

# Principles of Distributed Database Systems

M. Tamer Özsu • Patrick Valduriez

# Principles of Distributed Database Systems

Fourth Edition



Springer

M. Tamer Özsu  
Cheriton School of Computer Science  
University of Waterloo  
Waterloo, ON, Canada

Patrick Valduriez  
Inria and LIRMM  
University of Montpellier  
Montpellier, France

---

The first two editions of this book were published by: Pearson Education, Inc.

---

ISBN 978-3-030-26252-5      ISBN 978-3-030-26253-2 (eBook)  
<https://doi.org/10.1007/978-3-030-26253-2>

3<sup>rd</sup> edition: © Springer Science+Business Media, LLC 2011

© Springer Nature Switzerland AG 2020, corrected publication 2020

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.

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.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*To our families  
and our parents  
M.T.Ö. and P.V.*

# Preface

The first edition of this book appeared in 1991 when the technology was new and there were not too many products. In the Preface to the first edition, we had quoted Michael Stonebraker who claimed in 1988 that in the following 10 years, centralized DBMSs would be an “antique curiosity” and most organizations would move towards distributed DBMSs. That prediction has certainly proved to be correct, and a large proportion of the systems in use today are either distributed or parallel—commonly referred to as scale-out systems. When we were putting together the first edition, undergraduate and graduate database courses were not as prevalent as they are now; so the initial version of the book contained lengthy discussions of centralized solutions before introducing their distributed/parallel counterparts. Times have certainly changed on that front as well, and now, it is hard to find a graduate student who does not have at least some rudimentary knowledge of database technology. Therefore, a graduate-level textbook on distributed/parallel database technology needs to be positioned differently today. That was our objective in this edition while maintaining the many new topics we introduced in the third edition. The main revisions introduced in this fourth edition are the following:

1. Over the years, the motivations and the environment for this technology have somewhat shifted (Web, cloud, etc.). In light of this, the introductory chapter needed a serious refresh. We revised the introduction with the aim of a more contemporary look at the technology.
2. We have added a new chapter on big data processing to cover distributed storage systems, data stream processing, MapReduce and Spark platforms, graph analytics, and data lakes. With the proliferation of these systems, systematic treatment of these topics is essential.
3. Similarly, we addressed the growing influence of NoSQL systems by devoting a new chapter to it. This chapter covers the four types of NoSQL (key-value stores, document stores, wide column systems, and graph DBMSs), as well as NewSQL systems and polystores.
4. We have combined the database integration and multidatabase query processing chapters from the third edition into a uniform chapter on database integration.

5. We undertook a major revision of the web data management discussion that previously focused mostly on XML to refocus on RDF technology, which is more prevalent at this time. We now discuss, in this chapter, web data integration approaches, including the important issue of data quality.
6. We have revised and updated the peer-to-peer data management chapter and included a lengthy discussion of blockchain.
7. As part of our cleaning the previous chapters, we condensed the query processing and transaction management chapters by removing the fundamental centralized techniques and focused these chapters on distributed/parallel techniques. In the process, we included some topics that have since gained importance, such as dynamic query processing (eddis) and Paxos consensus algorithm and its use in commit protocols.
8. We updated the parallel DBMS chapter by clarifying the objectives, in particular, scale-up versus scale-out, and discussing parallel architectures that include UMA or NUMA. We also added a new section of parallel sorting algorithms and variants of parallel join algorithms to exploit large main memories and multicore processors that are prevalent today.
9. We updated the distribution design chapter by including a lengthy discussion of modern approaches that combine fragmentation and allocation. By rearranging material, this chapter is now central to data partitioning for both the distributed and parallel data management discussions in the remainder of the book.
- 10 Although object technology continues to play a role in information systems, its importance in distributed/parallel data management has declined. Therefore, we removed the chapter on object databases from this edition.

As is evident, the entire book and every chapter have seen revisions and updates for a more contemporary treatment. The material we removed in the process is not lost—they are included as online appendices and appear on the book’s web page: <https://cs.uwaterloo.ca/ddbs>. We elected to make these available online rather than in the print version to keep the size of the book reasonable (which also keeps the price reasonable). The web site also includes presentation slides that can be used to teach from the book as well as solutions to most of the exercises (available only to instructors who have adopted the book for teaching).

As in previous editions, many colleagues helped with this edition of the book whom we would like to thank (in no specific order). Dan Olteanu provided a nice discussion of two optimizations that can significantly reduce the maintenance time of materialized views in Chap. 3. Phil Bernstein provided leads for new papers on the multiversion transaction management that resulted in updates to that discussion in Chap. 5. Khuzaima Daudjee was also helpful in providing a list of more contemporary publications on distributed transaction processing that we include in the bibliographic notes section of that chapter. Ricardo Jimenez-Peris contributed text on high-performance transaction systems that is included in the same chapter. He also contributed a section on LeanXcale in the NoSQL, NewSQL, and polystores chapter. Dennis Shasha reviewed the new blockchain section in the P2P chapter. Michael Carey read the big data, NoSQL, NewSQL and

polystores, and parallel DBMS chapters and provided extremely detailed comments that improved those chapters considerably. Tamer's students Anil Pacaci, Khaled Ammar and postdoc Xiaofei Zhang provided extensive reviews of the big data chapter, and texts from their publications are included in this chapter. The NoSQL, NewSQL, and polystores chapter includes text from publications of Boyan Koley and Patrick's student Carlyna Bondiombouy. Jim Webber reviewed the section on Neo4j in that chapter. The characterization of graph analytics systems in that chapter is partially based on Minyang Han's master's thesis where he also proposes GiraphUC approach that is discussed in that chapter. Semih Salihoglu and Lukasz Golab also reviewed and provided very helpful comments on parts of this chapter. Alon Halevy provided comments on the WebTables discussion in Chap. 12. The data quality discussion in web data integration is contributed by Ihab Ilyas and Xu Chu. Stratos Idreos was very helpful in clarifying how database cracking can be used as a partitioning approach and provided text that is included in Chap. 2. Renan Souza and Fabian Stöter reviewed the entire book.

The third edition of the book introduced a number of new topics that carried over to this edition, and a number of colleagues were very influential in writing those chapters. We would like to, once again, acknowledge their assistance since their impact is reflected in the current edition as well. Renée Miller, Erhard Rahm, and Alon Halevy were critical in putting together the discussion on database integration, which was reviewed thoroughly by Avigdor Gal. Matthias Jarke, Xiang Li, Gottfried Vossen, Erhard Rahm, and Andreas Thor contributed exercises to this chapter. Hubert Naacke contributed to the section on heterogeneous cost modeling and Fabio Porto to the section on adaptive query processing. Data replication (Chap. 6) could not have been written without the assistance of Gustavo Alonso and Bettina Kemme. Esther Pacitti also contributed to the data replication chapter, both by reviewing it and by providing background material; she also contributed to the section on replication in database clusters in the parallel DBMS chapter. Peer-to-peer data management owes a lot to the discussions with Beng Chin Ooi. The section of this chapter on query processing in P2P systems uses material from the PhD work of Reza Akbarinia and Wenceslao Palma, while the section on replication uses material from the PhD work of Vidal Martins.

We thank our editor at Springer Susan Lagerstrom-Fife for pushing this project within Springer and also pushing us to finish it in a timely manner. We missed almost all of her deadlines, but we hope the end result is satisfactory.

Finally, we would be very interested to hear your comments and suggestions regarding the material. We welcome any feedback, but we would particularly like to receive feedback on the following aspects:

1. Any errors that may have remained despite our best efforts (although we hope there are not many);

2. Any topics that should no longer be included and any topics that should be added or expanded;
3. Any exercises that you may have designed that you would like to be included in the book.

Waterloo, Canada  
Montpellier, France  
June 2019

M. Tamer Özsu (tamer.ozsu@uwaterloo.ca)  
Patrick Valduriez (patrick.valduriez@inria.fr)



# Contents

<b>1</b>	<b>Introduction .....</b>	<b>1</b>
1.1	What Is a Distributed Database System?.....	1
1.2	History of Distributed DBMS .....	3
1.3	Data Delivery Alternatives .....	5
1.4	Promises of Distributed DBMSs .....	7
1.4.1	Transparent Management of Distributed and Replicated Data.....	7
1.4.2	Reliability Through Distributed Transactions.....	10
1.4.3	Improved Performance.....	11
1.4.4	Scalability .....	13
1.5	Design Issues .....	13
1.5.1	Distributed Database Design .....	13
1.5.2	Distributed Data Control.....	14
1.5.3	Distributed Query Processing.....	14
1.5.4	Distributed Concurrency Control.....	14
1.5.5	Reliability of Distributed DBMS .....	15
1.5.6	Replication.....	15
1.5.7	Parallel DBMSs .....	16
1.5.8	Database Integration .....	16
1.5.9	Alternative Distribution Approaches .....	16
1.5.10	Big Data Processing and NoSQL.....	16
1.6	Distributed DBMS Architectures .....	17
1.6.1	Architectural Models for Distributed DBMSs .....	17
1.6.2	Client/Server Systems.....	20
1.6.3	Peer-to-Peer Systems.....	22
1.6.4	Multidatabase Systems.....	25
1.6.5	Cloud Computing .....	27
1.7	Bibliographic Notes .....	31

<b>2</b>	<b>Distributed and Parallel Database Design</b>	33
2.1	Data Fragmentation	35
2.1.1	Horizontal Fragmentation	37
2.1.2	Vertical Fragmentation	52
2.1.3	Hybrid Fragmentation	65
2.2	Allocation	66
2.2.1	Auxiliary Information	68
2.2.2	Allocation Model	69
2.2.3	Solution Methods	72
2.3	Combined Approaches	72
2.3.1	Workload-Agnostic Partitioning Techniques	73
2.3.2	Workload-Aware Partitioning Techniques	74
2.4	Adaptive Approaches	78
2.4.1	Detecting Workload Changes	79
2.4.2	Detecting Affected Items	79
2.4.3	Incremental Reconfiguration	80
2.5	Data Directory	82
2.6	Conclusion	83
2.7	Bibliographic Notes	84
<b>3</b>	<b>Distributed Data Control</b>	91
3.1	View Management	92
3.1.1	Views in Centralized DBMSs	92
3.1.2	Views in Distributed DBMSs	95
3.1.3	Maintenance of Materialized Views	96
3.2	Access Control	102
3.2.1	Discretionary Access Control	103
3.2.2	Mandatory Access Control	106
3.2.3	Distributed Access Control	108
3.3	Semantic Integrity Control	110
3.3.1	Centralized Semantic Integrity Control	111
3.3.2	Distributed Semantic Integrity Control	116
3.4	Conclusion	123
3.5	Bibliographic Notes	123
<b>4</b>	<b>Distributed Query Processing</b>	129
4.1	Overview	130
4.1.1	Query Processing Problem	130
4.1.2	Query Optimization	133
4.1.3	Layers Of Query Processing	136
4.2	Data Localization	140
4.2.1	Reduction for Primary Horizontal Fragmentation	141
4.2.2	Reduction with Join	142
4.2.3	Reduction for Vertical Fragmentation	143
4.2.4	Reduction for Derived Fragmentation	145
4.2.5	Reduction for Hybrid Fragmentation	148

4.3	Join Ordering in Distributed Queries .....	149
4.3.1	Join Trees .....	149
4.3.2	Join Ordering .....	151
4.3.3	Semijoin-Based Algorithms .....	153
4.3.4	Join Versus Semijoin .....	156
4.4	Distributed Cost Model .....	157
4.4.1	Cost Functions .....	157
4.4.2	Database Statistics .....	159
4.5	Distributed Query Optimization .....	161
4.5.1	Dynamic Approach .....	161
4.5.2	Static Approach .....	165
4.5.3	Hybrid Approach .....	169
4.6	Adaptive Query Processing .....	173
4.6.1	Adaptive Query Processing Process .....	174
4.6.2	Eddy Approach .....	176
4.7	Conclusion .....	177
4.8	Bibliographic Notes .....	178
<b>5</b>	<b>Distributed Transaction Processing .....</b>	<b>183</b>
5.1	Background and Terminology .....	184
5.2	Distributed Concurrency Control .....	188
5.2.1	Locking-Based Algorithms .....	189
5.2.2	Timestamp-Based Algorithms .....	197
5.2.3	Multiversion Concurrency Control .....	203
5.2.4	Optimistic Algorithms .....	205
5.3	Distributed Concurrency Control Using Snapshot Isolation .....	206
5.4	Distributed DBMS Reliability .....	209
5.4.1	Two-Phase Commit Protocol .....	211
5.4.2	Variations of 2PC .....	217
5.4.3	Dealing with Site Failures .....	220
5.4.4	Network Partitioning .....	227
5.4.5	Paxos Consensus Protocol .....	231
5.4.6	Architectural Considerations .....	234
5.5	Modern Approaches to Scaling Out Transaction Management ....	236
5.5.1	Spanner .....	237
5.5.2	LeanXcale .....	237
5.6	Conclusion .....	239
5.7	Bibliographic Notes .....	241
<b>6</b>	<b>Data Replication .....</b>	<b>247</b>
6.1	Consistency of Replicated Databases .....	249
6.1.1	Mutual Consistency .....	249
6.1.2	Mutual Consistency Versus Transaction Consistency ...	251
6.2	Update Management Strategies .....	252
6.2.1	Eager Update Propagation .....	253
6.2.2	Lazy Update Propagation .....	254

6.2.3	Centralized Techniques .....	254
6.2.4	Distributed Techniques.....	255
6.3	Replication Protocols .....	255
6.3.1	Eager Centralized Protocols .....	256
6.3.2	Eager Distributed Protocols.....	262
6.3.3	Lazy Centralized Protocols .....	262
6.3.4	Lazy Distributed Protocols .....	268
6.4	Group Communication.....	269
6.5	Replication and Failures .....	272
6.5.1	Failures and Lazy Replication .....	273
6.5.2	Failures and Eager Replication .....	273
6.6	Conclusion.....	276
6.7	Bibliographic Notes .....	277
<b>7</b>	<b>Database Integration—Multidatabase Systems .....</b>	<b>281</b>
7.1	Database Integration .....	282
7.1.1	Bottom-Up Design Methodology .....	283
7.1.2	Schema Matching .....	287
7.1.3	Schema Integration.....	296
7.1.4	Schema Mapping .....	298
7.1.5	Data Cleaning .....	306
7.2	Multidatabase Query Processing .....	307
7.2.1	Issues in Multidatabase Query Processing .....	308
7.2.2	Multidatabase Query Processing Architecture.....	309
7.2.3	Query Rewriting Using Views .....	311
7.2.4	Query Optimization and Execution .....	317
7.2.5	Query Translation and Execution.....	329
7.3	Conclusion.....	332
7.4	Bibliographic Notes .....	334
<b>8</b>	<b>Parallel Database Systems .....</b>	<b>349</b>
8.1	Objectives.....	350
8.2	Parallel Architectures .....	352
8.2.1	General Architecture .....	353
8.2.2	Shared-Memory .....	355
8.2.3	Shared-Disk .....	357
8.2.4	Shared-Nothing.....	358
8.3	Data Placement .....	359
8.4	Parallel Query Processing.....	362
8.4.1	Parallel Algorithms for Data Processing .....	362
8.4.2	Parallel Query Optimization .....	369
8.5	Load Balancing .....	374
8.5.1	Parallel Execution Problems .....	374
8.5.2	Intraoperator Load Balancing.....	376
8.5.3	Interoperator Load Balancing.....	378
8.5.4	Intraquery Load Balancing .....	378

8.6	Fault-Tolerance .....	383
8.7	Database Clusters .....	384
8.7.1	Database Cluster Architecture .....	385
8.7.2	Replication.....	386
8.7.3	Load Balancing.....	386
8.7.4	Query Processing .....	387
8.8	Conclusion .....	390
8.9	Bibliographic Notes .....	390
<b>9</b>	<b>Peer-to-Peer Data Management.....</b>	<b>395</b>
9.1	Infrastructure .....	398
9.1.1	Unstructured P2P Networks .....	399
9.1.2	Structured P2P Networks .....	402
9.1.3	Superpeer P2P Networks .....	406
9.1.4	Comparison of P2P Networks .....	408
9.2	Schema Mapping in P2P Systems .....	408
9.2.1	Pairwise Schema Mapping.....	408
9.2.2	Mapping Based on Machine Learning Techniques .....	409
9.2.3	Common Agreement Mapping .....	410
9.2.4	Schema Mapping Using IR Techniques.....	411
9.3	Querying Over P2P Systems.....	411
9.3.1	Top-k Queries .....	412
9.3.2	Join Queries .....	424
9.3.3	Range Queries .....	425
9.4	Replica Consistency.....	428
9.4.1	Basic Support in DHTs .....	429
9.4.2	Data Currency in DHTs.....	431
9.4.3	Replica Reconciliation .....	432
9.5	Blockchain .....	436
9.5.1	Blockchain Definition.....	437
9.5.2	Blockchain Infrastructure .....	438
9.5.3	Blockchain 2.0.....	442
9.5.4	Issues.....	443
9.6	Conclusion.....	444
9.7	Bibliographic Notes .....	445
<b>10</b>	<b>Big Data Processing .....</b>	<b>449</b>
10.1	Distributed Storage Systems .....	451
10.1.1	Google File System .....	453
10.1.2	Combining Object Storage and File Storage.....	454
10.2	Big Data Processing Frameworks .....	455
10.2.1	MapReduce Data Processing .....	456
10.2.2	Data Processing Using Spark .....	466
10.3	Stream Data Management .....	470
10.3.1	Stream Models, Languages, and Operators .....	472
10.3.2	Query Processing over Data Streams.....	476
10.3.3	DSS Fault-Tolerance .....	483

10.4	Graph Analytics Platforms .....	486
10.4.1	Graph Partitioning .....	489
10.4.2	MapReduce and Graph Analytics .....	494
10.4.3	Special-Purpose Graph Analytics Systems .....	495
10.4.4	Vertex-Centric Block Synchronous .....	498
10.4.5	Vertex-Centric Asynchronous .....	501
10.4.6	Vertex-Centric Gather-Apply-Scatter .....	503
10.4.7	Partition-Centric Block Synchronous Processing .....	504
10.4.8	Partition-Centric Asynchronous .....	506
10.4.9	Partition-Centric Gather-Apply-Scatter .....	506
10.4.10	Edge-Centric Block Synchronous Processing .....	507
10.4.11	Edge-Centric Asynchronous .....	507
10.4.12	Edge-Centric Gather-Apply-Scatter .....	507
10.5	Data Lakes .....	508
10.5.1	Data Lake Versus Data Warehouse .....	508
10.5.2	Architecture .....	510
10.5.3	Challenges .....	511
10.6	Conclusion .....	512
10.7	Bibliographic Notes .....	512
<b>11</b>	<b>NoSQL, NewSQL, and Polystores .....</b>	<b>519</b>
11.1	Motivations for NoSQL .....	520
11.2	Key-Value Stores .....	521
11.2.1	DynamoDB .....	522
11.2.2	Other Key-Value Stores .....	524
11.3	Document Stores .....	525
11.3.1	MongoDB .....	525
11.3.2	Other Document Stores .....	528
11.4	Wide Column Stores .....	529
11.4.1	Bigtable .....	529
11.4.2	Other Wide Column Stores .....	531
11.5	Graph DBMSs .....	531
11.5.1	Neo4j .....	532
11.5.2	Other Graph Databases .....	535
11.6	Hybrid Data Stores .....	536
11.6.1	Multimodel NoSQL Stores .....	536
11.6.2	NewSQL DBMSs .....	537
11.7	Polystores .....	540
11.7.1	Loosely Coupled Polystores .....	540
11.7.2	Tightly Coupled Polystores .....	545
11.7.3	Hybrid Systems .....	549
11.7.4	Concluding Remarks .....	554
11.8	Conclusion .....	554
11.9	Bibliographic Notes .....	555

<b>12 Web Data Management .....</b>	<b>559</b>
12.1 Web Graph Management .....	560
12.2 Web Search .....	562
12.2.1 Web Crawling .....	563
12.2.2 Indexing .....	566
12.2.3 Ranking and Link Analysis .....	567
12.2.4 Evaluation of Keyword Search .....	568
12.3 Web Querying .....	569
12.3.1 Semistructured Data Approach .....	570
12.3.2 Web Query Language Approach .....	574
12.4 Question Answering Systems .....	580
12.5 Searching and Querying the Hidden Web .....	584
12.5.1 Crawling the Hidden Web .....	585
12.5.2 Metasearching .....	586
12.6 Web Data Integration .....	588
12.6.1 Web Tables/Fusion Tables .....	589
12.6.2 Semantic Web and Linked Open Data .....	590
12.6.3 Data Quality Issues in Web Data Integration .....	608
12.7 Bibliographic Notes .....	615
<b>Correction to: Principles of Distributed Database Systems .....</b>	<b>C1</b>
<b>A Overview of Relational DBMS .....</b>	<b>619</b>
<b>B Centralized Query Processing .....</b>	<b>621</b>
<b>C Transaction Processing Fundamentals .....</b>	<b>623</b>
<b>D Review of Computer Networks .....</b>	<b>625</b>
<b>References .....</b>	<b>627</b>
<b>Index .....</b>	<b>663</b>