

## 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](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9781107186125](http://www.cambridge.org/9781107186125)

DOI: 10.1017/9781316888773

© Wilfried Lemahieu, Seppe vanden Broucke, and Bart Baesens 2018

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

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://lcn.loc.gov/2018023251>

ISBN 978-1-107-18612-5 Hardback

Additional resources for this publication at [www.cambridge.org/Lemahieu](http://www.cambridge.org/Lemahieu)

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

## BRIEF CONTENTS

About the Authors	<i>page</i> xvii	<b>Part III Physical Data Storage, Transaction Management, and Database Access</b>	<b>349</b>
Preface	xix		
Sober: 1000% Driven by Technology	xxiv		
<b>Part I Databases and Database Design</b>	<b>1</b>	<b>12 Physical File Organization and Indexing</b>	<b>351</b>
<b>1 Fundamental Concepts of Database Management</b>	<b>3</b>	<b>13 Physical Database Organization</b>	<b>395</b>
<b>2 Architecture and Categorization of DBMSs</b>	<b>20</b>	<b>14 Basics of Transaction Management</b>	<b>430</b>
<b>3 Conceptual Data Modeling Using the (E)ER Model and UML Class Diagram</b>	<b>38</b>	<b>15 Accessing Databases and Database APIs</b>	<b>458</b>
<b>4 Organizational Aspects of Data Management</b>	<b>79</b>	<b>16 Data Distribution and Distributed Transaction Management</b>	<b>516</b>
<b>Part II Types of Database Systems</b>	<b>91</b>	<b>Part IV Data Warehousing, Data Governance, and (Big) Data Analytics</b>	<b>549</b>
<b>5 Legacy Databases</b>	<b>93</b>	<b>17 Data Warehousing and Business Intelligence</b>	<b>551</b>
<b>6 Relational Databases: The Relational Model</b>	<b>104</b>	<b>18 Data Integration, Data Quality, and Data Governance</b>	<b>590</b>
<b>7 Relational Databases: Structured Query Language (SQL)</b>	<b>146</b>	<b>19 Big Data</b>	<b>626</b>
<b>8 Object-Oriented Databases and Object Persistence</b>	<b>207</b>	<b>20 Analytics</b>	<b>664</b>
<b>9 Extended Relational Databases</b>	<b>231</b>	Appendix Using the Online Environment	731
<b>10 XML Databases</b>	<b>255</b>	Glossary	741
<b>11 NoSQL Databases</b>	<b>300</b>	Index	770

Cambridge University Press  
978-1-107-18612-5 — Principles of Database Management  
Wilfried Lemahieu , Seppe vanden Broucke , Bart Baesens  
Frontmatter  
[More Information](#)

---

# CONTENTS

About the Authors	page xvii	<b>2</b>	<b>Architecture and Categorization of DBMSs</b>	<b>20</b>
Preface	xix			
Sober: 1000% Driven by Technology	xxiv			
<b>Part I Databases and Database Design</b>	<b>1</b>			
<b>1 Fundamental Concepts of Database Management</b>	<b>3</b>			
1.1 Applications of Database Technology	3		2.1 Architecture of a DBMS	20
1.2 Key Definitions	4		2.1.1 Connection and Security Manager	21
1.3 File versus Database Approach to Data Management	5		2.1.2 DDL Compiler	22
1.3.1 The File-Based Approach	5		2.1.3 Query Processor	22
1.3.2 The Database Approach	6		2.1.3.1 DML Compiler	22
1.4 Elements of a Database System	8		2.1.3.2 Query Parser and Query Rewriter	25
1.4.1 Database Model versus Instances	8		2.1.3.3 Query Optimizer	25
1.4.2 Data Model	9		2.1.3.4 Query Executor	25
1.4.3 The Three-Layer Architecture	10		2.1.4 Storage Manager	25
1.4.4 Catalog	10		2.1.4.1 Transaction Manager	25
1.4.5 Database Users	11		2.1.4.2 Buffer Manager	26
1.4.6 Database Languages	12		2.1.4.3 Lock Manager	26
1.5 Advantages of Database Systems and Database Management	12		2.1.4.4 Recovery Manager	26
1.5.1 Data Independence	12		2.1.5 DBMS Utilities	26
1.5.2 Database Modeling	13		2.1.6 DBMS Interfaces	27
1.5.3 Managing Structured, Semi-Structured, and Unstructured Data	13		<b>2.2 Categorization of DBMSs</b>	<b>27</b>
1.5.4 Managing Data Redundancy	14		2.2.1 Categorization Based on Data Model	28
1.5.5 Specifying Integrity Rules	14		2.2.1.1 Hierarchical DBMSs	28
1.5.6 Concurrency Control	14		2.2.1.2 Network DBMSs	28
1.5.7 Backup and Recovery Facilities	15		2.2.1.3 Relational DBMSs	28
1.5.8 Data Security	15		2.2.1.4 Object-Oriented DBMSs	28
1.5.9 Performance Utilities	16		2.2.1.5 Object-Relational/Extended Relational DBMSs	29
Summary	16		2.2.1.6 XML DBMSs	29
Key Terms List	16		2.2.1.7 NoSQL DBMSs	30
Review Questions	17		2.2.2 Categorization Based on Degree of Simultaneous Access	30
Problems and Exercises	19		2.2.3 Categorization Based on Architecture	30
			2.2.4 Categorization Based on Usage	31
			Summary	33
			Key Terms List	33
			Review Questions	34
			Problems and Exercises	37

<b>3</b>	<b>Conceptual Data Modeling Using the (E)ER Model and UML Class Diagram</b>	<b>38</b>			
3.1	Phases of Database Design	38			
3.2	The Entity Relationship Model	40			
3.2.1	Entity Types	40			
3.2.2	Attribute Types	40			
3.2.3.1	Domains	41			
3.2.3.2	Key Attribute Types	42			
3.2.3.3	Simple versus Composite Attribute Types	42			
3.2.3.4	Single-Valued versus Multi-Valued Attribute Types	43			
3.2.3.5	Derived Attribute Type	43			
3.2.4	Relationship Types	43			
3.2.4.1	Degree and Roles	44			
3.2.4.2	Cardinalities	45			
3.2.4.3	Relationship Attribute Types	46			
3.2.5	Weak Entity Types	46			
3.2.6	Ternary Relationship Types	48			
3.2.7	Examples of the ER Model	50			
3.2.8	Limitations of the ER Model	51			
3.3	The Enhanced Entity Relationship (EER) Model	52			
3.3.1	Specialization/Generalization	52			
3.3.2	Categorization	54			
3.3.3	Aggregation	55			
3.3.4	Examples of the EER Model	55			
3.3.5	Designing an EER Model	56			
3.4	The UML Class Diagram	57			
3.4.1	Recap of Object Orientation	57			
3.4.2	Classes	58			
3.4.3	Variables	58			
3.4.4	Access Modifiers	59			
3.4.5	Associations	59			
3.4.5.1	Association Class	60			
3.4.5.2	Unidirectional versus Bidirectional Association	60			
3.4.5.3	Qualified Association	61			
3.4.6	Specialization/Generalization	62			
3.4.7	Aggregation	62			
3.4.8	UML Example	63			
3.4.9	Advanced UML Modeling Concepts	64			
3.4.9.1	Changeability Property	64			
3.4.9.2	Object Constraint Language (OCL)	64			
3.4.9.3	Dependency Relationship	66			
3.4.10	UML versus EER	66			
	Summary		67		
	Key Terms List		71		
	Review Questions		71		
	Problems and Exercises		75		
<b>4</b>	<b>Organizational Aspects of Data Management</b>		<b>79</b>		
4.1	Data Management		79		
4.1.1	Catalogs and the Role of Metadata		80		
4.1.2	Metadata Modeling		80		
4.1.3	Data Quality		81		
4.1.3.1	Data Quality Dimensions		82		
4.1.3.2	Data Quality Problems		84		
4.1.4	Data Governance		85		
4.2	Roles in Data Management		86		
4.2.1	Information Architect		86		
4.2.2	Database Designer		87		
4.2.3	Data Owner		87		
4.2.4	Data Steward		87		
4.2.5	Database Administrator		87		
4.2.6	Data Scientist		88		
	Summary		88		
	Key Terms List		89		
	Review Questions		89		
	Problems and Exercises		90		
<b>Part II Types of Database Systems</b>			<b>91</b>		
<b>5</b>	<b>Legacy Databases</b>		<b>93</b>		
5.1	The Hierarchical Model		93		
5.2	The CODASYL Model		97		
	Summary		102		
	Key Terms List		102		
	Review Questions		102		
	Problems and Exercises		103		
<b>6</b>	<b>Relational Databases: The Relational Model</b>		<b>104</b>		
6.1	The Relational Model		105		
6.1.1	Basic Concepts		105		
6.1.2	Formal Definitions		106		
6.1.3	Types of Keys		108		
6.1.3.1	Superkeys and Keys		108		
6.1.3.2	Candidate Keys, Primary Keys, and Alternative Keys		108		
6.1.3.3	Foreign Keys		109		
6.1.4	Relational Constraints		111		
6.1.5	Example Relational Data Model		111		



6.2 Normalization	111	7.1.1 Key Characteristics of SQL	147
6.2.1 Insertion, Deletion, and Update Anomalies in an Unnormalized Relational Model	112	7.1.2 Three-Layer Database Architecture	149
6.2.2 Informal Normalization Guidelines	114	7.2 SQL Data Definition Language	149
6.2.3 Functional Dependencies and Prime Attribute Type	114	7.2.1 Key DDL Concepts	150
6.2.4 Normalization Forms	115	7.2.2 DDL Example	151
6.2.4.1 First Normal Form (1NF)	115	7.2.3 Referential Integrity Constraints	154
6.2.4.2 Second Normal Form (2NF)	117	7.2.4 DROP and ALTER Command	155
6.2.4.3 Third Normal Form (3NF)	118	7.3 SQL Data Manipulation Language	156
6.2.4.4 Boyce–Codd Normal Form (BCNF)	119	7.3.1 SQL SELECT Statement	156
6.2.4.5 Fourth Normal Form (4NF)	120	7.3.1.1 Simple Queries	157
6.3 Mapping a Conceptual ER Model to a Relational Model	121	7.3.1.2 Queries with Aggregate Functions	161
6.3.1 Mapping Entity Types	121	7.3.1.3 Queries with GROUP BY/HAVING	163
6.3.2 Mapping Relationship Types	122	7.3.1.4 Queries with ORDER BY	165
6.3.2.1 Mapping a Binary 1:1 Relationship type	122	7.3.1.5 Join Queries	166
6.3.2.2 Mapping a Binary 1:N Relationship Type	124	7.3.1.6 Nested Queries	172
6.3.2.3 Mapping a Binary M:N Relationship Type	126	7.3.1.7 Correlated Queries	175
6.3.2.4 Mapping Unary Relationship Types	127	7.3.1.8 Queries with ALL/ANY	178
6.3.2.5 Mapping <i>n</i> -ary Relationship Types	129	7.3.1.9 Queries with EXISTS	181
6.3.3 Mapping Multi-Valued Attribute Types	130	7.3.1.10 Queries with Subqueries in SELECT/FROM	182
6.3.4 Mapping Weak Entity Types	131	7.3.1.11 Queries with Set Operations	183
6.3.5 Putting it All Together	132	7.3.2 SQL INSERT Statement	185
6.4 Mapping a Conceptual EER Model to a Relational Model	133	7.3.3 SQL DELETE Statement	185
6.4.1 Mapping an EER Specialization	133	7.3.4 SQL UPDATE Statement	186
6.4.2 Mapping an EER Categorization	136	7.4 SQL Views	188
6.4.3 Mapping an EER Aggregation	137	7.5 SQL Indexes	190
Summary	138	7.6 SQL Privileges	191
Key Terms List	139	7.7 SQL for Metadata Management	192
Review Questions	139	Summary	194
Problems and Exercises	143	Key Terms List	195
<b>7 Relational Databases: Structured   Query Language (SQL)</b>	<b>146</b>	Review Questions	196
7.1 Relational Database Management Systems and SQL	147	Problems and Exercises	205
		<b>8 Object-Oriented Databases and   Object Persistence</b>	<b>207</b>
		8.1 Recap: Basic Concepts of OO	208
		8.2 Advanced Concepts of OO	209
		8.2.1 Method Overloading	209
		8.2.2 Inheritance	210
		8.2.3 Method Overriding	211
		8.2.4 Polymorphism and Dynamic Binding	212
		8.3 Basic Principles of Object Persistence	214
		8.3.1 Serialization	214

8.4	OODBMS	216	10.1.2	Document Type Definition and XML Schema Definition	260
8.4.1	Object Identifiers	216	10.1.3	Extensible Stylesheet Language	263
8.4.2	ODMG Standard	217	10.1.4	Namespaces	266
8.4.3	Object Model	217	10.1.5	XPath	267
8.4.4	Object Definition Language (ODL)	218	10.2	Processing XML Documents	267
8.4.5	Object Query Language (OQL)	221	10.3	Storage of XML Documents	269
8.4.5.1	Simple OQL Queries	221	10.3.1	The Document-Oriented Approach for Storing XML Documents	270
8.4.5.2	SELECT FROM WHERE OQL Queries	221	10.3.2	The Data-Oriented Approach for Storing XML Documents	270
8.4.5.3	Join OQL Queries	222	10.3.3	The Combined Approach for Storing XML Documents	270
8.4.5.4	Other OQL Queries	222	10.4	Differences Between XML Data and Relational Data	271
8.4.6	Language Bindings	223	10.5	Mappings Between XML Documents and (Object-) Relational Data	272
8.5	Evaluating OODBMSs	225	10.5.1	Table-Based Mapping	272
	Summary	227	10.5.2	Schema-Oblivious Mapping	273
	Key Terms List	227	10.5.3	Schema-Aware Mapping	275
	Review Questions	228	10.5.4	SQL/XML	276
	Problems and Exercises	229	10.6	Searching XML Data	279
<b>9</b>	<b>Extended Relational Databases</b>	<b>231</b>	10.6.1	Full-Text Search	280
9.1	Limitations of the Relational Model	231	10.6.2	Keyword-Based Search	280
9.2	Active RDBMS Extensions	232	10.6.3	Structured Search With XQuery	280
9.2.1	Triggers	233	10.6.4	Semantic Search With RDF and SPARQL	282
9.2.2	Stored Procedures	234	10.7	XML for Information Exchange	284
9.3	Object-Relational RDBMS Extensions	236	10.7.1	Message-Oriented Middleware	284
9.3.1	User-Defined Types	236	10.7.2	SOAP-Based Web Services	285
9.3.1.1	Distinct Data Types	237	10.7.3	REST-Based Web Services	288
9.3.1.2	Opaque Data Types	238	10.7.4	Web Services and Databases	289
9.3.1.3	Unnamed Row Types	238	10.8	Other Data Representation Formats	290
9.3.1.4	Named Row Types	239		Summary	293
9.3.1.5	Table Data Types	240		Key Terms List	296
9.3.2	User-Defined Functions	240		Review Questions	297
9.3.3	Inheritance	242		Problems and Exercises	298
9.3.3.1	Inheritance at Data Type Level	242	<b>11</b>	<b>NoSQL Databases</b>	<b>300</b>
9.3.3.2	Inheritance at Table Type Level	243	11.1	The NoSQL Movement	301
9.3.4	Behavior	244	11.1.1	The End of the “One Size Fits All” Era?	301
9.3.5	Polymorphism	244			
9.3.6	Collection Types	245			
9.3.7	Large Objects	247			
9.4	Recursive SQL Queries	247			
	Summary	250			
	Key Terms List	251			
	Review Questions	252			
	Problems and Exercises	253			
<b>10</b>	<b>XML Databases</b>	<b>255</b>			
10.1	Extensible Markup Language	256			
10.1.1	Basic Concepts	256			

11.1.2 The Emergence of the NoSQL Movement	302	12.3.4.1 Key-to-Address Transformation	365
11.2 Key-Value Stores	304	12.3.4.2 Factors that Determine the Efficiency of Random File Organization	368
11.2.1 From Keys to Hashes	304	12.3.5 Indexed Sequential File Organization	370
11.2.2 Horizontal Scaling	305	12.3.5.1 Basic Terminology of Indexes	370
11.2.3 An Example: Memcached	306	12.3.5.2 Primary Indexes	371
11.2.4 Request Coordination	308	12.3.5.3 Clustered Indexes	373
11.2.5 Consistent Hashing	309	12.3.5.4 Multilevel Indexes	374
11.2.6 Replication and Redundancy	311	12.3.6 List Data Organization (Linear and Nonlinear Lists)	375
11.2.7 Eventual Consistency	312	12.3.6.1 Linear Lists	375
11.2.8 Stabilization	314	12.3.6.2 Tree Data Structures	377
11.2.9 Integrity Constraints and Querying	314	12.3.7 Secondary Indexes and Inverted Files	379
11.3 Tuple and Document Stores	315	12.3.7.1 Characteristics of Secondary Indexes	380
11.3.1 Items with Keys	316	12.3.7.2 Inverted Files	381
11.3.2 Filters and Queries	316	12.3.7.3 Multicolumn Indexes	382
11.3.3 Complex Queries and Aggregation with MapReduce	320	12.3.7.4 Other Index Types	383
11.3.4 SQL After All...	330	12.3.8 B-Trees and B <sup>+</sup> -Trees	384
11.4 Column-Oriented Databases	331	12.3.8.1 Multilevel Indexes Revisited	384
11.5 Graph-Based Databases	333	12.3.8.2 Binary Search Trees	385
11.5.1 Cypher Overview	335	12.3.8.3 B-Trees	386
11.5.2 Exploring a Social Graph	336	12.3.8.4 B <sup>+</sup> -Trees	388
11.6 Other NoSQL Categories	341	Summary	390
Summary	342	Key Terms List	391
Key Terms	344	Review Questions	392
Review Questions	345	Problems and Exercises	393
Problems and Exercises	347		
<b>Part III Physical Data Storage, Transaction Management, and Database Access</b>	<b>349</b>		
<b>12 Physical File Organization and Indexing</b>	<b>351</b>	<b>13 Physical Database Organization</b>	<b>395</b>
12.1 Storage Hardware and Physical Database Design	351	13.1 Physical Database Organization and Database Access Methods	396
12.1.1 The Storage Hierarchy	352	13.1.1 From Database to Tablespace	396
12.1.2 Internals of Hard Disk Drives	353	13.1.2 Index Design	398
12.1.3 From Logical Concepts to Physical Constructs	356	13.1.3 Database Access Methods	400
12.2 Record Organization	359	13.1.3.1 Functioning of the Query Optimizer	400
12.3 File Organization	361	13.1.3.2 Index Search (with Atomic Search Key)	402
12.3.1 Introductory Concepts: Search Keys, Primary, and Secondary File Organization	362	13.1.3.3 Multiple Index and Multicolumn Index Search	403
12.3.2 Heap File Organization	363	13.1.3.4 Index-Only Access	407
12.3.3 Sequential File Organization	363	13.1.3.5 Full Table Scan	408
12.3.4 Random File Organization (Hashing)	365		

13.1.4 Join Implementations	408	14.3.3 Media Recovery	438
13.1.4.1 Nested-Loop Join	409	14.4 Concurrency Control	439
13.1.4.2 Sort-Merge Join	410	14.4.1 Typical Concurrency Problems	439
13.1.4.3 Hash Join	410	14.4.1.1 Lost Update Problem	440
13.2 Enterprise Storage Subsystems		14.4.1.2 Uncommitted	
and Business Continuity	411	Dependency Problem	
13.2.1 Disk Arrays and RAID	411	(aka Dirty Read	
13.2.2 Enterprise Storage Subsystems	413	Problem)	440
13.2.2.1 Overview and		14.4.1.3 Inconsistent Analysis	
Classification	414	Problem	441
13.2.2.2 DAS (Directly		14.4.1.4 Other Concurrency-	
Attached Storage)	416	Related Problems	442
13.2.2.3 SAN (Storage Area		14.4.2 Schedules and Serial	
Network)	416	Schedules	442
13.2.2.4 NAS (Network		14.4.3 Serializable Schedules	442
Attached Storage)	417	14.4.4 Optimistic and Pessimistic	
13.2.2.5 NAS Gateway	418	Schedulers	443
13.2.2.6 iSCSI/Storage		14.4.5 Locking and Locking	
Over IP	419	Protocols	444
13.2.3 Business Continuity	421	14.4.5.1 Purposes of Locking	444
13.2.3.1 Contingency Planning,		14.4.5.2 The Two-Phase	
Recovery Point,		Locking Protocol	
and Recovery Time	421	(2PL)	446
13.2.3.2 Availability and		14.4.5.3 Cascading Rollbacks	447
Accessibility of		14.4.5.4 Dealing with	
Storage Devices	422	Deadlocks	448
13.2.3.3 Availability of		14.4.5.5 Isolation Levels	449
Database		14.4.5.6 Lock Granularity	450
Functionality	422	14.5 The ACID Properties of	
13.2.3.4 Data Availability	423	Transactions	452
Summary	426	Summary	453
Key Terms List	426	Key Terms List	453
Review Questions	427	Review Questions	454
Problems and Exercises	429	Problems and Exercises	456
<b>14 Basics of Transaction Management</b>	<b>430</b>	<b>15 Accessing Databases and</b>	
14.1 Transactions, Recovery, and		<b>Database APIs</b>	<b>458</b>
Concurrency Control	431	15.1 Database System Architectures	459
14.2 Transactions and Transaction		15.1.1 Centralized System	
Management	432	Architectures	459
14.2.1 Delineating Transactions		15.1.2 Tiered System Architectures	460
and the Transaction Lifecycle	432	15.2 Classification of Database APIs	462
14.2.2 DBMS Components Involved		15.2.1 Proprietary versus	
in Transaction Management	433	Universal APIs	463
14.2.3 The Logfile	435	15.2.2 Embedded versus Call-	
14.3 Recovery	436	Level APIs	464
14.3.1 Types of Failures	436	15.2.3 Early Binding versus Late	
14.3.2 System Recovery	436	Binding	465

15.3 Universal Database APIs	466	16.3.1 Vertical Fragmentation	520
15.3.1 ODBC	466	16.3.2 Horizontal Fragmentation (Sharding)	521
15.3.2 OLE DB and ADO	467	16.3.3 Mixed Fragmentation	521
15.3.3 ADO.NET	468	16.3.4 Replication	523
15.3.4 Java DataBase Connectivity (JDBC)	471	16.3.5 Distribution and Replication of Metadata	524
15.3.5 Intermezzo: SQL Injection and Access Security	477	16.4 Transparency	524
15.3.6 SQLJ	479	16.5 Distributed Query Processing	525
15.3.7 Intermezzo: Embedded APIs versus Embedded DBMSs	480	16.6 Distributed Transaction Management and Concurrency Control	528
15.3.8 Language-Integrated Querying	482	16.6.1 Primary Site and Primary Copy 2PL	529
15.4 Object Persistence and Object- Relational Mapping APIs	483	16.6.2 Distributed 2PL	529
15.4.1 Object Persistence with Enterprise JavaBeans	484	16.6.3 The Two-Phase Commit Protocol (2PC)	530
15.4.2 Object Persistence with the Java Persistence API	488	16.6.4 Optimistic Concurrency and Loosely Coupled Systems	532
15.4.3 Object Persistence with Java Data Objects	495	16.6.5 Compensation-Based Transaction Models	534
15.4.4 Object Persistence in Other Host Languages	498	16.7 Eventual Consistency and BASE Transactions	538
15.5 Database API Summary	502	16.7.1 Horizontal Fragmentation and Consistent Hashing	538
15.6 Database Access in the World Wide Web	504	16.7.2 The CAP Theorem	539
15.6.1 Introduction: the Original Web Server	504	16.7.3 BASE Transactions	540
15.6.2 The Common Gateway Interface: Toward Dynamic Web Pages	504	16.7.4 Multi-Version Concurrency Control and Vector Clocks	541
15.6.3 Client-Side Scripting: The Desire for a Richer Web	507	16.7.5 Quorum-Based Consistency	542
15.6.4 JavaScript as a Platform	508	Summary	544
15.6.5 DBMSs Adapt: REST, Other Web Services, and a Look Ahead	509	Key Terms	545
Summary	512	Review Questions	546
Key Terms List	513	Problems and Exercises	547
Review Questions	513	<b>Part IV Data Warehousing, Data   Governance, and (Big)   Data Analytics</b>	<b>549</b>
Problems and Exercises	515		
<b>16 Data Distribution and Distributed   Transaction Management</b>	<b>516</b>	<b>17 Data Warehousing and Business   Intelligence</b>	<b>551</b>
16.1 Distributed Systems and Distributed Databases	517	17.1 Operational versus Tactical/ Strategic Decision-Making	552
16.2 Architectural Implications of Distributed Databases	518	17.2 Data Warehouse Definition	553
16.3 Fragmentation, Allocation, and Replication	519	17.3 Data Warehouse Schemas	554
		17.3.1 Star Schema	555

17.3.2	Snowflake Schema	556	18.1.2.1	Data Consolidation: Extract, Transform, Load (ETL)	593
17.3.3	Fact Constellation	557	18.1.2.2	Data Federation: Enterprise Information Integration (EII)	595
17.3.4	Specific Schema Issues	557	18.1.2.3	Data Propagation: Enterprise Application Integration (EAI)	596
17.3.4.1	Surrogate keys	557	18.1.2.4	Data Propagation: Enterprise Data Replication (EDR)	597
17.3.4.2	Granularity of the Fact Table	558	18.1.2.5	Changed Data Capture (CDC), Near-Real- Time ETL, and Event Processing	598
17.3.4.3	Factless Fact Tables	559	18.1.2.6	Data Virtualization	598
17.3.4.4	Optimizing the Dimension Tables	559	18.1.2.7	Data as a Service and Data in the Cloud	599
17.3.4.5	Defining Junk Dimensions	560	18.1.3	Data Services and Data Flows in the Context of Data and Process Integration	601
17.3.4.6	Defining Outtrigger Tables	561	18.1.3.1	Business Process Integration	602
17.3.4.7	Slowly Changing Dimensions	561	18.1.3.2	Patterns for Managing Sequence Dependencies and Data Dependencies in Processes	604
17.3.4.8	Rapidly Changing Dimensions	563	18.1.3.3	A Unified View on Data and Process Integration	606
17.4	The Extraction, Transformation, and Loading (ETL) Process	565	18.2	Searching Unstructured Data and Enterprise Search	610
17.5	Data Marts	567	18.2.1	Principles of Full-Text Search	610
17.6	Virtual Data Warehouses and Virtual Data Marts	569	18.2.2	Indexing Full-Text Documents	611
17.7	Operational Data Store	571	18.2.3	Web Search Engines	613
17.8	Data Warehouses versus Data Lakes	571	18.2.4	Enterprise Search	616
17.9	Business Intelligence	572	18.3	Data Quality and Master Data Management	617
17.9.1	Query and Reporting	573	18.4	Data Governance	618
17.9.2	Pivot Tables	573	18.4.1	Total Data Quality Management (TDQM)	619
17.9.3	On-Line Analytical Processing (OLAP)	574	18.4.2	Capability Maturity Model Integration (CMMI)	619
17.9.3.1	MOLAP	574	18.4.3	Data Management Body of Knowledge (DMBOK)	620
17.9.3.2	ROLAP	575			
17.9.3.3	HOLAP	575			
17.9.3.4	OLAP Operators	575			
17.9.3.5	OLAP Queries in SQL	577			
	Summary	583			
	Key Terms List	584			
	Review Questions	585			
	Problems and Exercises	587			
<b>18</b>	<b>Data Integration, Data Quality, and Data Governance</b>	<b>590</b>			
18.1	Data and Process Integration	591			
18.1.1	Convergence of Analytical and Operational Data Needs	591			
18.1.2	Data Integration and Data Integration Patterns	593			



18.4.4	Control Objectives for Information and Related Technology (COBIT)	620	20.4.5	Outlier Detection and Handling	672
18.4.5	Information Technology Infrastructure Library	621	20.5	Types of Analytics	673
18.5	Outlook	621	20.5.1	Predictive Analytics	673
18.6	Conclusion	622	20.5.1.1	Linear Regression	673
	Key Terms List	622	20.5.1.2	Logistic Regression	675
	Review Questions	623	20.5.1.3	Decision Trees	677
	Problems and Exercises	625	20.5.1.4	Other Predictive Analytics Techniques	681
<b>19</b>	<b>Big Data</b>	<b>626</b>	20.5.2	Evaluating Predictive Models	682
19.1	The 5 Vs of Big Data	627	20.5.2.1	Splitting Up the Dataset	682
19.2	Hadoop	630	20.5.2.2	Performance Measures for Classification Models	684
19.2.1	History of Hadoop	630	20.5.2.3	Performance Measures for Regression Models	687
19.2.2	The Hadoop Stack	631	20.5.2.4	Other Performance Measures for Predictive Analytical Models	688
19.2.2.1	The Hadoop Distributed File System	631	20.5.3	Descriptive Analytics	689
19.2.2.2	MapReduce	635	20.5.3.1	Association Rules	689
19.2.2.3	Yet Another Resource Negotiator	641	20.5.3.2	Sequence Rules	691
19.3	SQL on Hadoop	643	20.5.3.3	Clustering	692
19.3.1	HBase: The First Database on Hadoop	644	20.5.4	Social Network Analytics	695
19.3.2	Pig	648	20.5.4.1	Social Network Definitions	696
19.3.3	Hive	649	20.5.4.2	Social Network Metrics	696
19.4	Apache Spark	652	20.5.4.3	Social Network Learning	699
19.4.1	Spark Core	653	20.6	Post-Processing of Analytical Models	700
19.4.2	Spark SQL	654	20.7	Critical Success Factors for Analytical Models	701
19.4.3	MLlib, Spark Streaming, and GraphX	656	20.8	Economic Perspective on Analytics	702
19.5	Conclusion	659	20.8.1	Total Cost of Ownership (TCO)	702
	Key Terms List	660	20.8.2	Return on Investment	702
	Review Questions	660	20.8.3	In- versus Outsourcing	704
	Problems and Exercises	662	20.8.4	On-Premises versus Cloud Solutions	705
<b>20</b>	<b>Analytics</b>	<b>664</b>	20.8.5	Open-Source versus Commercial Software	706
20.1	The Analytics Process Model	665			
20.2	Example Analytics Applications	667			
20.3	Data Scientist Job Profile	668			
20.4	Data Pre-Processing	669			
20.4.1	Denormalizing Data for Analysis	669			
20.4.2	Sampling	670			
20.4.3	Exploratory Analysis	671			
20.4.4	Missing Values	671			

20.9 Improving the ROI of Analytics	708	20.10.3.2 SQL Views	719
20.9.1 New Sources of Data	708	20.10.3.3 Label-Based Access Control	719
20.9.2 Data Quality	711	20.10.4 Privacy Regulation	721
20.9.3 Management Support	712	20.11 Conclusion	723
20.9.4 Organizational Aspects	712	Key Terms List	724
20.9.5 Cross-Fertilization	713	Review Questions	725
20.10 Privacy and Security	714	Problems and Exercises	729
20.10.1 Overall Considerations Regarding Privacy and Security	714	Appendix Using the Online Environment	731
20.10.2 The RACI Matrix	715	Glossary	741
20.10.3 Accessing Internal Data	716	Index	770
20.10.3.1 Anonymization	717		



## 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](http://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](http://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](http://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](http://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 Vandembulcke, 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.