Texts in Computer Science

Series Editors:
David Gries
Fred B. Schneider

For other titles published in this series, go to
http://www.springer.com/series/3191

Peter Revesz

Introduction to Databases

From Biological to Spatio-Temporal

Springer

Prof. Peter Revesz
University of Nebraska-Lincoln
Dept. Computer Science & Engineering
358 Avery Hall
Lincoln NE 68588
USA
revesz@cse.unl.edu

ISSN 1868-0941 e-ISSN 1868-095X
ISBN 978-1-84996-094-6 e-ISBN 978-1-84996-095-3
DOI 10.1007/978-1-84996-095-3
Springer London Heidelberg Dordrecht New York

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

Library of Congress Control Number: 2009942531

Cover design: SPi Publisher Services

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

In memory of

Paris C. Kanellakis

and

Alberto O. Mendelzon

I believe that there are a collection of planet-threatening problems, such as climate change and ozone depletion, that only scientists are in a position to solve. Hence, the sorry state of DBMS support...for this class of users is very troubling.

—Michael Stonebraker (2009), *CACM, 52(5):11.*

Preface

Introduced forty years ago, relational databases proved unusually successful and durable. However, relational database systems were not designed for modern applications and computers. As a result, specialized database systems now proliferate trying to capture various pieces of the database market. Database research is pulled into different directions, and specialized database conferences are created. Yet the current chaos in databases is likely only temporary because every technology, including databases, becomes standardized over time.

The history of databases shows periods of chaos followed by periods of dominant technologies. For example, in the early days of computing, users stored their data in text files in any format and organization they wanted. These early days were followed by information retrieval systems, which required some structure for text documents, such as a title, authors, and a publisher. The information retrieval systems were followed by database systems, which added even more structure to the data and made querying easier.

In the late 1990s, the emergence of the Internet brought a period of relative chaos and interest in unstructured and "semistructured data" as it was envisioned that every webpage would be like a page in a book. However, with the growing maturity of the Internet, the interest in structured data was regained because the most popular websites are, in fact, based on databases. The question is not whether future data stores need structure but what structure they need.

Today we see another period of relative chaos as specialized geographic, moving objects, biological, and other types of databases are unable to communicate with each other in spite of the need of many advanced applications. Database interoperability and data integration become important challenges in the current environment.

What are possible solutions to the above challenges? There are two observations that may be made. First, instead of expecting a radical movement away from relational data, future databases will probably provide various extensions of relational databases. Second, constraint databases, which are an extension of relational databases, appear to be suitable as a common standard for the various types of databases. In fact, geographic information systems sometimes convert internally their "vector data" into a constraint data to evaluate certain queries.

Purpose and Goals

This textbook provides comprehensive coverage of databases. The primary audience of the book is undergraduate computer science and non-computer science students with no or little prior exposure to databases. For them the extensive set of exercises at the end of each chapter will be useful. The text and the exercises assume as prerequisite only basic discrete mathematics, linear algebra, and programming knowledge. Many database experts will also find the bibliographic notes after each chapter a valuable reference for further reading. For both students and database experts the MLPQ constraint-relational database system is suggested for use. The MLPQ system, slides, solutions (for instructors), and other course aid is available free from the author's web page: `http://cse.unl.edu/~revesz`

Topics Coverage and Organization

This book has three parts:

- **Part I: Data and Queries:** The first part of the book defines databases in general (Chapter 1), describes eleven different types of databases (Chapters 2-12), and presents the MLPQ and the DISCO database systems (Chapters 13-14) that implement several different types of databases. The first part of the book is enough to start using a database system for simple applications.

- **Part II: Database Design and Applications:** The second part of the book describes database design (Chapter 15) and discusses the following advanced database application issues: database interoperability, which allows translating data from one database to another database (Chapter 16), data integration, which allows combining data from different sources (Chapter 17), interpolation and approximation, which allows a simple representation of complex data (Chapter 18), and prediction and data mining, which allows the discovery of new information (Chapter 19). The second part of the book enables the readers to develop more complex and rewarding database applications.

- **Part III: Query Evaluation, Safety, and Verification:** The third part of the book describes various issues related to the evaluation of database queries. The topics include indexing methods, which enable fast search in database tables (Chapter 20), data visualization, which is essential for a convenient interaction with the database system (Chapter 21), the safety of queries, that is, whether they are guaranteed to be evaluated precisely or approximately, or may enter an infinite loop and not terminate (Chapter 22), general evaluation algorithms (Chapter 23), the efficient implementation of the evaluation algorithms (Chapter 24), the complexity of the evaluation of different types of queries (Chapter 25). Computer software written in many programming languages can be translated into database queries. Software verification can be aided by translating the software to database queries, which are evaluated approximately by the above query evaluation algorithms (Chapter 26). The third part of the book gives students a deeper understanding and appreciation of the inside of database systems. It also enables students to read further scientific literature and begin research projects in database systems.

Figure 1 shows a dependency diagram of the chapters of the book. A special strength of the book is that it allows a flexible course design aided by the dependency diagram. We suggest the following courses.

- **Basic introductory course:** A basic introductory course, which is focused on a survey of various types of databases, would cover Chapters 1, 2, 3, and 4, any subset of Chapters 5-12, and the MLPQ system in Chapter 13. Such an introductory course would be suitable even for non-computer science majors. The topics could be easily fitted to the interest of the students. For example, geography majors may be interested in Chapters 5-8, while biology majors in Chapters 10-11. All of these students can be given a hands-on experience with a database system using the exercises in Chapter 13.

- **Enhanced introductory course:** An enhanced introductory course can add to the basic introductory course various applications-oriented chapters from the second part of the book (Chapters 15-19). These chapters enable more serious applications. For example, Chapter 18 discusses the triangulated irregular network (TIN) representation of surface data, which is important for geographic applications and enhances the material in Chapter 6. While Chapter 6 enables students to understand and use given geographic databases, Chapter 18 enables students to build geographic databases from measurement data. Many students may already have such measurement data from other projects outside of the database class.

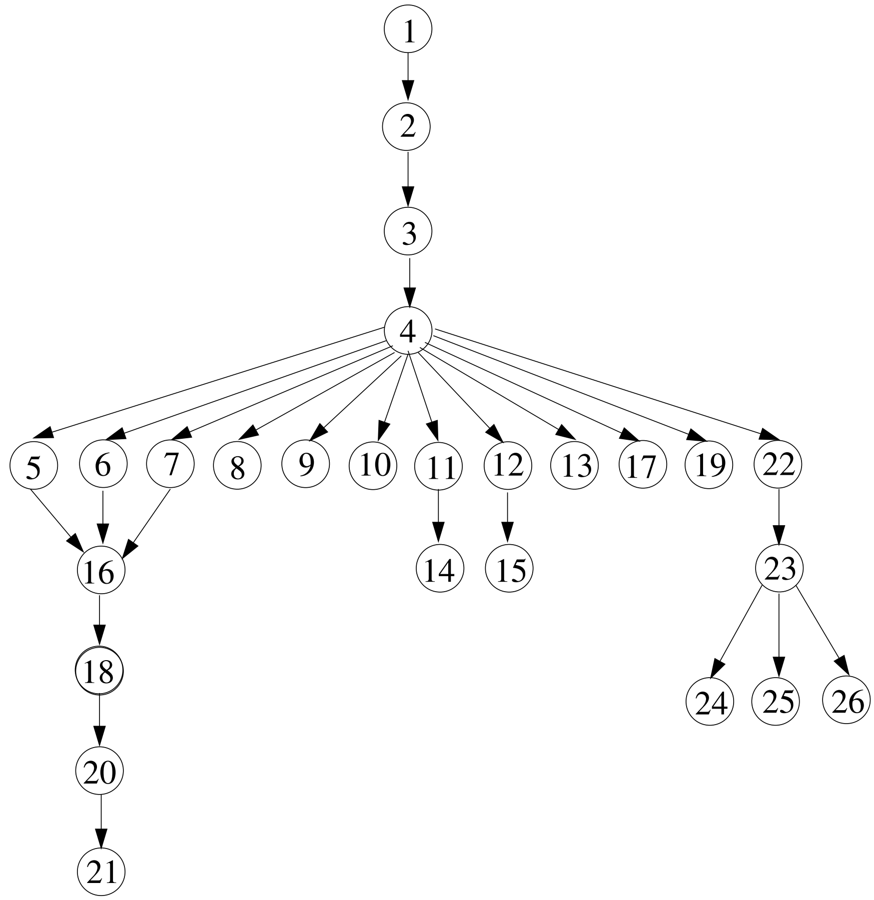

Figure 1: A dependency diagram of the chapters.

- **Advanced course:** The textbook also allows several types of advanced courses. An advanced course, focused on query evaluation, may cover interoperability (Chapter 16), indexing (Chapter 20), safety (Chapter 22), general evaluation algorithms (Chapter 23), and implementation methods (Chapter 24). These chapters already allow students to implement either individually or in group projects various simple query processors or propotype database systems. The coverage of the rest of the chapters depends on the interest of the students. Computer graphics students may study Chapter 21, theoretical computer science students may study Chapter 25, and software engineering students may study Chapter 26.

Acknowledgments

I'd like to thank for reviewing various parts of the book Berthe Choueiry, Manolis Koubarakis, Franz Mora, French Roy, and James Van Etten. I'd like to thank for introducing me to the database area the late Paris Kanellakis, my Ph.D. thesis advisor at Brown University, and the late Alberto Mendelzon, my postdoctoral advisor at the University of Toronto. I thank the following coauthors on books, journal articles or conference papers that contributed to the material in this book and from whom I learned much: András Benczur, Jan Chomicki, Floris Geerts, Sofie Haessevoets, Mark Griep, Gosta Grahne, Bart Kuijpers, Gabriel Kuper, Ram Narayanan, Agnes Novak, Robert Powers, James Pustejovsky, Raghu Ramakrishnan, Divesh Srivastava, Paolo Terenziani, and Jan Van den Bussche.

I would like to thank my former Ph.D. students for their comments on the text and other contributions to our database research group: Scot Anderson (Southern Adventist University), Mengchu Cai (IBM, DB2 Group), Rui Chen, Yi Chen, Ying Deng, Jun Gao (University of the Ozarks), Lixin Li (Georgia Southern University), Min Ouyang, Thomas Triplet (Concordia University), and Shasha Wu (Spring Arbor University). In addition, the following M.S. students also contributed significantly to the database research group: Brian Boon, Jo-Hag Byon, Vijay Eadala, Pradeep Gundavarapu, Pradip Kanjamala, Yiming Li, Yuguo Liu, Prashanth Nandavanam, Andrew Salamon, Yonghui Wang, and Lei Zhang.

The research, which forms several chapters of this book, was supported in part by the U.S. National Science Foundation under grants IRI-9625055, IRI-9632871, and EIA-0091530, by a NASA grant, by a Gallup Research Professorship, by an Alexander von Humboldt Research Fellowship, and by a J. William Fulbright Scholarship. The latter two awards enabled research leaves from the University of Nebraska-Lincoln to the Max Planck Institut für Informatik and the University of Athens, respectively. Andreas Podelski at the former and Manolis Koubarakis at the latter place served as great hosts during my visits.

I am grateful to the editors at Springer-Verlag, especially Wayne Wheeler, Senior Editor for Computer Science, and Simon Rees, Senior Editorial Assistant, for their careful editing and attention to several details in the production of the book.

Finally, I would like to thank Lilla, my wife, and our children, Sophie, Claire, and Gregory, for their enthusiasm and patience during the writing of this book.

Peter Revesz

Lincoln, Nebraska, USA
July 2009

Contents