MORGAN & CLAYPOOL PUBLISHERS

Instant Recovery with Write-Ahead Logging

Page Repair, System Restart, and Media Restore

Goetz Graefe
Wey Guy
Caetano Sauer

Synthesis Lectures on Data Management

Z. Meral Özsoyoğlu, *Series Editor*

Instant Recovery with Write-Ahead Logging

Page Repair, System Restart, and Media Restore

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.

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, and Media Restore
Goetz Graefe, Wey Guy, Caetano Sauer
December 2014

Similarity Joins in Relational Database Systems
Nikolaus Augsten, Michael H. Böhlen
November 2013

Information and Influence Propagation in Social Networks
Wei Chen, Laks V.S. Lakshmanan, Carlos Castillo
October 2013

Data Cleaning: A Practical Perspective
Venkatesh Ganti, Anish Das Sarma
September 2013

Data Processing on FPGAs
Jens Teubner , Louis Woods
June 2013

Perspectives on Business Intelligence

Raymond T. Ng , Patricia C. Arocena , Denilson Barbosa , Giuseppe Carenini , Luiz Gomes, Jr. , Stephan Jou , Rock Anthony Leung , Evangelos Milios , Renée J. Miller , John Mylopoulos , Rachel A. Pottinger , Frank Tompa , Eric Yu
April 2013

Semantics Empowered Web 3.0: Managing Enterprise, Social, Sensor, and Cloud-based Data and Services for Advanced Applications
Amit Sheth, Krishnaprasad Thirunarayan
December 2012

Data Management in the Cloud: Challenges and Opportunities
Divyakant Agrawal, Sudipto Das, Amr El Abbadi
December 2012

Query Processing over Uncertain Databases
Lei Chen, Xiang Lian
December 2012

Foundations of Data Quality Management
Wenfei Fan, Floris Geerts
July 2012

Incomplete Data and Data Dependencies in Relational Databases
Sergio Greco, Cristian Molinaro, Francesca Spezzano
July 2012

Business Processes: A Database Perspective
Daniel Deutch, 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, Clement T. Yu
June 2012

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

Query Answer Authentication

HweeHwa Pang, Kian-Lee Tan
February 2012

Declarative Networking
Boon Thau Loo, Wenchao Zhou
January 2012

Full-Text (Substring) Indexes in External Memory
Marina Barsky, Ulrike Stege, 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, Ramesh Jain
July 2011

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

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

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

Peer-to-Peer Data Management
Karl Aberer
May 2011

Probabilistic Ranking Techniques in Relational Databases
Ihab F. Ilyas, 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, Susan D. Urban
2010

Advanced Metasearch Engine Technology
Weiyi Meng, 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, Christian Thomsen
2010

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

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

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

Data Stream Management
Lukasz Golab , M. Tamer Özsu
2010

Access Control in Data Management Systems
Elena Ferrari
2010

An Introduction to Duplicate Detection
Felix Naumann, Melanie Herschel
2010

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

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

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, and Media Restore
Goetz Graefe, Wey Guy, Caetano Sauer
www.morganclaypool.com

ISBN: 9781627055543 print
ISBN: 9781627055550 ebook

DOI 10.2200/S00617ED1V01Y201411DTM039

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON DATA MANAGEMENT #39
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, and Media Restore

Goetz Graefe
HP Labs

Wey Guy

Caetano Sauer
University of Kaiserslautern

SYNTHESIS LECTURES ON DATA MANAGEMENT #39

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Traditional theory and practice of write-ahead logging and of database recovery techniques revolve around three failure classes: transaction failures resolved by rollback; system failures (typically 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 oppor-tuni-ties far beyond its 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 addition to these instant recovery techniques, the discussion introduces much faster offline restore operations without slowdown in backup operations and with hardly any slowdown in log archiving operations. The new restore techniques also render differential and incremental backups obsolete, complete backup commands on the database server practically instantly, and even permit taking full backups without imposing any load on the database server.

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

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.

In this first revision, the book describes techniques. At the time of this writing, software development efforts are underway and will yield functioning recovery techniques, deeper insights into implementation problems and solutions, and performance observations. We plan to include those in another revision.

Acknowledgments

Barb Peters and Arianna Lund encouraged combining all "instant recovery" techniques into a single article. Harumi Kuno participated in the research defining single-page failures and single-page recovery and later provided feedback on its applications.

CHAPTER 1

Introduction

Modern hardware differs from hardware of 25 years ago, when many of the database recovery techniques used today were designed. Current hardware includes high-capacity, high-density disks with single-page failures due to cross-track effects, e.g., in shingled or overlapping recording, semiconductor storage with single-page failures due to localized wear-out, large memory and large buffer pools with many pages and therefore many dirty pages and long restart recovery after system failures, and high-capacity storage devices and therefore long restore recovery after media failures.

For example,[1] some of today's servers have 1 TB (2^{40} B) of volatile memory, equal to over 100 million (2^{27}) pages of 8 KB (2^{13} B). If 3% (~2^{-5}) of these pages are dirty at the time of a system crash, "redo" recovery must inspect and recover several million (2^{22}) pages. Even if 16 (2^4) devices can each serve 250 (2^8) I/O operations per second, the "redo" phase alone of restart recovery takes about 15 minutes (2^{10} seconds). Fewer or slower devices or skew in the access pattern increase "redo" and "undo" times. In contrast, instant restart enables new transactions concurrently to "redo" and "undo" recovery, i.e., several minutes earlier. For a real-world example of the need for fast restart with large memory, some companies see themselves forced to invent special techniques even for clean shutdown and restart, specifically for software upgrade [GCG 14].

As another example, some of today's storage devices hold 4 TB (2^{42} B) of data, transfer data at 250 MB/s (2^{28} B/s), and support 250 (2^8) random I/O operations per second. Taking or restoring a full backup takes about 4½ hours (2^{14} seconds). Traditional media recovery starts with a full restore and then requires replaying the recovery log gathered since the last backup, often for many hours. In contrast, instant restore enables transaction processing during the entire restore operation, even permitting transactions to resume after a very short delay (a few seconds) rather than abort and eventually restart hours later.

This book covers techniques that seem more appropriate for contemporary hardware. It employs and builds on many proven techniques, in particular write-ahead logging, checkpoints, log archiving, and more. The foundations are two new ideas. First, single-page failures and single-page recovery [GK 12] enable incremental recovery fast enough to run on demand without imposing major delays in query and transaction processing. Second, log archiving not only compresses the log records with traditional techniques but also partially sorts the log archive, which enables multiple access patterns, all reasonably efficient. In other words, the contributions of "instant recovery" are

[1] This example, as well as the following one, relies on simple calculations and assumed system parameters. Their purpose is to illustrate orders of magnitude rather than precise values. Readers are welcome to repeat the calculations with alternative parameters or formulas more realistic and more accurate for their computer systems.

ubiquitous fine-grained on-demand recovery and a novel data organization of log archives that permits both efficient archiving, i.e., creation of the log archive, and efficient restore operations, i.e., usage of the log archive.[2]

These foundations are exploited for incremental recovery actions executing on demand, in particular after system failures (producing an impression of "instant restart") and after media failures ("instant restore"). In addition to incremental recovery, new techniques speed up offline backup and offline restore operations. In particular, full backups can be created efficiently without imposing any load on the active server process, differential and incremental backups become entirely obsolete, and restore times are reduced by the traditional times for restoring differential and incremental backups as well as by the time for log replay.

The problem of out-of-date recovery methods for today's hardware exists equally for file systems, databases, key-value stores, and contents indexes in information retrieval and internet search. Similarly, the techniques and solutions discussed below apply not only databases, even if they are often discussed using database terms, but also to file systems, key-value stores, and contents indexes. In other words, the problems, techniques, and solutions apply to practically all persistent digital storage.

Chapter 2 sketches the assumed system context and then reviews related prior work and its influence on instant recovery. Chapter 3 reviews the first step towards seemingly instantaneous recovery, i.e., single-page failures and single-page recovery, and then Chapter 4 introduces applications of single-page recovery after possibly deliberate introduction of single-page failures in the form of out-of-date page on persistent storage. Chapter 5 focuses on instant recovery after a system failure, i.e., restart after a software crash. Chapter 6 introduces new techniques for offline restore operations and Chapter 7 introduces applications of the new restore techniques, including "instant backup" techniques that prepare a full and current backup in seconds rather than hours. Chapter 8 introduces "instant restore" techniques for high-availability recovery from media failures. Chapter 9 considers multiple failure including media failures during system restart and system failures during media restore operations. Chapter 10 offers a summary, conclusions, and opportunities for future work.

[2] We use the term "instant" not in an absolute meaning but a relative one, i.e., in comparison to prior techniques. This is like instant coffee, which is not absolutely instantaneous but only relative to traditional techniques of coffee preparation. The reader's taste and opinion must decide whether instant coffee actually is coffee. Instant recovery, however, is true and reliable recovery from system and media failures, with guarantees as strong as those of traditional recovery techniques.