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

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 (eddies) 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 Kolev 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

The original version of this book was revised. The correction to this book is available at https://doi.org/10.1007/978-3-030-26253-2_13