The Berkeley DB Book

Himanshu Yadava

Apress®

The Berkeley DB Book

ISBN-13: 978-1-59059-672-2

ISBN-10: 1-59059-672-2

Printed and bound in the United States of America 9 8 7 6 5 4 3 2 1

Lead Editor: Jonathan Gennick
Technical Reviewer: Mike Olson
Editorial Board: Steve Anglin, Ewan Buckingham, Gary Cornell, Jonathan Gennick, Jason Gilmore, Jonathan Hassell, Chris Mills, Matthew Moodie, Jeffrey Pepper, Ben Renow-Clarke, Dominic Shakeshaft, Matt Wade, Tom Welsh
Project Manager: Sofia Marchant
Copy Editor: Nicole Abramowitz
Assistant Production Director: Kari Brooks-Copony
Production Editor: Lori Bring
Compositor: Susan Glinert Stevens
Proofreader: Linda Seifert
Indexer: Brenda Miller
Artist: April Milne
Cover Designer: Kurt Krames
Manufacturing Director: Tom Debolski

Distributed to the book trade worldwide by Springer-Verlag New York, Inc., 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax 201-348-4505, e-mail orders-ny@springer-sbm.com, or visit http://www.springeronline.com.

For information on translations, please contact Apress directly at 2855 Telegraph Avenue, Suite 600, Berkeley, CA 94705. Phone 510-549-5930, fax 510-549-5939, e-mail info@apress.com, or visit http://www.apress.com.

The source code for this book is available to readers at http://www.apress.com in the Source Code/Download section. You will need to answer questions pertaining to this book in order to successfully download the code.

Contents at a Glance

About the Author xiii
About the Technical Reviewer xv
Acknowledgments xvii
Introduction xix

CHAPTER 1 Introduction to Berkeley DB 1
CHAPTER 2 When to Use Berkeley DB 11
CHAPTER 3 Products, Compilation, and Installation 23
CHAPTER 4 Building a Simple Application Using Berkeley DB 31
CHAPTER 5 Introduction to Advanced Data Stores 61
CHAPTER 6 Advanced Operations 111
CHAPTER 7 A Real-World Data Store 159
CHAPTER 8 Replication 201
CHAPTER 9 Distributed Transactions and Data-Distribution Strategies 273
CHAPTER 10 Berkeley DB Utilities 323
CHAPTER 11 Berkeley DB Java APIs 359
CHAPTER 12 Berkeley DB C API 403

INDEX 431

Contents

About the Author xiii
About the Technical Reviewer xv
Acknowledgments xvii
Introduction xix

CHAPTER 1 Introduction to Berkeley DB 1

A Brief History 1
What Is Berkeley DB? 2
Relational Databases 2
Hierarchical Databases 2
Object Databases 3
Berkeley DB 3
Architecture of Berkeley DB 4
Berkeley DB vs. RDBMS 5
Why Berkeley DB Isn't As Popular As RDBMS 7
Why Berkeley DB May Become More Popular 8
Oracle Dual License 8
Summary 9

CHAPTER 2 When to Use Berkeley DB 11

What Berkeley DB Does and Doesn't Provide 11
Why Berkeley DB Doesn't Provide Everything 13
Threading 13
Communication Support 14
Query Support 14
Serialization Support 15
What Type of Data Store Do You Want? 16
Data-Access Patterns 16
Transaction and Recovery 16
Fault Tolerance and High Availability 16
Scalability 17
Data Organization 17
Berkeley DB Data Stores 17

A Checklist of Components to Be Built 18
Plan for the Future 19
Berkeley DB Licensing 20
Summary 21

CHAPTER 3 Products, Compilation, and Installation 23

Berkeley DB Product Family 23
Compiling and Installing 25
Berkeley DB Versioning 30
Summary 30

CHAPTER 4 Building a Simple Application Using Berkeley DB 31

Storing "Hello World" 31
Named, Unnamed, and In-Memory Databases 35
Error Reporting in Berkeley DB 35
Flags in Berkeley DB 38
`Dbt` Class 39
Basic Database Operations in Berkeley DB 40
Access Methods in Berkeley DB 41
Common Access Method Configuration Parameters 42
`Btree` Access Method 45
`Hash` Access Method 47
`Queue` Access Method 49
`Recno` Access Method 50
Berkeley DB Environment 51
Types of Shared Memory Regions 55
`DB_CONFIG` and `DB_HOME` 55
Remote File Systems and Berkeley DB 57
Access Control and Security 57
Viewing the Berkeley DB Environment State 58
Summary 59

CHAPTER 5 Introduction to Advanced Data Stores 61

Database Locking 61
Lock Granularity 63
Lock Types 64

Concurrent Data Store 65
A Note About the Code Example 65
Creating a CDS Data Store 68
Using `DB_THREAD` Correctly in CDS 73
Database Watchdog 74
Process Monitoring 75
Database Verification 75
Database Recovery 78
Watchdog for DS and CDS 78
An Introduction to TDS 78
Why ACID? 79
Transactions in Berkeley DB 80
Locking in TDS 83
Deadlocks 89
Why Deadlocks Are Created 90
Deadlock Detection 92
Deadlock Detection and Lock Time-Outs 92
Recovering from Deadlocks 99
How to Avoid Deadlocks 99
Nested Transactions 100
Database Recoverability 101
Checkpoints 101
Transaction Log Maintenance 103
Database Backups 105
Database Recovery 108
Summary 110

CHAPTER 6 Advanced Operations 111

Enhancing the Data Store 111
Endian Issues 123
Byte Order Reversal 123
`Btree` Comparison Function 125
Alignment Issues 126
Secondary Indices 129
Database Operations 135
Cursors 141
Duplicate Keys 143
Locking in Cursors 143
Long-Running Cursors 144
Cursors and Secondary Indices 144
Database Iterators 145

Equality Joins 151
Bulk Retrieval 155
Summary 158

CHAPTER 7 A Real-World Data Store 159

Constraints on Environment Usage 159
A Single Process With One Thread 161
A Single Process With Multiple Threads 161
Handling an Ungraceful Exit 167
Using DbEnv::failchk 170
Using DbEnv::set_thread_id 170
Using DbEnv::set_isalive 171
A Note on DS, CDS, and TDS 176
Multiple Processes 176
Groups of Cooperating Processes 177
Groups of Unrelated Processes 190
Database Configuration 198
Client/Server Configuration 198
Direct Linking 199
Hybrid Approach 199
Summary 200

CHAPTER 8 Replication 201

What Is Database Replication? 201
Single-Master Replication 202
Multimaster Replication 203
Hybrid Master-Client Replication 203
Berkeley DB Replication Architecture 204
Salient Features of Berkeley DB Replication 205
Replication Framework Components 206
Building the Replication Framework Using the Base API 208
Replication APIs 209
Opening the Environment 212
Managing Nodes 214
Designing the Communication Framework 216
Implementing the Communication Framework 219
Implementing the Replication Manager 225

Network Partitions 263
Recovering from a Network Partition 264
Preventing Network Partitions 264
The Degenerate Case 265
Transactional Guarantees 265
Replication Manager Interface 266
Tunable Parameters 268
`repmgr_start` 269
Limitations 270
Summary 271

CHAPTER 9 Distributed Transactions and Data-Distribution Strategies 273

Distributed Transactions 273
Properties of Distributed Transactions 274
Two-Phase Commit 274
Distributed Transactions in Berkeley DB 276
Building a Global Transaction Manager 281
Assumptions 281
Code Organization 282
Running the GTM 308
Disclaimer 311
Data-Distribution Strategies 311
Fault Tolerance 312
Load Balancing 316
The Hybrid Approach 319
Summary 321

CHAPTER 10 Berkeley DB Utilities 323

Introduction to the Utilities 323
`DB_HOME` Environment Variable 324
The Common Options 324
`db_stat` 326
Locking Subsystem 327
Detailed Information on the Locking Subsystem 329
Database Statistics 334
Logging Subsystem Statistics 336
Cache Statistics 337
Transaction Statistics 338
Replication Statistics 339

db_recover 342
Simple Recovery 342
Catastrophic Recovery 343
Preserving the Environment During Recovery 344
Recovering Up to a Timestamp 345
db_dump 346
Verification Tool 346
Data Salvage 348
db_load 349
db_checkpoint 350
db_deadlock 351
db_printlog 351
db_archive 355
db_hotbackup 355
db_verify 357
Summary 357

CHAPTER 11 Berkeley DB Java APIs 359

Understanding the Two APIs 359
The Java API 359
The Java Collections API 360
Berkeley DB Java API Packages 361
Compiling and Using the Java API 361
Opening the Environment 363
Opening the Database 364
Creating Database Records 365
Binding Data Types in the Java API 368
Binding for Primitive Types 368
Serial Binding for Complex Types 369
Custom Tuple Binding for Complex Types 371
Binding Data Types in the Collections API 374
Java Stored Collections 375
Entity Binding 376
Serializable Entity Class 382
Transactions 384
Transactions and Stored Collections API 385
Database Operations 386
Secondary Indices 387
Complete Code Example 388
Summary 402

CHAPTER 12 Berkeley DB C API 403

Compiling and Installing the C API 403
Basic Operations Using the C API 404
Opening the Environment 404
Creating a Database Record 406
Using Transactions 407
Opening the Database 408
Creating a Secondary Index 409
Simple Database Operations 410
Error Returns 417
Documented Errors 418
Handling Undocumented Errors 419
Complete Code Example 419
Summary 429

INDEX 431