Instant Recovery with Write-Ahead Logging

Page Repair, System Restart, Media Restore, and System Failover

Second Edition

Synthesis Lectures on Data Management

Synthesis Lectures on Data Management is edited by Meral Özsoyoğlu of Case Western Reserve University. The series publishes 80- to 150-page publications on topics pertaining to data management. Topics include query languages, database system architectures, transaction management, data warehousing, XML and databases, data stream systems, wide scale data distribution, multimedia data management, data mining, and related subjects.

Business Processes: A Database Perspective
Daniel Deutch and Tova Milo
July 2012

Data Protection from Insider Threats
Elisa Bertino
June 2012

Deep Web Query Interface Understanding and Integration
Eduard C. Dragut, Weiyi Meng, and Clement T. Yu
June 2012

P2P Techniques for Decentralized Applications
Esther Pacitti, Reza Akbarinia, and Manal El-Dick
April 2012

Query Answer Authentication
HweeHwa Pang and Kian-Lee Tan
February 2012

Declarative Networking
Boon Thau Loo and Wenchao Zhou
January 2012

Full-Text (Substring) Indexes in External Memory
Marina Barsky, Ulrike Stege, and Alex Thomo
December 2011

Spatial Data Management
Nikos Mamoulis
November 2011

Database Repairing and Consistent Query Answering
Leopoldo Bertossi
August 2011

Managing Event Information: Modeling, Retrieval, and Applications
Amarnath Gupta and Ramesh Jain
July 2011

Fundamentals of Physical Design and Query Compilation
David Toman and Grant Weddell
July 2011

Methods for Mining and Summarizing Text Conversations
Giuseppe Carenini, Gabriel Murray, and Raymond Ng
June 2011

Probabilistic Databases
Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch
May 2011

Peer-to-Peer Data Management
Karl Aberer
May 2011

Probabilistic Ranking Techniques in Relational Databases
Ihab F. Ilyas and Mohamed A. Soliman
March 2011

Uncertain Schema Matching
Avigdor Gal
March 2011

Fundamentals of Object Databases: Object-Oriented and Object-Relational Design
Suzanne W. Dietrich and Susan D. Urban
2010

Advanced Metasearch Engine Technology
Weiyi Meng and Clement T. Yu
2010

Web Page Recommendation Models: Theory and Algorithms
Şule Gündüz-Ögüdücü
2010

Multidimensional Databases and Data Warehousing
Christian S. Jensen, Torben Bach Pedersen, and Christian Thomsen
2010

Database Replication
Bettina Kemme, Ricardo Jimenez-Peris, and Marta Patino-Martinez
2010

Relational and XML Data Exchange
Marcelo Arenas, Pablo Barcelo, Leonid Libkin, and Filip Murlak
2010

User-Centered Data Management
Tiziana Catarci, Alan Dix, Stephen Kimani, and Giuseppe Santucci
2010

Data Stream Management
Lukasz Golab and M. Tamer Özsu
2010

Access Control in Data Management Systems
Elena Ferrari
2010

An Introduction to Duplicate Detection
Felix Naumann and Melanie Herschel
2010

Privacy-Preserving Data Publishing: An Overview
Raymond Chi-Wing Wong and Ada Wai-Chee Fu
2010

Keyword Search in Databases
Jeffrey Xu Yu, Lu Qin, and Lijun Chang
2009

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, Media Restore, and System Failover: Second Edition
Goetz Graefe, Wey Guy, Caetano Sauer

ISBN: 978-3-031-00729-3 print
ISBN: 978-3-031-01857-2 ebook

DOI 10.1007/978-3-031-01857-2

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MANAGEMENT #44
Series Editor: Z. Meral Özsoyoğlu, Case Western Reserve University
Founding Editor: M. Tamer Özsu, University of Waterloo

Series ISSN 2153-5418 Print 2153-5426 Electronic

Instant Recovery with Write-Ahead Logging

Page Repair, System Restart, Media Restore, and System Failover

Second Edition

Goetz Graefe
Hewlett Packard Labs

Wey Guy

Caetano Sauer
University of Kaiserslautern

SYNTHESIS LECTURES ON DATA MANAGEMENT #44

ABSTRACT

Traditional theory and practice of write-ahead logging and of database recovery focus on three failure classes: transaction failures (typically due to deadlocks) resolved by transaction rollback; system failures (typically power or software faults) resolved by restart with log analysis, "redo," and "undo" phases; and media failures (typically hardware faults) resolved by restore operations that combine multiple types of backups and log replay.

The recent addition of single-page failures and single-page recovery has opened new opportunities far beyond the original aim of immediate, lossless repair of single-page wear-out in novel or traditional storage hardware. In the contexts of system and media failures, efficient single-page recovery enables on-demand incremental "redo" and "undo" as part of system restart or media restore operations. This can give the illusion of practically instantaneous restart and restore: instant restart permits processing new queries and updates seconds after system reboot and instant restore permits resuming queries and updates on empty replacement media as if those were already fully recovered. In the context of node and network failures, instant restart and instant restore combine to enable practically instant failover from a failing database node to one holding merely an out-of-date backup and a log archive, yet without loss of data, updates, or transactional integrity.

In addition to these instant recovery techniques, the discussion introduces self-repairing indexes and much faster offline restore operations, which impose no slowdown in backup operations and hardly any slowdown in log archiving operations. The new restore techniques also render differential and incremental backups obsolete, complete backup commands on a database server practically instantly, and even permit taking full up-to-date backups without imposing any load on the database server.

Compared to the first version of this book, this second edition adds sections on applications of single-page repair, instant restart, single-pass restore, and instant restore. Moreover, it adds sections on instant failover among nodes in a cluster, applications of instant failover, recovery for file systems and data files, and the performance of instant restart and instant restore.

KEYWORDS

algorithms, databases, transactions, failures, recovery, availability, reliability, write-ahead logging, instant restart, log analysis, redo, undo, rollback, compensation, log replay, instant restore, single-pass restore, virtual backup, big data, file systems, key-value stores, clusters, log shipping, failover, elasticity, failover pool

Contents

Preface

It has been a pleasure developing and compiling this set of concepts and techniques in order to make them available to researchers and software developers around the world. While the foundation of the presented techniques is write-ahead logging as commonly found in database management systems, the techniques and their advantages apply similarly to key-value stores, file systems with journaling, etc.—in other words, to all storage management layers for important and big data. In all these systems, write-ahead logging can enable efficient single-page repair after a localized data loss, system restart after a software crash, and media restore after a failure in the storage hardware or firmware. Instead of copying each data page to multiple devices, as many file storage systems do today in order to achieve high availability, only a single copy is required, plus a log of changes.

Acknowledgments

Barb Peters and Arianna Lund encouraged combining all "instant recovery" techniques into a single book. Harumi Kuno participated in the research on single-page recovery and its applications.