

---

# **Undergraduate Topics in Computer Science**

Undergraduate Topics in Computer Science (UTiCS) delivers high-quality instructional content for undergraduates studying in all areas of computing and information science. From core foundational and theoretical material to final-year topics and applications, UTiCS books take a fresh, concise, and modern approach and are ideal for self-study or for a one- or two-semester course. The texts are all authored by established experts in their fields, reviewed by an international advisory board, and contain numerous examples and problems. Many include fully worked solutions.

For further volumes:

[www.springer.com/series/7592](http://www.springer.com/series/7592)

---

Peter Lake • Paul Crowther

# Concise Guide to Databases

A Practical Introduction

Foreword by Professor Richard Hill

 Springer

Peter Lake  
Sheffield Hallam University  
Sheffield, UK

Paul Crowther  
Sheffield Hallam University  
Sheffield, UK

*Series editor*  
Ian Mackie

*Advisory board*

Samson Abramsky, University of Oxford, Oxford, UK  
Karin Breitman, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil  
Chris Hankin, Imperial College London, London, UK  
Dexter Kozen, Cornell University, Ithaca, USA  
Andrew Pitts, University of Cambridge, Cambridge, UK  
Hanne Riis Nielson, Technical University of Denmark, Kongens Lyngby, Denmark  
Steven Skiena, Stony Brook University, Stony Brook, USA  
Iain Stewart, University of Durham, Durham, UK

ISSN 1863-7310

Undergraduate Topics in Computer Science

ISBN 978-1-4471-5600-0

DOI 10.1007/978-1-4471-5601-7

Springer London Heidelberg New York Dordrecht

ISSN 2197-1781 (electronic)

ISBN 978-1-4471-5601-7 (eBook)

Library of Congress Control Number: 2013955488

© Springer-Verlag London 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

*Dedicated to our mate Andy McEwan*

*Paul and Peter*

---

## Foreword

From tablets of stone through to libraries of parchments; from paper-based files to the electronic era, there is not one aspect of modern business that has avoided the need to collect, collate, organize and report upon data. The proliferation of databases and database technologies within modern times, has now been further secured by the use of the Internet to enable database integration on a massive scale.

In amongst the innovation, the basic concepts remain. The need to organize—a topic that Codd reminded us could be best done by relational models—is now being challenged, as processor power and storage space become cheap and utility-like with the advent of Cloud Computing infrastructure. But a glance at the past does much to inform future thinking, and this book serves to prepare the foundations of a mature approach to using database technologies in the 21st Century.

In many cases, both established and emerging database technologies are readily available and free to use. As such they may appear free to implement, which fuels rapid adoption of technology that may not have been proven sufficiently, without the formal governance that other business norms might impose. This creates an exciting, risky, domain where commercial models can make or lose money depending upon how they embrace and realize the potential of the technology. Conversely, there is also an opportunity to solve problems when things go awry—and the accelerated innovation that we now witness, presents more opportunities and pitfalls, if we do not possess the requisite understanding of how databases should serve our needs.

Proficiency in the field of databases is a combination of technical understanding, conceptual knowledge and business acumen. All of these traits are underpinned by education, and the need for professionals to continually update their knowledge. Since professionals not only face the challenge of when to introduce a technology, but also when *not* to adopt, it is important to understand the impact of failure as well as success. This book takes readers through the essential basics, before charting a path towards technical skill acquisition in the real-life context of business.

Head of Subject, Computing and Mathematics  
University of Derby, Derby, UK  
June 2013

Professor Richard Hill

**About Richard Hill:**

Richard Hill, PhD, is Professor of Intelligent Systems and Head of Department in the School of Computing and Mathematics, at the University of Derby, UK. Professor Hill has published widely in the areas of multi agent systems, computational intelligence, intelligent cloud computing and emerging technologies for distributed systems, and has organised a number of international conferences. Latterly, Professor Hill has edited and co-authored several book collections and textbooks, including 'Guide to Cloud Computing: Principles and Practice', published by Springer UK.

---

# Preface

---

## Overview and Goals

Databases are not new and there are many text books available which cover various database types, especially relational. What is changing, however, is that Relational Database Management Systems (RDBMS) are no longer the only database solution. In an era where Big Data is the current buzzword and Data Scientists are tomorrow's big earners, it is important to take a wider view of database technology.

Key objectives for this book include:

- Present an understanding of the key technologies involved in Database Systems in general and place those technologies in an historic context
- Explore the potential use of a variety of database types in a business environment
- Point out areas for further research in a fast moving domain
- Equip readers with an understanding of the important aspects of a database professional's job
- Provide some hands-on experience to further assist in the understanding of the technologies involved

---

## Organisation and Features

This book is organised into three parts:

- Part I introduces database concepts and places them in both a historic and business context;
- Part II provides insights into some of the major database types around today and also provides some hands-on tutorials in the areas concerned;
- Part III is devoted to issues and challenges which face Database Professionals.

---

## Target Audiences

This book has been written specifically to support the following audiences:

*Advanced undergraduate students and postgraduate students* should find the combination of theoretical and practical examples database usage of interest. We imagine this text would be of particular relevance for modern Computer Science,

Software Engineering, and Information Technology courses. However, any course that makes reference to databases, and in particular to the latest developments in computing will find this text book of use. As such, *University Instructors* may adopt the book as a core text.

Especially in Part II, this book adopts a learning-by-doing approach, with the extensive worked examples explaining how to use the variety of databases available to address today's business needs. Practising *Database Professionals*, and *Application Developers* will also be able to use this book to review the current state of the database domain.

---

## Suggested Uses

A Concise Guide to Databases can be used as a solid introduction to the concept of databases. The book is suitable as both a comprehensive introduction to databases, as well as a reference text as the reader develops their skills and abilities through practical application of the ideas. For *University Instructors*, we suggest the following programme of study for a twelve-week semester format:

- Weeks 1–3: Part I
- Weeks 4–8: Part II
- Weeks 9–12: Part III
- Week 12: Assessment

---

## Review Questions

Each chapter concludes with a set of review questions that make specific reference to the content presented in the chapter, plus an additional set of further questions that will require further research. The review questions are designed in such a way that the reader will be able to tackle them based on the chapter contents. They are followed by discussion questions, that often require research, extended reading of other material or discussion and collaboration. These can be used as classroom discussion topics by tutors or used as the basis of summative assignments.

---

## Hands-on Exercises

The technology chapters include extended hands-on exercises. Readers will then progressively engage in more complex activities, building skills and knowledge along the way. Such an approach ensures that a solid foundation is built before more advanced topics are undertaken. Some of the material here is Open Source, whilst some examples are Oracle specific, but even these latter can be applied to other SQL databases.

## Chapter Summary

A brief summary of each of the twelve chapters is as follows:

**Chapter 1:** Data is the lifeblood of all business systems and we place the use of data in its historical context and review some of the key concepts in handling data.

**Chapter 2:** Provides an examination of the way that data has been handled throughout history, using databases of a variety of types.

**Chapter 3:** Considers how we actually store data. Turning information into a series of 1s and 0s is at the heart of every current database system and so an understanding of issues like physical storage and distribution are important concepts to understand.

**Chapter 4:** The *de facto* standard database solution is, without doubt, the relational database. In this chapter we look at how RDBMS works and provide worked examples.

**Chapter 5:** The NoSQL movement is still relatively new. Databases which store data without schemas and which do not necessarily provide transactional security may seem like a bad idea to experienced relational database practitioners, but these tools do certainly have their place in today's data rich society. We review the area in general and then look at specific examples of a Column-based and a Document-based database, with hands-on tutorials for each.

**Chapter 6:** Look at many leading database vendors' web sites and you will see that we are in the Big Data era. We explore what this actually means and, using a tutorial, review one of the key concepts in this era—that of MapReduce.

**Chapter 7:** Object databases were once thought of as the next important design for databases. When used by developers using Object programming they can seem very appealing still. There are half-way house solutions also available—Oracle, for example, has an Object-Relational option. We explore this area with more tutorial material.

**Chapter 8:** Reading data from disk is far slower than reading from RAM. Computing technologies now exist that can allow databases to run entirely in memory, making for very rapid data processing. These databases may well become the norm as RAM becomes cheaper and hard disk technology becomes less able to improve in performance.

**Chapter 9:** Once you have designed your database, especially when supporting a web- or cloud-based solution, you need to be sure that it can grow if the business that the application supports is successful. Scalability is about ensuring that you can cope with many concurrent users, or huge amounts of data, or both.

**Chapter 10:** Once your system is built, you need to be able to have it available for use permanently (or as close to permanently as can be achieved within the financial resources at your disposal). We review key concepts such as back-up, recovery, and disaster recovery.

**Chapter 11:** For a DBA the dreaded phone call is “my report is running very slowly”. For a start, what is mean by slowly? What is the user used to? Then there is the problem of how you establish where the problem is—is it hardware related? Or Network related? At the Server or Client end? The solution may be indexes, or

partitions: we review a variety of performance related techniques. We include some tutorial material which explores some performance management tools.

**Chapter 12:** Data is one of an organisation's most important assets. It needs to be protected from people wanting to either take it, or bring the system down. We look at physical and software-related weaknesses and review approaches to making our databases secure.

Sheffield, UK

Peter Lake  
Paul Crowther

---

# Contents

## Part I Databases in Context

<b>1</b>	<b>Data, an Organisational Asset</b> . . . . .	3
1.1	Introduction . . . . .	3
1.2	In the Beginning . . . . .	4
1.3	The Rise of Organisations . . . . .	4
1.4	The Challenges of Multi-site Operation . . . . .	4
1.5	Internationalisation . . . . .	5
1.6	Industrialisation . . . . .	6
1.7	Mass Transport . . . . .	6
1.8	Communication . . . . .	7
1.9	Stocks and Shares . . . . .	9
1.10	Corporate Takeovers . . . . .	10
1.11	The Challenges of Multi National Operations . . . . .	11
1.12	The Data Asset . . . . .	12
1.13	Electronic Storage . . . . .	13
1.14	Big Data . . . . .	15
1.15	Assets in the Cloud . . . . .	16
1.16	Data, Data Everywhere . . . . .	17
1.17	Summary . . . . .	18
1.18	Exercises . . . . .	18
	1.18.1 Review Questions . . . . .	18
	1.18.2 Group Work Research Activities . . . . .	19
	References . . . . .	19
<b>2</b>	<b>A History of Databases</b> . . . . .	21
2.1	Introduction . . . . .	21
2.2	The Digital Age . . . . .	21
2.3	Sequential Systems . . . . .	22
2.4	Random Access . . . . .	23
2.5	Origins of Modern Databases . . . . .	24
2.6	Transaction Processing and ACID . . . . .	25
2.7	Two-Phase Commit . . . . .	26

---

2.8	Hierarchical Databases . . . . .	27
2.9	Network Databases . . . . .	27
2.10	Relational Databases . . . . .	28
2.11	Object Oriented Databases . . . . .	30
2.12	Data Warehouse . . . . .	30
2.13	The Gartner Hype Cycle . . . . .	32
2.14	Big Data . . . . .	33
2.15	Data in the Cloud . . . . .	33
2.16	The Need for Speed . . . . .	34
2.17	In-Memory Database . . . . .	34
2.18	NoSQL . . . . .	35
2.19	Spatial Databases . . . . .	35
2.20	Databases on Personal Computers . . . . .	36
2.21	Distributed Databases . . . . .	36
2.22	XML . . . . .	37
2.23	Temporal Databases . . . . .	38
2.24	Summary . . . . .	39
2.25	Exercises . . . . .	39
	2.25.1 Review Questions . . . . .	39
	2.25.2 Group Work Research Activities . . . . .	39
	References . . . . .	40
<b>3</b>	<b>Physical Storage and Distribution . . . . .</b>	<b>41</b>
3.1	The Fundamental Building Block . . . . .	41
3.2	Overall Database Architecture . . . . .	42
	3.2.1 In-Memory Structures . . . . .	42
	3.2.2 Walking Through a Straightforward Read . . . . .	43
	3.2.3 Server Processes . . . . .	45
	3.2.4 Permanent Structures . . . . .	46
3.3	Data Storage . . . . .	47
	3.3.1 Row Chaining and Migration . . . . .	52
	3.3.2 Non-relational Databases . . . . .	52
3.4	How Logical Data Structures Map to Physical . . . . .	52
3.5	Control, Redo and Undo . . . . .	52
3.6	Log and Trace Files . . . . .	54
3.7	Stages of Start-up and Shutdown . . . . .	54
3.8	Locking . . . . .	57
3.9	Moving Data . . . . .	58
3.10	Import and Export . . . . .	60
	3.10.1 Data Is Important . . . . .	61
3.11	Distributed Databases . . . . .	61
3.12	Summary . . . . .	64
3.13	Review Questions . . . . .	64
3.14	Group Work Research Activities . . . . .	64
	References . . . . .	65

**Part II Database Types**

- 4 Relational Databases . . . . . 69**
  - 4.1 Origins . . . . . 69
  - 4.2 Normalisation . . . . . 70
    - 4.2.1 First Normal Form (1NF) . . . . . 71
  - 4.3 Second Normal Form (2NF) . . . . . 72
  - 4.4 Third Normal Form (3NF) . . . . . 73
  - 4.5 Beyond Third Normal Form . . . . . 75
  - 4.6 Entity Modelling . . . . . 76
  - 4.7 Use Case Modelling . . . . . 76
  - 4.8 Further Modelling Techniques . . . . . 82
  - 4.9 Notation . . . . . 83
  - 4.10 Converting a Design into a Relational Database . . . . . 85
  - 4.11 Worked Example . . . . . 87
  - 4.12 Create the Tables . . . . . 87
  - 4.13 CRUDing . . . . . 89
  - 4.14 Populate the Tables . . . . . 90
  - 4.15 Retrieve Data . . . . . 90
  - 4.16 Joins . . . . . 91
  - 4.17 More Complex Data Retrieval . . . . . 93
  - 4.18 UPDATE and DELETE . . . . . 94
  - 4.19 Review Questions . . . . . 95
  - 4.20 Group Work Research Activity . . . . . 95
  - References . . . . . 96
  
- 5 NoSQL Databases . . . . . 97**
  - 5.1 Databases and the Web . . . . . 97
  - 5.2 The NoSQL Movement . . . . . 98
    - 5.2.1 What Is Meant by NoSQL? . . . . . 100
  - 5.3 Differences in Philosophy . . . . . 101
  - 5.4 Basically Available, Soft State, Eventually Consistent (BASE) . . . . . 103
  - 5.5 Column-Based Approach . . . . . 103
  - 5.6 Examples of Column-Based Using Cassandra . . . . . 104
    - 5.6.1 Cassandra’s Basic Building Blocks . . . . . 106
    - 5.6.2 Data Sources . . . . . 107
    - 5.6.3 Getting Started . . . . . 107
    - 5.6.4 Creating the Column Family . . . . . 110
    - 5.6.5 Inserting Data . . . . . 112
    - 5.6.6 Retrieving Data . . . . . 112
    - 5.6.7 Deleting Data and Removing Structures . . . . . 114
    - 5.6.8 Command Line Script . . . . . 115
    - 5.6.9 Shutdown . . . . . 116
  - 5.7 CQL . . . . . 116
    - 5.7.1 Interactive CQL . . . . . 118

---

5.7.2	IF You Want to Check How Well You Now Know	
Cassandra ...		119
5.7.3	Timings	120
5.8	Document-Based Approach	120
5.8.1	Examples of Document-Based Using MongoDB	122
5.8.2	Data Sources	122
5.8.3	Getting Started	122
5.8.4	Navigation	123
5.8.5	Creating a Collection	123
5.8.6	Simple Inserting and Reading of Data	125
5.8.7	More on Retrieving Data	127
5.8.8	Indexing	129
5.8.9	Updating Data	130
5.8.10	Moving Bulk Data into Mongo	130
5.9	IF You Want to Check How Well You Now Know	
MongoDB ...		130
5.9.1	Timings	131
5.10	Summary	132
5.11	Review Questions	132
5.12	Group Work Research Activities	132
5.12.1	Sample Solutions	133
5.12.2	MongoDB Crib	134
	References	134
<b>6</b>	<b>Big Data</b>	135
6.1	What Is Big Data?	135
6.2	The Datacentric View of Big Data	137
6.2.1	The Four Vs	138
6.2.2	The Cloud Effect	140
6.3	The Analytics View of Big Data	141
6.3.1	So Why Isn't Big Data Just Called Data Warehousing 2?	142
6.3.2	What Is a Data Scientist?	145
6.3.3	What Is Data Analysis for Big Data?	147
6.4	Big Data Tools	147
6.4.1	MapReduce	148
6.4.2	Hadoop	149
6.4.3	Hive, Pig and Other Tools	150
6.5	Getting Hands-on with MapReduce	151
6.6	Using MongoDB's db.collection.mapReduce() Method	152
6.6.1	And If You Have Time to Test Your MongoDB and JS Skills	156
6.6.2	Sample Solutions	156
6.7	Summary	158
6.8	Review Questions	158

6.9	Group Work Research Activities . . . . .	158
	References . . . . .	159
<b>7</b>	<b>Object and Object Relational Databases . . . . .</b>	<b>161</b>
7.1	Querying Data . . . . .	161
7.2	Problems with Relational Databases . . . . .	161
7.3	What Is an Object? . . . . .	163
7.4	An Object Oriented Solution . . . . .	164
7.5	XML . . . . .	166
7.6	Object Relational . . . . .	168
7.7	What Is Object Relational? . . . . .	169
7.8	Classes . . . . .	169
7.9	Pointers . . . . .	172
7.10	Hierarchies and Inheritance . . . . .	174
7.11	Aggregation . . . . .	176
7.12	Encapsulation and Polymorphism . . . . .	177
7.13	Polymorphism . . . . .	178
7.14	Support for Object Oriented and Object Relational Database Development . . . . .	179
7.15	Will Object Technology Ever Become Predominant in Database Systems? . . . . .	180
	7.15.1 Review Questions . . . . .	181
	7.15.2 Group Work Research Activities . . . . .	181
	References . . . . .	182
<b>8</b>	<b>In-Memory Databases . . . . .</b>	<b>183</b>
8.1	Introduction . . . . .	183
8.2	Origins . . . . .	183
8.3	Online Transaction Processing Versus Online Analytical Processing . . . . .	184
8.4	Interim Solution—Create a RAM Disk . . . . .	185
8.5	Interim Solution—Solid State Drive (SSD) . . . . .	186
8.6	In-Memory Databases—Some Misconceptions . . . . .	187
8.7	In-Memory Relational Database—The Oracle TimesTen Approach . . . . .	188
8.8	In-Memory Column Based Storage—The SAP HANA Approach . . . . .	190
8.9	In-Memory On-line Transaction Processing—The Starcounter Approach . . . . .	192
8.10	Applications Suited to In-Memory Databases . . . . .	193
8.11	In Memory Databases and Personal Computers . . . . .	193
8.12	Summary . . . . .	195
8.13	Review Questions . . . . .	195
8.14	Group Work Research Activities . . . . .	195
	References . . . . .	196

**Part III What Database Professionals Worry About**

- 9 Database Scalability . . . . . 201**
  - 9.1 What Do We Mean by Scalability? . . . . . 201
  - 9.2 Coping with Growing Numbers of Users . . . . . 202
    - 9.2.1 So Why Can't Access Cope with Lots of Users? . . . . 203
    - 9.2.2 Client/Server . . . . . 204
    - 9.2.3 Scalability Eras . . . . . 206
  - 9.3 Coping with Growing Volumes of Data . . . . . 208
    - 9.3.1 E-Commerce and Cloud Applications Need Scalable Solutions . . . . . 209
    - 9.3.2 Vertical and Horizontal Scaling . . . . . 210
    - 9.3.3 Database Scaling Issues . . . . . 210
    - 9.3.4 Single Server Solutions . . . . . 212
    - 9.3.5 Distributed RDBMS: Shared Nothing vs. Shared Disk . 213
    - 9.3.6 Horizontal Scaling Solutions . . . . . 215
    - 9.3.7 Scaling Down for Mobile . . . . . 217
  - 9.4 Summary . . . . . 218
  - 9.5 Review Questions . . . . . 218
  - 9.6 Group Work Research Activities . . . . . 218
  - References . . . . . 218
- 10 Database Availability . . . . . 221**
  - 10.1 What Do We Mean by Availability? . . . . . 221
  - 10.2 Keeping the System Running—Immediate Solutions to Short Term Problems . . . . . 222
    - 10.2.1 What Can Go Wrong? . . . . . 223
  - 10.3 Back-up and Recovery . . . . . 229
    - 10.3.1 MTBF and MTTR . . . . . 231
  - 10.4 Disaster Recovery (DR) . . . . . 235
    - 10.4.1 Business Impact . . . . . 236
    - 10.4.2 High Availability for Critical Systems . . . . . 237
    - 10.4.3 Trade-offs and Balances . . . . . 238
    - 10.4.4 The Challenge or Opportunity of Mobile . . . . . 238
  - 10.5 Summary . . . . . 239
  - 10.6 Review Questions . . . . . 239
  - 10.7 Group Work Research Activities . . . . . 239
  - References . . . . . 239
- 11 Database Performance . . . . . 241**
  - 11.1 What Do We Mean by Performance? . . . . . 241
  - 11.2 A Simplified RDBMS Architecture . . . . . 243
  - 11.3 Physical Storage . . . . . 245
    - 11.3.1 Block Size . . . . . 245
    - 11.3.2 Disk Arrays and RAID . . . . . 246
    - 11.3.3 Alternatives to the HDD . . . . . 246

---

11.3.4	Operating System (OS)	247
11.3.5	Database Server Processes	248
11.3.6	Archive Manager	248
11.3.7	Schema Level: Data Types, Location and Volumes	249
11.3.8	SQL Optimisation	250
11.3.9	Indexes	251
11.3.10	Network	255
11.3.11	Application	256
11.4	Tuning	258
11.4.1	What Is Database Tuning?	258
11.4.2	Benchmarking	259
11.4.3	Another Perspective	259
11.5	Tools	261
11.5.1	Tuning and Performance Tools	261
11.5.2	Using the Built-in Advisers	275
11.5.3	Over to You!	276
11.6	Summary	276
11.7	Review Questions	277
11.8	Group Work Research Activities	277
Appendix	Creation Scripts and Hints	277
A.1	Hints on the Over to You Section	279
References		282
<b>12</b>	<b>Security</b>	<b>283</b>
12.1	Introduction	283
12.2	Physical Security	284
12.3	Software Security—Threats	286
12.4	Privilege Abuse	286
12.5	Platform Weaknesses	291
12.6	SQL Injection	292
12.7	Weak Audit	293
12.8	Protocol Vulnerabilities	294
12.9	Authentication Vulnerabilities	295
12.10	Backup Data Exposure	295
12.11	Mobile Device Based Threats	296
12.12	Security Issues in Cloud Based Databases	296
12.13	Policies and Procedures	298
12.14	A Security Checklist	298
12.15	Review Questions	299
12.16	Group Work Research Activities	300
References		300
<b>Index</b>		<b>303</b>