Principles of Database Management

The Practical Guide to Storing, Managing and Analyzing Big and Small Data

Principles of Database Management provides students with the comprehensive database management information to understand and apply the fundamental concepts of database design and modeling, database systems, data storage and the evolving world of data warehousing, governance and more. Designed for those studying database management for information management or computer science, this illustrated textbook has a well-balanced theory–practice focus and covers the essential topics, from established database technologies up to recent trends like Big Data, NoSQL and analytics. On-going case studies, drill-down boxes that reveal deeper insights on key topics, retention questions at the end of every section of a chapter, and connections boxes that show the relationship between concepts throughout the text are included to provide the practical tools to get started in database management.

Key features include:

- Full-color illustrations throughout the text.
- Extensive coverage of important trending topics, including data warehousing, business intelligence, data integration, data quality, data governance, Big Data and analytics.
- An online playground with diverse environments, including MySQL for querying; MongoDB; Neo4j Cypher; and a tree structure visualization environment.
- Hundreds of examples to illustrate and clarify the concepts discussed that can be reproduced on the book's companion online playground.
- Case studies, review questions, problems and exercises in every chapter.
- Additional cases, problems and exercises in the appendix.

"Although there have been a series of classical textbooks on database systems, the new dramatic advances call for an updated text covering the latest significant topics, such as Big Data analytics, NoSQL and much more. Fortunately, this is exactly what this book has to offer. It is highly desirable for training the next generation of data management professionals."

– Jian Pei, *Simon Fraser University*

"I haven't seen an as up-to-date and comprehensive textbook for database management as this one in many years. *Principles of Database Management* combines a number of classical and recent topics concerning data modeling, relational databases, object-oriented databases, XML, distributed data management, NoSQL and Big Data in an unprecedented manner. The authors did a great job in stitching these topics into one coherent and compelling story that will serve as an ideal basis for teaching both introductory and advanced courses."

– Martin Theobald, *University of Luxembourg*

"This is a very timely book with outstanding coverage of database topics and excellent treatment of database details. It not only gives very solid discussions of traditional topics such as data modeling and relational databases, but also contains refreshing contents on frontier topics such as XML databases, NoSQL databases, Big Data and analytics. For those reasons, this will be a good book for database professionals, who will keep using it for all stages of database studies and works."

– J. Leon Zhao, *City University of Hong Kong*

"This accessible, authoritative book introduces the reader the most important fundamental concepts of data management, while providing a practical view of recent advances. Both are essential for data professionals today."

– Foster Provost, *New York University, Stern School of Business*

"This guide to big and small data management addresses both fundamental principles and practical deployment. It reviews a range of databases and their relevance for analytics. The book is useful to practitioners because it contains many case studies, links to open-source software, and a very useful abstraction of analytics that will help them choose solutions better. It is important to academics because it promotes database principles which are key to successful and sustainable data science."

– Sihem Amer-Yahia, *Laboratoire d'Informatique de Grenoble*; Editor-in-Chief, *The VLDB Journal (International Journal on Very Large DataBases)*

"This book covers everything you will need to teach in a database implementation and design class. With some chapters covering Big Data, analytic models/methods and NoSQL, it can keep our students up to date with these new technologies in data management-related topics."

– Han-fen Hu, *University of Nevada, Las Vegas*

Principles of Database Management

The Practical Guide to Storing, Managing and Analyzing Big and Small Data

Wilfried Lemahieu
KU Leuven, Belgium
Seppe vanden Broucke
KU Leuven, Belgium
Bart Baesens
KU Leuven, Belgium; University of Southampton, United Kingdom

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107186125
DOI: 10.1017/9781316888773

First published 2018

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Lemahieu, Wilfried, 1970– author. | Broucke, Seppe vanden, 1986– author. | Baesens, Bart, author.
Title: Principles of database management : the practical guide to storing, managing and analyzing big and small data / Wilfried Lemahieu, Katholieke Universiteit Leuven, Belgium, Seppe vanden Broucke, Katholieke Universiteit Leuven, Belgium, Bart Baesens, Katholieke Universiteit Leuven, Belgium.
Description: First edition. | New York, NY : Cambridge University Press, 2018. | Includes bibliographical references and index.
Identifiers: LCCN 2018023251 | ISBN 9781107186125 (hardback : alk. paper)
Subjects: LCSH: Database management.
Classification: LCC QA76.9.D3 L454 2018 | DDC 005.74–dc23
LC record available at https://lccn.loc.gov/2018023251

ISBN 978-1-107-18612-5 Hardback

Additional resources for this publication at www.cambridge.org/Lemahieu

BRIEF CONTENTS

CONTENTS

ABOUT THE AUTHORS

Bart was born in Bruges (Belgium). He speaks West-Flemish, Dutch, French, a bit of German, some English, and can order a beer in Chinese. Besides enjoying time with his family, he is also a diehard Club Brugge soccer fan. Bart is a foodie and amateur cook and loves a good glass of wine overlooking the authentic red English phone booth in his garden. Bart loves traveling; his favorite cities are San Francisco, Sydney, and Barcelona. He is fascinated by World War I and reads many books on the topic. He is not a big fan of being called "Professor Baesens", shopping, vacuuming, long meetings, phone calls, admin, or students chewing gum during their oral exam on database management. He is often praised for his sense of humor, although he is usually more modest about this.

Bart is a professor of Big Data and analytics at KU Leuven (Belgium) and a lecturer at the University of Southampton (United Kingdom). He has done extensive research on Big Data and analytics, credit risk modeling, fraud detection, and marketing analytics. He has written more than 200 scientific papers and six books. He has received various best paper and best speaker awards. His research is summarized at www.dataminingapps.com.

Seppe was born in Jette (Brussels, Belgium), but has lived most of his life in Leuven. Seppe speaks Dutch, some French, English, understands German, and can order a beer in Chinese (and unlike Bart he can do so in the right intonation, having studied Mandarin for three years). He is married to Xinwei Zhu (which explains the three years of Mandarin). Besides spending time with family, Seppe enjoys traveling, reading (Murakami to Bukowski to Asimov), listening to music (Booka Shade to Miles Davis to Claude Debussy), watching movies and series, gaming, and keeping up with the news. He is not a fan of any physical activity other than walking way too fast through Leuven. Seppe does not like vacuuming (this seems to be common with database book authors), bureaucracy, meetings, public transportation (even though he has no car) or Windows updates that start when he is teaching or writing a book chapter.

Seppe is an assistant professor at the Faculty of Economics and Business, KU Leuven, Belgium. His research interests include business data mining and analytics, machine learning, process management, and process mining. His work has been published in well-known international journals and presented at top conferences. Seppe's teaching includes advanced analytics, Big Data, and information management courses. He also frequently teaches for industry and business audiences. See www.seppe.net for further details.

Wilfried was born in Turnhout, Belgium. He speaks Dutch, English, and French, and can decipher some German, Latin, and West-Flemish. Unable to order a beer in Chinese, he has perfected a "looking thirsty" facial expression that works in any language. He is married to Els Mennes, and together they produced three sons – Janis, Hannes, and Arne – before running out of boys' names. Apart from family time, one of Wilfried's most cherished pastimes is music. Some would say he is stuck in the eighties, but his taste ranges from Beethoven to Hendrix and from Cohen to The Cure. He also likes traveling, with fond memories of Alaska, Bali, Cuba, Beijing, the Swiss Alps, Rome, and Istanbul. He enjoys many different genres of movies, but is somewhat constrained by his wife's bias toward tearful-kiss-and-make-up-at-the-airport scenes. His sports watch contains data (certainly no Big Data!) on erratic attempts at running, swimming, biking, and skiing. Wilfried has no immediate aversion to vacuuming, although his fellow household members would claim that his experience with the matter is mainly theoretical.

Wilfried is a full professor at the Faculty of Economics and Business (FEB) of KU Leuven, Belgium. He conducts research on (big) data storage, integration, and analytics; data quality; business process management and service orchestration, often in collaboration with industry partners. Following his position of Vice Dean for Education at FEB, he was elected as Dean in 2017. See www.feb.kuleuven.be/wilfried.lemahieu for further details.

PREFACE

Congratulations! By picking up this book, you have made the first step in your journey through the wonderful world of databases. As you will see in this book, databases come in many different forms – from simple spreadsheets or other file-based attempts and hierarchical structures, to relational, object-oriented, and even graph-oriented ones – and are used across the world throughout a variety of industry sectors to manage, store, and analyze data.

This book is the result of having taught an undergraduate database management class and a postgraduate advanced database management class for more than ten years. Throughout the years we have found no textbook that covers the material in a comprehensive way without becoming flooded by theoretical detail and losing focus. Hence, after we teamed up together, we decided to start writing a book ourselves. This work aims to offer a complete and practical guide covering all the governing principles of database management, including:

- end-to-end coverage, starting with legacy technologies to emerging trends such as Big Data, NoSQL databases, analytics, data governance, etc.;
- a unique perspective on how lessons learned from past data management could be relevant in today's technology setting (e.g., navigational access and its perils in CODASYL and XML/OO databases);
- a critical reflection and accompanying risk management considerations when implementing the technologies considered, based on our own experiences participating in data and analytics-related projects with industry partners in a variety of sectors, from banking to retail and from government to the cultural sector;
- a solid balance between theory and practice, including various exercises, industry examples and case studies originating from diverse and complementary business practices, scientific research, and academic teaching experience.

The book also includes an appendix explaining our "online playground" environment, where you can try out many concepts discussed in the book. Additional appendices, including an exam bank containing several cross-chapter questions and references to our YouTube lectures, are provided online as well.

We hope you enjoy this book and that you, the reader, will find it a useful reference and trusted companion in your work, studies, or research when storing, managing, and analyzing small or Big Data!

Who This Book is For

We have tried to make this book complete and useful for both novice and advanced database practitioners and students alike. No matter whether you're a novice just beginning to work with

database management systems, a versed SQL user aiming to brush up your knowledge of underlying concepts or theory, or someone looking to get an update on newer, more modern database approaches, this book aims to familiarize you with all the necessary concepts. Hence, this book is well suited for:

- under- or postgraduate students taking courses on database management in BSc and MSc programs in information management and/or computer science;
- business professionals who would like to refresh or update their knowledge on database management; and
- information architects, database designers, data owners, data stewards, database administrators, or data scientists interested in new developments in the area.

Thanks to the exercises and industry examples throughout the chapters, the book can also be used by tutors in courses such as:

- principles of database management;
- database modeling;
- database design;
- database systems;
- data management;
- data modeling;
- data science.

It can also be useful to universities working out degrees in, for example, Big Data and analytics.

Topics Covered in this Book

This book is organized in four main parts. Chapters 1–4 address preliminary and introductory topics regarding databases and database design, starting with an introduction to basic concepts in Chapter 1, followed by a description of common database management system types and their architecture in Chapter 2. Chapter 3 discusses conceptual data modeling, and Chapter 4 provides a management overview of the different roles involved in data management and their responsibilities.

Part II (Chapters 5–11) then takes a dive into the various types of databases, from legacy pre-relational and relational database management systems into more recent approaches such as object-oriented, object-relational, and XML-based databases in Chapters 8–10, ending with a solid and up-to-date overview of NoSQL technologies in Chapter 11. This part also includes a comprehensive overview of the Structured Query Language (SQL) in Chapter 7.

In Part III, physical data storage, transaction management, and database access are discussed in depth. Chapter 12 discusses physical file organization and indexing, whereas Chapter 13 elaborates on physical database organization and business continuity. This is followed by an overview on the basics of transaction management in Chapter 14. Chapter 15 introduces database access mechanisms and various database application programming interfaces (APIs). Chapter 16 concludes this part of the book by zooming in on data distribution and distributed transaction management.

Chapters 17–20 form the last part of the book. Here, we zoom out and elaborate on data warehousing and emerging interest areas such as data governance, Big Data, and analytics. Chapter 17 discusses data warehouses and business intelligence in depth; Chapter 18 covers

managerial concepts such as data integration, data quality, and data governance; Chapter 19 provides an in-depth overview of Big Data and shows how a solid database set-up can form the cornerstone of a modern analytical environment. Chapter 20 concludes this part and the book by examining different types of analytics.

By the end of the book, you will have gained a strong knowledge of all aspects that make up a database management system. You will be able to discern the different database systems, and to contrast their advantages and disadvantages. You will be able to make the best (investment) decisions through conceptual, logical, and physical data modeling, all the way to Big Data and analytical applications. You'll have gained a strong understanding of SQL, and will also understand how database management systems work at the physical level – including transaction management and indexing. You'll understand how database systems are accessed from the outside world and how they can be integrated with other systems or applications. Finally, you'll also understand the various managerial aspects that come into play when working with databases, including the roles involved, data integration, quality, and governance aspects, and you will have a clear idea on how the concept of database management systems fits in the Big Data and analytics story.

How to Read this Book

This book can be used as both a reference manual for more experienced readers wishing to brush up their skills and knowledge regarding certain aspects, as well as an end-to-end overview on the whole area of database management systems. Readers are free to read this book cover to cover, or to skip certain chapters and start directly with a topic of interest. We have separated the book clearly into different parts and chapters so readers should have little trouble understanding the global structure of the book and navigating to the right spot. Whenever a topic is expanded upon in a later chapter or re-uses concepts introduced in an earlier chapter, we include clear "Connections" boxes so readers can (re-)visit earlier chapters for a quick refresher before moving on, or move ahead to other places in the book to continue their learning trail.

The following overview provides some common "reading trails", depending on your area of interest:

- Newcomers wishing to get up to speed quickly with relational database systems and SQL: start with Part I (Chapters 1–4), then read Chapters 6–9.
- Experienced users wishing to update their knowledge on recent trends: read Chapter 11, and then Chapters 15–20.
- Daily database users wishing to have high-level knowledge about database systems: Part I (Chapters 1–4) is for you.
- Managers wishing to get a basic overview on fundamental concepts and a broad idea of managerial issues: start with Part I (Chapters 1–4), then move on to Chapters 17, 18, 19, and 20.
- Professors teaching an undergraduate course in database management: Parts I and II.
- Professors teaching a postgraduate course in advanced database management: Parts III and IV.

The recommended chapters for each of these profiles, together with some others (which will be discussed in Chapter 4), are summarized in the table.

Chapter	Newcomers	Experienced users	Database users	Managers	Professor (undergraduate course)	Professor (postgraduate course)	Information architect	Database designer	Database administrator	Data scientist
1	X		X	X	X		X	X	X	
2	X		X	X	X		X	X	X	
3	X		X	X	X		X	X	X	
4	X		X	X	X		X	X	X	
5					X			X	X	
6	X				X			X	X	
7	X				X			X	X	
8	X				X			X	X	
9	X				X			X	X	
10					X			X	X	
11		X			X			X	X	X
12						X		X	X	
13						X		X	X	
14						X			X	
15		X				X				
16		X				X			X	
17		X		X		X			X	X
18		X		X		X			X	X
19		X		X		X			X	X
20		X		X		X				X

Every chapter aims to strike a balance between theory and practice, so theoretical concepts are often alternated with examples from industry in small "Drill Down" boxes that provide more background knowledge or an interesting story to illustrate a concept. We also include theoretical discussions on pros and cons of a specific technique or technology. Each chapter closes with a set of exercises to test your understanding. Both multiple-choice and open questions have been included.

Cross-Chapter Case Study: Sober

Throughout the book we use an encompassing case (about a fictional self-driving car taxi company called "Sober") that will be revisited and expanded in each chapter. When reading the book from cover to cover you'll therefore be able to learn together with the people at Sober, experiencing how their database management system evolves from a simple small-scale system toward a more modern and robust set-up as they continue to grow. This way, the different chapters also form a cohesive whole from a practical perspective, and you'll see how all the technologies and concepts fit together.

Additional Material

We are also happy to refer you to our book website at www.pdbmbook.com. The site includes additional information such as updates, PowerPoint slides, video lectures, additional appendices, and a Q&A section. It also features a hands-on, online environment where readers can play around

with a MySQL relational database management system using SQL, explore NoSQL database systems, and other small examples without having to install anything. You'll find a guide in the Appendix that will set you on your way.

Acknowledgments

It is a great pleasure to acknowledge the contributions and assistance of various colleagues, friends, and fellow database management lovers in the writing of this book. This book is the result of many years of research and teaching in database management.

We first would like to acknowledge our publisher, Cambridge University Press, for accepting our book proposal about two years ago. We would like to thank Lauren Cowles for supervising the entire process. We first met Lauren in August 2016 in San Francisco, discussing the book details during dinner (crab cakes paired with Napa white) while overlooking an ensemble of sunbathing seals. This turned out to be the perfect setting for initiating a successful partnership. We are also thankful to everyone at Cambridge University Press for their help in the editing, production, and marketing processes.

Gary J. O'Brien deserves a special mention as well. His careful proofreading of the text proved invaluable. Although opening a Word document with Gary's comments sometimes felt like being thrown in the ocean knowing sharks had been spotted, the mix of to-the-point remarks with humorous notes made the revision a truly enjoyable experience.

We would like to thank professor Jacques Vandenbulcke, who was the first to introduce us to the magical world of database management. Jacques' exquisite pedagogical talent can only be surpassed by his travel planning skills. His legacy runs throughout the entire book, not only in terms of database concepts and examples, but also travel experiences (e.g., the Basilica Cistern on the front cover, Meneghetti wine).

We would also like to acknowledge the direct and indirect contributions of the many colleagues, fellow professors, students, researchers, business contacts, and friends with whom we collaborated during the past years. We are grateful to the active and lively database management community for providing various user fora, blogs, online lectures, and tutorials that proved very helpful.

Last but not least, we are grateful to our partners, kids, parents, and families for their love, support, and encouragement! We trust they will read this book from the first page to the last, which will yield ample topics for lively and interesting discussions at the dinner table.

We have tried to make this book as complete, accurate, and enjoyable as possible. Of course, what really matters is what you, the reader, think of it. Please share your views by getting in touch. The authors welcome all feedback and comments, so do not hesitate to let us know your thoughts.

Front cover: The cover picture represents the Basilica Cistern, an immense subterranean water storage facility built in the sixth century by the Romans in Istanbul. Why this picture? Well, overall it is a spectacular location in a truly magnificent city, which throughout its history has been a meeting point of cultures, civilizations, and, literally, continents. However, more to the point, it is definitely a storage infrastructure organized as rows and columns, which even involves replication and mirroring, not to mention historical data. In addition, it contained one of the most famous primary keys ever: 007, as it featured prominently in the James Bond movie *From Russia With Love*.

Sober

1000‰ Driven by Technology

Sober is a new taxi company deploying self-driving cars to provide cab services. Although it operates its own fleet of self-driving cabs, people can also register their cars as Sober cabs and have them provide taxi services whenever they are not using their cars. For the latter, Sober also wants to keep track of the car owners.

Sober offers two types of taxi services: ride-hailing and ride-sharing. Ride-hailing is a service whereby customers can hail a taxi so they can be picked up and driven to their destination for a time- and distance-based fee. The hailing is an immediate, on-demand service and requests can be made with the Sober App. With just one tap on the screen, a customer can request a cab from anywhere, receive an estimated wait time, and a notification when the car has arrived. Besides the Sober App, users can also hail Sober cabs by hand-waving them as they see them pass, in which case Sober's deep-learning based image recognition system identifies the wave gesture as a cab request. For each use of the ride-hail service, Sober wants to store the time of pick-up and drop-off, the location of pick-up and drop-off, the ride duration, the distance, the number of passengers, the fee, the type of request (via Sober App or hand-waving) and the number and name of the lead customer (the one who pays). The maximum number of passengers for a ride-hail service is six.

Ride-sharing is another service offered by Sober, which requires more careful planning. It can also be referred to as carpooling and aims at reducing costs, traffic congestion, and the carbon footprint. Because of the planning, both Sober and its customers can negotiate the fee whereby more customers per cab means a lower fee per customer (flexible pricing). To provide an eco-friendly incentive, Sober pledges to plant a tree for each customer who books 20 uses of the Sober ride-sharing service. For each ride-share service, Sober wants to store the time of pick-up and drop-off, the location of pick-up and drop-off, the ride duration, the distance, the number and names of all customers, and the upfront negotiated fee. The maximum number of passengers for a ride-share service is ten.

Due to the novelty of the self-driving car technology, accidents cannot be 100% ruled out. Sober also wants to store information about accident dates, location, and damage amounts per car.