated fact tables, slowly changing dimensions, and other dimensional modeling related topics. This chapter also presents examples of ER modeled/normalized data warehouses. Exercises, mini cases, and free software (ERDPlus—*star schema feature*) reinforce the concepts that have been introduced. This chapter is designed to be covered in depth by first relating the data warehouse and data mart modeling concepts to the students and then reinforcing the concepts through multiple hands-on exercises.

Chapter 9: Data Warehouse Implementation and Use

MAIN TOPICS This chapter gives an overview of the ETL process, including the creation of infrastructure and procedures for the tasks of extracting analytically useful data from the operational sources, transforming such data so that it conforms to the structure of the target data warehouse model, ensuring the quality of the transformed data through processes such as data cleansing or scrubbing, and loading the transformed and quality assured data into the target data warehouse. This chapter defines the terms “online analytical processing (OLAP)” and “business intelligence (BI)," which are commonly used to refer to front-end use of analytical databases. It also presents functionalities that are common for all OLAP/BI tools.

ADDITIONAL NOTES Additional notes at the end of this chapter discuss different database models for OLAP/BI tools and different OLAP/BI architectures.

PEDAGOGY This chapter is example-driven and presents examples of concepts that have been introduced. The companion Web site (*dbtextbook.com*) gives instructions for obtaining free unlimited access to state-of-the-art OLAP/BI software, data sets, and exercises. This chapter is designed to provide quick but meaningful coverage of the most fundamental data warehouse implementation and use issues.

Chapter 10: Overview of DBMS Functionalities and Database Administration

MAIN TOPICS This chapter gives an informative overview of the DBMS functionalities and components. It also gives an overview of database administration issues, such as data security, backup, recovery, performance, and optimization.

PEDAGOGY This chapter presents a quick overview of the introduced topics. It is designed to familiarize readers with DBMS functionalities and database administration topics without covering them in a great level of detail.

Appendices: Overview of Additional Topics

MAIN TOPICS The appendices give an overview of additional database-related topics, including enhanced ER modeling (EER), further normal forms (beyond 3NF), enterprise resource planning (ERP), data governance and master data management, object-oriented databases, distributed databases, parallel databases, cloud computing, data mining, XML, NoSQL databases, and Big Data.

PEDAGOGY The topics in the appendices are covered in the form of short notes and examples. They are designed to familiarize readers with additional database-related topics without covering them in a great level of detail.

ACKNOWLEDGMENTS

We wish to thank the following reviewers for their thoughtful comments and suggestions during the development of this book:

Gary Baram, Temple University
Jeff Hassett, University of Utah
Emily Kelly, Loyola University Chicago
Barry King, Butler University
Mark Llewellyn, University of Central Florida
Brian Mennecke, Iowa State University
Sree Nilakanta, Iowa State University
Janet Renwick, University of Arkansas
John Russo, Wentworth Institute of Technology
Julian Scher, New Jersey Institute of Technology
Linda Zhou, University of Maryland

We are truly grateful for the generous contribution of time, effort, advice, and expertise by a number of individuals. We wish to thank industry experts Benjamin Korallus (PricewaterhouseCoopers), Robin Edison and Douglas Marcis (OptumInsight), Stan Ozenbaugh and Kevin Little (Teradata), Mark Watson and Abhishek Sharma (Federal Reserve Bank of Kansas City), Ryane Bohm (General Electric), Alexander Rukhostkiy (Boeing), Creighton Lang (NTT DATA, Inc.), Zachary Relli (Booz Allen Hamilton), and Gregory Roszczybiuk (Leo Burnett). We are also thankful to our colleagues Stefan Lessmann (University of Hamburg), Mary Malliaris and Faruk Guder (Loyola University Chicago), Michael Goul (Arizona State University), Hugh Watson (University of Georgia), Boris Jukić (Clarkson University), and Miguel Velasco (University of Minnesota). Our thanks also go to the following students at Loyola University Chicago: Melanie Ruiz, Jordan Braun, Subash Pant, Sarah Shea, Adam Chorazy, Anton Dokov, and Karan Desai.

Special thanks to world-class DBA/database designer/programmer Zoran Avramovic for persistently keeping us grounded by providing forthright and detailed scrutiny of the manuscript, and to Ivan Bare for providing extensive logistical and technical support during the preparation of this manuscript.

Nenad would also like to express his gratitude to the excellent information technology teachers he was fortunate to have in his life: Professor Zoran Vlašić, Professor Damir Kalpić, Professor Allen Parrish, and Professor Susan Vrbsky. In the process of writing this book we received kind and thoughtful advice and encouragement from Dr. Paul Gray, who passed away in 2012. Paul was one of the founding pioneers of information systems education and research, and he is truly missed.

Nenad Jukić

Susan Vrbsky

Svetlozar Nestorov

ABOUT THE AUTHORS

Nenad Jukić is a professor of information systems and the director of the graduate certificate program in Business Data Analytics at the Quinlan School of Business at Loyola University Chicago.

Dr. Jukić has been teaching undergraduate, graduate, and executive education classes in the Information Systems and Operations Management Department at the Quinlan School of Business since 1999. Between 2005 and 2007, Dr. Jukić was also a visiting professor of information systems in the Beijing International MBA Program at the China Center for Economic Research at Peking University, Beijing, China. Between 1997 and 1999, he taught at the School of Computing and Information Systems at Grand Valley State University in Allendale, Michigan.

Dr. Jukić received his undergraduate degree in computer science and electrical engineering from the School of Computing and Electrical Engineering at the University of Zagreb in Croatia. He received his M.S. and Ph.D. in computer science from the University of Alabama in Tuscaloosa, Alabama.

Dr. Jukić conducts active research in various information technology–related areas, including database modeling and management, data warehousing, business intelligence, data mining, e-business, and IT strategy. His work has been published in numerous management information systems and computer science academic journals, conference publications, and books. In addition to his academic work, his engagements include providing expertise to database, data warehousing, and business intelligence projects for corporations and organizations that vary from startups to *Fortune* 500 companies to U.S. government and military agencies.

Susan V. Vrbsky is an associate professor and graduate program director of computer science at the University of Alabama in Tuscaloosa, Alabama.

Dr. Vrbsky has been teaching undergraduate and graduate courses in the Department of Computer Science at the University of Alabama since 1992. She is the director of the Cloud and Cluster Computing Laboratory at the University of Alabama. She also taught in the Computer Science Department at Southern Illinois University from 1983 to 1986.

Dr. Vrbsky received her B.A. from Northwestern University in Evanston, IL, and her M.S. in computer science from Southern Illinois University in Carbondale, IL. She received her Ph.D. in computer science from the University of Illinois, Urbana-Champaign.

Dr. Vrbsky's research is in the area of databases and cloud computing, including data intensive computing, real-time databases, database security, mobile databases, and green computing. She has co-authored over 100 peer-review publications in computer science academic journals, conference publications, and books. She has received funding from such sources as the National Science Foundation.

Svetlozar Evtimov Nestorov is an assistant professor of information systems at the Quinlan School of Business at Loyola University Chicago.

Previously, he was a senior research associate at the Computation Institute at the University of Chicago, and he was an assistant professor in computer science at the University of Chicago, where he taught databases and computer systems to undergraduate and graduate students. While on leave, he co-founded *Mobissimo*, a venture-backed travel search engine that was chosen as one of the 50 coolest Web sites by *Time* magazine in 2004.

Dr. Nestorov received his undergraduate degrees in computer science and mathematics from Stanford University and his M.S. in computer science from Stanford University. He received his Ph.D. in computer science from Stanford University with a dissertation titled "Data Mining Techniques for Structured and Semistructured Data." His advisor was Professor Jeffrey Ullman.

Svetlozar led the design and development of the data warehouse project at the Nielsen Data Center at the Kilts Center for Marketing, which is part of the Chicago Booth School of Business. His research interests also include data mining, high-performance computing, and Web technologies.