Database Internals

A Deep Dive into How Distributed Data Systems Work

Alex Petrov

Beijing · Boston · Farnham · Sebastopol · Tokyo

O'REILLY®

Database Internals
by Alex Petrov

Copyright © 2019 Oleksandr Petrov. All rights reserved.

Printed in the United States of America.

Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O’Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Mike Loukides
Development Editor: Michele Cronin
Production Editor: Christopher Faucher
Copyeditor: Kim Cofer
Proofreader: Sonia Saruba

Indexer: Judith McConville
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

October 2019: First Edition

Revision History for the First Edition
2019-09-12: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781492040347* for release details.

The O’Reilly logo is a registered trademark of O’Reilly Media, Inc. *Database Internals*, the cover image, and related trade dress are trademarks of O’Reilly Media, Inc.

The views expressed in this work are those of the author, and do not represent the publisher’s views. While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-492-04034-7

[MBP]

Table of Contents

Preface. xiii

Part I. Storage Engines

1. Introduction and Overview. 7
   - DBMS Architecture 8
   - Memory- Versus Disk-Based DBMS 10
     - Durability in Memory-Based Stores 11
   - Column- Versus Row-Oriented DBMS 12
     - Row-Oriented Data Layout 13
     - Column-Oriented Data Layout 14
     - Distinctions and Optimizations 15
     - Wide Column Stores 15
   - Data Files and Index Files 17
     - Data Files 18
     - Index Files 18
     - Primary Index as an Indirection 20
   - Buffering, Immutability, and Ordering 21
   - Summary 22

2. B-Tree Basics. 25
   - Binary Search Trees 26
     - Tree Balancing 27
     - Trees for Disk-Based Storage 28
   - Disk-Based Structures 29
     - Hard Disk Drives 30
     - Solid State Drives 30

On-Disk Structures 32
Ubiquitous B-Trees 33
B-Tree Hierarchy 35
Separator Keys 36
B-Tree Lookup Complexity 37
B-Tree Lookup Algorithm 38
Counting Keys 38
B-Tree Node Splits 39
B-Tree Node Merges 41
Summary 42

3. File Formats. 45
Motivation 46
Binary Encoding 47
Primitive Types 47
Strings and Variable-Size Data 49
Bit-Packed Data: Booleans, Enums, and Flags 49
General Principles 50
Page Structure 52
Slotted Pages 52
Cell Layout 54
Combining Cells into Slotted Pages 56
Managing Variable-Size Data 57
Versioning 58
Checksumming 59
Summary 60

4. Implementing B-Trees. 61
Page Header 61
Magic Numbers 62
Sibling Links 62
Rightmost Pointers 63
Node High Keys 64
Overflow Pages 65
Binary Search 67
Binary Search with Indirection Pointers 67
Propagating Splits and Merges 68
Breadcrumbs 69
Rebalancing 70
Right-Only Appends 71
Bulk Loading 72
Compression 73

Vacuum and Maintenance 74
Fragmentation Caused by Updates and Deletes 75
Page Defragmentation 76
Summary 76

5. Transaction Processing and Recovery. 79
Buffer Management 81
Caching Semantics 83
Cache Eviction 83
Locking Pages in Cache 84
Page Replacement 85
Recovery 88
Log Semantics 90
Operation Versus Data Log 91
Steal and Force Policies 91
ARIES 92
Concurrency Control 93
Serializability 94
Transaction Isolation 95
Read and Write Anomalies 95
Isolation Levels 96
Optimistic Concurrency Control 98
Multiversion Concurrency Control 99
Pessimistic Concurrency Control 99
Lock-Based Concurrency Control 100
Summary 108

6. B-Tree Variants. 111
Copy-on-Write 112
Implementing Copy-on-Write: LMDB 113
Abstracting Node Updates 113
Lazy B-Trees 114
WiredTiger 114
Lazy-Adaptive Tree 116
FD-Trees 117
Fractional Cascading 118
Logarithmic Runs 119
Bw-Trees 120
Update Chains 121
Taming Concurrency with Compare-and-Swap 121
Structural Modification Operations 122
Consolidation and Garbage Collection 123

Cache-Oblivious B-Trees 124
van Emde Boas Layout 125
Summary 127

7. Log-Structured Storage. 129
LSM Trees 130
LSM Tree Structure 132
Updates and Deletes 136
LSM Tree Lookups 137
Merge-Iteration 137
Reconciliation 140
Maintenance in LSM Trees 141
Read, Write, and Space Amplification 143
RUM Conjecture 144
Implementation Details 145
Sorted String Tables 145
Bloom Filters 146
Skiplist 148
Disk Access 150
Compression 151
Unordered LSM Storage 152
Bitcask 153
WiscKey 154
Concurrency in LSM Trees 155
Log Stacking 157
Flash Translation Layer 157
Filesystem Logging 159
LLAMA and Mindful Stacking 160
Open-Channel SSDs 161
Summary 162

Part I Conclusion. 165

Part II. Distributed Systems

8. Introduction and Overview. 171
Concurrent Execution 171
Shared State in a Distributed System 173
Fallacies of Distributed Computing 174
Processing 175
Clocks and Time 176

State Consistency 177
Local and Remote Execution 178
Need to Handle Failures 178
Network Partitions and Partial Failures 179
Cascading Failures 180
Distributed Systems Abstractions 181
Links 182
Two Generals' Problem 187
FLP Impossibility 189
System Synchrony 190
Failure Models 191
Crash Faults 191
Omission Faults 192
Arbitrary Faults 193
Handling Failures 193
Summary 193

9. Failure Detection. 195
Heartbeats and Pings 196
Timeout-Free Failure Detector 197
Outsourced Heartbeats 198
Phi-Accural Failure Detector 199
Gossip and Failure Detection 200
Reversing Failure Detection Problem Statement 201
Summary 202

10. Leader Election. 205
Bully Algorithm 207
Next-In-Line Failover 208
Candidate/Ordinary Optimization 209
Invitation Algorithm 210
Ring Algorithm 211
Summary 212

11. Replication and Consistency. 215
Achieving Availability 216
Infamous CAP 216
Use CAP Carefully 217
Harvest and Yield 218
Shared Memory 219
Ordering 221
Consistency Models 222

Strict Consistency 223
Linearizability 223
Sequential Consistency 227
Causal Consistency 229
Session Models 233
Eventual Consistency 234
Tunable Consistency 235
Witness Replicas 236
Strong Eventual Consistency and CRDTs 238
Summary 240

12. Anti-Entropy and Dissemination. 243
Read Repair 245
Digest Reads 246
Hinted Handoff 246
Merkle Trees 247
Bitmap Version Vectors 248
Gossip Dissemination 250
Gossip Mechanics 251
Overlay Networks 251
Hybrid Gossip 253
Partial Views 254
Summary 255

13. Distributed Transactions. 257
Making Operations Appear Atomic 258
Two-Phase Commit 259
Cohort Failures in 2PC 261
Coordinator Failures in 2PC 262
Three-Phase Commit 264
Coordinator Failures in 3PC 265
Distributed Transactions with Calvin 266
Distributed Transactions with Spanner 268
Database Partitioning 270
Consistent Hashing 271
Distributed Transactions with Percolator 272
Coordination Avoidance 275
Summary 277

14. Consensus. 279
Broadcast 280
Atomic Broadcast 281

Virtual Synchrony 282
Zookeeper Atomic Broadcast (ZAB) 283
Paxos 285
Paxos Algorithm 286
Quorums in Paxos 287
Failure Scenarios 288
Multi-Paxos 291
Fast Paxos 292
Egalitarian Paxos 293
Flexible Paxos 296
Generalized Solution to Consensus 297
Raft 300
Leader Role in Raft 302
Failure Scenarios 304
Byzantine Consensus 305
PBFT Algorithm 306
Recovery and Checkpointing 309
Summary 309

Part II Conclusion. 313

A. Bibliography. 317

Index. 337