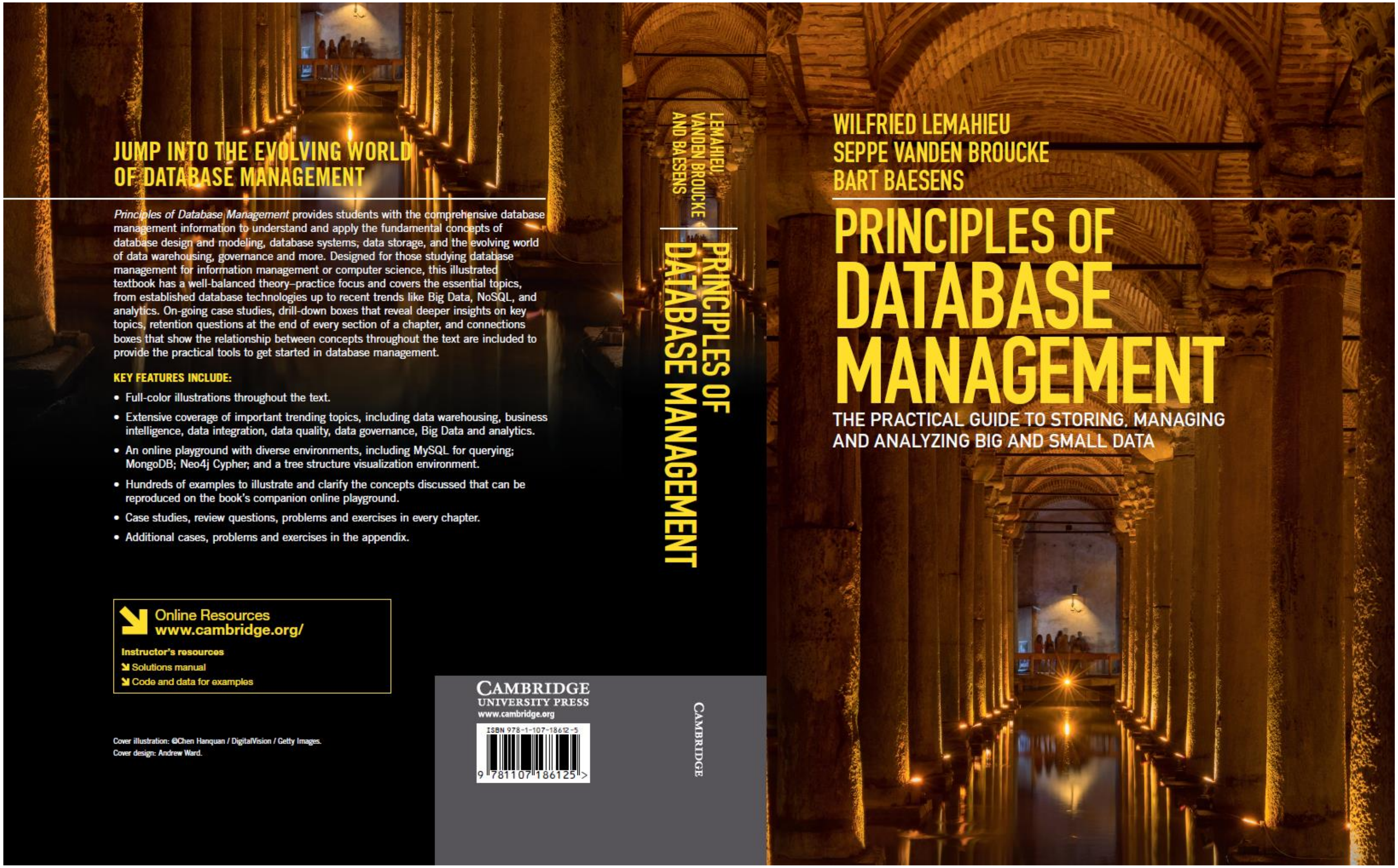
Organizational Aspects of Data Management



JUMP INTO THE EVOLVING WORLD OF DATABASE MANAGEMENT

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.

Online Resources
www.cambridge.org/

Instructor's resources
Solutions manual
Code and data for examples

Cover illustration: ©Chen Hanquan / DigitalVision / Getty Images.
Cover design: Andrew Ward.

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org
ISBN 978-1-107-18612-5

LEMAHIEU
VANDEN BROUCKE
AND BAESENS

PRINCIPLES OF DATABASE MANAGEMENT

CAMBRIDGE

WILFRIED LEMAHIEU
SEPPE VANDEN BROUCKE
BART BAESENS

PRINCIPLES OF DATABASE MANAGEMENT

THE PRACTICAL GUIDE TO STORING, MANAGING AND ANALYZING BIG AND SMALL DATA

www.pdbmbook.com

Introduction

- Data Management
- Roles in Data Management

Data Management

- Catalogs and the Role of Metadata
- Metadata Modeling
- Data Quality
- Data Governance

Catalogs and the Role of Metadata

- Just as raw data, also metadata is data that needs to be properly modeled, stored and managed
- Concepts of data modeling should also be applied to metadata
- In a DBMS approach, metadata is stored in a catalog (a.k.a. data dictionary, data repository), which constitutes the heart of the database system
  - can be part of a DBMS or standalone component

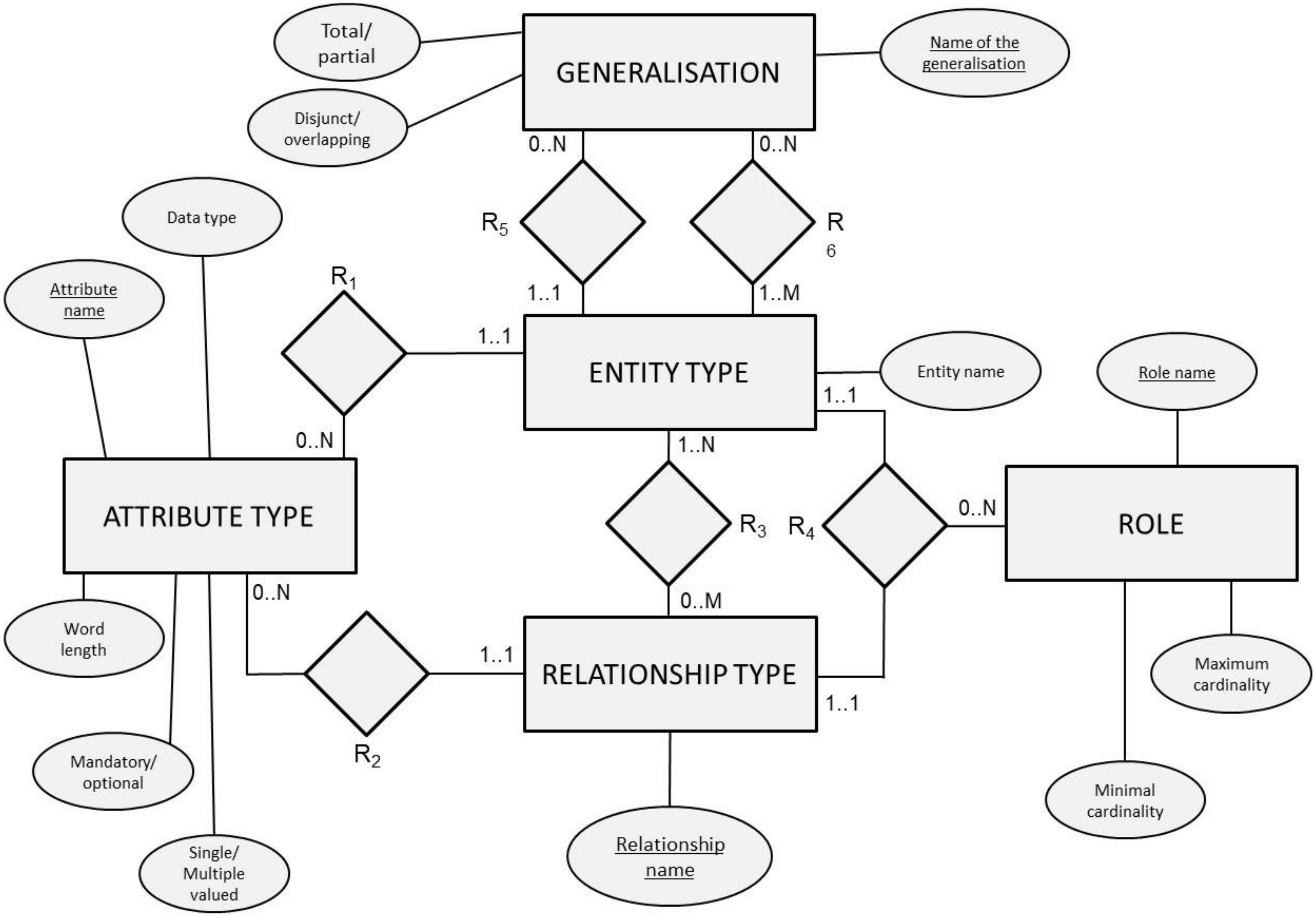
Catalogs and the Role of Metadata

- The catalog provides an important source of information for end users, application developers, as well as the DBMS itself
- Catalog should provide:
  - an extensible metamodel
  - import/export facilities
  - support for maintenance and re-use of metadata
  - monitoring of integrity rules
  - facilities for user access
  - statistics about the data and its usage for the DBA and query optimizer

Metadata Modelling

- A metamodel is data model for metadata
- A database design process can be used to design a database storing metadata
- Design a conceptual model of the metadata: EER model or UML model

Metadata Modelling



Data Quality

- Data quality (DQ) is often defined as ‘fitness for use’
  - data of acceptable quality in one decision context may be perceived to be of poor quality in another
- Data quality determines the intrinsic value of the data to the business
  - GIGO: Garbage In, Garbage Out
  - E.g., obsolete addresses
- Poor DQ impacts organizations in many ways
  - operational versus strategic level

Data Quality

- DQ is a multi-dimensional concept in which each dimension represents a single aspect or construct, comprising both objective and subjective perspectives
- A DQ framework categorizes the different dimensions of data quality
- Example: Wang et al. (1996)
  - 4 categories: intrinsic, contextual, representation, access

Data Quality

Category	DQ dimensions	Definitions
Intrinsic	Accuracy	The extent to which data is certified, error-free, correct, flawless and reliable
	Objectivity	The extent to which data is unbiased, unprejudiced, based on facts and impartial
	Reputation	The extent to which data is highly regarded in terms of its sources or content

Data Quality

Category	DQ dimensions	Definitions
Contextual	Completeness	The extent to which data is not missing and covers the needs of the tasks and is of sufficient breadth and depth of the task at hand
	Appropriate-amount	The extent to which the volume of data is appropriate for the task at hand
	Value-added	The extent to which data is beneficial and provides advantages from its use
	Relevance	The extent to which data is applicable and helpful for the task at hand
	Timeliness	The extent to which data is sufficiently up-to-date for the task at hand

Data Quality

Category	DQ dimensions	Definitions
Representation	Interpretable	The extent to which data is in appropriate languages, symbols and the definitions are clear
	Easily-understandable	The extent to which data is easily comprehended
	Consistency	The extent to which data is continuously presented in the same format
	Concisely-represented (CR)	The extent to which data is compactly represented, well-presented, well-organized, and well-formatted
	Alignment	The extent to which data is reconcilable (compatible)

Data Quality

Category	**DQ dimensions**	**Definitions**
Access	Accessibility	The extent to which data is available, or easily and swiftly retrievable
	Security	The extent to which access to data is restricted appropriately to maintain its security
	Traceability	The extent to which data is traceable to the source

Data Quality

- Accuracy refers to whether the data values stored for an object are the correct values
  - often correlated with other DQ dimensions
- Completeness can be viewed from at least 3 perspectives:
  - schema completeness: refers to the degree to which entity types and attribute types are missing from the schema
  - column completeness: refers to the degree to which there exist missing values in a column of a table
  - population completeness: refers degree to which the necessary members of a population are present or not

Data Quality

- The consistency dimension can also be viewed from several perspectives:
  - consistency of redundant or duplicated data in one table or in multiple tables
  - consistency between two related data elements
  - consistency of format for the same data element used in different tables

Data Quality

- The accessibility dimension reflects the ease of retrieving the data from the underlying data sources
  - often involves a trade-off with security

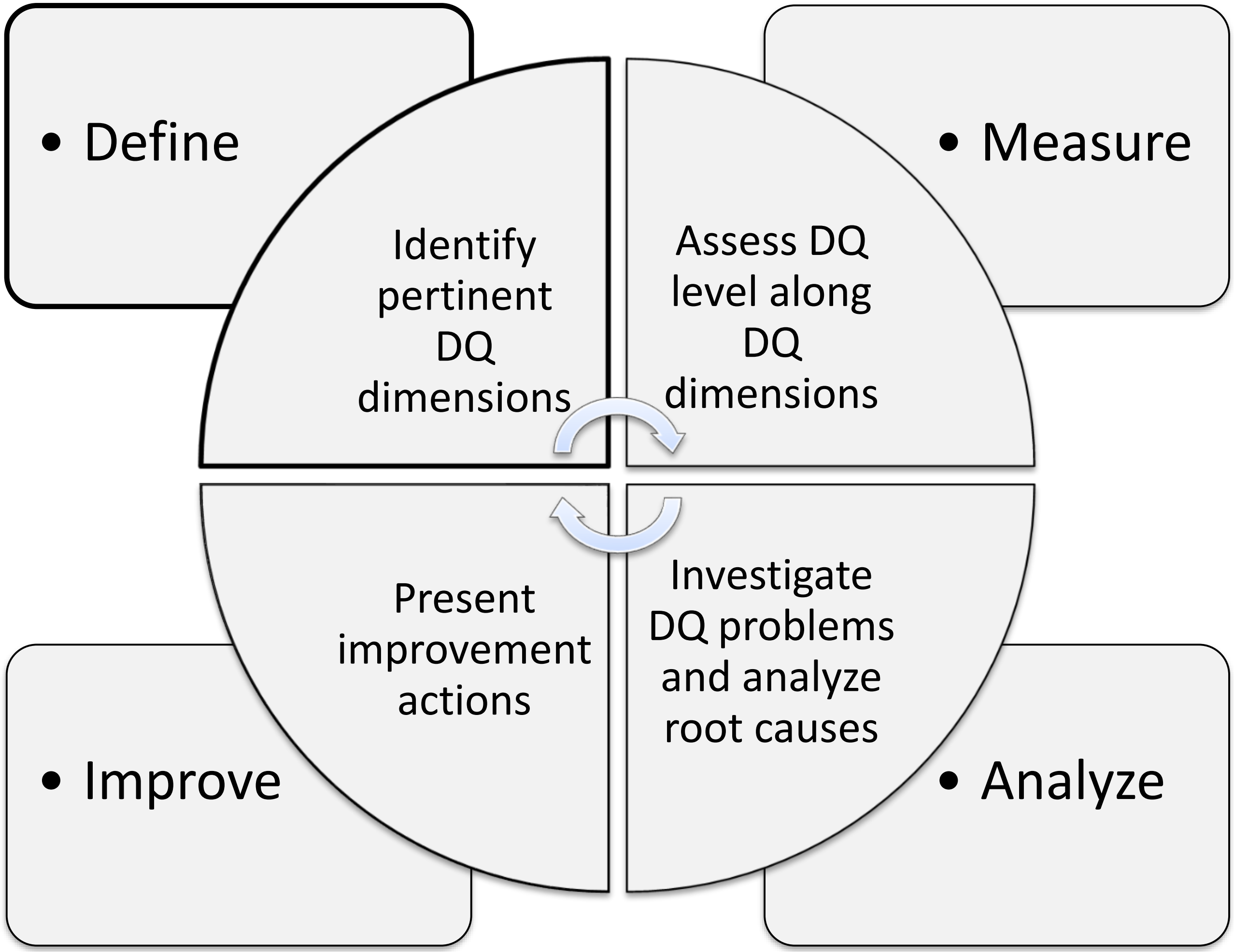
Data Quality

- Common causes of bad data quality are:
  - multiple data sources: multiple sources with the same data may produce duplicates; a problem of consistency.
  - subjective judgment in data production: data production using human judgment can result in biased information; a problem of objectivity.
  - limited computing resources: lack of sufficient computing resources may limit the accessibility of relevant data; a problem of accessibility.
  - volume of data: Large volumes of stored data make it difficult to access needed information in a reasonable time; a problem of accessibility.
  - changing data needs: data requirements change on an ongoing basis; a problem of relevance.
  - different processes updating the same data; a problem of consistency.
- Decoupling of data producers and consumers contributes to data quality problems

Data Governance

- To manage and safeguard data quality, a data governance culture should be put in place assigning clear roles and responsibilities
  - manage data as an asset rather than a liability
- Different frameworks have been introduced for data quality management and data quality improvement
  - examples: Total Data Quality Management (TDQM), Total Quality Management (TQM), Capability Maturity Model Integration (CMMI), ISO 9000, Control Objectives for Information and Related Technology (CobiT), Data Management Body of Knowledge (DMBOK), Information Technology Infrastructure Library (ITIL) and Six Sigma

Data Governance



- Define

Identify pertinent DQ dimensions

- Measure

Assess DQ level along DQ dimensions

- Analyze

Investigate DQ problems and analyze root causes

- Improve

Present improvement actions

Wang (1998)

Data Governance

- Annotate the data with data quality metadata as a short term solution
  - can be stored in the catalog
  - E.g., credit risk models could incorporate an additional risk factor to account for uncertainty in the data
- Unfortunately, many data governance efforts (if any) are mostly reactive and ad-hoc

Roles in Data Management

- Information Architect
- Database Designer
- Data owner
- Data steward
- Database Administrator
- Data Scientist

Roles in Data Management

- Information Architect (a.k.a. Information Analyst)
  - responsible for designing the conceptual data model
  - bridges the gap between the business processes and the IT environment
  - closely collaborates with the database designer who may assist in choosing the type of conceptual data model (e.g. EER or UML) and the database modeling tool

Roles in Data Management

- Database Designer
  - translates the conceptual data model into a logical and internal data model
  - also assists the application developers in defining the views of the external data model
  - defines company-wide uniform naming conventions when creating the various data models

Roles in Data Management

- Data owner
  - has the authority to ultimately decide on the access to, and usage of, the data
  - could be the original producer of the data, one of its consumers, or a third party
  - should be able to insert or update data
  - can be requested to check or complete the value of a field

Roles in Data Management

- Data steward
  - DQ experts in charge of ensuring the quality of both the actual business data and the metadata
  - perform extensive and regular data quality checks
  - can initiate corrective measures or deeper investigation into root causes of data quality issues
  - can help design preventive measures (e.g. modifications to operational information systems, integrity rules)

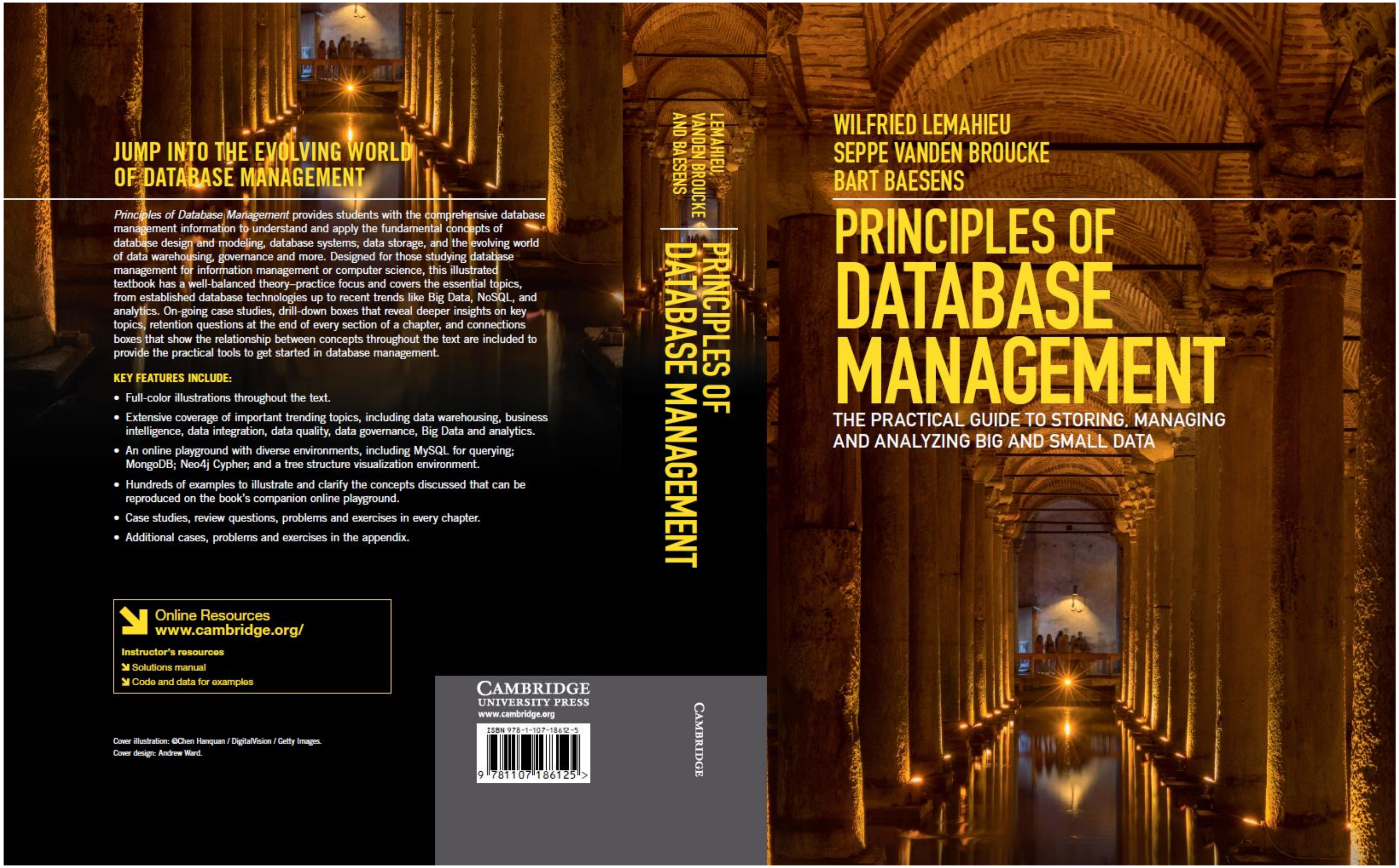
Roles in Data Management

- Database Administrator (DBA)
  - responsible for the implementation and monitoring of the database
  - closely collaborates with network and system managers
  - also interacts with database designers
- Data scientist
  - responsible for analyzing data using state-of-the-art analytical techniques to provide new insights into e.g. customer behavior
  - has a multidisciplinary profile combining ICT skills with quantitative modeling, business understanding, communication, and creativity

Conclusions

- Data Management
- Roles in Data Management

More information?



JUMP INTO THE EVOLVING WORLD OF DATABASE MANAGEMENT

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.

Online Resources
www.cambridge.org/

Instructor's resources
- Solutions manual
- Code and data for examples

Cover illustration: ©Chen Hanquan / DigitalVision / Getty Images.
Cover design: Andrew Ward.

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-107-18612-5
9 781107 186125 >

LEMAHIEU
VANDEN BROUCKE
AND BAESENS

PRINCIPLES OF
DATABASE MANAGEMENT

CAMBRIDGE

WILFRIED LEMAHIEU
SEPPE VANDEN BROUCKE
BART BAESENS

PRINCIPLES OF
DATABASE
MANAGEMENT

THE PRACTICAL GUIDE TO STORING, MANAGING
AND ANALYZING BIG AND SMALL DATA

www.pdbmbook.com