Hasso Plattner

A Course in In-Memory Data Management

The Inner Mechanics of In-Memory Databases

Springer

A Course in In-Memory Data Management

Hasso Plattner

A Course in In-Memory Data Management

The Inner Mechanics of In-Memory Databases

Springer

Hasso Plattner
Hasso Plattner Institute
Potsdam, Brandenburg
Germany

ISBN 978-3-642-36523-2 ISBN 978-3-642-36524-9 (eBook)
DOI 10.1007/978-3-642-36524-9
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013932332

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Why We Wrote This Book

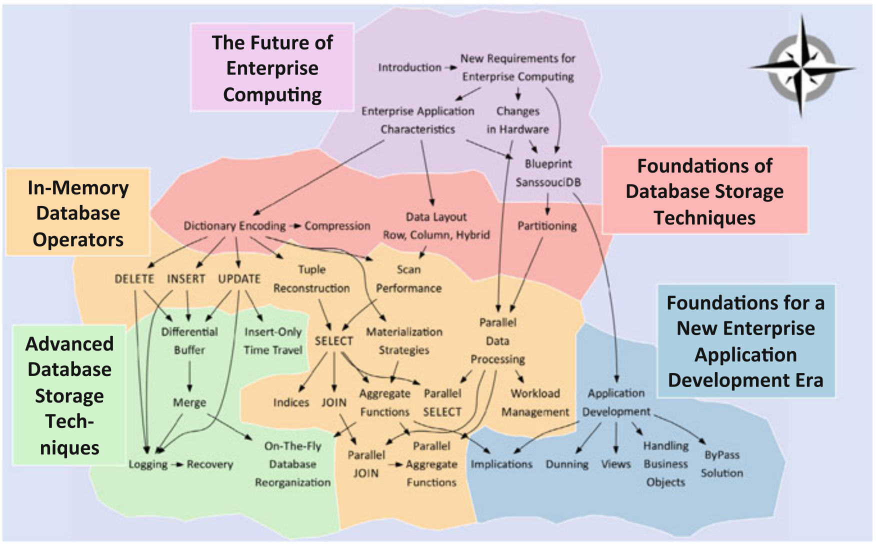
Our research group at the HPI has conducted research in the area of in-memory data management for enterprise applications since 2006. The ideas and concepts of a dictionary-encoded column-oriented in-memory database gained much traction due to the success of SAP HANA as the cutting-edge industry product and from followers trying to catch up. As this topic reached a broader audience, we felt the need for proper education in this area. This is of utmost importance as students and developers have to understand the underlying concepts and technology in order to make use of it.

At our institute, we have been teaching in-memory data management in a Master's course since 2009. When I learned about the current movement towards the direction of Massive Open Online Courses, I immediately decided that we should offer our course about in-memory data management to the public. On September 3, 2012 we started our online education with the new online platform http://www.openHPI.de. We granted 2,137 graded certificates to the 13,126 participating learners of the first iteration of the online course. Please feel free to register at openHPI.de to be informed about upcoming lectures.

Several thousand people have already used our material in order to study for the homework assignments and final exam of our online course. This book is based on the reading material that we provided to the online community. In addition to that, we incorporated many suggestions for improvement as well as self-test questions and explanations. As a result, we provide you with a textbook teaching you the inner mechanics of a dictionary-encoded column-oriented in-memory database.

Navigating the Chapters

When giving a lecture, content is typically taught in a one-dimensional sequence. You have the advantage that you can read the book according to your interests. To this end, we provide a learning map, which also reappears in the introduction to make sure that all readers notice it. The learning map shows all chapters of this book, also referred to as learning units, and shows which topics are prerequisites for which other topics. For example, the learning unit “Differential Buffer” (Chap. 25) is referred to relatively late in the book. Nevertheless, you might already read it earlier. The prerequisites are that you understood the concepts of how “DELETEs”, “INSERTs”, and “UPDATEs” are conducted without a differential buffer.


The Future of Enterprise Computing
Introduction
New Requirements for Enterprise Computing
Enterprise Application Characteristics
Changes in Hardware
Blueprint SanssouciDB
Foundations of Database Storage Techniques
In-Memory Database Operators
Dictionary Encoding
Compression
Data Layout Row, Column, Hybrid
Partitioning
DELETE
INSERT
UPDATE
Tuple Reconstruction
Scan Performance
Foundations for a New Enterprise Application Development Era
Advanced Database Storage Tech-niques
Differential Buffer
Insert-Only Time Travel
SELECT
Materialization Strategies
Parallel Data Processing
Merge
Indices
JOIN
Aggregate Functions
Parallel SELECT
Workload Management
Application Development
Logging
Recovery
On-The-Fly Database Reorganization
Parallel JOIN
Parallel Aggregate Functions
Implications
Dunning
Views
Handling Business Objects
ByPass Solution

The last section of each chapter contains self-test questions. You also find the questions including the solutions and explanations in Sect. 34.3.

The Development Process of the Book

I want to thank the team of our research chair “Enterprise Platform and Integration Concepts” at the Hasso Plattner Institute at the University of Potsdam in Germany. This book would not exist without this team.

Special thanks go to our online lecture core team consisting of *Ralf Teusner*, *Martin Grund*, *Anja Bog*, *Jens Krüger*, and *Jürgen Müller*.

During the preparation of the online lecture as well as during the online lecture itself, the whole research group took care that no email remained unanswered and all reported bugs in the learning material were fixed. Thus, I want to thank the research assistants *Martin Faust*, *Franziska Häger*, *Thomas Kowark*, *Martin Lorenz*, *Stephan Müller*, *Jan Schaffner*, *Matthieu Schapranow*, *David Schwalb*, *Christian Schwarz*, *Christian Tinnefeld*, *Arian Treffer*, *Johannes Wust*, as well as our team assistant *Andrea Lange* for their commitment.

During the development process, several HPI bachelor students (Frank Blechschmidt, Maximilian Grundke, Jan Lindemann, Lars Rückert) and HPI master students (Sten Ächtner, Martin Boissier, Ekaterina Gavrilova, Martin Köppelmann, Paul Möller, Michael Wolowyk) supported us during the online lecture preparations. Special thanks go to *Martin Boissier*, *Maximilian Grundke*, *Jan Lindemann*, and *Jasper Schulz*, who worked on all the corrections and adjustments that have to be made when teaching material is enhanced in order to print a book.

Help Improving This Book

We are continuously seeking to improve the learning material provided in this book. If you identify any flaws, please do not hesitate to contact me at hasso.plattner@hpi.uni-potsdam.de.

So far, we received bug reports that resulted in improvements in the learning material from the following attentive readers: Shakir Ahmed, Heiko Betzler, Christoph Birkenhauer, Jonas Bränzel, Dmitry Bondarenko, Christian Butzlaff, Peter Dell, Michael Dietz, Michael Max Eibl, Roman Ganopolskyi, Christoph Gilde, Hermann Grahm, Jan Grasshoff, Oliver Hahn, Ralf Hubert, Katja Huschle, Jens C. Ittel, Alfred Jockisch, Ashutosh Jog, Gerold Kasemir, Alexander Kirov, Jennifer Köenig, Stephan Lange, Francois-David Lessard, Verena Lommatsch, Clemens Müller, Hendrik Müller, Debanshu Mukherjee, Holger Pallak, Jelena Perfiljeva, Dieter Rieblinger, Sonja Ritter, Veronika Rodionova, Viacheslav Rodionov, Yannick Rödl, Oliver Roser, Alice-Rosalind Schell, Wolfgang Schill, Leo Schneider, Jürgen Seitz, David Siegel, Markus Steiner, Reinhold Thurner, Florian Tönjes, Wolfgang Weinmann, Bert Wunderlich, and Dieter Zürn.

We are thankful for any kind of feedback and hope that the learning material will be further improved by the in-memory database community.

Hasso Plattner

Contents

Abbreviations

ATP	Available-to-Promise
BI	Business Intelligence
ccNUMA	Cache-Coherent Non-Uniform Memory Architecture
CPU	Central Processing Unit
DML	Data Manipulation Language
DPL	Data Prefetch Logic
DRAM	Dynamic Random Access Memory
DW	Data Warehouse
e.g.	For example
EPIC	Enterprise Platform and Integration Concepts
ERP	Enterprise Resource Planning
et al.	And others
etc.	Et cetera
ETL	Extract Transform Load
FSB	Front Side Bus
HDD	Hard Disk Drive
HPI	Hasso-Plattner-Institut
HR	Human resources
i.e.	That is
IMC	Integrated Memory Controller
IMDB	In-Memory Database
IO	Index Offsets
IP	Index Positions
MDX	Multidimensional Expression
MIPS	Million Instructions Per Second
NUMA	Non-Uniform Memory Architecture
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
ORM	Object-Relational Mapping
PADD	Parallel Add
PDA	Personal Digital Assistant
QPI	Quick Path Interconnect
RAM	Random Access Memory

RISC	Reduced Instruction Set Computing
SADD	Scalar Add
SIMD	Single Instruction Multiple Data
SRAM	Static Random Access Memory
SSE	Streaming SIMD Extensions
TLB	Translation Lookaside Buffer
UMA	Uniform Memory Architecture

Figures

Tables

Chapter 1
Introduction

This book *A Course in In-Memory Data Management* focuses on the technical details of in-memory columnar databases. In-memory databases, and especially column-oriented databases, are a recently vastly researched topic [BMK09, KNF+, Pla09]. With modern hardware technologies and increasing main memory capacities, groundbreaking new applications are becoming viable.

1.1 Goals of the Lecture

Everybody who is interested in the future of databases and enterprise data management should benefit from this course, regardless whether one is still studying, already working, or perhaps even developing software in the affected fields. The primary goal of this course is to achieve a deep understanding of column-oriented, dictionary-encoded in-memory databases and the implications of those for enterprise applications. This learning material does not include introductions into Structured Query Language (SQL) or similar basics; these topics are expected to be prior knowledge. However, even if you do not yet have solid SQL knowledge, we encourage you to follow the course, since most examples with relation to SQL will be understandable from the context.

With new applications and upcoming hardware improvements, fundamental changes will take place in enterprise applications. The participants ought to understand the technical foundation of next generation database technologies and get a feeling for the difference between in-memory databases and traditional databases on disk. In particular, you will learn why and how these new technologies enable performance improvements by factors of up to 100,000.

1.2 The Idea

The foundation for the learning material is an idea that professor Hasso Plattner and his "Enterprise Platform and Integration Concepts" (EPIC) research group came up with in a discussion in 2006. At this time, lectures about Enterprise

DOI: 10.1007/978-3-642-36524-9_1, © Springer-Verlag Berlin Heidelberg 2013

Resource Planning (ERP) systems were rather dry with no intersections to modern technologies as used by Google, Twitter, Facebook, and several others.

The team decided to start a new radical approach for ERP systems. To start from scratch, the particular enabling technologies and possibilities of upcoming computer systems had to be identified. With this foundation, they designed a completely new system based on two major trends in hardware technologies:

- Massively parallel systems with an increasing number of Central Processing Units (CPUs) and CPU-cores
- Increasing main memory volumes

To leverage the parallelism of modern hardware, substantial changes had to be made. Current systems were already parallel in respective to their ability to handle thousands of concurrent users. However, the underlying applications were not exploiting parallelism.

Exploiting hardware parallelism is difficult. Hennessy et al. [PH12] discuss what changes have to be made to make an application run in parallel, and explain why it is often very hard to change sequential applications to use multiple cores efficiently.

For the first prototypes, the team decided to look more closely into accounting systems. In 2006, computers were not yet capable of keeping big companies' data completely in memory. So, the decision was made to concentrate on rather small companies in the first place. It was clear that the progress in hardware development would continue and that the advances will automatically enable the systems to keep bigger volumes of data in memory.

Another important design decision was the complete removal of materialized aggregates. In 2006, ERP systems were highly depending on pre-computed aggregates. With the computing power of upcoming systems, the new design was not only capable of increasing the granularity of aggregates, but of completely removing them.

As the new system keeps every bit of the processed information in memory, disks are only used for archiving, backup, and recovery. The primary persistence is the Dynamic Random Access Memory (DRAM), which is accomplished by increased capacities and data compression.

To evaluate the new approach, several bachelor projects and master projects implemented new applications using in-memory database technology over the next several years. Ongoing research focuses on the most promising findings of these projects as well as completely new approaches to enterprise computing with an enhanced user experience in mind.

1.3 Learning Map

The learning map (see Fig. 1.1) gives a brief overview over the parts of the learning material and the respective chapters in these parts. In this graph, you can easily see what the prerequisites for a chapter are and which contents will follow.

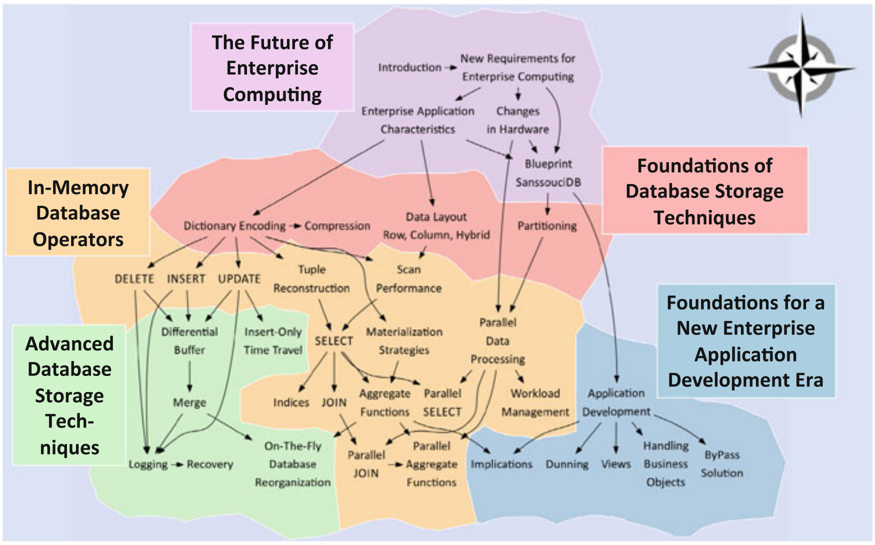

The Future of Enterprise Computing
Introduction
New Requirements for Enterprise Computing
Enterprise Application Characteristics
Changes in Hardware
Blueprint SanssouciDB
Foundations of Database Storage Techniques
In-Memory Database Operators
Dictionary Encoding
Compression
Data Layout Row, Column, Hybrid
Partitioning
DELETE
INSERT
UPDATE
Tuple Reconstruction
Scan Performance
Foundations for a New Enterprise Application Development Era
Advanced Database Storage Tech-niques
Differential Buffer
Insert-Only Time Travel
SELECT
Materialization Strategies
Parallel Data Processing
Merge
Indices
JOIN
Aggregate Functions
Parallel SELECT
Workload Management
Application Development
Logging
Recovery
On-The-Fly Database Reorganization
Parallel JOIN
Parallel Aggregate Functions
Implications
Dunning
Views
Handling Business Objects
ByPass Solution

Fig. 1.1 Learning map

1.4 Self Test Questions

1. **Rely on Disks**
   Does an in-memory database still rely on disks?
   (a) Yes, because disk is faster than main memory when doing complex calculations
   (b) No, data is kept in main memory only
   (c) Yes, because some operations can only be performed on disk
   (d) Yes, for archiving, backup, and recovery

References

[BMK09] P.A. Boncz, S. Manegold, M.L. Kersten, Database Architecture Evolution: Mammals Flourished long before Dinosaurs became Extinct. PVLDB **2**(2), 1648–1653 (2009)

[KNF+12] A. Kemper, T. Neumann, F. Funke, V. Leis, H. Mühe, Hyper: adapting columnar main-memory data management for transactional and query processing. IEEE Data Eng. Bull. **35**(1), 46–51 (2012)

[PH12] D.A. Patterson, J.L. Hennessy, in *Computer Organization and Design—The Hardware / Software Interface, (Revised 4th edn.)*. The Morgan Kaufmann Series in Computer Architecture and Design (Academic Press, San Francisco, CA, USA, 2012)

[Pla09] H. Plattner, in *A common database approach for OLTP and OLAP using an in-memory column database*, ed. by U. Çetintemel, S. Zdonik, D. Kossmann. SIGMOD Conference (ACM, Newyork, 2009), pp. 1–2

Part I
The Future of Enterprise Computing

Chapter 2
New Requirements for Enterprise Computing

When thinking about developing a completely new database management system for enterprise computing, the question whether there is a need for a new database management system arises. And the answer is yes! Modern companies have changed dramatically. Nowadays companies are more data-driven than ever before. For example, during manufacturing a much higher amount of data is produced, e.g. by assembly line sensors or manufacturing robots. Furthermore, companies process data at a much larger scale, e.g. competitor behavior, price trends, etc. to support management decisions. And data volumes will continue to grow in the future. There are two major requirements for a modern database management system:

- Data from various sources have to be combined in a single database management system, and
- This data has to be analyzed in real-time to support interactive decision taking.

The following sections outline use cases for modern enterprises and derive associated requirements for a completely new enterprise data management system.

2.1 Processing of Event Data

Event data influences enterprises today more and more. Event data is characterized by the following aspects:

- Each event dataset itself is small (some bytes or kilobytes) compared to the size of traditional enterprise data, such as all data contained in a single sales order, and
- The number of generated events for a specific entity is high compared to the amount of entities, e.g. hundreds or thousand events are generated for a single product.

In the following, use cases of event data in modern enterprises are outlined.

DOI: 10.1007/978-3-642-36524-9_2, © Springer-Verlag Berlin Heidelberg 2013

2.1.1 Sensor Data

Sensors are used to supervise the function of more and more systems today. One example is the tracking and tracing of sensitive goods, such as pharmaceuticals, clothes, or spare parts. Hereby packages are equipped with Radio-Frequency Identification (RFID) tags or two-dimensional bar codes, the so-called data matrix. Each product is virtually represented by an Electronic Product Code (EPC), which describes the manufacturer of a product, the product category, and a unique serial number. As a result, each product can be uniquely identified by its EPC code. In contrast, traditional one-dimensional bar codes can only be used for identification of classes of products due to their limited domain set. Once a product passes through a reader gate, a reading event is captured. The reading event consists of the current reading location, timestamp, the current business step, e.g. receiving, unpacking, repacking or shipping, and further related details. All events are stored in decentralized event repositories.

Real-Time Tracking of Pharmaceuticals

For example, approx. 15 billion prescription-based pharmaceuticals are produced in Europe. Tracking any of them results in approx. 8,000 read event notifications per second. These events build the basis for anti-counterfeiting techniques. For example, the route of a specific pharmaceutical can be reconstructed by analyzing all relevant reading events. The in-memory technology enables tracing of 10 billion events in less than 100 ms.

Formula One Racing Cars

Formula one racing cars are also generating excessive sensor data. These sports cars are equipped with up to 600 individual sensors, each recording tens to hundreds of events per second. Capturing sensor data for a 2 h race produces giga- or even terabytes of sensor data depending on their granularity. The challenge is to capture, process, and analyze the acquired data during the race to optimize the car parameters instantly, e.g. to detect part faults, optimize fuel consumption or top speed.

2.1.2 Analysis of Game Events

Personalized content in online games is a success factor for the gaming industry. The German company Bigpoint is a provider of browser games with more than 200 million active users.[1] Their browser games generate a steady stream of more than

[1] Bigpoint GmbH—http://www.bigpoint.net/

10,000 events per second, such as current level, virtual goods, time spent in the game, etc. Bigpoint tracks more than 800 million events per day. Traditional databases do not support processing of these huge amounts of data in an interactive way, e.g. join and full table scans require complex index structures or data warehouse systems optimized to return some selected aspects in a very fast way. However, individual and flexible queries from developers or marketing experts cannot be answered interactively.

Gamers tend to spend money when virtual goods or promotions are provided in a critical game state, e.g. a lost adventure or a long-running level that needs to be passed. In-game trade promotion management needs to analyze the user data, the current in-game events, and external details, e.g. current discount prices.

In-memory database technology is used to conduct in-game trade promotions and, at the same time, conduct A/B testing. To this end, the gamers are divided into two segments. The promotion is applied to one group. Since the feedback of the users is analyzed in real-time, the decision to roll-out a huge promotion can be taken within seconds after the small test group accepted the promotion.

Furthermore, in-memory technology improves discovery of target groups and testing of beta features, real-time prediction, and evaluation of advert placement.

2.2 Combination of Structured and Unstructured Data

Firstly, we want to understand structured data as any kind of data that is stored in a format, which is automatically processed by computers. Examples for structured data are ERP data stored in relational database tables, tree structures, arrays, etc. Secondly, we want to understand partially or mostly unstructured data, which cannot easily be processed automatically, e.g. all data that is available as raw documents, such as videos or photos. In addition, any kind of unformatted text, such as freely entered text in a text field, document, spreadsheet or database, is considered as unstructured data unless a data model for its interpretation is available, e.g. a possible semantic ontology.

For years, enterprise data management focused on structured data only. Structured data is stored in a relational database format using tables with specific attributes. However, many documents, papers, reports, web sites, etc. are only available in an unstructured format, e.g. text documents. Information within these documents is typically identified via the document's meta data. However, a detailed search within the content of these documents or the extraction of specific facts is not possible by using the meta data. As a result, there is a need to harvest information buried within unstructured enterprise data. Searching any kind of data—structured or unstructured—needs to be equally flexible and fast.

2.2.1 Patient Data

In the course of the patient treatment process, e.g. in hospitals, structured and unstructured data is generated. Examples of unstructured data are diagnosis reports, histologies, and tumor documentations. Examples of structured data are results of the erythrogram, blood pressure, temperature measurements, or the patient's gender. The in-memory technology enables the combination of both classes of patient data with additional external sources, such as clinical trials, pharmacological combinations or side-effects. As a result, physicians can prove their hypotheses by interactively combing data and reduce necessary manual and time-consuming searches. Physicians are able to access all relevant patient data and to take their decision on latest available patient details.

Due to their high fluctuation of unexpected events, such as emergencies or delayed surgeries, the daily time schedule of physicians is very time-optimized. In addition to certain technical requirements of their tools, they have also very strict response time requirements. For example, the HANA Oncolyzer, an application for physicians and researchers was designed for mobile devices. The mobile application supports the use-as-you-go factor, i.e., the required patient data is available at any location on the hospital campus and the physician is no longer forced to go to a certain desktop computer for checking a certain aspect. In addition, if the required detail is not available in real-time for the physician, she/he will no longer use the application. Thus, all analyses performed by the in-memory database are running on a server landscape in the IT department while the mobile application is the remote user interface for it.

Having the flexibility to request arbitrary analyses and getting the results within milliseconds back to the mobile application makes in-memory technology a perfect technology for the requirements of physicians. Furthermore, the mobility aspect bridges the gap between the IT department where the data are stored and the physician that visits multiple work places throughout the hospital every day.

2.2.2 Airplane Maintenance Reports

Airport maintenance logs are documented during exchange of any spare parts at Boeing. These reports contain structured data, such as date and time of the replacement or order number of the spare part, and unstructured data, e.g. kind of damage, location, and observations in the spacial context of the part. By combining structured and unstructured data, in-memory technology supports the detection of correlations, e.g. how often a specific part was replaced in a specific aircraft or location. As a result, maintenance managers are able to discover risks for damages before a certain risk for human-beings occurs.

2.3 Social Networks and the Web

Social networks are very popular today. Meanwhile, the time when they were only used to update friends about current activities are long gone. Nowadays, they are also used by enterprises for global branding, marketing and recruiting.

Additionally, they generate a huge amount of data, e.g. Twitter deals with one billion new tweets in five days. This data is analyzed, e.g. to detect messages about a new product, competitor activities, or to prevent service abuses. Combining social media data with external details, e.g. sales campaigns or seasonal weather details, market trends for certain products or product classes can be derived. These insights are valuable, e.g. for marketing campaigns or even to control the manufacturing rate.

Another example for extracting business relevant information from the Web is monitoring search terms. The search engine Google analyzes regional and global search trends. For example, searches for "influenza" and flu related terms can be interpreted as a indicator for a spread out of the influenza disease. By combining location data and search terms, Google is able to draw a map of regions that might be affected from an influenza epidemic.

2.4 Operating Cloud Environments

Operating software systems in the cloud requires a perfect data integration strategy. Assume, you process all your company's human resources (HR) tasks in an on-demand HR system provided by provider A. Consider a change of the provider to cloud provider B. Of course, a standardized data format for HR records can be used to export data from A and import it at B. However, what happens if there is no compatible standard for your application? Then, the data exported from A needs to be migrated, respectively remodeled, before it can imported by B. Data transformation is a complex and time-consuming task which often has to be done manually due to the required knowledge about source and target formats and many exceptions which have to be solved separately.

In-memory technology provides a transparent view concept. Views deta scribe how input values are transformed to the desired output format. The required transformations are performed automatically when the view is called. For example, consider the attributes `first name` and `last name` that need to be transformed into a single attribute `contact name`. A possible view `contact name` performs the concatenation of both attributes by performing `concat(first name, last name)`.

Thus, in-memory technology does not change the input data, while offering the required data formats by transparent processing of the view functions. This enables a transparent data integration compared to the traditional Extract Transform and Load (ETL) process used for (BI) systems.

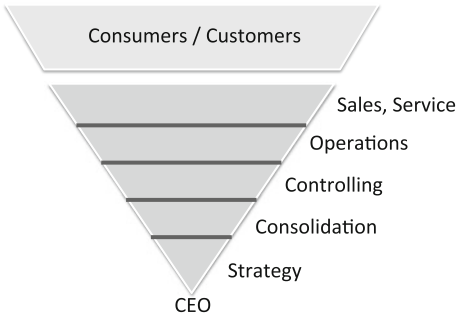


Fig. 2.1 Inversion of corporate structures

2.5 Mobile Applications

The wide-spread of mobile applications fundamentally changed the way enterprises process information. First BI systems were designed to provide detailed business insights for CEOs and controllers only. Nowadays, every employee is getting insights by the use of BI systems. However, for decades information retrieval was bound to stationary desktop computers. With the wide-spread of mobile devices, e.g. PDAs, smartphones, etc., even field workers are able to analyze sales reports or retrieve the latest sales funnel for a certain product or region.

Figure 2.1 depicts the new design of BI systems, which is no longer top-down but bottom-up. Modern BI systems provide all required information to sales representatives directly talking to customers. Thus, customers and sales representatives build the top of the pyramid.

In-memory databases build the foundation for this new corporate structure. On mobile devices, people are eager to get a response within a few seconds [Oul05, OTRK05, RO05]. With the ability to perform complex and freely formulated queries with sub-second responds, in-memory databases can revolutionize the way employees communicate with customers. An example of the radical improvements through in-memory databases is the dunning run. A traditional dunning process took 20 min on an average SAP system, but by rewriting the dunning run on in-memory technology it now takes less than 1 s.

2.6 Production and Distribution Planning

Two further prominent use cases for in-memory databases are complex and long-running processes such as production planning and availability checking.

2.6.1 Production Planning

Production planning identifies the current demand for certain products and consequently adjusts the production rate. It analyzes several indicators, such as the users' historic buying behavior, upcoming promotions, stock levels at manufacturers and whole-sellers. Production planning algorithms are complex due to required calculations, which are comparable to those found in BI systems. With an in-memory database, these calculations are now performed directly on latest transactional data. Thus, algorithms are more accurate with respect to current stock levels or production issues, allowing faster reactions to unexpected incidents.

2.6.2 Available to Promise Check

The Available-to-Promise (ATP) check validates the availability of certain goods. It analyzes whether the amount of sold and manufactured goods are in balance. With raising numbers of products and sold goods, the complexity of the check increases. In certain situations it can be advantageous to withdraw already agreed goods from certain customers and reschedule them to customers with a higher priority. ATP checks can also take additional data into account, e.g. fees for delayed or canceled deliveries or costs for express delivery if the manufacturer is not able to sent out all goods in time.

Due to the long processing time, ATP checks are executed on top of pre-aggregated totals, e.g. stock level aggregates per day. Using in-memory databases enables ATP checks to be performed on the latest data without using pre-aggregated totals. Thus, manufacturing and Scheduling rescheduling decisions can be taken on real-time data. Furthermore, removing aggregates simplifies the overall system architecture significantly, while adding flexibility.

2.7 Self Test Questions

1. **Compression Factor**
   What is the average compression factor for accounting data in an in-memory column-oriented database?

   (a) 100x
   (b) 10x
   (c) 50x
   (d) 5x

2. **Data explosion**
   Consider the formula 1 race car tracking example, with each race car having 512 sensors, each sensor records 32 events per second whereby each event is 64 byte in size.
   How much data is produced by a F1 team, if a team has two cars in the race and the race takes 2 h?
   For easier calculation, assume 1,000 byte = 1 kB, 1,000 kB = 1 MB, 1,000 MB = 1 GB.

   (a) 14 GB
   (b) 15.1 GB
   (c) 32 GB
   (d) 7.7 GB

References

[OTRK05] A. Oulasvirta, S. Tamminen, V. Roto, J. Kuorelahti, Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile hci, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '05* (ACM, New York, 2005), pp. 919–928

[Oul05] A. Oulasvirta, The fragmentation of attention in mobile interaction, and what to do with it. Interactions **12**(6), 16–18 (2005)

[RO05] V. Roto, A. Oulasvirta, Need for non-visual feedback with long response times in mobile hci, in *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, WWW '05* (ACM, New York, 2005), pp. 775–781

Chapter 3
Enterprise Application Characteristics

3.1 Diverse Applications

An enterprise data management system should be able to handle data coming from several different source types.

- **Transactional data** is coming from different applications, e.g. Enterprise Resource Planning (ERP) systems.
- The sources for **event processing and stream data** are machines and sensors, typically high volume systems.
- **Real-time analytics** usually leverage structured data for transactional reporting, classical analytics, planning, and simulation.
- Finally, **text analytics** is typically based on unstructured data coming from the web, social networks, log files, support systems, etc.

3.2 OLTP Versus OLAP

An enterprise data management system should be able to handle transactional and analytical query types, which differ in several dimensions. Typical queries for **Online Transaction Processing (OLTP)** can be the creation of sales orders, invoices, accounting data, the display of a sales order for a single customer, or the display of customer master data. **Online Analytical Processing (OLAP)** consists of analytical queries. Typical OLAP-style queries are dunning (payment reminder), cross selling (selling additional products or services to a customer), operational reporting, or analyzing history-based trends.

Because it has always been considered that these query types are significantly different, it was argued to split the data management system into two separate systems handling OLTP and OLAP queries separately. In the literature, it is claimed that OLTP workloads are write-intensive, whereas OLAP-workloads are read-only and that the two workloads rely on "Opposing Laws of Database Physics" [Fre95].

Yet, research in current enterprise systems showed that this statement is not true [KGZP10, KKG+11]. The main difference between systems that handle these query types is that OLTP systems handle more queries with a single select or queries that are highly selective returning only a few tuples, whereas OLAP systems calculate aggregations for only a few columns of a table, but for a large number of tuples.

For the synchronization of the analytical system with the transactional system(s), a cost-intensive ETL (Extract-Transform-Load) process is required. The ETL process takes a lot of time and is relatively complex, because all changes have to be extracted from the outside source or sources if there are several, data is transformed to fit analytical needs, and it is loaded into the target database.

3.3 Drawbacks of the Separation of OLAP from OLTP

While the separation of the database into two systems allows for specific workload optimizations in both systems, it also has a number of drawbacks:

- The OLAP system does not have the latest data, because the latency between the systems can range from minutes to hours, or even days.Consequently, many decisions have to rely on stale data instead of using the latest information.
- To achieve acceptable performance, OLAP systems work with predefined, materialized aggregates which reduce the query flexibility of the user.
- Data redundancy is high. Similar information is stored in both systems, just differently optimized.
- The schemas of the OLTP and OLAP systems are different, which introduces complexity for applications using both of them and for the ETL process synchronizing data between the systems.

3.4 The OLTP Versus OLAP Access Pattern Myth

The workload analysis of multiple real customer systems reveals that OLTP and OLAP systems are not as different as expected. For OLTP systems, the lookup rate is only 10 % higher than for OLAP systems. The number of inserts is a little higher on the OLTP side. However, the OLAP systems are also faced with inserts, as they have to permanently update their data. The next observation is that the number of updates in OLTP systems is not very high [KKG+11]. In the high-tech companies it is about 12 %. It means that about 88 % of all tuples saved in the transactional database are never updated. In other industry sectors, research showed even lower update rates, e.g., less than 1 % in banking and discrete manufacturing [KKG+11].

This fact leads to the assumption that updating as such or alternatively deleting the old tuple and inserting the new one and keeping track of changes in a "side note" like it is done in current systems is no longer necessary. Instead, changed or deleted tuples can be inserted with according time stamps or invalidation flags. The additional benefit of this **insert-only approach** is that the complete transactional data history and a tuple's life cycle are saved in the database automatically. More details about the insert-only approach will be provided in Chap. 26.

The further fact that workloads are not that different after all leads to the vision of reuniting the two systems and to combine OLTP and OLAP data in one system.

3.5 Combining OLTP and OLAP Data

The main benefit of the combination is that both, transactional and analytical queries can be executed on the same machine using the same set of data as a "single source of truth". ETL-processing becomes obsolete.

Using modern hardware, pre-computed aggregates and materialized views can be eliminated as data aggregation can be executed on-demand and views can be provided virtually. With the expected response time of analytical queries below one second, it is possible to do the analytical query processing on the transactional data directly anytime and anywhere. By dropping the pre-computation of aggregates and materialization of views, applications and data structures can be simplified, as management of aggregates and views (building, maintaining, and storing them) is not necessary any longer.

A mixed workload combines the characteristics of OLAP and OLTP workloads. The queries in the workload can have full row operations or retrieve only a small number of columns. Queries can be simple or complex, pre-determined or ad hoc. This includes analytical queries that now run on latest transactional data and are able to see the real-time changes.

3.6 Enterprise Data Characteristics

By analyzing enterprise data, special data characteristics were identified. Most interestingly, many attributes of a table are not used at all while table can be very wide. 55 % of columns are unused on average per company and tables with up to hundreds of columns exist. Many columns that are used have a low cardinality of values, i.e., there are very few distinct values. Further, in many columns NULL or default values are dominant, so the entropy (information containment) of these column is very low (near zero).

These characteristics facilitate the efficient use of compression techniques, resulting in lower memory consumption and better query performance as will be seen in later chapters.

3.7 Self Test Questions

1. **OLTP OLAP Separation Reasons**
   Why was OLAP separated from OLTP?

   (a) Due to performance problems
   (b) For archiving reasons; OLAP is more suitable for tape-archiving
   (c) Out of security concerns
   (d) Because some customers only wanted either OLTP or OLAP and did not want to pay for both.

References

[Fre95] C.D. French, "One size fits all" database architectures do not work for DSS. SIGMOD Rec. **24**(2), 449–450 (1995)

[KGZP10] J. Krueger, M. Grund, A. Zeier, H. Plattner, Enterprise application-specific data management, in *EDOC*, 2010, pp. 131–140

[KKG+11] J. Krueger, C. Kim, M. Grund, N. Satish, D. Schwalb, J. Chhugani, H. Plattner, P. Dubey, A. Zeier, Fast updates on read-optimized databases using multi-core CPUs, in *PVLDB*, 2011

Chapter 4
Changes in Hardware

This chapter deals with hardware and lays the foundations to understand how the changing hardware impacts software and application development and is partly taken from [SKP12].

In the early 2000s multi-core architectures were introduced, starting a trend introducing more and more parallelism. Today, a typical board has eight CPUs and 8–16 cores per CPU. So each one has between 64 and 128 cores. A board is a pizza-box sized server component and it is called blade or node in a multi-node system. Each of those blades offers a high level of parallel computing for a price of about $50,000.

Despite the introduction of massive parallelism, the disk totally dominated all thinking and performance optimizations not long ago. It was extremely slow, but necessary to store the data. Compared to the speed development of CPUs, the development of disk performance could not keep up. This resulted in a complete distortion of the whole model of working with databases and large amounts of data. Today, the large amounts of main memory available in servers initiate a shift from disk based systems to in-memory based systems. In-memory based systems keep the primary copy of their data in main memory.

4.1 Memory Cells

In early computer systems, the frequency of the CPU was the same as the frequency of the memory bus and register access was only slightly faster than memory access. However, CPU frequencies did heavily increase in the last years following Moore's Law[1] [Moo65], but frequencies of memory buses and latencies of memory chips did not grew with the same speed. As a result, memory access gets more expensive, as more CPU cycles are wasted while stalling for memory access. This development is not due to the fact that fast memory can not be built, it is an economical decision as memory which is as fast as current CPUs would be

[1] Moore's Law is the assumption that the number of transistors on integrated circuits doubles every 18–24 months. This assumption still holds till today.

orders of magnitude more expensive and would require extensive physical space on the boards. In general, memory designers have the choice between Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM).

SRAM cells are usually built out of six transistors (although variants with only four do exist but have disadvantages [MSMH08]) and can store a stable state as long as power is supplied. Accessing the stored state requires raising the word access line and the state is immediately available for reading.

In contrast, DRAM cells can be constructed using a much simpler structure consisting of only one transistor and a capacitor. The state of the memory cell is stored in the capacitor while the transistor is only used to guard the access to the capacitor. This design is more economical compared to SRAM. However, it introduces a couple of complications. First, the capacitor discharges over time and while reading the state of the memory cell. Therefore, today's systems refresh DRAM chips every 64 ms [CJDM01] and after every read of the cell in order to recharge the capacitor. During the refresh, no access to the state of the cell is possible. The charging and discharging of the capacitor takes time, which means that the current can not be detected immediately after requesting the stored state, therefore limiting the speed of DRAM cells.

In a nutshell, SRAM is fast but requires a lot of space whereas DRAM chips are slower but allow larger chips due to their simpler structure. For more details regarding the two types of RAM and their physical realization the interested reader is referred to [Dre07].

4.2 Memory Hierarchy

An underlying assumption of the memory hierarchy of modern computer systems is a principle known as *data locality* [HP03]. Temporal data locality indicates that data which is accessed is likely to be accessed again soon, whereas spatial data locality indicates that data which is stored together in memory is likely to be accessed together. These principles are leveraged by using caches, combining the best of both worlds by leveraging the fast access to SRAM chips and the sizes made possible by DRAM chips. Figure 4.1 shows a hierarchy of memory on the example of the Intel Nehalem architecture. Small and fast caches close to the CPUs built out of SRAM cells cache accesses to the slower main memory built out of DRAM cells. Therefore, the hierarchy consists of multiple levels with increasing storage sizes but decreasing speed. Each CPU core has its private L1 and L2 cache and one large L3 cache shared by the cores on one socket. Additionally, the cores on one socket have direct access to their local part of main memory through an Integrated Memory Controller (IMC). When accessing other parts than their local memory, the access is performed over a Quick Path Interconnect (QPI) controller coordinating the access to the remote memory.

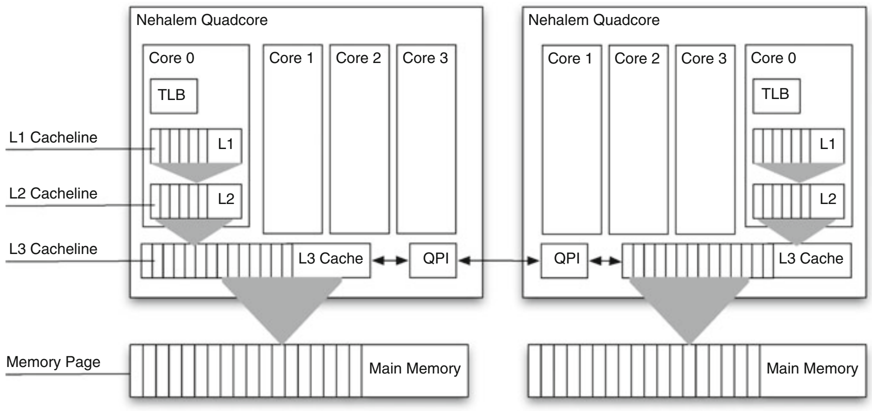


Fig. 4.1 Memory hierarchy on Intel Nehalem architecture

The first level are the actual registers inside the CPU, used to store inputs and outputs of the processed instructions. Processors usually only have a small amount of integer and floating point registers which can be accessed extremely fast. When working with parts of the main memory, their content has to be first loaded and stored in a register to make it accessible for the CPU. However, instead of accessing the main memory directly the content is first searched in the L1 cache. If it is not found in L1 cache it is requested from L2 cache. Some systems even make use of a L3 cache.

4.3 Cache Internals

Caches are organized in cache lines, which are the smallest addressable unit in the cache. If the requested content cannot be found in any cache, it is loaded from main memory and transferred down the hierarchy. The smallest transferable unit between each level is one *cache line*. Caches, where every cache line of level i is also present in level $i + 1$ are called *inclusive caches* otherwise the model is called *exclusive caches*. All Intel processors implement an inclusive cache model. This inclusive cache model is assumed for the rest of this text.

When requesting a cache line from the cache, the process of determining whether the requested line is already in the cache and locating where it is cached is crucial. Theoretically, it is possible to implement fully associative caches, where each cache line can cache any memory location. However, in practice this is only realizable for very small caches as a search over the complete cache is necessary when searching for a cache line. In order to reduce the search space, the concept of a *n-way set associative cache* with associativity A_i divides a cache with C_i bytes in $C_i/B_i/A_i$ sets and restricts the number of cache lines which can hold a copy of a certain memory

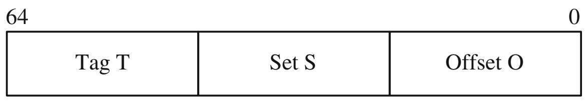


Fig. 4.2 Parts of a memory address

address to one set or A_i cache lines. Thus, when determining if a cache line is already present in the cache only one set with A_i cache lines has to be searched.

A requested address from main memory is split into three parts for determining if the address is already cached as shown by Fig. 4.2. The first part is the offset O, which size is determined by the cache line size of the cache. So with a cache line size of 64 bytes, the lower 6 bits of the address would be used as the offset into the cache line. The second part identifies the cache set. The number s of bits used to identify the cache set is determined by the cache size C_i, the cache line size $B_i$and the associativity A_i of the cache by $s = \log_2(C_i/B_i/A_i)$. The remaining $64 - o - s$ bits of the address are used as a tag to identify the cached copy. Therefore, when requesting an address from main memory, the processor can calculate S by masking the address and then search the respective cache set for the tag T. This can be easily done by comparing the tags of the A_i cache lines in the set in parallel.

4.4 Address Translation

The operating system provides each process a dedicated continuous address space, containing an address range from 0 to 2^x. This has several advantages as the process can address the memory through virtual addresses and does not have to bother about the physical fragmentation. Additionally, memory protection mechanisms can control the access to memory, restricting programs to access memory which was not allocated by them. Another advantage of virtual memory is the use of a paging mechanism which allows a process to use more memory than is physically available by paging pages in and out and saving them on secondary storage.

The continuous virtual address space of a process is divided into pages of size p, which is on most operating system 4 KB. Those virtual pages are mapped to physical memory. The mapping itself is saved in a so called page table, which resides in main memory itself. When the process accesses a virtual memory address, the address is translated by the operating system into a physical address with help of the memory management unit inside the processor.

We do not go into details of the translation and paging mechanisms. However, the address translation is usually done by a multi-level page table, where the virtual address is split into multiple parts which are used as an index into the page directories resulting in a physical address and a respective offset. As the page table is kept in main memory, each translation of a virtual address into a physical address would require additional main memory accesses or cache accesses in case the page table is cached.

In order to speed up the translation process, the computed values are cached in the so called , which is a small and fast cache. When accessing a virtual address, the respective tag for the memory page is calculated by masking the virtual address and the TLB is searched for the tag. In case the tag is found, the physical address can be retrieved from the cache. Otherwise, a TLB miss occurs and the physical address has to be calculated, which can be quite costly. Details about the address translation process, TLBs and paging structure caches for Intel 64 and IA-32 architectures can be found in [Int08].

The costs introduced by the address translation scale linearly with the width of the translated address [HP03, CJDM99], therefore making it hard or impossible to built large memories with very small latencies.

4.5 Prefetching

Modern processors try to guess which data will be accessed next and initiate loads before the data is accessed in order to reduce the incurring access latencies. Good prefetching can completely hide the latencies so that the data is already in the cache when accessed. However, if data is loaded which is not accessed later it can also evict data which would be accessed later and thereby induce additional misses by loading this data again. Processors support software and hardware prefetching. Software prefetching can be seen as a hint to the processor, indicating which addresses are accessed next. Hardware prefetching automatically recognizes access patterns by utilizing different prefetching strategies. The Intel Nehalem architecture contains two second level cache prefetchers—the L2 streamer and data prefetch logic (DPL) [Int11]. The prefetching mechanisms only work inside the page boundaries of 4 KB, in order to avoid triggering expensive TLB misses.

4.6 Memory Hierarchy and Latency Numbers

The memory hierarchy can be viewed as pyramid of storage mediums. The slower a medium is, the cheaper it gets. This also means that the amount of storage on the lower levels increases, because it is simply more affordable. The hierarchy levels of nowadays hardware are outlined by Fig. 4.1. This also means that the amount of storage offered by a lower medium can be higher, as outlined in Fig. 4.3.

At the very bottom is the hard disk. It is cheap, offers large amounts of storage and replaces tapes as the slowest storage medium necessary.

The next medium is flash. It is faster than disk, but it is still regarded as disk from a software perspective because of its persistence and its usage characteristics. This means that the same block oriented input and output methods which were developed more than 20 years ago for disks are still in place for flash. In order to

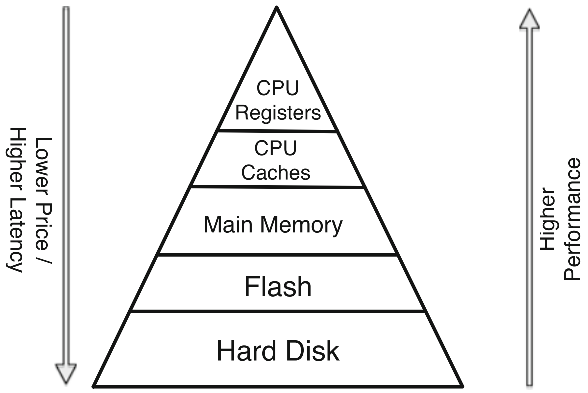


Fig. 4.3 Conceptual view of the memory hierarchy

fully utilize the speed of flash based storage the interfaces and drivers have to be adapted accordingly.

On top of flash is the main memory, which is directly accessible. The next level are the CPU caches—L3, L2, L1—with different characteristics. Finally, the top level of the memory hierarchy are the registers of the CPU where things like calculations are happening.

When accessing data from a disk, there are usually four layers between the accessed disk and the registers of the CPU which only transport information. In the end, every operation takes place inside the CPU and in turn the data has to be in the registers.

Table 4.1 shows some of the latencies, which come into play regarding the memory hierarchy. Latency is the time delay experienced by the system to load the data from the storage medium until it is available in a CPU register. The L1 cache latency is 0.5 ns. In contrast, accessing a main memory reference takes 100 ns and a simple disk seek is taking 10 ms.

In the end, there is nothing special about "in-memory" computing and all computing ever done was in memory, because it can only take place in the CPU.

Table 4.1 Latency numbers

Action	Time in nanoseconds (ns)	Time
L1 cache reference (cached data word)	0.5	
Branch mispredict	5	
L2 cache reference	7	
Mutex lock / unlock	25	
Main memory reference	100	0.1 µs
Send 2,000 byte over 1 Gb/s network	20,000	20 µs
SSD random read	150,000	150 µs
Read 1 MB sequentially from memory	250,000	250 µs
Disk seek	10,000,000	10 ms
Send packet CA to Netherlands to CA	150,000,000	150 ms

Assuming a bandwidth-bound application, the performance is determined by how fast the data can be transferred through the hierarchy to the CPU. In order to estimate the runtime of an algorithm, it is possible to roughly estimate the amount of data which has to be transferred to the CPU. A very simple operation that a CPU can do is a comparison like filtering for an attribute. Let us assume a calculation speed of 2 MB per millisecond for this operation using one core. So one core of a CPU can digest 2 MB per millisecond. This number scales with the amount of cores and if there are ten cores, they can scan 20 GB per second. If there are ten nodes with ten cores each, then that is already 200 GB in per second.

Considering a large multi-node system like that, having ten nodes and 40 CPUs per node where the data is distributed across the nodes, it is hard to write an algorithm which needs more than 1 s. This includes large amounts of data. The previously mentioned 200 GB are highly compressed data. So it is a much higher amount of plain character data. To sum this up, the number to remember is 2 MB per millisecond per core. If an algorithm shows a completely different result it is worth looking into it as there is probably something going wrong. This could be an issue in SQL, like a too complicated join or a loop in a loop.

4.7 Non-Uniform Memory Architecture

As the development in modern computer systems goes from multi-core to many-core systems and the amount of main memory continues to increase, the in Uniform Memory Architecture (UMA) systems becomes a bottleneck and introduces heavy challenges in hardware design to connect all cores and memory.

Non-Uniform Memory Architectures (NUMA) attempt to solve this problems by introducing local memory locations which are cheap to access for local processors. Figure 4.4 pictures an overview of an UMA and a NUMA system. In an UMA

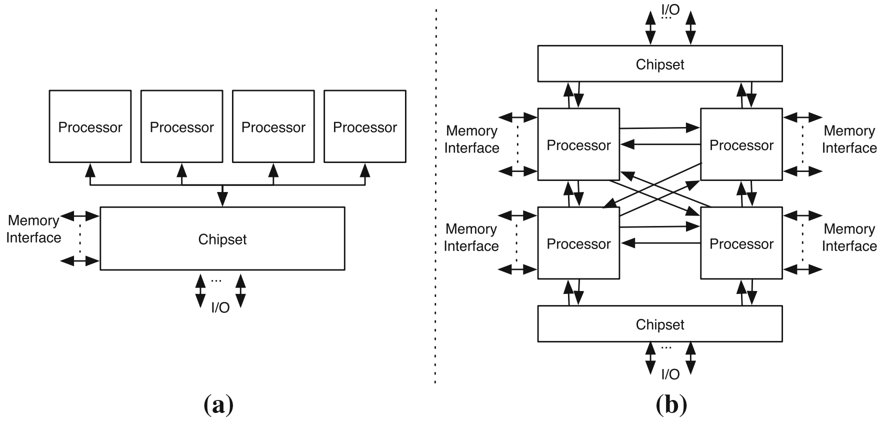


Fig. 4.4 (**a**) Shared FSB, (**b**) Intel quick path interconnect [Int09]

system every processor observes the same speeds when accessing an arbitrary memory address as the complete memory is accessed through a central memory interface as shown in Fig. 4.4a. In contrast, in NUMA systems, every processor has its primary used local memory as well as remote memory supplied from the other processors. This setup is shown in Fig. 4.4b. The different kinds of memory from the processors point of view introduce different memory access times between local memory (adjacent slots) and remote memory that is adjacent to the other processing units.

Additionally, systems can be classified into cache-coherent NUMA (ccNUMA) and non cache-coherent NUMA systems. ccNuma systems provide each CPU cache the same view to the complete memory and enforce coherency by a hardware implemented protocol. Non cache-coherent NUMA systems require software layers to handle memory conflicts accordingly. Although non ccNUMA hardware is easier and cheaper to build, most of todays available standard hardware provides ccNUMA, since non ccNUMA hardware is more difficult to program.

To fully exploit the potentials of NUMA, applications have to be made aware of the different memory latencies and should primarily load data from the locally attached memory slots of a processor. Memory-bound applications may suffer a degradation of up to 25 % of their maximal performance if remote memory is accessed instead of local memory.

By introducing NUMA, the central bottleneck of the FSB can be avoided and memory bandwidth can be increased. Benchmark results have shown that a throughput of more than 72 GB per second is possible on an Intel XEON 7560 (Nehalem EX) system with four processors [Fuj10].

4.8 Scaling Main Memory Systems

An example system that consists of multiple nodes can be seen in Fig. 4.5. One node has eight CPUs with eight cores, so each system has 64 cores, and there are four nodes. Each of them has a terabyte of RAM and SSDs for persistence. Everything which is below DRAM is for logging, archiving, and for emergency reconstruction of data, which means reloading the data after the power supply was turned off.

The networks which connect the nodes are continuously increasing in speed. In the example shown in Fig. 4.5, a 10 Gb Ethernet network connects the four nodes. Computers with 40 Gb Infiniband are already on the market and switch manufacturers are talking about 100 Gb switches which even have logic allowing smart switching. This is another location where an optimization can take place—on a low level and very effective for applications. It can be leveraged to improve joins, where calculations often go across multiple nodes.

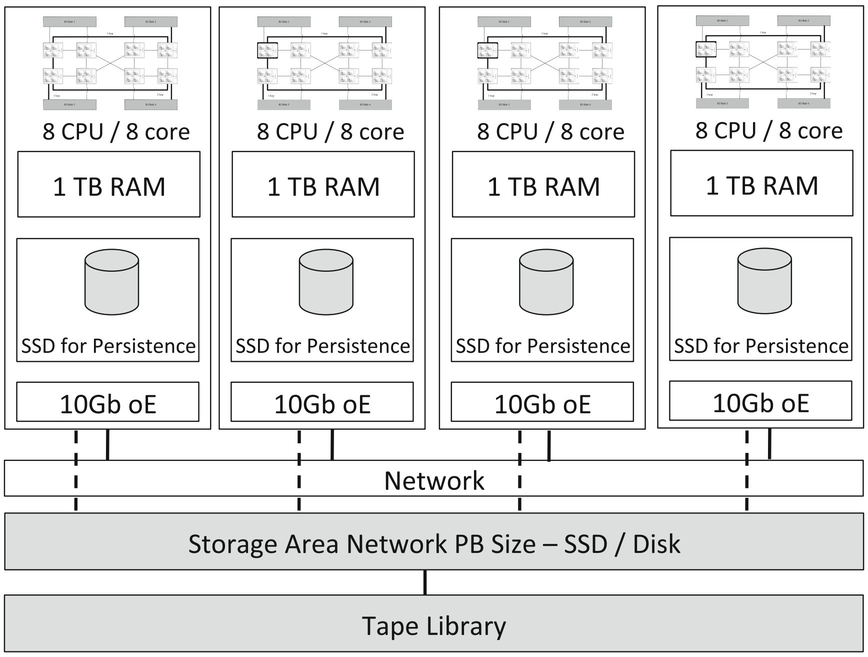


Fig. 4.5 A system consisting of multiple blades

4.9 Remote Direct Memory Access

Shared memory is another interesting way to directly access memory between multiple nodes. The nodes are connected with the network via Infiniband and create a shared memory region. The main idea is to automatically access data which is on a different node without explicitly shipping the data. In turn, there is direct access without shipping and processing it on the other side. Research has been done at Stanford University in cooperation with the HPI using a RAM cluster. It is very promising as it could basically offer direct access to a seemingly unlimited amount of memory from a program's perspective.

4.10 Self Test Questions

1. **Speed per Core**
   What is the speed of a single core when processing a simple scan operation (under optimal conditions)?

   (a) 2 GB/ms/core
   (b) 2 MB/ms/core

(c) 2 MB/s/core
(d) 200 MB/s/core

2. **Latency of Hard Disk and Main Memory**
Which statement concerning latency is wrong?

(a) The latency of main memory is about 100 ns
(b) A disk seek takes an average of 0.5 ms
(c) Accessing main memory is about 100,000 times faster than a disk seek
(d) 10 ms is a good estimation for a disk seek.

References

[CJDM99] V. Cuppu, B. Jacob, B. Davis, T. Mudge, A performance comparison of contemporary DRAM architectures, in *Proceedings of the 26th Annual International Symposium on Computer Architecture* (1999)

[CJDM01] V. Cuppu, B. Jacob, B. Davis, T. Mudge, High-performance DRAMs in workstation environments. IEEE Trans. Comput. **50**(11), 1133–1153 (2001)

[Dre07] U. Drepper, What every programmer should know about memory. http://people.redhat.com/drepper/cpumemory.pdf (2007)

[Fuj10] Fujitsu, Speicher-performance Xeon 7500 (Nehalem EX) basierter Systeme (2010)

[HP03] J. Hennessy, D. Patterson, *Computer Architecture: A Quantitative Approach* (Morgan Kaufmann, San Francisco, 2003)

[Int08] Intel Inc. TLBs, Paging-structure caches, and their invalidation (2008)

[Int09] Intel, An introduction to the Intel QuickPath Interconnect (2009)

[Int11] Intel Inc., Intel 64 and IA-32 architectures optimization reference manual (2011)

[Moo65] G. Moore, Cramming more components onto integrated circuits. Electronics **38**, 114 ff. (1965)

[MSMH08] A.A. Mazreah, M.R. Sahebi, M.T. Manzuri, S.J. Hosseini, A novel zero-aware four-transistor SRAM cell for high density and low power cache application, in *International Conference on Advanced Computer Theory and Engineering, ICACTE '08*, pp. 571–575 (2008)

[SKP12] D. Schwalb, J. Krueger, H. Plattner, Cache conscious column organization in in-memory column stores. Technical Report 60, Hasso-Plattner-Institute, December 2012

Chapter 5
A Blueprint of SanssouciDB

SanssouciDB is a prototypical database system for unified analytical and transactional processing. The concepts of SanssouciDB build on prototypes developed at the HPI and an existing SAP database system. SanssouciDB is an SQL database and it contains similar components as other databases such as a query builder, a plan executer, meta data, a transaction manager, etc.

5.1 Data Storage in Main Memory

In contrast to most other databases data is kept permanently in main memory. Main memory is the primary persistence for data, yet logging and recovery still need the disk as non-volatile data storage. All operators, e.g., find, join, or aggregation, can anticipate that data resides in main memory. Thus, operators can be programmed differently, free of any hassles coming from optimizing for disk access. Using main memory as the primary persistence leads to a different organization of data that only works if data is always available in memory. If this is the case, pointer arithmetic and following pointer sequences is all that is necessary to retrieve data.

5.2 Column-Orientation

Another concept used in SanssouciDB was invented more than two decades ago, that is, storing data column-wise [CK85] instead of row-wise. In column-orientation, complete columns are stored in adjacent blocks. This can be contrasted with row-oriented storage where complete tuples (rows) are stored in adjacent blocks. Column-oriented storage, in contrast to row-oriented storage, is well suited for reading consecutive entries from a single column. This can be useful for aggregation and column scans. More details on column-orientation and its differences to row-orientation can be found in Chap. 8. To minimize the amount of data that needs to be transferred between storage and processor, SanssouciDB uses several different data compression techniques, which will be discussed in Chap. 7.

5.3 Implications of Column-Orientation

Column-oriented storage has become widespread in database systems specifically developed for OLAP, as the advantage of column-oriented storage is clear in case of quasi-sequential scanning of single attributes and set processing thereof. If not all fields of a table are queried, column-orientation can be exploited as well in transactional processing (avoiding "SELECT *"). An analysis of enterprise applications showed that there is actually no application that uses all fields of a given tuple. For example, in dunning only 17 attributes are necessary out of a table that contains 300 attributes. If only the 17 needed attributes are queried instead of the full tuple representation of all 300 attributes, an instant advantage of factor eight to 20 for data to be scanned can be achieved.

As disk is not the bottleneck any longer, but access to main memory has to be considered, an important aspect is to work on a minimal set of data. So far, application programmers were fond of "SELECT *" statements. The difference in runtime between selecting specific fields or all fields in row-oriented storage is insignificant and in case changes to an application need more fields, the data was already there (which besides is a weak argument for using SELECT * and retrieving unnecessary data). However, in case of column-orientation, the penalty for "SELECT *" statements grows with table width. Especially if tables are growing in width during productive usage, actual runtimes of applications cannot be anticipated during programming.

With the column-store approach, the number of indices can be significantly reduced. In a column store, every attribute can be used as an index. Because all data is available in memory and the data of a column is stored consecutively, the scanning speed is high enough that a full sequential scan of an attribute is sufficient in most cases. If this is not fast enough, dedicated indices can still be used in addition for further speedup.

Storing data in columns instead of rows is more complicated for workloads with write access, so the concept of a differential store was introduced. New entries are written to a differential store first. In contrast to the main store, the differential store is optimized for inserts. At a later point in time and depending on thresholds, e.g. the frequency of changes and new entries, the data in the differential store is merged into the main store. More details about the differential buffer and the merge process will provided later in Chaps. 25 and 27.

5.4 Active and Passive Data

The data in SanssouciDB is separated into active data (data of business processes that are not yet completed) and passive data (data of business processes that are closed/completed and will not be changed any more). Active data is stored in main memory. Passive data can be moved to slower storage as it is queried less frequently. Separating passive data from active data reduces the amount of main memory needed to store the entire data set of an enterprise.

Whenever new data is written to the database or existing data is changed, logging to non-volatile storage is needed. During the merge of the differential store to the main store, snapshots are taken and stored in non-volatile memory, as well. Logs and snapshots are necessary to restore the database in case of failure.

The largest advantage so far is that main memory access depends on time-deterministic processes in contrast to seek-times of HDDs that depend on mechanical parts. Thus, runtimes of in-memory processing can be calculated (although it might be complicated). Observations from using in-memory databases show that response times are smooth—always the same—and not varying like it is the case with disks and their response time variations due to disk seeks.

5.5 Architecture Overview

The architecture shown in Fig. 5.1 grants an overview of the components of SanssouciDB.

SanssouciDB is split in three different logical layers fulfilling specific tasks inside the database system. The “Distribution Layer” handles the communication to applications, creates query execution plans, stores meta data contains the logic for database transactions. Inside the main memory of a specific machine the main working set of SanssouciDB is located. That working set is accessed during query execution and is stored either in row, column or hybrid-oriented data layout, depending on the specific type of queries sends to the database tables. The non-volatile memory is used for logging and recovery purposes, as well as for data aging and time travel.

All those concepts will be described in the subsequent sections.

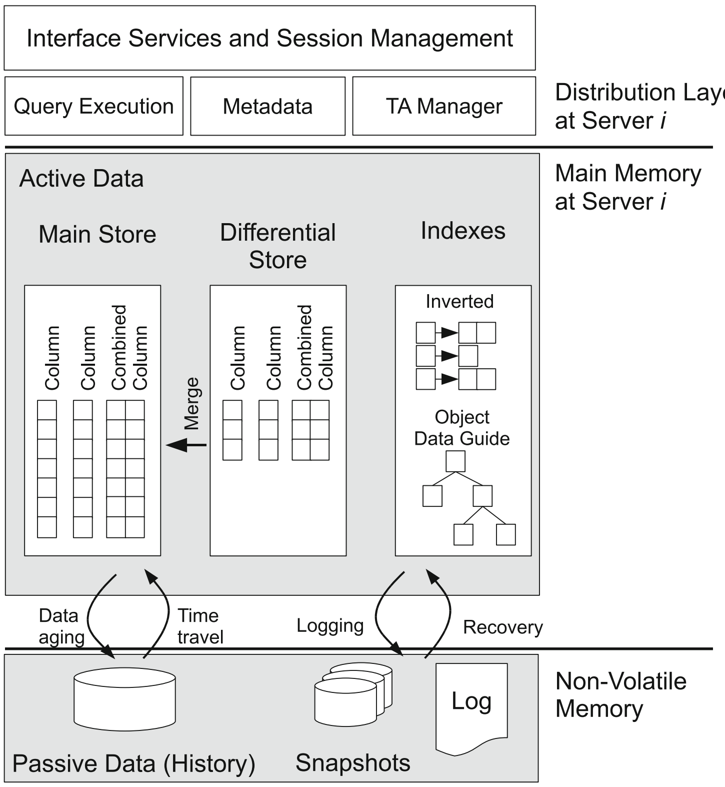


Fig. 5.1 Schematic architecture of SanssouciDB

5.6 Self Test Questions

1. **New Bottleneck**
   What is the new bottleneck of SanssouciDB that data access has to be optimized for?
   (a) Disk
   (b) The ETL process
   (c) Main memory
   (d) CPU

2. **Indexes**
   Can indexes still be used in SanssouciDB?
   (a) No, because every column can be used as an index
   (b) Yes, they can still be used to increase performance
   (c) Yes, but only because data is compressed
   (d) No, they are not even possible in columnar databases.

Reference

[CK85] G.P. Copeland, S.N. Khoshafian, A decomposition storage model. SIGMOD Rec. **14**(4), 268–279 (1985)

Part II
Foundations of Database Storage Techniques

Chapter 6
Dictionary Encoding

Since memory is the new bottleneck, it is required to minimize access to it. Accessing a smaller number of columns can do this on the one hand; so only required attributes are queried. On the other hand, decreasing the number of bits used for data representation can reduce both memory consumption and memory access times.

Dictionary encoding builds the basis for several other compression techniques (see Chap. 7) that might be applied on top of the encoded columns. The main effect of dictionary encoding is that long values, such as texts, are represented as short integer values.

Dictionary encoding is relatively simple. This means not only that it is easy to understand, but also it is easy to implement and does not have to rely on complex multilevel procedures, which would limit or lessen the performance gains. First, we will explain the general algorithm how original values are translated to integers using the example presented in Fig. 6.1.

Dictionary encoding works column-wise. In the example, every distinct value in the first name column "fname" is replaced by a distinct integer value. The position of a text value (e.g. Mary) in the dictionary is the representing number for that text (here: "24" for Mary). Until now, we have not saved any storage space. The benefits come to effect with values appearing more than once in a column. In our tiny example, the value "John" can be found twice in the column "fname", namely on position 39 and 42. Using dictionary encoding, the long text value (we assume 49 Byte per entry in the first name column) is represented by the short integer value (23 bit are needed to encode the 5 million different first names we assume to exist in the world). The more often identical values appear, the greater the benefits. As we noted in Sect. 3.6, enterprise data has low entropy. For this, dictionary encoding is well suited and grants a good compression ratio. A calculation for the complete first name and gender columns in our world-population example will exemplify the effects.

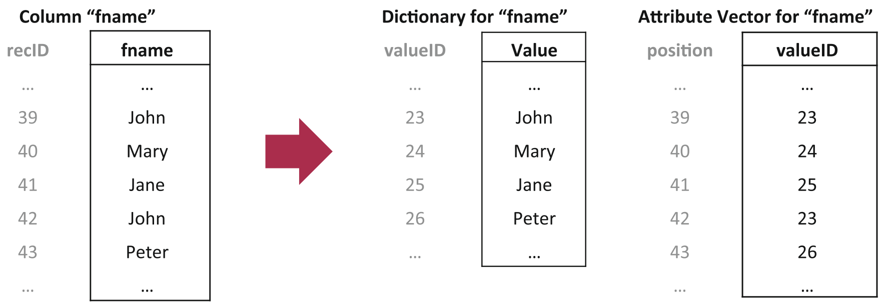


Fig. 6.1 Dictionary encoding example

6.1 Compression Example

Given is the *world population* table with 8 billion rows, 200 Byte per row:

Attribute	# of Distinct Values	Size
First name	5 million	49 Byte
Last name	8 million	50 Byte
Gender	2	1 Byte
Country	200	49 Byte
City	1 million	49 Byte
Birthday	40 000	2 Byte
	Sum	200 Byte

The complete amount of data is:

$$8\,\text{billion}\,\textit{rows} \cdot 200\,\text{Byte}\,\textit{per row} = 1.6\,\text{TB}$$

Each column is split into a dictionary and an attribute vector. Each dictionary stores all distinct values along with their implicit positions, i.e. valueIDs.

In a dictionary-encoded column, the attribute vectors now only store valueIDs, which correspond to the valueIDs in the dictionary. The recordID (row number) is stored implicitly via the position of an entry in the attribute vector. To sum up, via dictionary encoding, all information can be stored as integers instead of other, usually larger, data types.

6.1.1 Dictionary Encoding Example: First Names

How many bits are required to represent all 5 million distinct values of the first name column “fname”?

$$\lceil log_2(5,000,000) \rceil = 23$$

Therefore, 23 bits are enough to represent all distinct values for the required column. Instead of using

$$8\,\text{billion} \cdot 49\,\text{Byte} = 392\,\text{billion}\,\text{Byte} = 365.1\,\text{GB}$$

for the first name column, the attribute vector itself can be reduced to the size of

$$8\,\text{billion} \cdot 23\,\text{bit} = 184\,\text{billion}\,\text{bit} = 23\,\text{billion}\,\text{Byte} = 21.4\,\text{GB}$$

and an additional dictionary is introduced, which needs

$$49\,\text{Byte} \cdot 5\,\text{million} = 245\,\text{million}\,\text{Byte} = 0.23\,\text{GB}.$$

The achieved compression factor can be calculated as follows:

$$\frac{\textit{uncompressed size}}{\textit{compressed size}} = \frac{365.1\,\text{GB}}{21.4\,\text{GB} + 0.23\,\text{GB}} \approx 17$$

That means we reduced the column size by a factor of 17 and the result only consumes about 6 % of the initial amount of main memory.

6.1.2 Dictionary Encoding Example: Gender

Let us look on another example with the gender column. It has only 2 distinct values. For the gender representation without compression for each value (“m” or “f”) 1 Byte is required. So, the amount of data without compression is:

$$8\,\text{billion} \cdot 1\,\text{Byte} = 7.45\,\text{GB}$$

If compression is used, then 1 bit is enough to represent the same information. The attribute vector takes:

$$8\,\text{billion} \cdot 1\,\text{bit} = 8\,\text{billion}\,\text{bit} = 0.93\,\text{GB}$$

The dictionary needs additionally:

$$2 \cdot 1\,\text{Byte} = 2\,\text{Byte}$$

This concludes to a compression factor of:

$$\frac{uncompressed\ size}{compressed\ size} = \frac{7.45\text{GB}}{0.93\,\text{GB} + 2\,\text{Byte}} \approx 8$$

The compression rate depends on the size of the initial data type as well as on the column's entropy, which is determined by two cardinalities:

- Column cardinality, which is defined as the number of distinct values in a column, and
- Table cardinality, which is the total number of rows in the table or column

Entropy is a measure which shows how much information is contained in a column. It is calculated as

$$entropy = \frac{column\,cardinality}{table\,cardinality}$$

The smaller the entropy of the column, the better the achievable compression rate.

6.2 Sorted Dictionaries

The benefits of dictionary encoding can be further enhanced if sorting is applied to the dictionary. Retrieving a value from a sorted dictionary speeds up the lookup process from $O(n)$, which means a full scan through the dictionary, to $O(log(n))$, because values in the dictionary can be found using binary search. Sadly, this optimization comes at a cost: Every time a new value is added to the dictionary which does not belong at the end of the sorted sequence of the existing values, the dictionary has to be re-sorted. Even the insertion of only one value somewhere except the end of the dictionary causes a re-sorting, since the position of already present values behind the inserted value has to be moved one position up. While sorting the dictionary is not that costly, updating the corresponding attribute vector is. In our example, about 8 billion values have to be checked or updated if a new first name is added to the dictionary.

6.3 Operations on Encoded Values

The first and most important effect of dictionary encoding is that all operations concerning the table data are now done via attribute vectors, which solely consist of integers. This causes an implicit speedup of all operations, since a CPU is designed to perform operations on numbers, not on characters. When explaining dictionary encoding, a question often asked is: "But isn't the process of looking up all values via an additional data structure more costly than the actual savings? We understand the benefits concerning main memory, but what about the processor"—First, it has to be stated that the question is deemed appropriate. The processor has to take additional load, but this is acceptable, given the fact that our bottleneck is memory and bandwidth, so a slight shift of pressure in the direction of the processor is not only accepted but also welcome. Second, the impact of retrieving the actual values for the encoded columns is actually rather small. When selecting tuples, only the corresponding values from the query have to be looked up in the dictionary for the column scan. Generally, the result set is small compared to the total table size, so the lookup of all other selected columns to materialize the query result is not that expensive. Carefully written queries also only select those columns that are really needed, which not only saves bandwidth but also further reduces the number of necessary lookups. Finally, several operations such as COUNT or NOT NULL can even be performed without retrieving the real values at all.

6.4 Self Test Questions

1. **Lossless Compression**
   For a column with few distinct values, how can dictionary encoding significantly reduce the required amount of memory without any loss of information?
   (a) By mapping values to integers using the smallest number of bits possible to represent the given number of distinct values
   (b) By converting everything into full text values. This allows for better compression techniques, because all values share the same data format.
   (c) By saving only every second value
   (d) By saving consecutive occurrences of the same value only once
2. **Compression Factor on Whole Table**
   Given a population table (50 millions rows) with the following columns:
   - name (49 bytes, 20, 000 distinct values)
   - surname (49 bytes, 100, 000 distinct values)
   - age (1 byte,128 distinct values)
   - gender (1 byte, 2 distinct values)

   What is the compression factor (uncompressed size/compressed size) when applying dictionary encoding?
   (a) ≈ 20
   (b) ≈ 90
   (c) ≈ 10
   (d) ≈ 5

3. **Information in the Dictionary**
   What information is saved in a dictionary in the context of dictionary encoding?

   (a) Cardinality of a value
   (b) All distinct values
   (c) Hash of a value of all distinct values
   (d) Size of a value in bytes

4. **Advantages through Dictionary Encoding**
   What is an advantage of dictionary encoding?

   (a) Sequentially writing data to the database is sped up
   (b) Aggregate functions are sped up
   (c) Raw data transfer speed between application and database server is increased
   (d) INSERT operations are simplified

5. **Entropy**
   What is entropy?

   (a) Entropy limits the amount of entries that can be inserted into a database. System specifications greatly affect this key indicator.
   (b) Entropy represents the amount of information in a given dataset. It can be calculated as the number of distinct values in a column (column cardinality) divided by the number of rows of the table (table cardinality).
   (c) Entropy determines tuple lifetime. It is calculated as the number of duplicates divided by the number of distinct values in a column (column cardinality).
   (d) Entropy limits the attribute sizes. It is calculated as the size of a value in bits divided by number of distinct values in a column the number of distinct values in a column (column cardinality).

Chapter 7
Compression

As discussed in Chap. 5, SanssouciDB is a database architecture designed to run transactional and analytical workloads in enterprise computing. The underlying data set can easily reach a size of several terabytes in large companies. Although memory capacities of commodity servers are growing, it is still expensive to process those huge data sets entirely in main memory. Therefore, SanssouciDB and most modern in-memory storage engines use compression techniques on top of the initial dictionary encoding to decrease the total memory requirements. The columnar storage of data, as applied in SanssouciDB, is well suited for compression techniques, as data of the same type and domain is stored consecutively.

Another advantage of compression is that it reduces the amount of data that needs to be shipped between main memory and CPUs, thereby increasing the performance of query execution. We discuss this in more detail in Chap. 16 on materialization strategies.

This chapter introduces several lightweight compression techniques, which provide a good trade-off between compression rate and additional CPU-cycles needed for encoding and decoding. There are also a large number of so-called heavyweight compression techniques. They achieve much higher compression rates, but encoding and decoding is prohibitively expensive for their usage in our context. An in-depth discussion of many compression techniques can be found in [AMF06].

7.1 Prefix Encoding

In real-world databases, we often find the case that a column contains one predominant value and the remaining values have low redundancy. In this case, we would store the same value very often in an uncompressed format. Prefix encoding is the simplest way to handle this case more efficiently. To apply prefix encoding, the data sets need to be sorted by the column with the predominant value and the attribute vector has to start with the predominant value.

To compress the column, the predominant value should not be stored explicitly every time it occurs. This is achieved by saving the number of occurrences of the predominant value and one instance of the value itself in the attribute vector. Thus, a prefix-encoded attribute vector contains the following information:

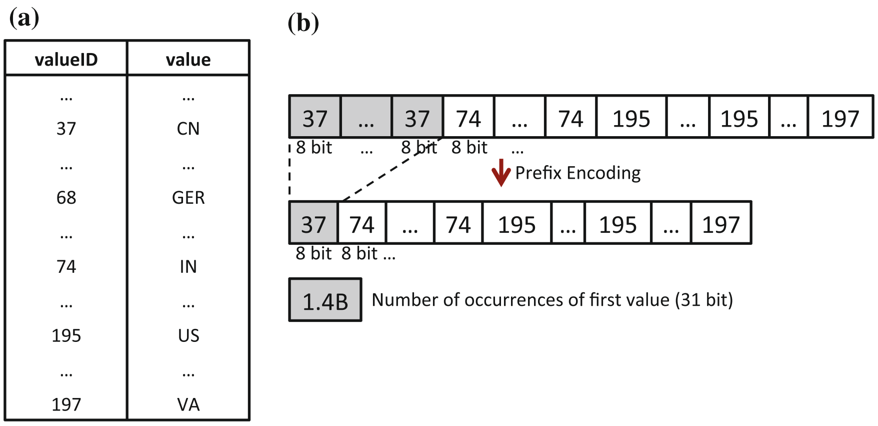
- number of occurrences of the predominant value
- valueID of the predominant value from the dictionary
- valueIDs of the remaining values.

7.1.1 Example

Given is the attribute vector of the country column from the world population table, which is sorted by population of countries in descending order. Thus, the 1.4 billion Chinese citizens are listed at first, then Indian citizens and so on. The valueID for China, which is situated at position 37 in the dictionary (see Fig. 7.1a), is stored 1.4 billion times at the beginning of the attribute vector in uncompressed format. In compressed format, the valueID 37 will be written only once, followed by the remaining valueIDs for the other countries as before. The number of occurrences "1.4 billion" for China will be stored explicitly. Figure 7.1b depicts examples of the uncompressed and compressed attribute vectors.

The following calculation illustrates the compression rate. First of all the number of bits required to store all 200 countries is calculated as $log_2(200)$ which results in 8 bit.

Without compression the attribute vector stores the 8 bit for each valueID 8 billion times:



(a)

valueID	value
...	...
37	CN
...	...
68	GER
...	...
74	IN
...	...
195	US
...	...
197	VA

Fig. 7.1 Prefix encoding example. (**a**) Dictionary. (**b**) Dictionary-encoded attribute vector (*top*) and prefix-encoded dictionary-encoded attribute vector (*bottom*)

$$8\,\text{billion} \cdot 8\,\text{bit} = 8\,\text{billion Byte} = 7.45\,\text{GB}$$

If the country column is prefix-encoded, the valueID for China is stored only once in 8 bit instead of 1.4 billion times 8 bit. An additional 31 bit field is added to store the number of occurrences ($\lceil log_2\ (1.4\,\text{billion})\ \rceil = 31\,\text{bit}$). Consequently, instead of storing 1.4 billion times 8 bit, only $31\,\text{bit} + 8\,\text{bit} = 39\,\text{bit}$ are really necessary. The complete storage space for the compressed attribute vector is now:

$$(8\,\text{billion} - 1.4\,\text{billion}) \cdot 8\,\text{bit} + 31\,\text{bit} + 8\,\text{bit} = 6.15\,\text{GB}$$

Thus, 1.3 GB, i.e., 17 % of storage space is saved. Another advantage of prefix encoding is direct access with row number calculation. For example, to find all male Chinese the database engine can determine that only tuples with row numbers from 1 until 1.4 billion should be considered and then filtered by the gender value.

Although we see that we have reduced the required amount of main memory, it is evident that we still store much redundant information for all other countries. Therefore, we introduce run-length encoding in the next section.

7.2 Run-Length Encoding

Run-length encoding is a compression technique that works best if the attribute vector consists of few distinct values with a large number of occurrences. For maximum compression rates, the column needs to be sorted, so that all the same values are located together. In run-length encoding, value sequences with the same value are replaced with a single instance of the value and

(a) either its number of occurrences or
(b) its starting position as offsets.

Figure 7.2 provides an example of run-length encoding using the starting positions as offsets. Storing the starting position speeds up access. The address of a specific value can be read in the column directly instead of computing it from the beginning of the column, thus, providing direct access.

7.2.1 *Example*

Applied to our example of the country column sorted by population, instead of storing all 8 billion values (7.45 GB), we store two vectors:

- one with all distinct values: 200 times 8 bit

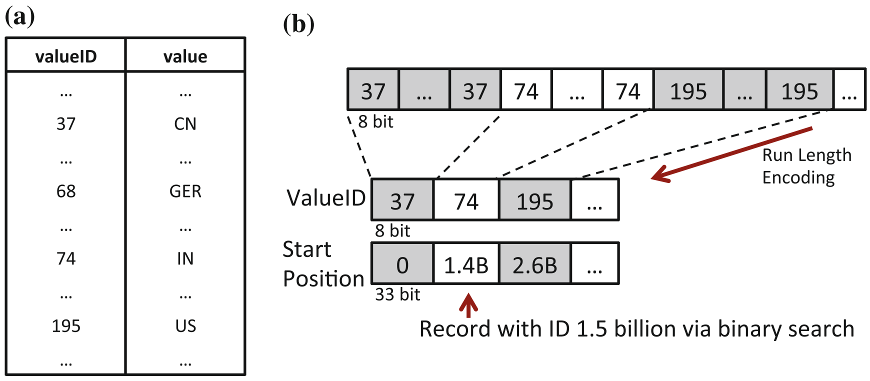


Fig. 7.2 Run-length encoding example. (**a**) Dictionary. (**b**) Dictionary-encoded attribute vector (*top*) and compressed dictionary-encoded attribute vector (*bottom*)

- the other with starting positions: 200 times 33 bit with 33 bit necessary to store the offsets up to 8 billion ($\lceil log_2\ (8\,\text{billion})\rceil = 33\,\text{bit}$). An additional 33 bit field at the end of this vector stores the number of occurrences for the last value.

Hence, the size of the attribute vector can be significantly reduced to approximately 1 KB without any loss of information:

$$200 \cdot (33\,\text{bit} + 8\,\text{bit}) + 33\,\text{bit} \approx 1\,\text{KB}$$

If the number of occurrences is stored in the second vector, one field of 33 bit can be saved with the disadvantage of losing the direct access possibility via binary search. Losing direct access results in longer response times, which is no option for enterprise data management.

7.3 Cluster Encoding

Cluster encoding works on equal-sized blocks of a column. The attribute vector is partitioned into N blocks of fixed size (typically 1024 elements). If a cluster contains only a single value, it is replaced by a single occurrence of this value. Otherwise, the cluster remains uncompressed. An additional bit vector of length N indicates which blocks have been replaced by a single value (1 if replaced, 0 otherwise). For a given row, the index of the corresponding block is calculated by integer division of the row number and the block size N. Figure 7.3 depicts an example for cluster encoding with the uncompressed attribute vector on the top and the compressed attribute vector on the bottom. Here, the blocks only contain four elements for simplicity.

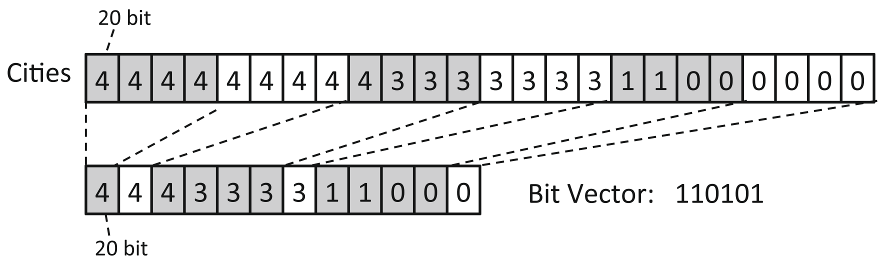


Fig. 7.3 Cluster encoding example

7.3.1 Example

Given is the city column (1 million different cities) from the world population table. The whole table is sorted by country and city. Hence, cities, which belong to the same country, are stored next to each other. Consequently, the occurrences of the same city values are stored next to each other, as well. 20 bit are needed to represent 1 million city valueIDs ($\lceil log_2$ (1 million) $\rceil = 20$ bit). Without compression, the city attribute vector requires 18.6 GB (8 billion times 20 bit).

Now, we compute the size of the compressed attribute vector illustrated in Fig. 7.3. With a cluster size of 1024 elements the number of blocks is 7.8 million ($\frac{8\,billion\,rows}{1024\,elements\,per\,block}$). In the worst case every city has 1 incompressible block. Thus, the size of the compressed attribute vector is computed from the following sizes:

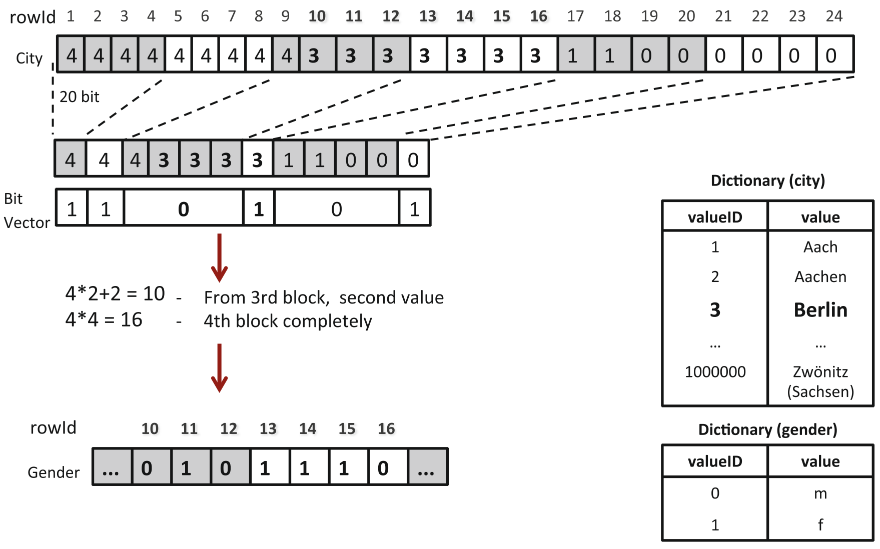
$$\begin{aligned}
&incompressible\ blocks + compressible\ blocks + \text{bit}\ vector\\
&= 1\,\text{million} \cdot 1024 \cdot 20\,\text{bit} + (7.8 - 1)\,\text{million} \cdot 20\,\text{bit} + 7.8\,\text{million} \cdot 1\,\text{bit}\\
&= 2441\,\text{MB} + 16\,\text{MB} + 1\,\text{MB}\\
&\approx 2.4\,\text{GB}
\end{aligned}$$

With a resulting size of 2.4 GB, a compression rate of 87 % (16.2 GB less space required) can be achieved.

Cluster encoding does not support direct access to records. The position of a record needs to be computed via the bit vector. As an example, consider the query that counts how many men and women live in Berlin (for simplicity, we assume that only one city with the name "Berlin" exists and the table is sorted by city):

```
SELECT gender, COUNT(gender)
FROM world_population
WHERE city = 'Berlin'
GROUP BY gender;
```

To find the recordIDs for the result set, we look up the valueID for "Berlin" in the dictionary. In our example, illustrated in Fig. 7.4, this valueID is 3. Then, we scan the cluster-encoded city attribute vector for the first appearance of valueID 3.



rowId 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
City 4 4 4 4 4 4 4 4 4 3 3 3 3 3 3 3 1 1 0 0 0 0 0 0
20 bit
4 4 4 3 3 3 3 1 1 0 0 0
Bit Vector 1 1 0 1 0 1
4*2+2 = 10 - From 3rd block, second value
4*4 = 16 - 4th block completely
rowId 10 11 12 13 14 15 16
Gender ... 0 1 0 1 1 1 0 ...
Dictionary (city)
valueID value
1 Aach
2 Aachen
3 Berlin
... ...
1000000 Zwönitz (Sachsen)
Dictionary (gender)
valueID value
0 m
1 f

Fig. 7.4 Cluster encoding example: no direct access possible

While scanning the cluster-encoded vector, we need to maintain the corresponding position in the bit vector, as each position in the vector is mapped to either one value (if the cluster is compressed) or four values (if the cluster is uncompressed) of the cluster-encoded city attribute vector. In Fig. 7.4, this is illustrated by stretching the bit vector to the corresponding value or values of the cluster-encoded attribute vector. After the position is found, a bit vector lookup is needed to check whether the block(s) containing this valueID are compressed or not to determine the recordID range containing the value "Berlin". In our example, the first block containing "Berlin" is uncompressed and the second one is compressed. Thus, we need to analyze the first uncompressed block to find the first occurrence of valueID 3, which is the second position, and can calculate the range of recordIDs with valueID 3, in our example 10 to 16. Having determined the recordIDs that match the city predicate, we can use these recordID to access the corresponding gender records and aggregate according to the gender values.

7.4 Indirect Encoding

Similar to cluster encoding, indirect encoding operates on blocks of data with N elements (typically 1024). Indirect Encoding can be applied efficiently if data blocks hold a few numbers of distinct values. It is often the case if a table is sorted by another column and a correlation between these two columns exists (e.g., name column if table is sorted by countries).

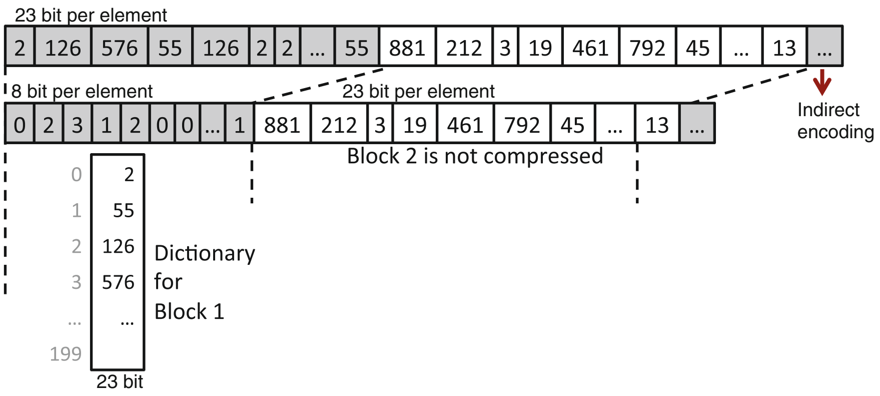


Fig. 7.5 Indirect encoding example

Besides a global dictionary used by dictionary encoding in general, additional local dictionaries are introduced for those blocks that contain only a few distinct values. A local dictionary for a block contains all (and only those) distinct values that appear in this specific block. Thus, mapping to even smaller valueIDs can save space. Direct access is still possible, however, an indirection is introduced because of the local dictionary. Figure 7.5 depicts an example for indirect encoding with a block size of 1024 elements. The upper part shows the dictionary-encoded attribute vector, the lower part shows the compressed vector. The first block contains only 200 distinct values and is compressed. The second block is not compressed.

7.4.1 Example

Given is the dictionary-encoded attribute vector for the first name column (5 million distinct values) of the world population table that is sorted by country. The number of bits required to store 5 million distinct values is 23 bit ($\lceil log_2\ (5\,\text{million})\ \rceil = 23\,\text{bit}$). Thus, the size of this vector without additional compression is 21.4 GB ($8\,\text{billion} \cdot 23\,\text{bit}$).

Now we split up the attribute vector into blocks of 1024 elements resulting in 7.8 million blocks ($\frac{8\,billion\,rows}{1024\,elements}$). For our calculation and for simplicity, we assume that each set of 1024 people of the same country contains on average 200 different first names and all blocks will be compressed. The number of bits required to represent 200 different values is 8 bit ($\lceil log_2(200)\rceil = 8\,\text{bit}$). As a result, the elements in the compressed attribute vector need only 8 bit instead of 23 bit when using local dictionaries.

Dictionary sizes can be calculated from the (average) number of distinct values in a block (200) multiplied by the size of the corresponding old valueID (23 bit) being the value in the local dictionary. For the reconstruction of a certain row, a pointer to the local dictionary for the corresponding block is stored (64 bit). Thus, the runtime for accessing a row is constant. The total amount of memory necessary for the compressed attribute vector is calculated as follows:

$$\begin{aligned}&\textit{local dictionaries} + \textit{compressed attribute vector}\\&= (200 \cdot 23\,\text{bit} + 64\,\text{bit}) \cdot 7.8\,\text{million } \textit{blocks} + 8\,\text{billion} \cdot 8\,\text{bit}\\&= 4.2\,\text{GB} + 7.6\,\text{GB}\\&\approx 11.8\,\text{GB}\end{aligned}$$

Compared to the 21.4 GB for the dictionary-encoded attribute vector, a saving of 9.6 GB (44 %) can be achieved. The following example query that selects the birthdays of all people named “John” in the “USA” shows that indirect encoding allows for direct access:

```
SELECT birthday
FROM world_population
WHERE first_name = ‘John’ AND country = ‘USA’
```

Listing 7.4.1: Birthdays for all residents of the USA with first name John

As the table is sorted by country, we can easily identify the recordIDs of the records with country=“USA”, and determine the corresponding blocks to scan the “first_name” column by dividing the first and last recordID by the cluster size.

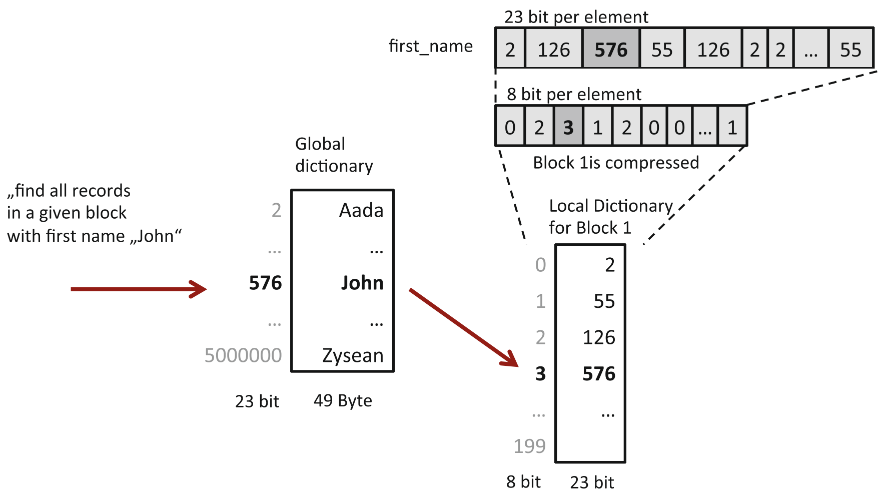


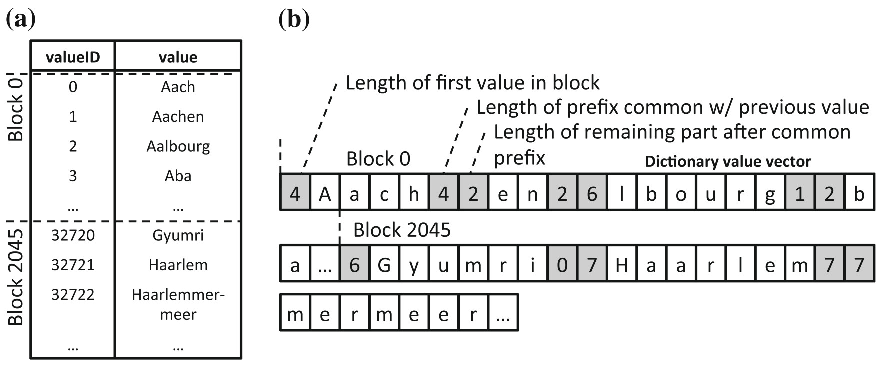
Fig. 7.6 Indirect encoding example: direct access

Then, the valueID for "John" is retrieved from the global dictionary and, for each block, the global valueID is translated into the local valueID by looking it up in the local dictionary. This is illustrated in Fig. 7.6 for a single block. Then, the block is scanned for the local valueID and corresponding recordIDs are returned for the birthday projection. In most cases, the starting and ending recordID will not match the beginning and the end of a block. In this case, we only consider the elements between the first above found recordID in the starting block up to the last found recordID for the value "USA" in the ending block.

7.5 Delta Encoding

The compression techniques covered so far reduce the size of the attribute vector. There are also some compression techniques to reduce the data amount in the dictionary as well. Let us assume that the data in the dictionary is sorted alpha-numerically and we often encounter a large number of values with the same prefixes. Delta encoding exploits this fact and stores common prefixes only once.

Delta encoding uses a block-wise compression like in previous sections with typically 16 strings per block. At the beginning of each block, the length of the first string, followed by the string itself, is stored. For each following value, the number of characters used from the previous prefix, the number of characters added to this prefix and the characters added are stored. Thus, each following string can be composed of the characters shared with the previous string and its remaining part. Figure 7.7 shows an example of a compressed dictionary. The dictionary itself is shown in Fig. 7.7a. Its compressed counterpart is provided in Fig. 7.7b.



(a)

	valueID	value
Block 0	0	Aach
	1	Aachen
	2	Aalbourg
	3	Aba
	…	…
Block 2045	32720	Gyumri
	32721	Haarlem
	32722	Haarlemmer-meer
	…	…

(b)
Length of first value in block
Length of prefix common w/ previous value
Length of remaining part after common prefix
Block 0
Dictionary value vector
4 A a c h 4 2 e n 2 6 l b o u r g 1 2 b
Block 2045
a … 6 G y u m r i 0 7 H a a r l e m 7 7
m e r m e e r …

Fig. 7.7 Delta encoding example. **a** Dictionary. **b** Compressed dictionary

7.5.1 Example

Given is a dictionary for the city column sorted alpha-numerically. The size of the uncompressed dictionary with 1 million cities, each value using 49 Byte (we assume the longest city name has 49 letters), is 46.7 MB.

For compression purposes, the dictionary is separated into blocks of 16 values. Thus, the number of blocks is 62,500 $(\frac{1\,million\,cities}{16})$. Furthermore, we assume the following data characteristics to calculate the required size in memory:

- average length of city names is 7
- average overlap of 3 letters
- the longest city name is 49 letters $(\lceil log_2(49) \rceil = 6\,\text{bit})$.

The size of the compressed dictionary is now calculated as follows:

$$\begin{aligned}
&block\ size \cdot number\ of\ blocks \\
&= encoding\ lengths + 1st\ city + 15\ other\ cities \cdot number\ of\ blocks \\
&= ((1 + 15 \cdot 2) \cdot 6\,\text{bit} + 7 \cdot 1\,\text{Byte} + 15 \cdot (7 - 3) \cdot 1\,\text{Byte}) \cdot 62{,}500 \\
&\approx 5.4\,\text{MB}
\end{aligned}$$

Compared to the 46.7 MB without compression the saving is 42.2 MB (90 %).

7.6 Limitations

What has to be kept in mind is that most compression techniques require sorted sets to tap their full potential, but a database table can only be sorted by one column or cascading. Furthermore, some compression techniques do not allow direct access. This has to be carefully considered with regard to response time requirements of queries.

7.7 Self Test Questions

1. **Sorting Compressed Tables**
   Which of the following statements is correct?
   (a) If you sort a table by the amount of data for a row, you achieve faster read access
   (b) Sorting has no effect on possible compression algorithms
   (c) You can sort a table by multiple columns at the same time
   (d) You can sort a table only by one column.

2. **Compression and OLAP / OLTP**
   What do you have to keep in mind if you want to bring OLAP and OLTP together?

   (a) You should not use any compression techniques because they increase CPU load
   (b) You should not use compression techniques with direct access, because they cause major security concerns
   (c) Legal issues may prohibit to bring certain OLTP and OLAP datasets together, so all entries have to be reviewed
   (d) You should use compression techniques that give you direct positional access, since indirect access is too slow.

3. **Compression Techniques for Dictionaries**
   Which of the following compression techniques can be used to decrease the size of a sorted dictionary?

   (a) Cluster Encoding
   (b) Prefix Encoding
   (c) Run-Length Encoding
   (d) Delta Encoding.

4. **Indirect Access Compression Techniques**
   Which of the explained compression techniques does not support direct access?

   (a) Run-Length Encoding
   (b) Prefix Encoding
   (c) Cluster Encoding
   (d) Indirect Encoding.

5. **Compression Example Prefix Encoding**
   Suppose there is a table where all 80 million inhabitants of Germany are assigned to their cities. Germany consists of about 12,200 cities, so the valueID is represented in the dictionary via 14 bit. The outcome of this is that the attribute vector for the cities has a size of 140 MB. We compress this attribute vector with Prefix Encoding and use Berlin, which has nearly 4 million inhabitants, as the prefix value. What is the size of the compressed attribute vector? Assume that the needed space to store the amount of prefix values and the prefix value itself is neglectable, because the prefix value only consumes 22 bit to represent the number of citizens in Berlin and additional 14 bit to store the key for Berlin once. Further assume the following conversions: $1\,\text{MB} = 1000\,\text{kB},\ 1\,\text{kB} = 1000\,\text{B}$

   (a) 0.1 MB
   (b) 133 MB
   (c) 63 MB
   (d) 90 MB

6. **Compression Example Run-Length Encoding Germany**
Suppose there is a table where all 80 million inhabitants of Germany are assigned to their cities. The table is sorted by city. Germany consists of about 12,200 cities (represented by 14 bit). Using Run-Length Encoding with a start position vector, what is the size of the compressed city vector? Always use the minimal number of bits required for any of the values you have to choose. Further assume the following conversions: $1\,\text{MB} = 1000\,\text{kB},\ 1\,\text{kB} = 1000\,\text{B}$

   (a) 1.2 MB
   (b) 127 MB
   (c) 5.2 KB
   (d) 62.5 kB

7. **Compression Example Cluster Encoding**
Assume the world population table with 8 billion entries. This table is sorted by countries. There are about 200 countries in the world. What is the size of the attribute vector for countries if you use Cluster Encoding with 1,024 elements per block assuming one block per country can not be compressed? Use the minimum required count of bits for the values. Further assume the following conversions: $1\,\text{MB} = 1000\,\text{kB},\ 1\,\text{kB} = 1000\,\text{B}$

   (a) $\approx$ 9 MB
   (b) $\approx$ 4 MB
   (c) $\approx$ 0.5 MB
   (d) $\approx$ 110 MB

8. **Best Compression Technique for Example Table**
Find the best compression technique for the name column in the following table. The table lists the names of all inhabitants of Germany and their cities, i.e. there are two columns: first_name and city. Germany has about 80 million inhabitants and 12,200 cities. The table is sorted by the city column. Assume that any subset of 1,024 citizens contains at most 200 different first names.

   (a) Run-Length Encoding
   (b) Indirect Encoding
   (c) Prefix Encoding
   (d) Cluster Encoding.

Reference

[AMF06] D. Abadi, S. Madden, M. Ferreira, Integrating compression and execution in column-oriented database systems, in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, SIGMOD '06* (ACM, New York, 2006), pp. 671–682

Chapter 8
Data Layout in Main Memory

In this chapter, we address the question how data is organized in memory. Relational database tables have a two-dimensional structure but main memory is organized unidimensional, providing memory addresses that start at zero and increase serially to the highest available location. The database storage layer has to decide how to map the two-dimensional table structures to the linear memory address space.

We will consider two ways of representing a table in memory, called row and columnar layout, and a combination of both ways, a hybrid layout.

8.1 Cache Effects on Application Performance

In order to understand the implications introduced by row-based and column-based layouts, a basic understanding of memory access performance is essential. Due to the different available types of memory as described in Sect. 4.1, modern computer systems leverage a so-called memory hierarchy as described in Sect. 4.2. These caching mechanisms plus techniques like the Translation Lookaside Buffer (TLB, see Sect. 4.4) or hardware prefetching (see Sect. 4.5) introduce various performance implications, which will be outlined in this section that is based on [SKP12].

The described caching and virtual memory mechanisms are implemented as transparent systems from the viewpoint of an actual application. However, knowing the used system with its characteristics and optimizing applications based on this knowledge can have crucial implications on application performance.

The following two sections describe two small experiments, outlining performance differences when accessing main memory. These experiments are for the interested reader and will not be relevant for the exam.

8.1.1 The Stride Experiment

As the name random access memory suggests, the memory can be accessed randomly and one would expect constant access costs. In order to test this assumption, we run a simple benchmark accessing a constant number (4,096) of addresses with an increasing stride, i.e. distance, between the accessed addresses.

We implemented this benchmark by iterating through an array chasing a pointer. The array is filled with structs so that following the pointer of the elements creates a circle through the complete array. Structs are data structures, which allow to create user-defined aggregate data types that group multiple individual variables together. The structs consist of a pointer and an additional data attribute realizing the padding in memory, resulting in a memory access with the desired stride when following the pointer chained list.

```
struct element {
  struct element *pointer;
  size_t padding[PADDING];
}
```

In case of a sequential array, the pointer of element i points to element $i + 1$ and the pointer of the last element references the first element so that the circle is closed. In case of a random array, the pointer of each element points to a random element of the array while ensuring that every element is referenced exactly once. Figure 8.1 outlines the created sequential and random array.

If the assumption holds and random memory access costs are constant, then the size of the padding in the array and the array layout (sequential or random) should make no difference when iterating over the array. Figure 8.2 shows the result for iterating through a list with 4,096 elements, while following the pointers inside the elements and increasing the padding between the elements. As we can clearly see, the access costs are not constant and increase with an increasing stride. We also see multiple points of discontinuity in the curves, e.g. the random access times increase heavily up to a stride of 64 bytes and continue increasing with a smaller slope.

Figure 8.3 indicates that an increasing number of cache misses is causing the increase in access times. The first point of discontinuity in Fig. 8.2 is quite exactly the size of the cache lines of the test system. The strong increase is due to the fact, that with a stride smaller than 64 bytes, multiple list elements are located on one cache line and the overhead of loading one line is amortized over the multiple elements.

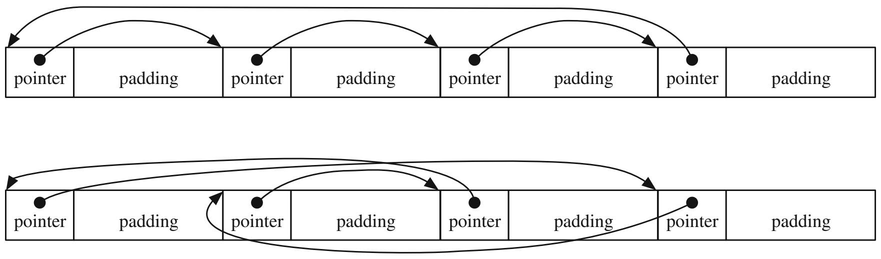


Fig. 8.1 Sequential versus random array layout

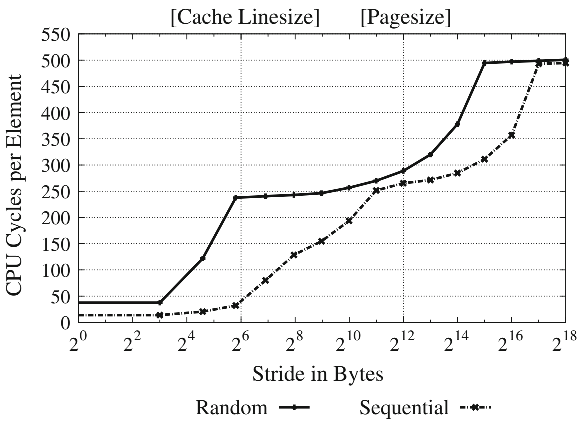


Fig. 8.2 Cycles for cache accesses with increasing stride

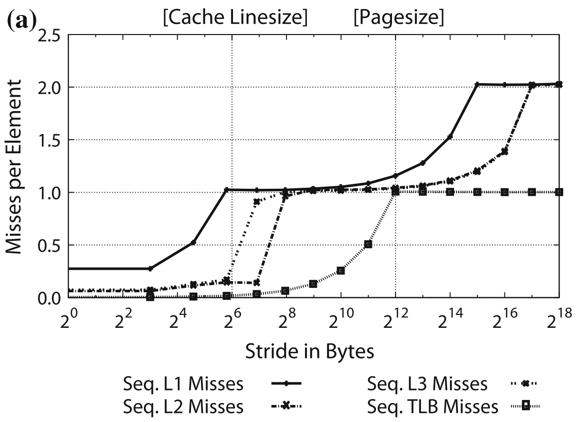
For strides greater than 64 bytes, we would expect a cache miss for every single list element and no further increase in access times. However, as the stride gets larger the array is placed over multiple pages in memory and more TLB misses occur, as the virtual addresses on the new pages have to be translated into physical addresses. The number of TLB cache misses increases up to the page size of 4 KB and stays at its worst case of one miss per element. With strides greater as the page size, the TLB misses can induce additional cache misses when translating the virtual to a physical address. These cache misses are due to accesses to the paging structures which reside in main memory [BCR10, BCR10, SS95].

To summarize, the performance of main memory accesses can largely differ depending on the access patterns. In order to improve application performance, main memory access should be optimized in order to exploit the usage of caches.

8.1.2 The Size Experiment

In a second experiment, we access a constant number of addresses in main memory with a constant stride of 64 bytes and vary the size of the working set size or accessed area in memory. A run with n memory accesses and a working set size of s bytes would iterate $\frac{n}{s \cdot 64}$ times through the array, which is created as described earlier in the stride experiment in Sect. 8.1.1.

Figure 8.4a shows that the access costs differ up to a factor of 100, depending on the working set size. The points of discontinuity correlate with the sizes of the caches in the system. As long as the working set size is smaller than the size of the L1 Cache, only the first iteration results in cache misses and all other accesses can be answered out of the cache. As the working set size increases, the accesses in one iteration start to evict the earlier accessed addresses, resulting in cache misses in the next iteration.



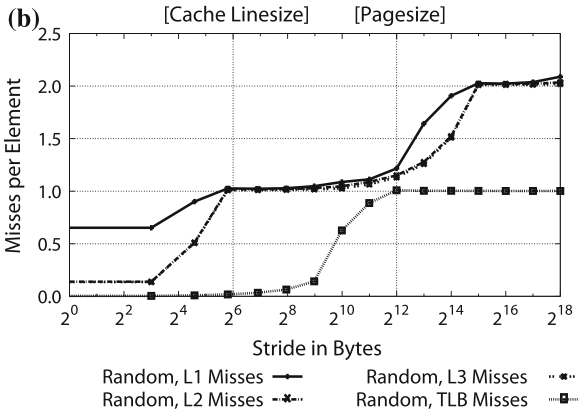


Fig. 8.3 Cache misses for cache accesses with increasing stride. (**a**) Sequential Access. (**b**) Random Access

Figure 8.4b shows the individual cache misses with increasing working set sizes. Up to working sets of 32 KB, the misses for the L1 cache go up to one per element, the L2 cache misses reach their plateau at the L2 cache size of 256 KB and the L3 cache misses at 12 MB.

As we can see, the larger the accessed area in main memory, the more capacity cache misses occur, resulting in poorer application performance. Therefore, it is advisable to process data in cache friendly chunks if possible.

8.2 Row and Columnar Layouts

Let us consider a simple example to illustrate the two mentioned approaches for representing a relational table in memory. For simplicity, we assume that all values

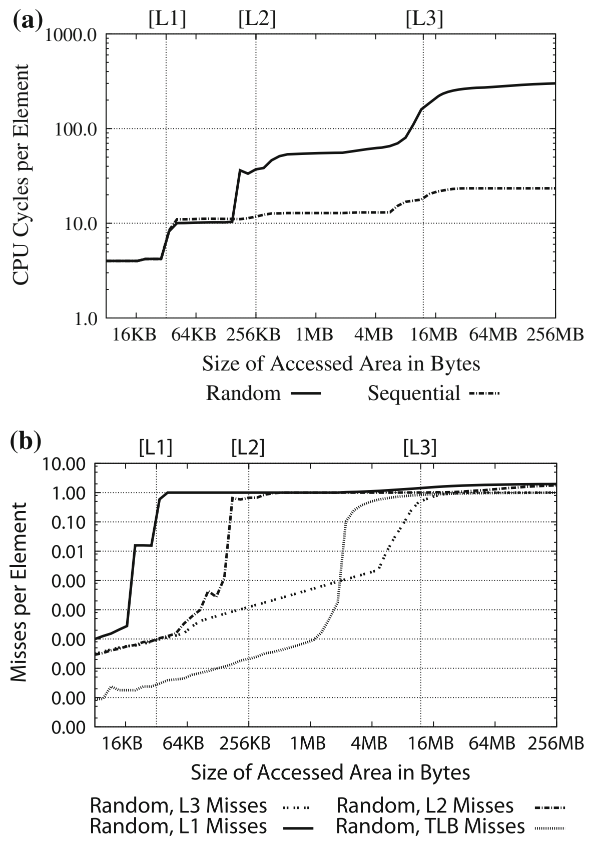


Fig. 8.4 Cycles and cache misses for cache accesses with increasing working sets. (**a**) Sequential Access. (**b**) Random Access

are stored as strings directly in memory and that we do not need to store any additional data. As an example, let us look at the simple world population example:

Id	Name	Country	City
1	Paul Smith	Australia	Sydney
2	Lena Jones	USA	Washington
3	Marc Winter	Germany	Berlin

As discussed above, the database must transform its two-dimensional table into a one-dimensional series of bytes for the operating system to write them to memory. The classical and obvious approach is a row- or record-based layout. In this case, all attributes of a tuple are stored consecutively and sequentially in memory. In other words, the data is stored tuple-wise. Considering our example table, the data

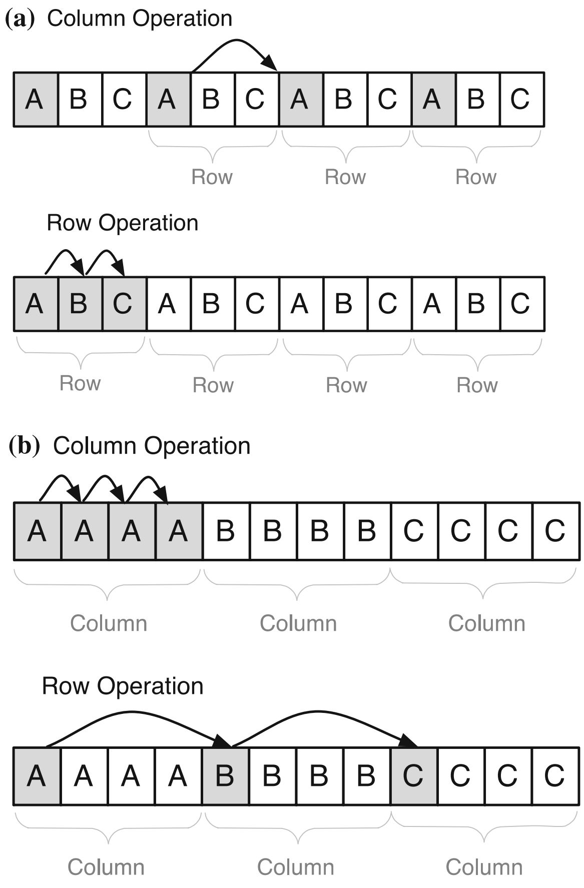


Fig. 8.5 Illustration of memory accesses for row-based and column-based operations on row and columnar data layouts.

would be stored as follows: "`1`, `Paul Smith`, `Australia`, `Sydney; 2`, `Lena Jones`, `USA`, `Washington; 3`, `Marc Winter`, `Germany`, `Berlin`".

On the contrary, in a columnar layout , the values of one column are stored together, column by column. The resulting layout in memory for our example would be: "`1`, `2`, `3; Paul Smith`, `Lena Jones`, `Marc Winter; Australia`, `USA`, `Germany; Sydney`, `Washington`, `Berlin`".

The columnar layout is especially effective for set-based reads. In other words, it is useful for operations that work on many rows but only on a notably smaller subset of all columns, as the values of one column can be read sequentially, e.g. when performing aggregate calculations. However, when performing operations on single tuples or for inserting new rows, a row-based layout is beneficial. The different access patterns for row-based and column-based operations are illustrated in Fig. 8.5.

Currently, row-oriented architectures are widely used for OLTP workloads while column stores are widely utilized in OLAP scenarios like data warehousing, which typically involve a smaller number of highly complex queries over the complete data set.

8.3 Benefits of a Columnar Layout

As mentioned above, there are use cases where a row-based table layout can be more efficient. Nevertheless, many advantages speak in favor of the usage of a columnar layout in an enterprise scenario.

First, when analyzing the workloads enterprise databases are facing, it turns out that the actual workloads are more read-oriented and dominated by set processing [KKG+11].

Second, despite the fact that hardware technology develops very rapidly and the size of available main memory constantly grows, the use of efficient compression techniques is still important in order to (a) keep as much data in main memory as possible and to (b) minimize the amount of data that has to be read from memory to process queries as well as the data transfer between non-volatile storage mediums and main memory.

Using column-based table layouts enables the use of efficient compression techniques leveraging the high data locality in columns (see Chap. 7). They mainly use the similarity of the data stored in a column. Dictionary encoding can be applied to row-based as well as column-based table layout, whereas other techniques like prefix encoding, run-length encoding, cluster encoding or indirect encoding directly leverage the benefits of columnar table layouts.

Third, using columnar table layouts enables very fast column scans as they can sequentially scan the memory, allowing e.g. on the fly calculations of aggregates. Consequently, storing pre-calculated aggregates in the database can be avoided, thus minimizing redundancy and complexity of the database.

8.4 Hybrid Table Layouts

As stated above, set processing operations are dominating enterprise workloads. Nevertheless, each concrete workload is different and might favor a row-based or a column-based layout. Hybrid table layouts combine the advantages of both worlds, allowing to store single attributes of a table column oriented while grouping other attributes into a row-based layout [GKP+11]. The actual optimal combination highly depends on the actual workload and can be calculated by layouting algorithms.

As an illustrating example, think about attributes, which inherently belong together in commercial applications, e.g. quantity and measuring unit or payment conditions in accounting. The idea of the hybrid layout is that if the set of attributes are processed together, it makes sense from a performance point of view to physically store them together. Considering the example table provided in Sect. 8.2 and assuming the fact that the attributes *Id* and *Name* are often processed together, we can outline the following hybrid data layout for the table: "`1, Paul Smith; 2, Lena Jones; 3, Marc Winter; Australia, USA, Germany; Sydney, Washington, Berlin`". This hybrid layout may decrease the number of cache misses caused by the expected workload, resulting in increased performance.

The usage of hybrid layouts can be beneficial but also introduces new questions like how to find the optimal layout for a given workload or how to react on a changing workload.

8.5 Self Test Questions

1. When DRAM can be accessed randomly with the same costs, why are consecutive accesses usually faster than stride accesses?
   (a) With consecutive memory locations, the probability that the next requested location has already been loaded in the cache line is higher than with randomized/strided access. Furthermore is the memory page for consecutive accesses probably already in the TLB
   (b) The bigger the size of the stride, the higher the probability, that two values are both in one cache line
   (c) Loading consecutive locations is not faster, since the CPU performs better on prefetching random locations, than prefetching consecutive locations
   (d) With modern CPU technologies like TLBs, caches and prefetching, all three access methods expose the same performance.

References

[BCR10] T.W. Barr, A.L. Cox, S. Rixner, Translation caching: skip, don't walk (the Page Table). ACM SIGARCH Comput Arch. News **38**(3), 48–59 (2010)

[BT09] V. Babka, P. Tůma, Investigating cache parameters of x86 family processors. Comput. Perform. Eval. Benchmarking. 77–96 (2009)

[GKP+11] M. Grund, J. Krueger, H. Plattner, A. Zeier, S. Madden, P. Cudre-Mauroux, HYRISE - A hybrid main memory storage engine, in *VLDB* (2011)

[KKG+11] J. Krueger, C. Kim, M. Grund, N. Satish, D. Schwalb, J. Chhugani, H. Plattner, P. Dubey, A. Zeier, Fast updates on read-optimized databases using multi-core CPUs, in *PVLDB* (2011)

[SKP12] D. Schwalb, J. Krueger, H. Plattner, Cache conscious column organization in in-memory column stores. Technical Report 60, Hasso-Plattner-Institute, December 2012.

[SS95] R.H. Saavedra, A.J. Smith, Measuring cache and TLB performance and their effect on benchmark runtimes. IEEE Trans. Comput. **44**(10), 1223–1235 (1995)

Chapter 9
Partitioning

9.1 Definition and Classification

Partitioning is the process of dividing a logical database into distinct independent datasets. Partitions are database objects itself and can be managed independently. The main reason to apply data partitioning is to achieve data-level parallelism. Data-level parallelism enables performance gains, a classic example for that is to use a multi-core CPU to process several distinct data areas in parallel, whereas each core works on a separate partition. Since partitioning is applied as a technical step to increase the query speed, it should be transparent[1] to the user. In order to ensure the transparency of the applied partitioning for the end user, a view showing the complete table as a union of all query results from all involved partitions is required. With data-level parallelism it is possible to increase performance, availability, or manageability of datasets. Which of these sometimes contradicting goals is favored usually depends on the actual use case. Two short examples are given in Sect. 9.4. Because data partitioning is a classical NP-complete[2] problem, finding the best partition is a complicated task, even if the desired goal has been clearly outlined [Kar72]. There are mainly two types of data partitioning: horizontal and vertical partitioning, which will be covered in detail in the following.

9.2 Vertical Partitioning

Vertical partitioning results in splitting the data into attribute groups with replicated primary keys. These groups are then distributed across two (or more) tables (Fig. 9.1). Attributes that are usually accessed together should be in the same table, in order to increase join performance. Such optimizations can only be applied if

1 Transparent in IT means that something is completely invisible to the user, not that the user can inspect the implementation through the cover. Except of their effects like improvements in speed or usability, transparent components should not be noticeable at all.

2 NP-complete means that the problem can not be solved in polynomial time.

DOI: 10.1007/978-3-642-36524-9_9, © Springer-Verlag Berlin Heidelberg 2013

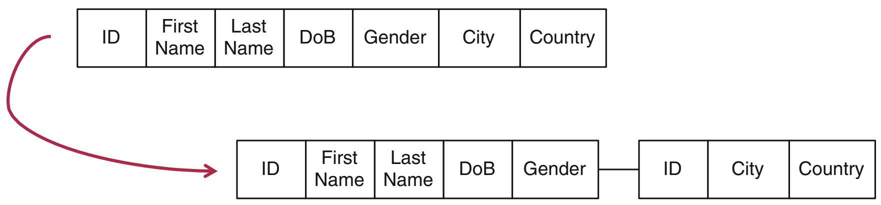


Fig. 9.1 Vertical partitioning

actual usage data exists, which is one point why application development should always be based on real customer data and workloads.

In row-based databases, vertical partitioning is possible in general. However, it is not a common approach because it is hard to establish, because the underlying concept of values tuple wise is contradicted when separating parts of the attributes. Column-based databases implicitly support vertical partitioning, since each column can be regarded as a possible partition.

9.3 Horizontal Partitioning

Horizontal Partitioning is used more often in classic row-oriented databases. To apply this partitioning, the table is split into disjoint tuple groups by some condition. There are several sub-types of horizontal partitioning:

The first partitioning approach we present here is *range partitioning*, which separates tables into partitions by a predefined partitioning key, which determines how individual data rows are distributed to different partitions. The partition key can consist of a single key column or multiple key columns. For example, customers could be partitioned based on their date of birth. If one is aiming for a number of four partitions, each partition would cover a range of about 25 years (Fig. 9.2).[3] Because the implications of the chosen partition key depend on the workload, it is not trivial to find the optimal solution.

The second horizontal partitioning type is *round robin partitioning*. With round robin, a partitioning server does not use any tuple information as partitioning criteria, so there is no explicit partition key. The algorithm simply assigns tuples turn by turn to each partition, which automatically leads to an even distribution of entries and should support load-balancing to some extent (Fig. 9.3).

However, since specific entries might be accessed way more often than others, an even workload distribution can not be guaranteed. Improvements from intelligent data co-location or appropriate data-placement are not leveraged, because the data distribution is not dependent on the data, but only on the insertion order.

[3] Based on the assumption that the companies' customers mainly live nowadays and are between 0 and 100 years old.

Partition 1

ID	First Name	Last Name	DoB	Gender	City	Country

Partition 2

ID	First Name	Last Name	DoB	Gender	City	Country
1	John	Dillan	1943/05/12	m	Berlin	Germany

Partitioning along the age:

Partition 1:	76 – 100
Partition 2:	51 – 75
Partition 3:	26 – 50
Partition 4:	0 – 25

Partition 3

ID	First Name	Last Name	DoB	Gender	City	Country
3	Nina	Burg	1952/12/12	w	London	UK

Partition 4

ID	First Name	Last Name	DoB	Gender	City	Country
2	Peter	Black	1982/06/02	m	Austin	USA
4	Lucy	Sehan	1990/01/20	w	Jerusalem	Israel
5	Ariel	Shiva	1984/07/18	w	Tokio	Japan
6	Sharon	Lokida	1982/02/24	m	Madrid	Spain

Fig. 9.2 Range partitioning

Partition 1

ID	First Name	Last Name	DoB	Gender	City	Country
1	John	Dillan	1943/05/12	m	Berlin	Germany
5	Ariel	Shiva	1984/07/18	w	Tokio	Japan

Partition 2

ID	First Name	Last Name	DoB	Gender	City	Country
2	Peter	Black	1982/06/02	m	Austin	USA
6	Sharon	Lokida	1982/02/24	m	Madrid	Spain

Partition 3

ID	First Name	Last Name	DoB	Gender	City	Country
3	Nina	Burg	1952/12/12	w	London	UK

Partition 4

ID	First Name	Last Name	DoB	Gender	City	Country
4	Lucy	Sehan	1990/01/20	w	Jerusalem	Israel

Fig. 9.3 Round robin partitioning

Partition 1

ID	First Name	Last Name	DoB	Gender	City	Country	hash(Country)
4	Lucy	Sehan	1990/01/20	w	Jerusalem	Israel	0x00

Partition 2

ID	First Name	Last Name	DoB	Gender	City	Country	hash(Country)
1	John	Dillan	1943/05/12	m	Berlin	Germany	0x01

Partition 3

ID	First Name	Last Name	DoB	Gender	City	Country	hash(Country)
3	Nina	Burg	1952/12/12	w	London	UK	0x03

Partition 4

ID	First Name	Last Name	DoB	Gender	City	Country	hash(Country)
2	Peter	Black	1982/06/02	m	Austin	USA	0x02
5	Ariel	Shiva	1984/07/18	w	Tokio	Japan	0x02

Fig. 9.4 Hash-based partitioning

The third horizontal partitioning type is *hash-based partitioning*. Hash partitioning uses a hash function[4] to specify the partition assignment for each row (Fig. 9.4).

The main challenge for hash-based partitioning is to choose a good hash function, that implicitly achieves locality or access improvements.

The last partitioning type is *semantic partitioning*. It uses knowledge about the application to split the data. For example, a database can be partitioned according

---

[4] A hash function maps a potentially large amount of data with often variable length to a smaller value of fixed length. In the figurative sense, hash functions generate a digital fingerprint of the input data.

to the life-cycle of a sales order. All tables required for the sales order represent one or more different life-cycle steps, such as creation, purchase, release, delivery, or dunning of a product. One possibility for suitable partitioning is to put all tables that belong to a certain life-cycle step into a separate partition.

9.4 Choosing a Suitable Partitioning Strategy

There are number of different optimization goals to be considered while choosing a suitable partitioning strategy. For instance, when optimizing for *performance*, it makes sense to have tuples of different tables, that are likely to be joined for further processing, on one server. This way the join can be done much faster due to optimal *data locality*, because there is no delay for transferring the data across the network. In contrast, for statistical queries like counts, tuples from one table should be distributed across as many nodes as possible in order to benefit from parallel processing.

To sum up, the best partitioning strategy depends very much on the specific use case.

9.5 Self Test Questions

1. **Partitioning Types**
   Which partitioning types do really exist and are mentioned in the course?
   (a) Selective Partitioning
   (b) Syntactic Partitioning
   (c) Range Partitioning
   (d) Block Partitioning.

2. **Partitioning Type for Given Query**
   Which partitioning type fits best for the column ‘birthday’ in the world population table, when we assume that the main workload is caused by queries like ‘SELECT first_name, last_name FROM population WHERE birthday $> 01.01.1990$ AND birthday $< 31.12.2010$ AND country $=$ ‘England’? Assume a non-parallel setting, so we can not scan partitions in parallel. The only parameter that is changed in the query is the country.
   (a) Round Robin Partitioning
   (b) All partitioning types will show the same performance
   (c) Range Partitioning
   (d) Hash Partitioning.

3. **Partitioning Strategy for Load Balancing**
   Which partitioning type is suited best to achieve fair load-balancing if the values of the column are non-uniformly distributed?
   (a) Partitioning based on the number of attributes used modulo the number of systems
   (b) Range Partitioning
   (c) Round Robin Partitioning
   (d) All partitioning types will show the same performance.

Reference

[Kar72] R. Karp, Reducibility among combinatorial problems, in *Complexity of Computer Computations*, eds. by R. Miller, J. Thatcher (Plenum Press, 1972), pp. 85–103

Part III
In-Memory Database Operators

Chapter 10
Delete

The delete operation terminates the validity of a given tuple. It stores the information in the database that a certain item is no longer valid This operation can either be of *physical* or *logical* nature. A physical delete operation removes an item from the database so that it is no longer physically accessible. In contrast, a *logical* delete operation only terminates the validity of an item in the dataset, but keeps the tuple still available for temporal queries [Pla09].

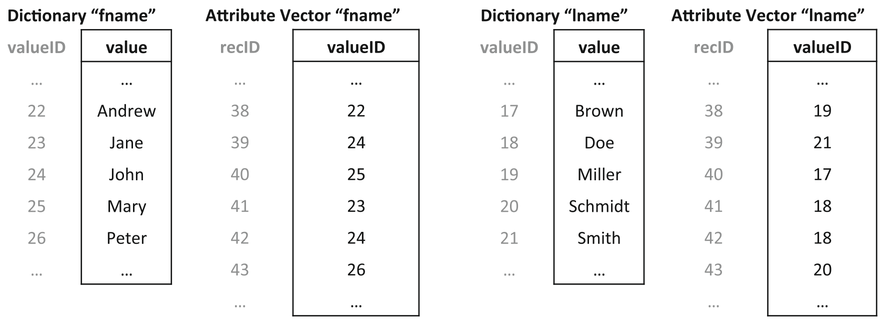
The simplified SQL-Syntax for a delete statement looks like the following, where the predicate may select a single or multiple tuples.

```
DELETE FROM table_name WHERE condition
```

Listing 10.1: Delete syntax

10.1 Example of Physical Delete

In the following example, all persons with the name *'Jane Doe'* are supposed to be removed from a database table storing first and last names. Based on the applied dictionary encoding (see Chap. 6), the table consists of two dictionaries and two value attribute vectors.



Dictionary "fname"

valueID	value
…	…
22	Andrew
23	Jane
24	John
25	Mary
26	Peter
…	…

Attribute Vector "fname"

recID	valueID
…	…
38	22
39	24
40	25
41	23
42	24
43	26
…	…

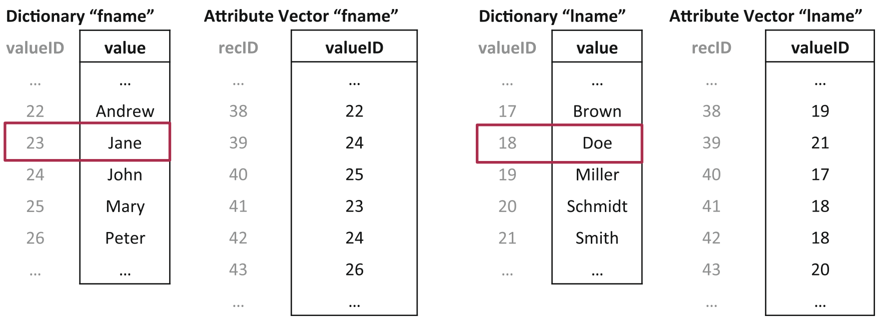
Dictionary "lname"

valueID	value
…	…
17	Brown
18	Doe
19	Miller
20	Schmidt
21	Smith
…	…

Attribute Vector "lname"

recID	valueID
…	…
38	19
39	21
40	17
41	18
42	18
43	20
…	…

First, the valueIDs for the first and last name need to be identified. *Jane* corresponds to valueID 23 and *Doe* to valueID 18, according to their respective dictionary.



Dictionary “fname”

valueID	value
...	...
22	Andrew
23	Jane
24	John
25	Mary
26	Peter
...	...

Attribute Vector “fname”

recID	valueID
...	...
38	22
39	24
40	25
41	23
42	24
43	26
...	...

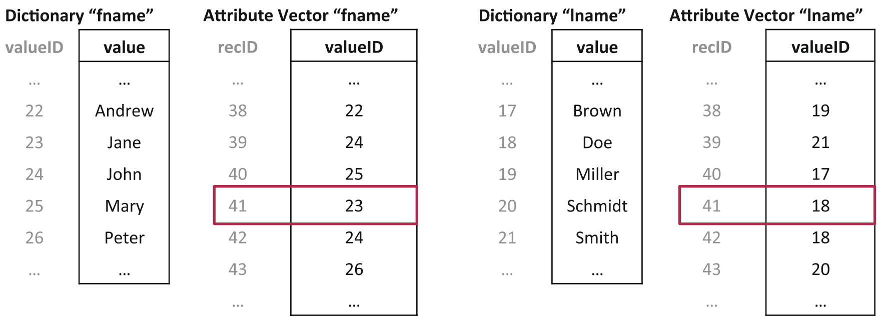
Dictionary “lname”

valueID	value
...	...
17	Brown
18	Doe
19	Miller
20	Schmidt
21	Smith
...	...

Attribute Vector “lname”

recID	valueID
...	...
38	19
39	21
40	17
41	18
42	18
43	20
...	...

Next, we scan through the attribute vectors and find the appropriate positions, which means we look up the recordIDs for these values. In our example, there is only one tuple with that combination of first and last name.



Dictionary “fname”

valueID	value
...	...
22	Andrew
23	Jane
24	John
25	Mary
26	Peter
...	...

Attribute Vector “fname”

recID	valueID
...	...
38	22
39	24
40	25
41	23
42	24
43	26
...	...

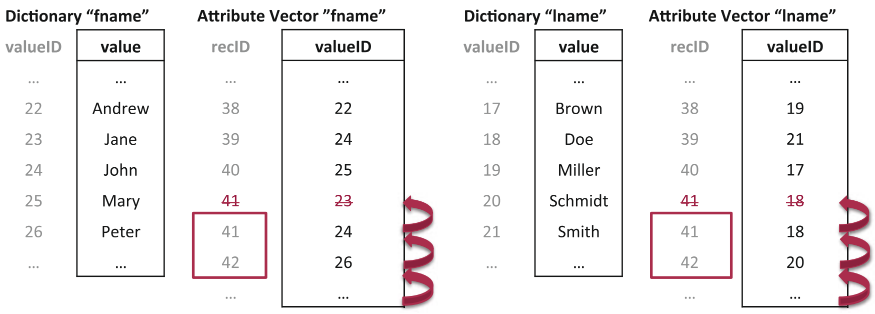
Dictionary “lname”

valueID	value
...	...
17	Brown
18	Doe
19	Miller
20	Schmidt
21	Smith
...	...

Attribute Vector “lname”

recID	valueID
...	...
38	19
39	21
40	17
41	18
42	18
43	20
...	...

When finally deleting the two values from the attribute vectors, all subsequent tuples need to be adjusted to maintain a sequence without gaps and they are moved to preserve a sequential memory area. This implementation alternative of the delete operation is therefore very expensive in terms of performance. In Chap. 26, later during the course, the *insert-only* approach is presented as a better alternative to implement deletion in typical enterprise use cases. This approach is of *logical* nature.



Dictionary “fname”

valueID	value
...	...
22	Andrew
23	Jane
24	John
25	Mary
26	Peter
...	...

Attribute Vector ”fname”

recID	valueID
...	...
38	22
39	24
40	25
~~41~~	~~23~~
41	24
42	26
...	...

Dictionary “lname”

valueID	value
...	...
17	Brown
18	Doe
19	Miller
20	Schmidt
21	Smith
...	...

Attribute Vector “lname”

recID	valueID
...	...
38	19
39	21
40	17
~~41~~	~~18~~
41	18
42	20
...	...

10.2 Self Test Questions

1. **Delete Implementations**
   Which two possible delete implementations are mentioned in the course?
   (a) White box and black box delete
   (b) Physical and logical delete
   (c) Shifted and liquid delete
   (d) Column and row deletes
2. **Arrays to Scan for Specific Query with Dictionary Encoding**
   When applying a delete with two predicates, e.g. firstname = 'John' AND lastname = 'Smith', how many logical blocks in the IMDB are being looked at during determination which tuples to delete (all columns are dictionary encoded)?
   (a) 1
   (b) 2
   (c) 4
   (d) 8
3. **Fast Delete Execution**
   Assume a physical delete implementation and the following two SQL statements on our world population table:
   (A) DELETE FROM world_population WHERE country = 'China';
   (B) DELETE FROM world_population WHERE country = 'Ireland'; Which query will execute faster? Please only consider the concepts learned so far.
   (a) Equal execution time
   (b) A
   (c) Depends on the ordering of the dictionary
   (d) B

Reference

[Pla09] H. Plattner, in *A common database approach for OLTP and OLAP using an in-memory column database*, ed. by U. Çetintemel, S. Zdonik, D. Kossmann. SIGMOD Conference (ACM, Newyork, 2009), pp. 1–2

Chapter 11
Insert

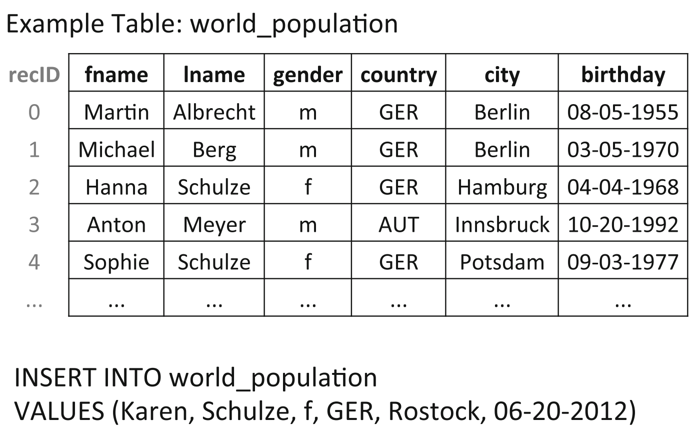
This chapter outlines what happens when inserting a new tuple into a table (execution of an insert statement). Compared to a row-based database, the insert in a column store is a bit more complicated. For a row-oriented database, the new tuple is simply appended to the end of the table, i.e., the tuple is stored as one piece. SanssouciDB uses column-orientation to store the data physically. A detailed description of the differences between row store and column store is given in Chap. 8. So, adding a new tuple to the database means to add a new entry to every column that the table comprises of. Internally, every column consists of a dictionary and an attribute vector (see Chap. 6). Adding a new entry to a column means to check the dictionary and adding a new value if necessary. Afterwards, the respective value of the dictionary entry is added to the attribute vector of the column. Since the dictionary is sorted, adding a new entry to a column results in three different scenarios:

1. Without a new dictionary entry
2. With a new dictionary entry, without resorting the dictionary
3. With a new dictionary entry, with resorting the dictionary

In this chapter, we will give a step by step explanation of the three different scenarios.

11.1 Example

In this example, we insert the data of a new person into the *world_population* table (see Fig. 11.1) that we used before.The example outlines what happens for the column *lname*, representing the last name of a person, and *fname*, representing the first name of a person.



Example Table: world_population

recID	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Sophie	Schulze	f	GER	Potsdam	09-03-1977
...	...	...	...	...	...	...

```
INSERT INTO world_population
VALUES (Karen, Schulze, f, GER, Rostock, 06-20-2012)
```

Fig. 11.1 Example database table named *world_population*

11.1.1 *INSERT without New Dictionary Entry*

To demonstrate a scenario were we have an insert without a new entry to the dictionary, we will look at the insert of the last name attribute to the *lname* column of our *world_population* table. Attribute vector and dictionary of the *lname* column are initially filled as displayed in Fig. 11.2.

To add the string *Schulze* to the column, we need to look up whether it already exists in the dictionary. Since there is another person named *Sophie Schulze* (recordID four of the world_population table) in the database, the dictionary for the *lname* column already contains an entry with the string *Schulze*. As one can see from Fig. 11.3, the dictionary position of *Schulze* is "3".

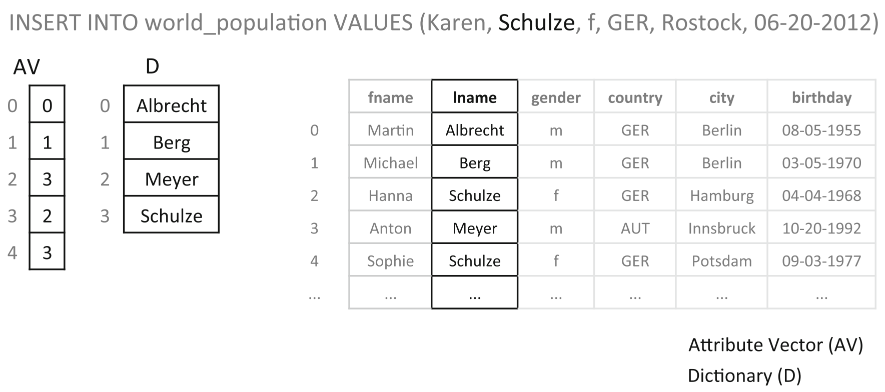
Since *Schulze* is on position 3 of the dictionary, we append 3 to the end of the attribute vector (see Fig. 11.4).

11.1.2 *INSERT with New Dictionary Entry*

When inserting the first name, the first name dictionary is scanned for the string *Karen*. As shown in Fig. 11.5, this name is not present in the dictionary, yet.

Therefore, the name is appended to the end of the first name dictionary (see Fig. 11.6).

As outlined in Chap. 6, the dictionary needs to be kept sorted. After appending *Karen* to the end of the dictionary, the dictionary needs to be resorted. Therefore, as shown in Fig. 11.7, a new dictionary is created with sorted order. In the new dictionary most of the dictionaryIDs changed. For instance, the valueID for *Michael* is changed from 3 to 4.



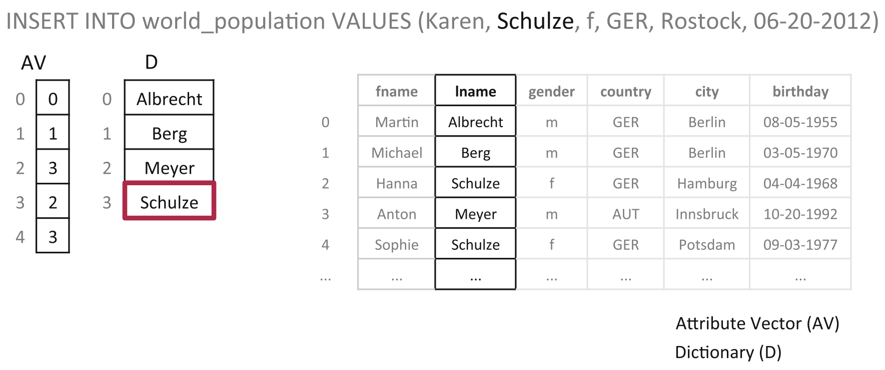
INSERT INTO world_population VALUES (Karen, Schulze, f, GER, Rostock, 06-20-2012)

	AV
0	0
1	1
2	3
3	2
4	3

	D
0	Albrecht
1	Berg
2	Meyer
3	Schulze

	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Sophie	Schulze	f	GER	Potsdam	09-03-1977
...	...	...	...	...	...	...

Fig. 11.2 Initial status of the *lname* column



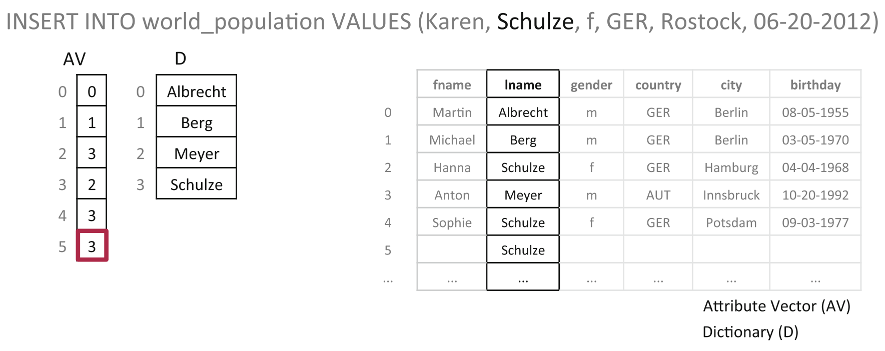
INSERT INTO world_population VALUES (Karen, Schulze, f, GER, Rostock, 06-20-2012)

	AV
0	0
1	1
2	3
3	2
4	3

	D
0	Albrecht
1	Berg
2	Meyer
3	Schulze

	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Sophie	Schulze	f	GER	Potsdam	09-03-1977
...	...	...	...	...	...	...

Fig. 11.3 Position of the string *Schulze* in the dictionary of the *lname* column



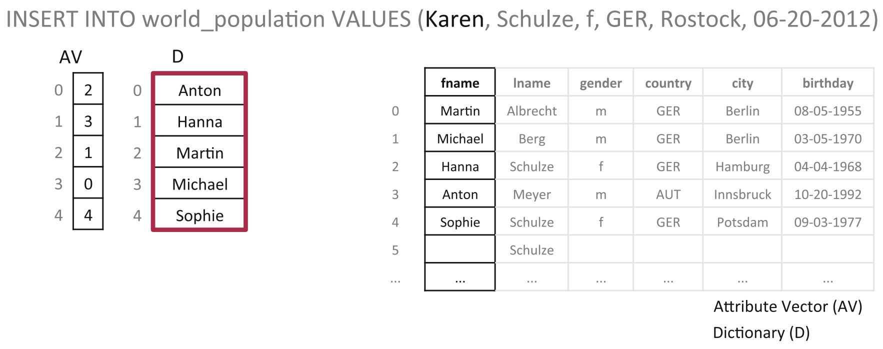
INSERT INTO world_population VALUES (Karen, Schulze, f, GER, Rostock, 06-20-2012)

	AV
0	0
1	1
2	3
3	2
4	3
5	3

	D
0	Albrecht
1	Berg
2	Meyer
3	Schulze

	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Sophie	Schulze	f	GER	Potsdam	09-03-1977
5		Schulze				
...	...	...	...	...	...	...

Fig. 11.4 Appending dictionary position of *Schulze* to the end of the attribute vector

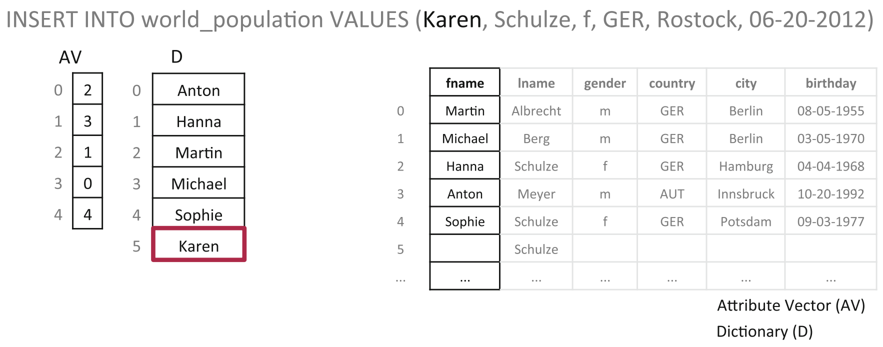

INSERT INTO world_population VALUES (Karen, Schulze, f, GER, Rostock, 06-20-2012)

	AV
0	2
1	3
2	1
3	0
4	4

	D
0	Anton
1	Hanna
2	Martin
3	Michael
4	Sophie

	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Sophie	Schulze	f	GER	Potsdam	09-03-1977
5		Schulze				
...	...	...	...	...	...	...

Fig. 11.5 Dictionary for first name column

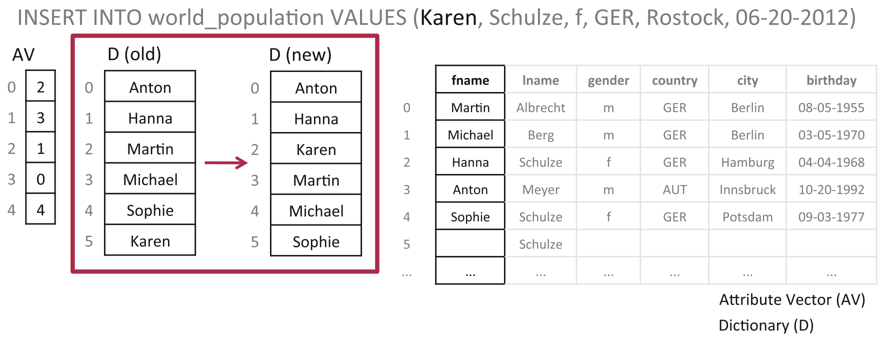

INSERT INTO world_population VALUES (Karen, Schulze, f, GER, Rostock, 06-20-2012)

	AV
0	2
1	3
2	1
3	0
4	4

	D
0	Anton
1	Hanna
2	Martin
3	Michael
4	Sophie
5	Karen

	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Sophie	Schulze	f	GER	Potsdam	09-03-1977
5		Schulze				
...	...	...	...	...	...	...

Fig. 11.6 Addition of *Karen* to *fname* dictionary


INSERT INTO world_population VALUES (Karen, Schulze, f, GER, Rostock, 06-20-2012)

	AV
0	2
1	3
2	1
3	0
4	4

	D (old)		D (new)
0	Anton	0	Anton
1	Hanna	1	Hanna
2	Martin	2	Karen
3	Michael	3	Martin
4	Sophie	4	Michael
5	Karen	5	Sophie

	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Sophie	Schulze	f	GER	Potsdam	09-03-1977
5		Schulze				
...	...	...	...	...	...	...

Fig. 11.7 Resorting the *fname* dictionary

Based on the changed valueIDs of the new first name dictionary, all valueIDs of the first name attribute vector need to be updated as well. Figure 11.8 shows the changes to the attribute vector. For instance at position 1, the valueID for *Michael* is changed from 3 to 4.

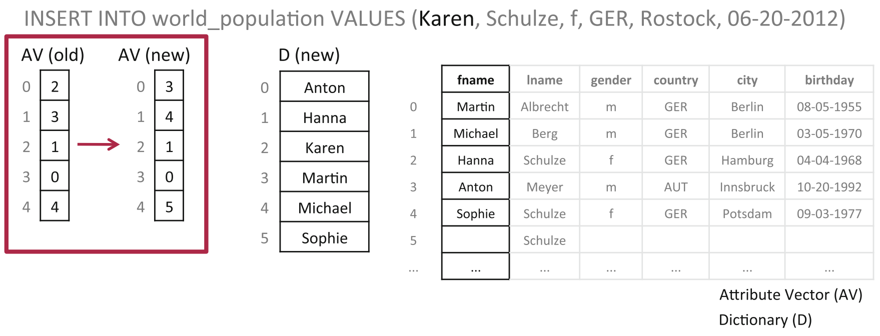


Fig. 11.8 Rebuilding the *fname* attribute vector

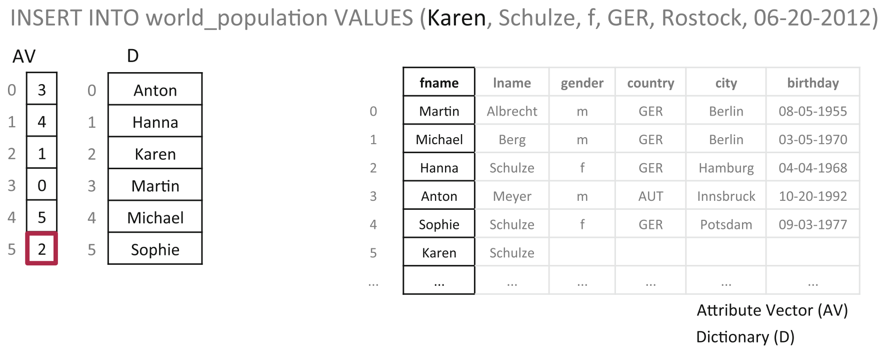


Fig. 11.9 Appending the valueID representing *Karen* to the attribute vector

In case the newly added dictionary value is inserted at the end based on the sorting order of the dictionary, those two steps are omitted. The dictionary does not need to be resorted and therefore the attribute vector does not need to be rebuild.

Finally the valueID 2, representing the dictionary position of the string *Karen*, is appended to the attribute vector (see Fig. 11.9).

11.2 Performance Considerations

When thinking of the *world_population* example, there are about 8 billion people and 5 million unique first names. Every new entry to the dictionary may cause an overhead regarding resorting of the dictionary and reorganization of the respective attribute vector. Triggering resorting and reorganization at every single insert would lead to a performance penalty, which compromises the overall performance of the system.Therefore, an additional insert layer needs to be added, the *differential buffer*. Chapter 25 explains in detail how write performance is kept at a high level using periodic merges of the differential buffer and the main store.

The vulnerability of a column to reorganization heavily depends on the column cardinality (the number of distinct values in a dictionary). When the dictionary only has a few entries, it is most likely that a column needs to be reorganized with a new insert. However, especially with attributes of low column cardinality, e.g., gender or country, the likelihood of reorganization decreases over time, since most of the possible values for the respective column have been inserted into the dictionary already. In real world applications, the dictionary only changes occasionally after it has reached a certain size. The additional steps necessary for new unique dictionary entries will occur less frequent and therefore expensive reorganization becomes less frequent.

11.3 Self Test Questions

1. Access Order of Structures During Insert
   When doing an insert, what entity is accessed first?
   (a) The attribute vector
   (b) The dictionary
   (c) No access of either entity is needed for an insert
   (d) Both are accessed in parallel in order to speed up the process.
2. **New Value in Dictionary**
   Given the following entities:
   Old dictionary: ape, dog, elephant, giraffe
   Old attribute vector: 0, 3, 0, 1, 2, 3, 3
   Value to be inserted: lamb
   What value is the lamb mapped to in the new attribute vector?
   (a) 1
   (b) 2
   (c) 3
   (d) 4
3. **Insert Performance Variation Over Time**
   Why might real world productive column stores experience faster insert performance over time?
   (a) Because the dictionary reaches a state of saturation and, thus, rewrites of the attribute vector become less likely.
   (b) Because the hardware will run faster after some run-in time.
   (c) Because the column is already loaded into main-memory and does not have to be loaded from disk.
   (d) An increase in insert performance should not be expected.

4. **Resorting Dictionaries of Columns**
   Consider a dictionary encoded column store (without a differential buffer) and the following SQL statements on an initially empty table:
   INSERT INTO students VALUES('Daniel', 'Bones', 'USA');
   INSERT INTO students VALUES('Brad', 'Davis', 'USA');
   INSERT INTO students VALUES('Hans', 'Pohlmann', 'GER');
   INSERT INTO students VALUES('Martin', 'Moore', 'USA');
   How many complete attribute vector rewrites are necessary?

   (a) 2
   (b) 3
   (c) 4
   (d) 5

5. **Insert Performance**
   Which of the following use cases will have the worst insert performance when all values will be dictionary encoded?

   (a) A city resident database, that store all the names of all the people from that city
   (b) A database for vehicle maintenance data which stores failures, error codes and conducted repairs
   (c) A password database that stores the password hashes
   (d) An inventory database of a company storing the furnature for each room.

Chapter 12
Update

The “UPDATE” is part of SQL’s data manipulation language (DML) and is used for changing one or more tuples in a table. The UPDATE statement has the following general form:

```
UPDATE TABLE table_name
SET column_name = value
[WHERE condition]
```

Listing 12.1: Update syntax

The optional WHERE condition restricts the update to tuples that match the given condition. If no WHERE condition is specified, then all tuples in the table are updated. Logically, i.e., in relational algebra, an UPDATE statement is equivalent to a DELETE statement followed by an INSERT statement.

12.1 Update Types

Three different types of updates can be found in a typical enterprise application [Pla09]:

- *Aggregate update*: The attributes are accumulated values as part of materialized views. From our experience in enterprise systems, typically between 1 and 5 materialized aggregates are maintained for each accounting line item.
- *Status update*: Binary change of a status variable, typically with timestamps
- *Value update*: The value of an attribute changes by replacement.

12.1.1 Aggregate Updates

Most of the updates taking place in financial applications apply to complete records, containing e.g. account number, legal organization, year, etc. The system contains aggregates for these records, e.g., by account, by project, or by region. Directly reading these aggregates is faster than computing them on the fly.

12.1.2 Status Updates

Status variables (e.g. unpaid, paid) typically use a predefined set of a small number values and thus create no problem when performing an in-place update since the column cardinality does not change. It is advisable that compression of sequences (e.g. run-length encoding) in the columns is not allowed for status fields. If the automatic recording of status changes is preferable for the application, we can also use the insert-only approach, which will be discussed in Chap. 26, for these changes. In case the status variable has only two states, a null value and a time stamp can be used as values to note if the status has been set. Thus, an in-place update is fully transparent even considering temporal queries.

12.1.3 Value Updates

Since the change of an attribute in an enterprise application in most cases has to be recorded (log of changes), the insert-only approach seems to be the appropriate answer. On average only 5 % of the tuples of a financial accounting system are actually changed over a longer period of time [KKG+11]. The extra load for the differential buffer (the write-optimized store in a column store database, which handles updates and inserts) and the extra consumption of main memory are acceptable. With insert-only, we also capture the change history including time and origin of the change.

Despite the fact that typical enterprise systems are not update-intensive, by using insert-only and by not maintaining totals, we can even further reduce the number of updates, which also reduces locking issues.

12.2 Update Example

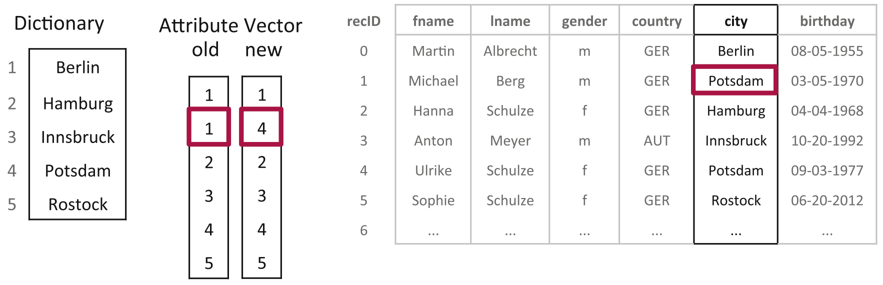
Given is the world population table. Michael Berg moves from Berlin to Potsdam. So the following query should be executed:

```
UPDATE world_population SET city = 'Potsdam'
WHERE fname = 'Michael' AND
      lname = 'Berg' AND
      city = 'Berlin';
```

Listing 12.2: Michael Berg moves from Berlin to Potsdam

recID	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Ulrike	Schulze	f	GER	Potsdam	09-03-1977
5	Sophie	Schulze	f	GER	Rostock	06-20-2012
...	...	...	...	...	...	...
$8{\times}10^9$	Zacharias	Perdopolus	m	GRE	Athen	03-12-1979

Fig. 12.1 The *world_population* table before updating



recID	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Potsdam	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Ulrike	Schulze	f	GER	Potsdam	09-03-1977
5	Sophie	Schulze	f	GER	Rostock	06-20-2012
6	...	...	...	...	...	...

Fig. 12.2 Dictionary, old and new attribute vector of the *city* column, and state of the *world_population* table after updating

Figure 12.1 shows the table before the update is executed.

Because the value “Potsdam” already exists in dictionary, the query executor can simply look up the dictionary key for the value and update the attribute vector accordingly. This is shown in Fig. 12.2.

Now, assume that Hanna Schulze moves from Hamburg to Bamberg:

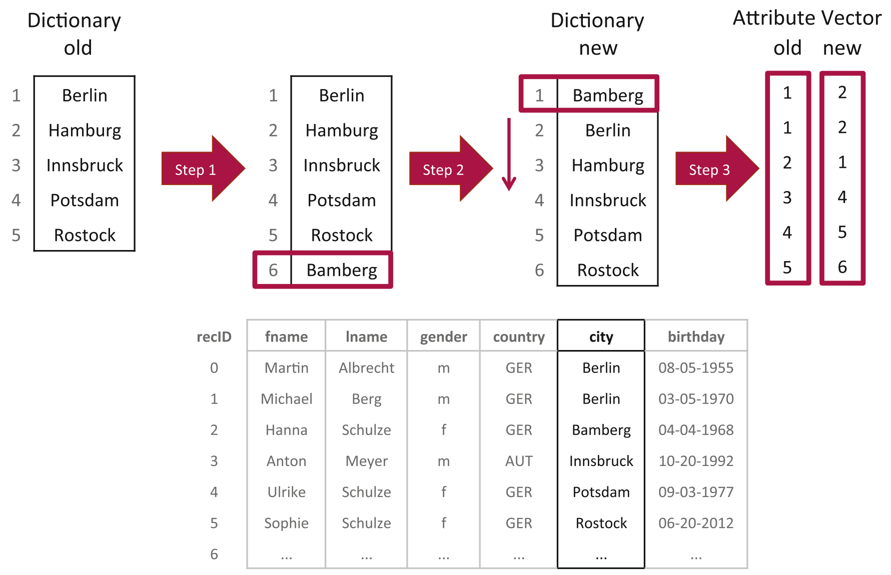
```
UPDATE world_population SET city = 'Bamberg'
WHERE fname = 'Hanna' AND
      lname = 'Schulze' AND
      city = 'Hamburg';
```

Listing 12.3: Hanna Schulze moves from Hamburg to Bamberg

This time, the value “Bamberg” is not yet in the dictionary.

The query executor performs the following actions:

1. The value “Bamberg” is appended at the end of the dictionary.
2. The dictionary is reorganized in order to maintain its sort order, which is required for fast binary search on the dictionary.
3. Every value in the attribute vector is potentially updated (i.e. replaced with the new dictionary value representing the actual value). Depending on the position



recID	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Bamberg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Ulrike	Schulze	f	GER	Potsdam	09-03-1977
5	Sophie	Schulze	f	GER	Rostock	06-20-2012
6	...	...	...	...	...	...

Fig. 12.3 Updating the *world_population* table with a value that is not yet in the dictionary

of the new value in the new-sorted dictionary, this step becomes very expensive. In our example, the complete attribute vector must be rewritten since "Bamberg" is the first item in the new-sorted dictionary.

Figure 12.3 illustrates this process.

12.3 Self Test Questions

1. **Status Update Realization**
   How do we want to realize status updates for binary status variables?
   (a) Single status field: "false" means state 1, "true" means state 2
   (b) Two status fields: "true/false" means state 1, "false/true" means state 2
   (c) Single status field: "null" means state 1, a timestamp signifies transition to state 2
   (d) Single status field: timestamp 1 means state 1, timestamp 2 means state 2.

2. **Value Updates**
   What is a "value update"?
   (a) Changing the value of an attribute
   (b) Changing the value of a materialized aggregate
   (c) The addition of a new column
   (d) Changing the value of a status variable.

3. **Attribute Vector Rewriting after Updates**
   Consider the world population table (first name, last name) that includes all people in the world: Angela Mueller marries Friedrich Schulze and becomes Angela Schulze. Should the complete attribute vector for the last name column be rewritten?

   (a) No, because 'Schulze' is already in the dictionary and only the valueID in the respective row will be replaced
   (b) Yes, because 'Schulze' is moved to a different position in the dictionary
   (c) It depends on the position: All values after the updated row need to be rewritten
   (d) Yes, because after each update, all attribute vectors affected by the update are rewritten.

References

[KKG+11] J. Krüger, C. Kim, M. Grund, N. Satish, D. Schwalb, J. Chhugani, H. Plattner, P. Dubey, A. Zeier, Fast updates on read-optimized databases using multi-core cpus. PVLDB **5**(1), 61–72 (2011)

[Pla09] H. Plattner, in *A common database approach for OLTP and OLAP using an in-memory column database*, ed. by U. Çetintemel, S. Zdonik, D. Kossmann. SIGMOD Conference (ACM, Newyork, 2009), pp. 1–2

Chapter 13
Tuple Reconstruction

13.1 Introduction

As mentioned earlier, data with matrix characteristics can be stored in linear memory either column by column (columnar layout) or row by row (row layout). The impacts were already discussed in Chap. 8 in more detail. The columnar layout is optimized for analytical set-based operations that work on many rows but for a notably smaller subset of all columns of data. The row layout shows a better performance for select operations on few single tuples. In this chapter, we discuss the operations needed for tuple reconstruction in detail and explain the influence of the different layouts on the performance of these operations. Tuple reconstruction is a typical functionality in OLTP applications. It is executed whenever more than one column is requested from the database, for example when the user in an ERP system calls the "show" or "edit" transactions for the master data object or for a document.

To explain the influence of the main memory layout organization on the performance of the tuple reconstruction operation, we have to consider the notion of the cache access and the size of the cache line. A CPU cache is a cache used by the central processing unit of a computer to reduce the average time to access memory. The cache is a smaller, faster memory which stores copies of the data from the most frequently used main memory locations. Memory cache is organized in 32 or 64 byte long cache lines. Even when reading just one byte from the memory, the CPU reads a complete cache line and places it into the cache. This characteristic of a cache will help us to estimate the response time for the tuple reconstruction operations for both layouts.

13.2 Tuple Reconstruction in Row-Oriented Databases

First, let us consider an example using the row layout. Let us assume, we need to reconstruct the tuple knowing the position of the tuple. As a first example, we take into account the following properties of the tuple:

- the size of one tuple is 200 byte;
- the number of attributes in the tuple is 6.

To estimate the result, we also need the following parameters:

- speed of the read operation from main memory: 2 MB/ms/core;
- we consider 64 byte long cache lines;
- all calculations will be done for one core per CPU. If we consider more cores, the performance will increase appropriately.

Let us calculate how much time the read operation for the tuple reconstruction will take in this case considering that the data is organized using row layout. The operation is executed relatively fast, as all attributes are stored sequentially. Considering a size of 200 bytes per tuple, we will need 4 cache accesses ($\lceil\frac{200}{64}\rceil = 4$) to read the whole tuple from main memory. The CPU reads a bit more than the size of a tuple (200 byte) in this case, because it will read a complete cache line for every cache access (in case of a row layout, the CPU will load some data of the following tuple to the cache). Thus, we read 256 byte from main memory. Considering the speed 2 MB/ms/core, we can calculate the time as described below:

$$\begin{aligned} \textit{Tuple reconstruction response time (row layout)} &= \frac{256\,\text{byte}}{2{,}000{,}000\,\text{byte/ms/core}} \\ &= 0.128\,\mu\text{s} \end{aligned}$$

13.3 Tuple Reconstruction in Column-Oriented Databases

Now let us estimate the processing time for the same operation and tuples with the same characteristics but taking into account that the data is organized in a columnar layout. The data is stored attribute-wise in this case. To reconstruct the tuple, the CPU cannot just sequentially read data from memory. It needs to do cache accesses for every attribute of the tuple required for the tuple reconstruction. Therefore, knowing the implicit recordID of the tuple to be reconstructed, it will "jump" between the attributes of the tuple to collect the values. Let us calculate how much time the read operation for the tuple reconstruction will take in this case. Considering that the reconstructed tuple has 6 attributes and that for a complete read of every attribute one cache access is required, we will need 6 cache accesses to read all attributes of the tuple from main memory. Taking into account a cache line size of 64 byte, the CPU needs to read: $64\,\text{byte} \cdot 6 = 384$ byte from main memory. The CPU reads more than the size of a tuple (200 byte) in this case, because it will read a complete cache line for every cache access (in case of a columnar layout, a CPU will load some additional attributes' values of the following tuples). Considering the speed 2 MB/ms/core, we can calculate the time as described below:

$$\begin{aligned} \textit{Tuple reconstruction response time (column layout)} &= \frac{384\,\text{byte}}{2,000,000\ \text{byte/ms/core}} \\ &= 0.192\,\mu\text{s} \end{aligned}$$

In this simple example, performance of the tuple reconstruction operation on the column layout is not significantly worse in comparison with the row layout. Nevertheless, the difference in the response time can be more significant if we consider an example for a tuple with a larger number of attributes.

13.4 Further Examples and Discussion

In reality, the number of attributes in the tables of business applications is much larger. As an example, let us calculate the response time for tuple reconstruction with the following characteristics:

- The size of one tuple is 3,200 byte. For the response time of the column layout calculation, we also consider that for every attribute of the tuple, one cache access is enough to read the whole attribute of the tuple.
- The number of attributes in the tuple is 100.

Let us calculate response times for the tuple reconstruction operation for both layouts considering the same CPU characteristics that were described in the example above.

Row layout: 50 cache accesses are required for a CPU to read the whole tuple: $50 \cdot 64\,\text{byte} = 3{,}200$ byte.

$$\begin{aligned} \textit{Tuple reconstruction response time (row layout)} &= \frac{3{,}200\,\text{byte}}{2{,}000{,}000\,\text{byte/ms/core}} \\ &= 1.6\,\mu\text{s} \end{aligned}$$

Columnar layout:

100 cache accesses are required in case of the columnar layout to read the attributes of the tuple: $100 \cdot 64\,\text{byte} = 6{,}400$ byte;

$$\begin{aligned} \textit{Tuple reconstruction response time (row layout)} &= \frac{6{,}400\,\text{byte}}{2{,}000{,}000\,\text{byte/ms/core}} \\ &= 3.2\,\mu\text{s} \end{aligned}$$

This example shows how the number of attributes of the tuple can influence the response time for both layouts. The performance for tuple reconstruction of the columnar layout will become progressively worse in comparison to the row store when we increase the number of a tuple's attributes and request all attributes.

We can conclude that it is important to select only the necessary fields of a tuple. This way, the potential disadvantage of a columnar layout can be reduced to a minimum. As you can imagine, queries that really need 100 columns are seldom.

13.5 Self Test Questions

1. **Tuple Reconstruction on the Row Layout: Performance**
   Given a table with the following characteristics:

   - Physical storage in rows
   - The size of each field is 34 byte
   - The number of attributes is 9
   - A cache line has 64 byte
   - The CPU processes 2 MB per millisecond.

   Calculate the time required for reconstructing a full row. Please assume the following conversions: $1\,\text{MB} = 1{,}000\,\text{kB}$, $1\,\text{kB} = 1{,}000\,\text{B}$

   (a) $\approx$ 0.1 µs
   (b) $\approx$ 0.275 µs
   (c) $\approx$ 0.16 µs
   (d) $\approx$ 0.416 µs

2. **Tuple Reconstruction on the Column Layout: Performance**
   Given a table with the following characteristics:

   - Physical storage in columns
   - The size of each field is 34 byte
   - The number of attributes is 9
   - A cache line has 64 byte
   - The CPU processes 2 MB per millisecond.

   Calculate the time required for reconstructing a full row. Please assume the following conversions: $1\text{MB} = 1{,}000\,\text{KB}, 1\,\text{kB} = 1{,}000\,\text{B}$

   (a) $\approx$ 0.16 µs
   (b) $\approx$ 0.145 µs
   (c) $\approx$ 0.288 µs
   (d) $\approx$ 0.225 µs

3. **Tuple Reconstruction in Hybrid Layout**
   A table containing product stock information has the following attributes:

   Warehouse (4 byte); Product Id (4 byte); Product Name Short (20 byte); Product Name Long (40 byte); Self Production (1 byte); Production Plant (4 byte); Product Group (4 byte); Sector (4 byte); Stock Volume (8 byte); Unit of Measure (3 byte); Price (8 byte); Currency (3 byte); Total Stock Value (8 byte); Stock Currency (3 byte)

The size of a full tuple is 114 byte.
The size of a cache-line is 64 byte.

The table is stored in main memory using a hybrid layout. The following fields are stored together:

- Stock Volume and Unit of Measure;
- Price and Currency;
- Total Stock Value and Stock Currency;

All other fields are stored column-wise.
Calculate and select from the list below the time required for reconstructing a full tuple using a single CPU core. Please assume the following conversions: $1\,\text{MB} = 1{,}000\,\text{kB}, 1\,\text{kB} = 1{,}000\,\text{B}$

(a) ≈ 0.352 μs
(b) ≈ 0.020 μs
(c) ≈ 0.061 μs
(d) ≈ 0.427 μs

4. **Comparison of Performance of the Tuple Reconstruction on Different Layouts**
A table containing product stock information has the following attributes: Warehouse (4 byte); Product Id (4 byte); Product Name Short (20 byte); Product Name Long (40 byte); Self Production (1 byte); Production Plant (4 byte); Product group (4 byte); Sector (4 byte); Stock Volume (8 byte); Unit of Measure (3 byte); Price (8 byte); Currency (3 byte); Total Stock Value (8 byte); Stock Currency (3 byte)

The size of a full tuple is 114 byte.
The size of a cache-line is 64 byte.

Which of the following statements are true?

(a) If the table is physically stored in column layout, the reconstruction of a single full tuple consumes ≈0.192 μs using a single CPU core
(b) If the table is physically stored in row layout, the reconstruction of a single full tuple consumes ≈128 ns using a single CPU core
(c) If the table is physically stored in column layout, the reconstruction of a single full tuple consumes ≈448 nanoseconds using a single CPU core
(d) If the table is physically stored in row layout, the reconstruction of a single full tuple consumes ≈0.64 μs using a single CPU core.

Chapter 14
Scan Performance

14.1 Introduction

In this chapter, we discuss the performance of the scan operation. Scan operations require the values of a single attribute or a small set of attributes but go through the whole dataset. Scan operations search one or more attributes for a certain value. Unless the table is sorted by the attribute to scan, a scan has to iterate over all lines and returns those lines, which fulfill the search predicate (e.g. "SELECT * FROM world_population WHERE lastname = 'Smith'"). As in Chap. 13, we will discuss the influence of different layouts (row and column) and different approaches on the performance of the scan operation. We compare the following three approaches:

- full table scan in row layout
- stride access for the selected attributes in row layout
- full column scan in columnar layout

In the following examples, the world population table already known from previous chapters is scanned. To recap, the table has the following properties:

- 8 billion tuples
- tuple size of 200 byte
- table size of 8 billions $\cdot$ 200 byte $=$ 1.6 TB
- attributes: first name, last name, gender, country, city, birthday
- all attributes have a fixed length

In addition, the previous assumptions for the response time calculations are used:

- bandwidth of read operations from main memory: 2 MB/ms/core
- cache line size of 64 byte

In the first example, the scan operation will help to answer the question: "How many women are in the world?".

The target column that has to be scanned to answer this question is "Gender", which has two possible distinct values. For simplicity, the calculations of the scan performance are done using a single core. When performing the scan operation, each row of the table is independent from all other rows. Consequently the scan operation can be efficiently parallelized and scales nearly linearly.

14.2 Row Layout: Full Table Scan

Having the data organized in a row layout, the first and most obvious approach to find the exact number of women in the world is to scan sequentially through all rows to read the gender attribute. We have seen this behavior for software that uses Object-Relational Mapping (ORM) and does calculations on the application side. Having to retrieve whole data sets in order to create the needed objects to interact with, results in a full table scan. During this operation, the CPU will read 1.6 TB from main memory. Taking into account the scan speed of 2 MB/ms per core, we can calculate the runtime on one core as follows:

$$\textit{Full table scan response time on } 1 \textit{ core} = \frac{1.6\text{ TB}}{2\text{ MB/ms}} = 800\,\text{s}$$

We would have to wait for more then 10 min to get the answer to our question. In order to achieve a better performance we have to look for optimizations. An obvious and simple solution is to compute the question in parallel on multiple cores and CPUs. We could do a vertical partitioning of the table and let the processing units execute the scan operation on the table parts in parallel.

Let us have a quick look on an example for a quad core CPU for that. The scan speed for a quad core CPU can be calculated as follows: 4 cores $\cdot$ 2 MB/ms/core = 8 MB/ms. The full table scan response time is 1.6 TB/8 MB/ms = 200 s. Even with four cores the query execution takes several minutes.

Another approach that could help to increase the performance of the scan operation is to take advantage of the in-memory database and read the gender fields with direct access. On disk based databases, only pages instead of single attributes are usually directly accessible (see Sect. 4.4). The results for this approach are calculated and discussed in the next Sect. 14.3.

14.3 Row Layout: Stride Access

Assume we still use row layout to store the data in memory. But now, instead of scanning the whole table and reading all table fields from main memory, we read the target field via direct access. To scan all gender fields the CPU does 8 billions cache accesses, one access for each tuple. Assuming the cache line size is 64 byte, and considering the fact that a CPU will read exactly 64 byte during each cache access independently of the gender field size, we can calculate the data size that is read from main memory during the whole scan operation:

$$data\ volume \ = \ 8\ billion \cdot 64\ byte \approx 512\,\mathrm{GB}$$

Taking into account the scan speed, we get the following the response time for one core:

$$Stride\ access\ response\ time\ on\ 1\ core = 512\ \mathrm{GB}\ /\ 2\ \mathrm{MB/ms} = 256\,\mathrm{s}$$

The result is better than that for the full table scan, but answering the question still takes several minutes. However, there are further opportunities to optimize the scan speed for our initial question.

The next Sect. 14.4 will discuss the effects of using a columnar data layout.

14.4 Columnar Layout: Full Column Scan

When using a columnar layout, the data is stored attribute-wise in main memory. This fact leads us to the following conclusions:

- As attributes of the same types are stored together, effective compression algorithms can be used to reduce the data volume that is stored in memory and that has to be transferred between main memory and CPU.
- As values of the same attribute are stored consecutively, the probability that the next accessed item has already been loaded as part of the same cache line, is relatively high depending on the length of the compressed values. The shorter the values, the higher the probability.

Consequently, two aspects of columnar layouts can be leveraged: scanning only target fields and reading compressed values. Both aspects reduce the data volume transferred between main memory and CPU and consequently reduce response times. In our example, the CPU will scan through the “Gender” field.

Considering this column is dictionary-encoded as described in Chap. 6, only one bit is necessary to encode the two possible values ‘m’ and ‘f’. As before, we can calculate the data volume to be read from main memory using the size of the attribute and the number of tuples: 8 billion $\cdot$ 1 bit $\approx$ 1 GB, which leads to a full column scan response time of 1 GB/2 MB/ms/core $=$ 0.5 s on one core.

The result shows a significant difference in the performance in comparison with both presented approaches for the row layout. Further taking into account the opportunity to use several cores and to execute the scan operation in parallel, using the column layout we can speed up the answer to our example-question even more. While our example query, which is using only one attribute and is posed against a vast number of tuples might not be the common use case, it can be stated that in analytical workloads the general circumstances are favorable for this approach, as we have already seen in Chap. 3. Queries against huge data volumes that operate on a small number of columns are characteristic for analytical and transactional enterprise applications.

14.5 Additional Examples and Discussion

In our previous example, we considered an almost "perfect" case. With 1-bit length, the gender attribute was compressed to the minimum for dictionary-encoded values. This fact decreased the data volume to be transferred between CPU and main memory. Of course, the outcome of the performance calculation depends on the size of the scanned fields. For larger fields, the CPU will need to scan through a higher data volume and less values will fit into one cache line.

To compare the results with another attribute, let us take the same table from the first example and calculate the response time for the full column scan operation on the column "Birthday". This column has more distinct values than the "Gender" field.

Considering that every value (i.e. valueID of a compressed value in the "Birthday" column) has a size of 2 byte, we can calculate the transferred data volume and appropriate response time as follows:

- data volume to be read from main memory = 8 billion $\cdot$ 2 byte $\approx$ 16 GB
- full column scan response time = 16 GB/ 2 MB/ms/core = 8 s (with one core)

To summarize the calculations performed above, we can conclude, that the following parameters of a CPU and a scanned table influence scan performance:

- cache utilization
- memory bandwidth
- number of processing units
- number of tuples in the table (table cardinality)
- used compression
- used layout: column or row layout

The example calculations in this chapter show a significant speed up of the scan performance when switching from row to dictionary-encoded column layout. While the columnar layout with its higher data density better utilizes the CPU caches, we would also like to note that it enables further optimizations, e.g. the usage of SIMD/SSE operations (see Sect. 17.1.2).

14.6 Self Test Questions

1. **Loading Dictionary-Encoded Row-Oriented Tuples**
   Consider the example in Sect. 14.2 with dictionary-encoded tuples. In this example, each tuple has a size of 32 byte. What is the time that a single core processor needs to scan the whole world_population table if all data is stored in a dictionary-encoded row layout?

   (a) 128 s
   (b) 256 s
   (c) 64 s
   (d) 96 s

Chapter 15
Select

In this chapter, we describe how an application can extract data that was once stored in the database (execution of the `SELECT` statement).

The `SELECT` statement is a combination of multiple relational operations, mainly selection, projection, and Cartesian product. We focus on the implications of SanssouciDB's column-orientated data layout.

15.1 Relational Algebra

Three different basic operations of the relational algebra can be used to create SQL's `SELECT` statement. These are the Cartesian product, the projection and the selection.

15.1.1 Cartesian Product

The Cartesian product (or cross product) is a binary operation, taking two relations R_1 and R_2 to produce the result $R_1 \times R_2$. Those are having n_{R_1} and n_{R_2} attributes and a cardinality of $|R_1|$ and $|R_2|$. As a result, a new relation R_3 with $n_{R_3} = n_{R_1} + n_{R_2}$ and $|R_3| = |R_1| \cdot |R_2|$ tuples is returned. After both relations were combined, projections and selections can be applied to reduce the size of the result set. Database systems tend to use join operations to reduce the size of intermediate results, as described in Chap. 19.

15.1.2 Projection

Projection is used to delete or permute the attributes of its input relation. Looking at the logical layout of a table, projection is a "vertical" operator. It can be written as $\pi_{j_1,\ldots,j_n}(R)$, with j_1 to j_n being an ordered sequence representing the ordered sequence of attributes of R contained in the projections result. Using a column-oriented data layout, only columns that are part of the projection (and those attributes used in predicates, which are not necessarily projected) need to be read by the database. Thus, query processing consumes fewer resources if only a subset of the entire set of attributes needs to be touched.

15.1.3 Selection

When the data stored within a relation needs to be filtered by some criteria, the selection is used. The selection, written as σ, is a "horizontal" operator. It evaluates an expression (predicate) consisting of a and b that are combined via a binary operation θ. While a and b can be attribute names, specified, or calculated values, θ represents any binary operation (e.g., equals, greater, smaller) that evaluates to "true" or "false". Only tuples of the relation with a positive evaluation ("true") of θ are included into the result set.

15.2 Data Retrieval

In most applications, SELECT is a commonly used command.
The typical SQL SELECT statement can be defined as

$$\begin{array}{ll} \text{SELECT} & \pi_{j_1,\ldots,j_n}(R) \\ \text{FROM} & R \\ \text{WHERE} & \sigma_{a\theta b}(R) \end{array}$$

Because SQL presents a declarative description of the result requested from the database, an ordered set of execution steps is required to extract the data from the database, a so-called query execution plan. For each SQL query, multiple execution plans can exist that deliver the same results with differing performance. Query optimizers are used to calculate the cost of different query execution plans. Relying on cost models and heuristics used within the optimizer an effective plan is chosen. The goal is to reduce the size of the result set as early as possible, e.g., by

- applying selections as early as possible
- ordering sequential selections so that the most restrictive ones are executed first
- ordering joins corresponding to their tables' cardinalities (smallest tables are used first)

id	fname	lname	country	gender
2394	Gianluigi	Buffon	Italy	m
3010	Lena	Gercke	Germany	f
3040	Mario	Balotelli	Italy	m
3949	Manuel	Neuer	Germany	m
4902	Lukas	Podolski	Germany	m
20102	Klaas-Jan	Huntelaar	Netherlands	m

Fig. 15.1 Example database table *world_population*

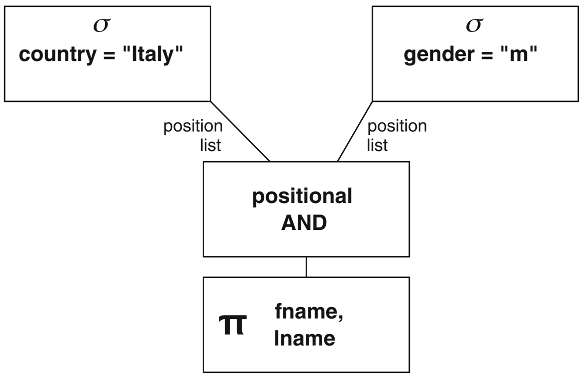


Fig. 15.2 Example query execution plan for SELECT statement

As a concrete example we use the table shown in Fig. 15.1 and execute the following SELECT statement that retrieves the first names and last names of male Italians from the world population table:

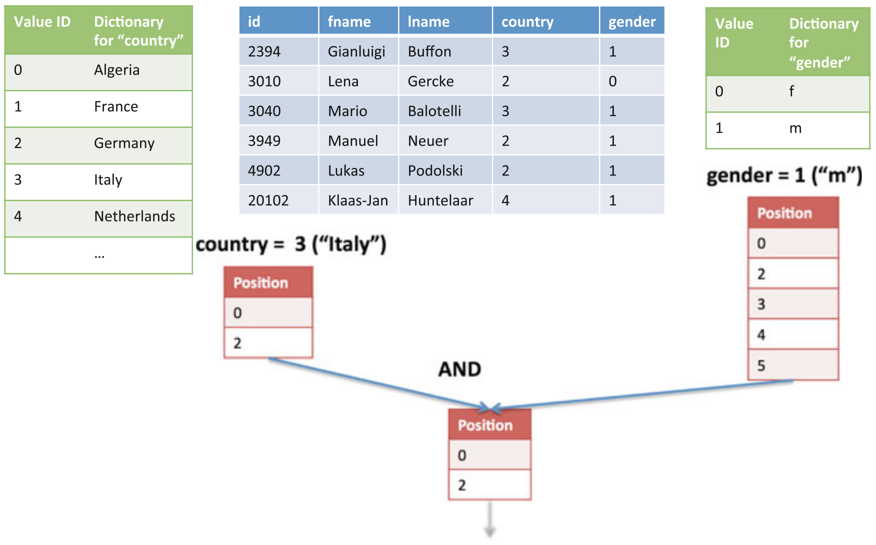
Retrieve first and last names for male Italiens

```
SELECT   fname, lname
FROM     world_polulation
WHERE    country = 'Italy'AND gender = 'm'
```

The corresponding query execution plan for that particular SQL query could look like shown in Fig. 15.2.

The query plan would than be executed in the database, as shown in Fig. 15.3. Database operations with independent inputs can be executed in parallel.

Because of SanssouciDB's dictionary encoding, a dictionary lookup is used to find the valueIDs for "Italy" and "m", in our example 3 and 1. Afterwards the attribute vectors of country and gender are scanned and position lists identifying valid tuples are created. Those lists are intersected, resulting in a new list containing the positions of all tuples fulfilling the two selections.



Value ID	Dictionary for "country"
0	Algeria
1	France
2	Germany
3	Italy
4	Netherlands
	...

id	fname	lname	country	gender
2394	Gianluigi	Buffon	3	1
3010	Lena	Gercke	2	0
3040	Mario	Balotelli	3	1
3949	Manuel	Neuer	2	1
4902	Lukas	Podolski	2	1
20102	Klaas-Jan	Huntelaar	4	1

Value ID	Dictionary for "gender"
0	f
1	m

Fig. 15.3 Execution of the created query plan

15.3 Self Test Questions

1. **Table Size**
   What is the table size if it has 8 billion tuples and each tuple has a total size of 200 byte?
   (a) ≈ 12.8 TB
   (b) ≈ 12.8 GB
   (c) ≈ 2 TB
   (d) ≈ 1.6 TB
2. **Optimizing SELECT**
   How could the performance of SELECT statements be improved?
   (a) Reduce the number of indices
   (b) By using the FAST SELECT keyword
   (c) Order multiple sequential select statements from low selectivity to high selectivity
   (d) Optimizers try to keep intermediate result sets large for maximum flexibility during query processing.

3. **Selection Execution Order**
   Given is a query that selects the names of all German women born after January 1, 1990 from the world_population table (contains data about all people in the world). In which order should the query optimizer execute the selections? Assume a sequential query execution plan.

   (a) country first, birthday second, gender last
   (b) country first, gender second, birthday last
   (c) gender first, country second, birthday last
   (d) birthday first, gender second, country last.

4. **Selectivity Calculation**
   Given is the query to select the names from German men born after January 1, 1990 and before December 31, 2010 from the world population table (8 billion people). Calculate the selectivity.

   Selectivity = number of tuples selected / number of tuples in the table

   Assumptions:

   - there are about 80 million Germans in the table
   - males and females are equally distributed in each country
   - there is an equal distribution between all generations from 1910 until 2010

   (a) 0.001
   (b) 0.005
   (c) 0.1
   (d) 1

5. **Execution Plans**
   For any one SELECT statement...

   (a) there always exist exactly two execution plans, which mirror each other
   (b) exactly one execution plan exists
   (c) several execution plans with the same result set, but differing performance may exist
   (d) several executions plans may exist that deliver differing result sets.

Chapter 16
Materialization Strategies

SQL is the most common language to interact with databases. Users are accustomed to the table-oriented output format of SQL. To provide the same data interfaces as known from row stores in column stores, the returned results have to be transformed into tuples in row format. The process of transforming encoded columnar data into row-oriented tuples is called materialization.

Especially for column-oriented databases with lightweight compression, an appropriate materialization strategy is essential. Abadi et al. [AMDM07] analyzed different materialization strategies for column-oriented databases. Depending on the storage technique (e.g. compressed vs. uncompressed data, dictionary encoding vs. no dictionary encoding), different materialization strategies can be superior. Grund et al. [GKK+11] analyzed database operators and the impact of materialization strategies for intermediate results, in particular for dictionary-encoded columnar data structures.

16.1 Aspects of Materialization

Abadi et al. [AMDM07] divide the topic of materialization into two aspects, the execution of materialization and the time of materialization. The execution can be divided into parallel and pipelined materialization. The advantages and disadvantages of both approaches are discussed in detail in [GKK+11] and are not part of this learning material. All the following examples use a non-pipelined execution, where each operator is independent from the others.

There are two different strategies concerning the time aspect of materialization: early and late materialization. Early materialization describes the strategy, where data is decoded early (using dictionary lookups) during the query execution. For example, consider a dictionary-encoded string column. It contains the attribute vector of integer values and the sorted dictionary of strings. Here, the actual string replaces the positional integer value representing the corresponding dictionary position early. Hence, a row-oriented tuple representation is created early on.

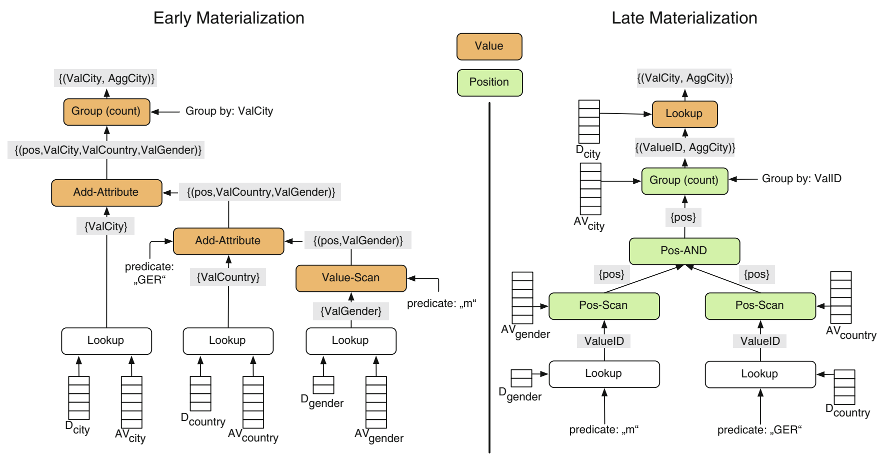


Fig. 16.1 Example comparison between early and late materialization

With the late materialization strategy, column-orientation and the positional information instead of the actual value are used as long as possible during query execution. Ideally, the row-oriented tuple will be materialized in the very last step before returning the result to the user.

Figure 16.1 shows in an example where actual values and positions are used in early and late materialization.

In many cases, late materialization can improve the performance for column stores, especially when light-weight compression techniques are used [AMDM07]. The following sections will discuss both strategies based on an example query.

16.2 Example

To discuss the difference between early and late materialization, we will examine the query “List the number of male inhabitants per city in Germany”, see SQL query in Listing 16.1.

```
SELECT city , COUNT(*)
FROM world_population
WHERE gender = "m"
    AND country = "GER"
GROUP BY city
```

Listing 16.1: Example query

In both following examples, one strategy will be used throughout the whole query execution for exemplary purposes, even though a combination is often advantageous in real world situations. Example data of the *World Population Table* which is used in the query is shown in Fig. 16.2.

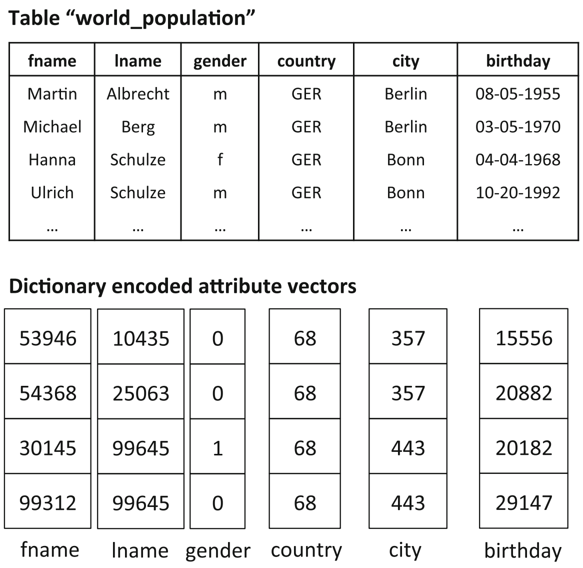


Table "world_population"

fname	lname	gender	country	city	birthday
Martin	Albrecht	m	GER	Berlin	08-05-1955
Michael	Berg	m	GER	Berlin	03-05-1970
Hanna	Schulze	f	GER	Bonn	04-04-1968
Ulrich	Schulze	m	GER	Bonn	10-20-1992
...	...	...	...	...	...

Dictionary encoded attribute vectors

fname	lname	gender	country	city	birthday
53946	10435	0	68	357	15556
54368	25063	0	68	357	20882
30145	99645	1	68	443	20182
99312	99645	0	68	443	29147

Fig. 16.2 Example data of table "*world_population*"

16.3 Early Materialization

When early materialization is used as the materialization strategy throughout the complete query, all required columns are materialized first. In our case, required columns are all columns that are used as predicates in the query (i.e., *country* and *gender*), as well as all columns that are part of the result (i.e., *city*). Dictionary lookups are performed for each of these columns using the valueIDs in the corresponding attribute vectors. For the gender column, the result of these lookups is the vector `{ValGender}` with the actual values (see Fig. 16.3a).

The next step is to scan the intermediate vector `{ValGender}` for the gender predicate 'm'. To all qualifying lines the corresponding position is added and copied to the intermediate vector `{(pos, ValGender)}` (see Fig. 16.3b).

In the next step, the columns are combined as shown in Fig. 16.4. Hereby, the `{ValCountry}` vector is added to the intermediate result `{(pos, ValGender)}` while scanning for the predicate value 'GER'.

The final step is to aggregate and return the requested data of the SQL query. For that the intermediate result `{(pos, ValGender‘ ValCountry‘ ValCity)}` is grouped by ValCity and aggregated. The result is `{(ValCity‘ AggCity)}`, as shown in Fig. 16.5.

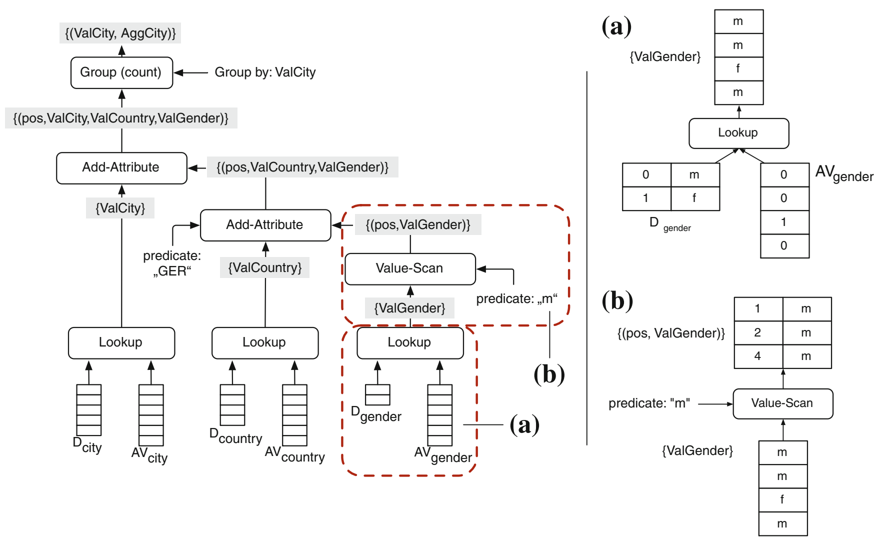


Fig. 16.3 Early materialization: materializing column via dictionary lookups and scanning for predicate

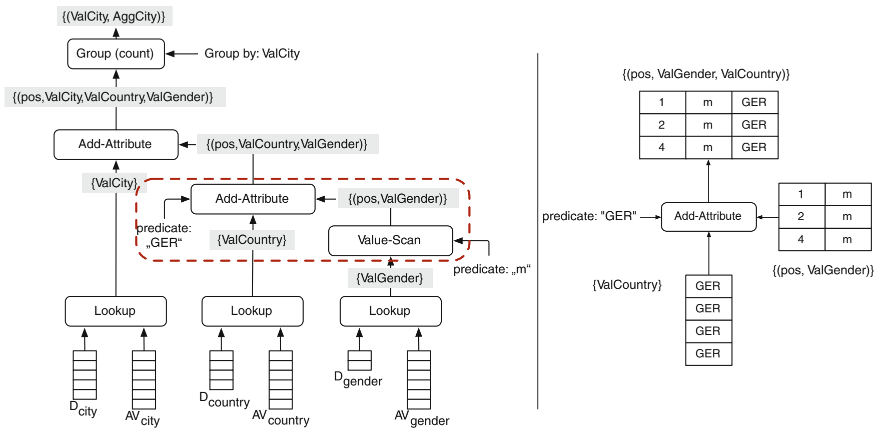


Fig. 16.4 Early materialization: scan for constraint and addition to intermediate result

16.4 Late Materialization

Instead of materializing the values of the dictionary lookup early (as done in the early materialization strategy), the dictionary-encoded value (`valueID`) contained in the attribute vector is being used. Ideally, the lookup into the dictionary for materialization is performed in the very last step before returning the result.

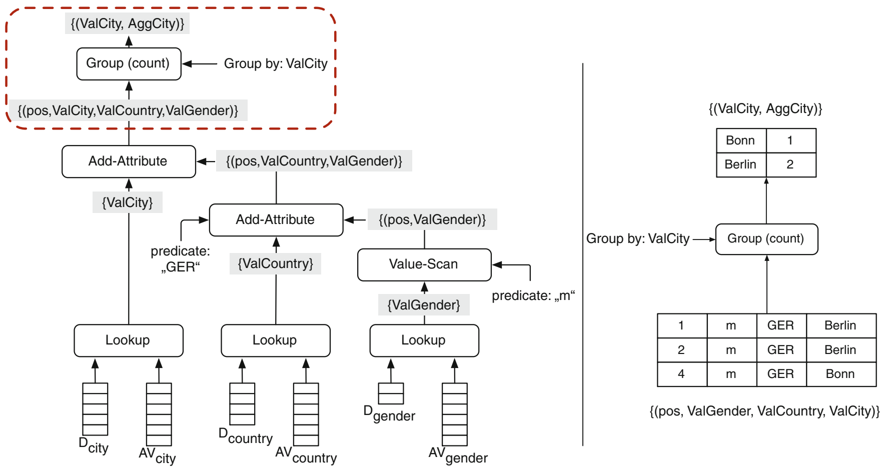


Fig. 16.5 Early materialization: group by *ValCity* and aggregation

Figure 16.6 shows the first step. Here, the predicates *gender* = ‘*m*’ and *country* = ‘*GER*’ are used for the lookup using the corresponding dictionaries. The outcome is a vector of dictionary positions (valueIDs) per column that qualify for the given predicates. Notice that the dictionary for the column *city* is not accessed, since it is not required for the actual processing of the query right now. Only the `valueID` of the columns *gender* and *country* are looked up, as they are required for the succeeding scan operation.

Even though the visualization of the late materialization strategy implies a parallel execution of the lookups, the execution can also be done sequentially. Actually, with

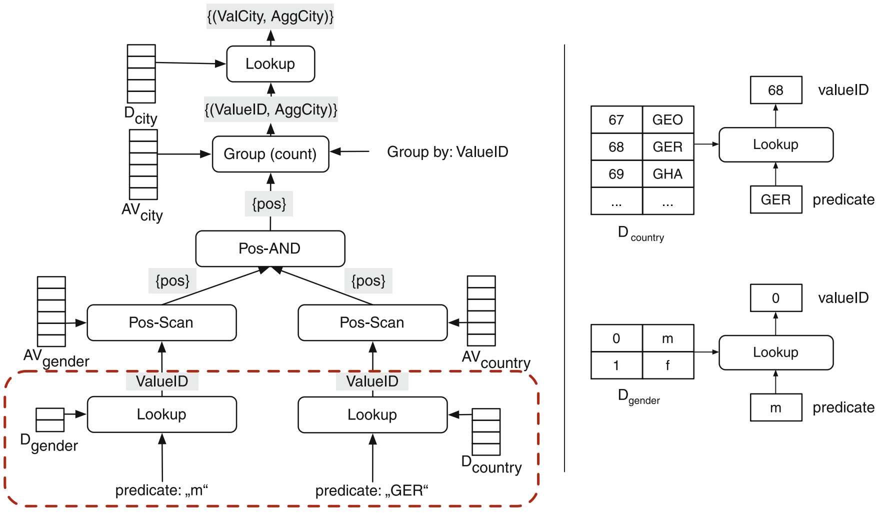


Fig. 16.6 Late materialization: lookup predicate values in dictionary

a predicate as *country* = '*GER*', for which less than 2 % of the world population qualify, a sequential execution is advantageous (see Chap. 15 for more details).

Figure 16.7a shows the scan phase. With the valueIDs from the first step, now the attribute vectors are scanned. The position of each matching valueID in the attribute vector is added to the output vector of this step ({*pos*}). The merge of these positional lists is shown in Fig. 16.7b. Here, each value that is existent in both vectors is appended to the result vector of this step.

Figure 16.8a shows the group by operation. Hereby, the intermediate vectors are taken to group the positions in {*pos*} by the valueIDs in the city attribute vector and add the count of each city to the output vector. In the last step the actual lookup of the city valueIDs is performed, as shown in Fig. 16.8b.

Compared to the early materialization strategy, the late materialization strategy might have to perform an additional lookup, e.g. when the gender would also be part of the result. This penalty can diminish the advantages, for example when many columns have to be materialized (consequently many dictionary lookups, what typically occurs when using `'SELECT*'`) or when the result set is very large (i.e., many output rows).

In general, the question to which extend—and even if—late materialization is in favor of early materialization depends on many variables like the used query operations and selectivity, among others [GKK+11].

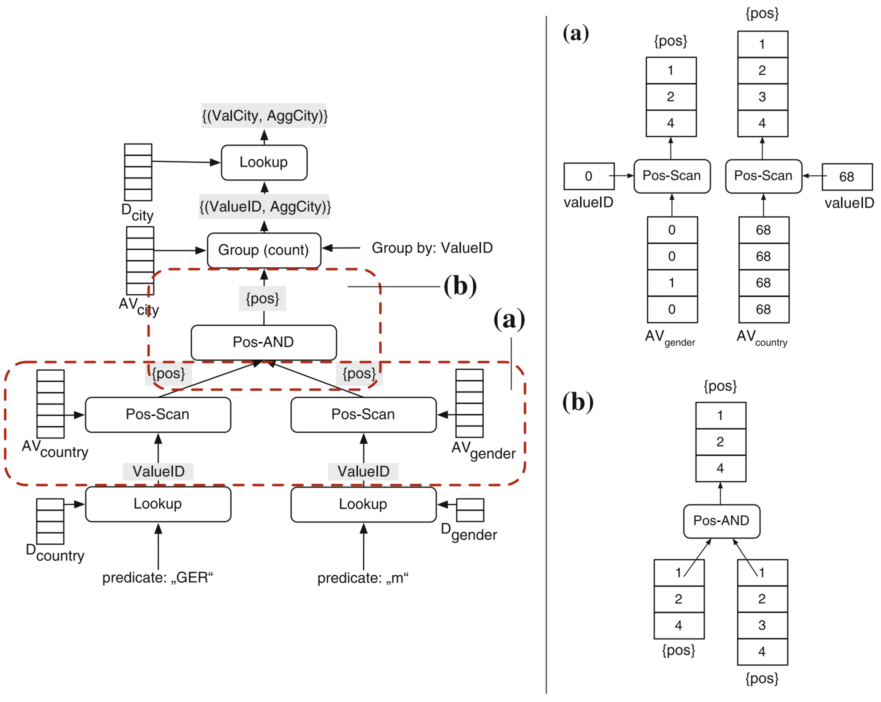


Fig. 16.7 Late materialization: scan and logical *AND*

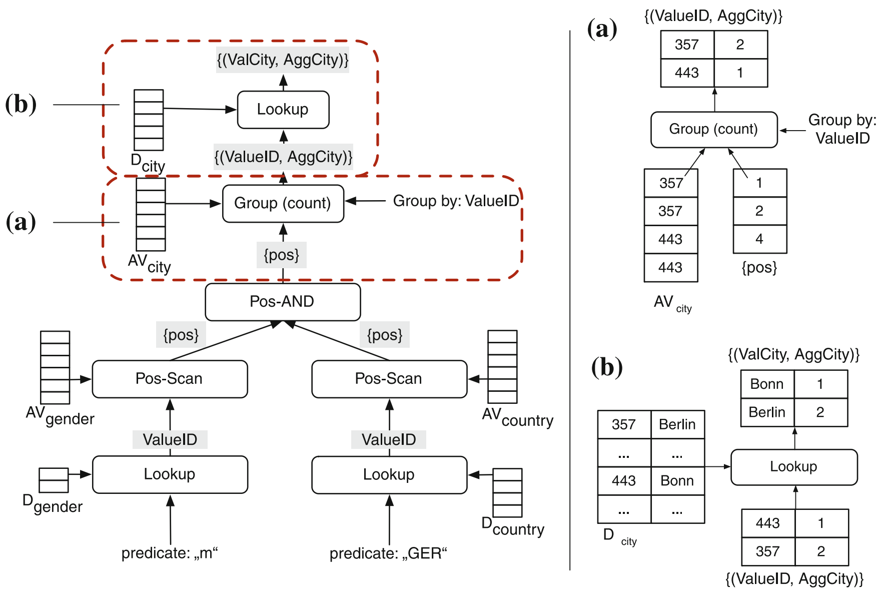


Fig. 16.8 Late materialization: filtering of attribute vector and dictionary lookup

16.5 Self Test Questions

1. **Which Strategy is Faster?**
   Which materialization strategy—late or early materialization—provides the better performance?
   (a) Early materialization
   (b) Late materialization
   (c) Depends on the characteristics of the executed query
   (d) Late and early materialization always provide the same performance.
2. **Disadvantages of Early Materialization**
   Which of the following statements is true?
   (a) The execution of an early materialized query plan can not be parallelized
   (b) Whether late or early materialization is used is determined by the system clock
   (c) Early materialization requires lookups into the dictionary, which can be very expensive and are not required when using late materialization
   (d) Depending on the persisted value types of a column, using positional information instead of actual values can be advantageous (e.g. in terms of cache usage or SIMD execution).

References

[AMDM07] D.J. Abadi, D.S. Myers, D.J. DeWitt, S. Madden, Materialization strategies in a column-oriented dbms, in *ICDE*, ed. by R. Chirkova, A. Dogac, M.T. Ã-zsu, T.K. Sellis (IEEE, New York, 2007), pp. 466–475 Url: http://dblp.uni-trier.de/db/conf/icde/icde2007.html#AbadiMDM07

[GKK+11] M. Grund, J. Krueger, M. Kleine, A. Zeier, H. Plattner, Optimal Query Operator Materialization Strategy for Hybrid Databases, in *DBKDA* (IARIA, Cancun, 2011), pp. 169–174

Chapter 17
Parallel Data Processing

In the following, we discuss how to achieve parallelism in in-memory and traditional database management systems. Pipelined parallelism and data parallelism are two approaches to speed up query processing.

In pipelined parallelism, the next operator already starts while the current operator is not finished but has already produced partial results. Thus, the execution time of operators partly overlaps. For example, consider a `JOIN` and `SORT` operator involving the evaluation of a predicate. Each operator performs its tasks and if first results become available, they are used by the next operator until the end of the pipeline is reached as depicted in Fig. 17.1 on the left.

In data parallelism, the data set is partitioned so that the operators of a query work on individual parts of the data set in parallel. Afterwards, the results of the parallel streams are merged to the complete result set (see Fig. 17.1 on the right). The query plan becomes more complex since all operators are executed on each data partition individually and a merge operation is added.

In database management systems, further aspects can be considered for parallelization. We distinguish between intra-query and inter-query parallelism. Intra-query parallelism addresses parallelization of operators within a query, i.e., the query looks like a single operation, but it is parallelized internally, e.g., by spawning multiple threads and using data parallelism. Inter-query parallelism addresses the aspect to schedule multiple queries to execute them in parallel. This may also results in parallel data access if queries use the same data.

17.1 Hardware Layer

Parallel processing of data is an essential aspect of achieving high performance for in-memory database systems. But what are the reasons for using parallelization instead of a single CPU core running at a tremendous high frequency such as 1 PHz? We want to discuss this question in the remainder of this section.

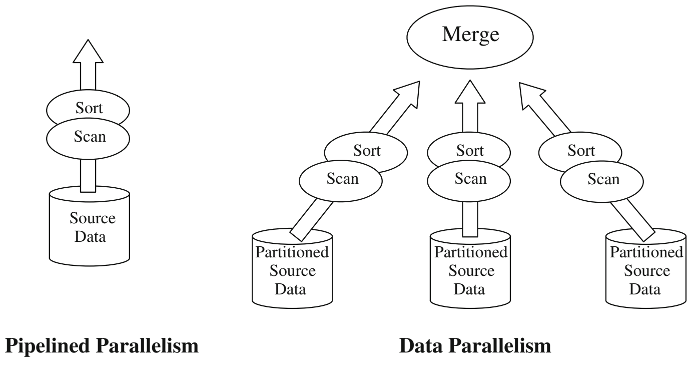


Fig. 17.1 Pipelined and data parallelism

17.1.1 Multi-Core CPUs

Ideally, a modern computer would consist of a single CPU core running at 1 PHz and huge persistent integrated main memory as shown in Fig. 17.2. Reality looks different, though. Nowadays, we typically have multiple CPU cores on one CPU die. Furthermore, modern server systems consist of multiple CPUs. This multiplies the number of cores.

The reasons for that multi-core development are buried in the hardware developments of the last decade. The assumption that the number of transistors doubles every 18 month, known as Moore's law, is still valid [Moo65]. However, the operating frequency of the transistors cannot be increased infinitely. For example, with increasing frequency the ratio of heat loss increases. As a result, the

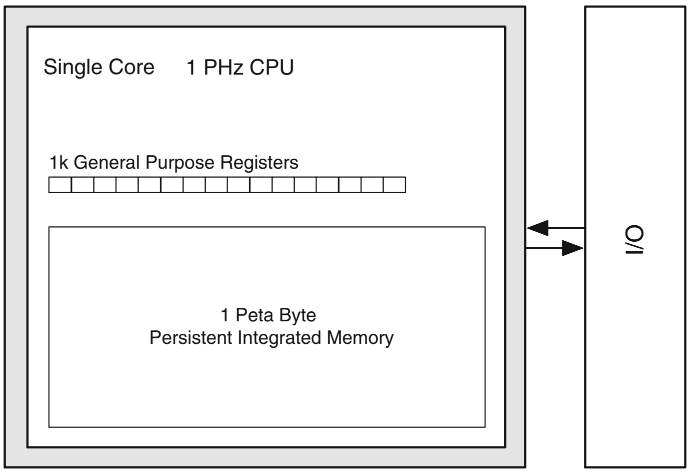


Fig. 17.2 The ideal hardware?

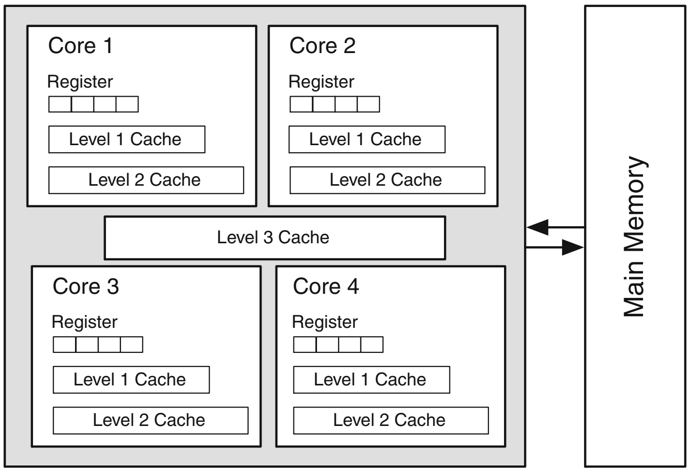


Fig. 17.3 A multi-core processor consisting of 4 cores

energy efficiency degrades while additional power is required to cool transistors. Hardware vendors proved that using multiple CPU cores operated at a lower frequency, e.g., 2.4–2.7 GHz, increases efficiency while keeping cooling requirements at an adequate level. For example, Fig. 17.3 depicts the conceptual architecture of single CPU consisting of four cores. Combining multiple CPUs within a single server is shown in Fig. 17.4 and the combination of multiple servers to form a more powerful data processing system is depicted in Fig. 17.5.

17.1.2 Single Instruction Multiple Data

The foundation of parallelization can directly be found within the CPU, i.e., data processing can be parallelized using the Single instruction multiple data (SIMD) paradigm. In contrast to traditional Reduced Instruction Set Computing (RISC) CPUs, SIMD parallelization builds on the use of so-called vectorized operators. These operators are directly implemented in the CPU to perform operations on multiple data words in specialized CPU registers in parallel. Computer graphics makes use of Streaming SIMD Extensions (SSE) instructions that operate on either 128 or 256 bit wide registers. For example, in one 128 bit register you can store two 64 bit values to perform a Parallel add (PADD) as depicted in Fig. 17.6. Thus, two calculations can be processed within one instruction step instead of Scalar add (SADD) where one calculation is performed at a time.

For instance, let us consider the aggregation of outstanding items. Using PADD reduces the time to sum up the individual items dramatically by summing up multiple items in a single instruction.

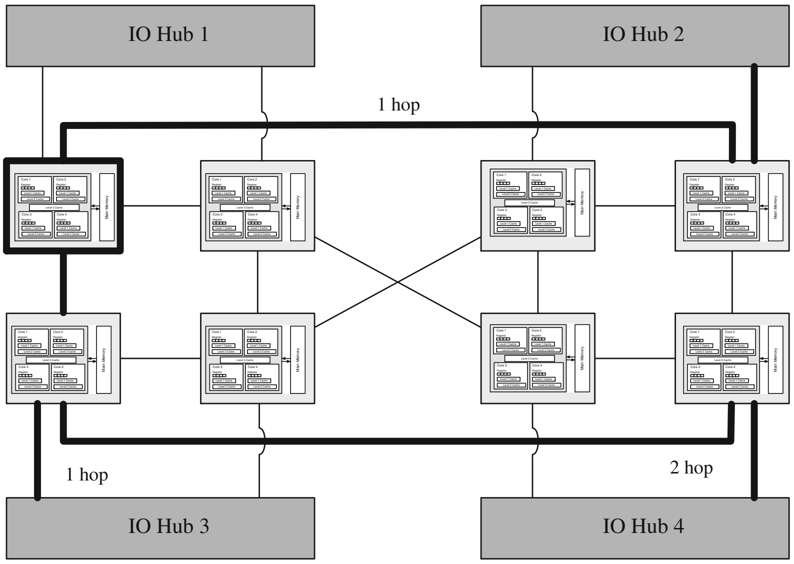


Fig. 17.4 A server consisting of multiple processors

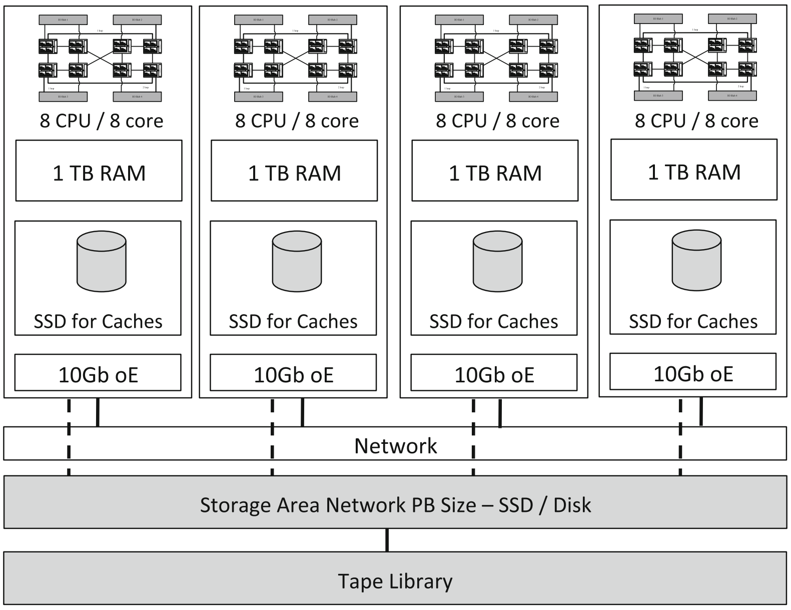


Fig. 17.5 A system consisting of multiple servers

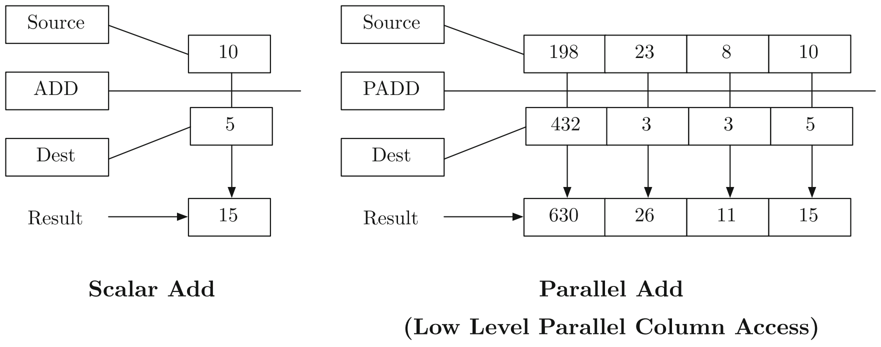


Fig. 17.6 Single instruction multiple data parallelism

Let us assume the following example: the gender attribute values can be stored in a single bit. Using SIMD you can process the gender of 128 persons in a single CPU instruction step. For example, if you encode male as 1 and female as 0, calculating the ratio of male and female persons of this group of 128 persons is performed within a single instruction by performing a PADD. For comparison, modern processor families are able to perform 100,000 Million instructions per second (MIPS) and more [HP11]. SIMD is the lowest level of parallelization on a computer system.

In-memory database management systems differ from traditional database management systems in how they address the performance topic. For example, in-memory databases incorporate data partitioning to improve performance. Furthermore, query result sets are aligned to fit the CPU cache lines, i.e., the number of cache misses is minimized so that additional data loading from main memory is minimized.

17.2 Software Layer

In addition to hardware parallelism, we consider software-level parallelism in the following section.

17.2.1 Amdahl's Law

Gene Amdahl conducted fundamental considerations about software-level parallelism. He defined that the maximum speedup of executing a piece of code in parallel is limited by the time needed to process the longest sequential fraction of the code. This is nowadays known as Amdahl's law [Amd67].

$$\text{max. speedup(N)} = \frac{1}{(1-\text{P}) + \frac{P}{N}} \tag{17.1}$$

Equation (17.1) defines Amdahl's law with P giving the fraction of the code that can be processed in parallel and N giving the level of parallelism, e.g. the number of CPU cores.

Let us assume the following example: the ratio of parallel and the sequential part are 3:1. If the execution time of the parallel part is decreased, e.g. by increasing the number of cores, the maximum speedup cannot exceed four as can be seen in Eq. (17.2).

$$\lim_{N\to\infty} \text{speedup(N)} = \frac{1}{\left(1-\frac{3}{4}\right) + \frac{3}{4N}} = \frac{1}{\frac{1}{4}} = 4 \tag{17.2}$$

Amdahl's law assumes that the there is a fixed size of the solution space, i.e. a tasks generates repeatable a finite number of results. In contrast, Gustafson assumes that there is a maximum acceptable response time while the solution space is not known beforehand [Gus88]. Equation (17.3) defines Gustafson's law with C defining the number of cores and α defining the non-parallelizable fraction of the program code.

$$\text{max. speedup(C)} = \text{C} - \alpha(\text{C} - 1) \tag{17.3}$$

17.2.2 Shared Memory

In a shared memory system [Li86], data that is stored in the shared memory segment is accessible by all processors in an uniform way. Special programming concepts, such as mutexes and semaphores, are used to avoid conflicting data access in the shared memory segment, e.g., simultaneous write access. Although shared memory is an easy way to share data across processes or CPU cores, it comes with the problem of scaling.

Shared memory systems suffer from scalability issue since the maximum size of the shared memory segment is limited by available memory size. The total memory size of a single system is small compared to the total main memory size formed by multiple servers.

17.2.3 Message Passing

Message passing is a very powerful paradigm to improve the processing of algorithmic problems [GLS94]. Instead of sharing memory between all threads, only messages are passed between individual processing threads. This paradigm is widely used for number crunching tasks, such as prediction of meteorology, earthquakes, and other kinds of simulations.

All processors can perform tasks independently while processors depending on results of each other exchange messages for coordination. In comparison to the shared memory approach, message passing can easily scale out, because processors are independent of shared memory. Thus, they can perform their tasks individually while exchanging messages, e.g., via network links. However, if the sum of exchanged messages exceeds the network capacity, the network becomes a bottleneck for this parallelism paradigm.

17.2.4 MapReduce

The data parallel paradigm aims at identifying a portion of data so that each portion can be processed in parallel. Thus, each processing job is performing the same task on an individual partition of the complete data. Examples of data parallel paradigm are the MapReduce framework and the OpenMP library [DG08, DM98].

MapReduce consists of two specific functions: the map and the reduce function. The former operates on individual data partitions in parallel and produced partial results $r_1..r_n$ for its assigned partition $p_1..p_n$. The reduce function forms one overall result r_{all} by merging all partial result $r_1..r_n$. Map and reduce steps can be chained to produce arbitrary results for complex tasks.

The canonical example for MapReduce is counting the number of occurrences of a specific word in a defined set of text documents. Each map function processes an individual text document or a part of it. It counts the number of occurrences for a specific word within this document. Since map functions are executed in parallel, multiple text documents are scanned for the desired word simultaneously. The following reduce function calculates the total number of occurrences for the specific word by summing up the individual results. Google is also using MapReduce for indexing of and searching in Web sites and text documents.

MapReduce requires the developer to define the "how" and the "what", i.e., if your algorithm does not scale efficiently, the overall response time will not be reduced. This direct control may also be a disadvantage for some tasks since you only want to define the "what". For example, in a database management system you expect an optimizer to generate the proper code—the "how"—to retrieve the desired data —the "what".

Thus, MapReduce does not address all problems efficiently. It is designed for parallel processing of a batch job, e.g., word counting. However, interactive analytical queries require flexible access to data. For example, exploring overdue payers requires subsequent analyses of data subsets, which makes it hard to have all possible map function available.

17.3 Self Test Questions

1. **Shared Memory**
   What limits the use of shared memory?
   (a) The number of workers, which share the same resources and the limited memory itself
   (b) The caches of each CPU
   (c) The operation frequency of the processor
   (d) The usage of SSE instructions.

References

[Amd67] G.M. Amdah, Validity of the single processor approach to achieving large scale computing capabilities. In: *Proceedings of the April 18–20, 1967, Spring Joint computer Conference, AFIPS '67 (Spring).* (ACM, New York,1967), pp. 483–485

[DG08] Jeffrey Dean, Sanjay Ghemawat, Mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)

[DM98] Leonardo Dagum, Ramesh Menon, Openmp: an industry-standard api for shared-memory programming. IEEE Comput. Sci. Eng. **5**(1), 46–55 (1998)

[GLS94] William Gropp, Ewing Lusk, Anthony Skjellum, *Using MPI: portable parallel programming with the message-passing interface* (MIT Press, Cambridge, 1994)

[Gus88] J.L. Gustafson, Reevaluating amdahl's law. Commun. ACM **31**(5), 532–533 (1988)

[HP11] J.L. Hennessy, D.A. Patterson, Computer Architecture: a quantitative approach 5th edn. (Elsevier Science, Burlington, 2011)

[Li86] K. Li, Shared virtual memory on loosely coupled multiprocessors. Ph.D. thesis. (New Haven, 1986), AAI8728365

[Moo65] G. Moore, Cramming more components onto integrated circuits. Electroni. **38**, p 114 ff. (1965)

Chapter 18
Indices

18.1 Indices: A Query Optimization Approach

Usually, applications work only with a subset of records at a time. Therefore, before processing the portion, it must be located within the database. Hence, records should be stored in a manner that makes it possible to locate them efficiently whenever they are needed. The process of locating a specific set of records is determined by the predicates that are used to characterize these records.

SanssouciDB organizes its records in columns (see Chap. 8). To determine the records, it is necessary to perform a scan on all the columns, which are used as filter criteria. In main memory column-oriented databases, which store column values continuously, i.e. in adjacent memory blocks, searching for a value with a full scan (by iterating through all items placed in memory sequentially) can be done by orders of magnitude faster than in row-oriented databases. Therefore, the usefulness of index structures in such databases is limited. Nevertheless, because the complexity of a full column scan is linear, it is just a matter of data volume that will make the speed advantage of indices relevant to main memory column-oriented databases.

In this chapter, we discuss the topic of inverted indices in the context of main memory databases in more detail.

18.2 Technical Considerations

Let us look at the *world_population* table from Chap. 11 again. For the readers convenience we repeat (see Fig. 18.1). Let us assume that we want to locate the records of all the people from Berlin. The dictionary and the attribute vector of the column are shown in Fig. 18.2.

To determine the set of *berlin* records, we need to set the filter criterion on the *city* attribute of the table. The respective SQL query could look like depicted in Listing 18.1.

Example Table: world_population

recID	fname	lname	gender	country	city	birthday
0	Martin	Albrecht	m	GER	Berlin	08-05-1955
1	Michael	Berg	m	GER	Berlin	03-05-1970
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992
4	Sophie	Schulze	f	GER	Potsdam	09-03-1977
...	...	...	...	...	...	...

```
INSERT INTO world_population
VALUES (Karen, Schulze, f, GER, Rostock, 06-20-2012)
```

Fig. 18.1 Example database table named *world_population*

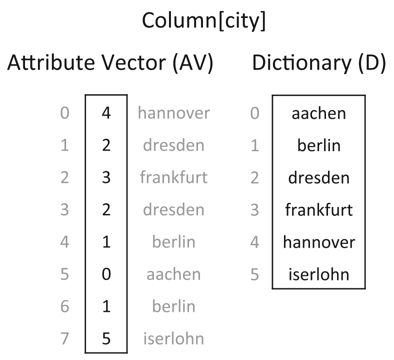


Fig. 18.2 City column of the *world_population* table

```
SELECT * FROM world_population WHERE city = 'berlin';
```

Listing 18.1: Query to select all people from Berlin

Now, let us do a back of the envelope calculation. Assuming the table contains 8 billion records, our CPU is able to process 2 GB per second per core, and the city column is encoded using 20 bit. The memory footprint of the *city* column's attribute vector can be calculated using

$$8\,\text{billion} \cdot 20\,\text{bit} = 160\,\text{billion bit} = 20\,\text{billion Byte} \approx 18.6\,\text{GB}$$

The time it takes a single core to process that amount of data can be calculated using

$$18.6\,\text{GB} \div 2\,\text{GB/sec} \approx 9.3\,\text{sec}$$

This calculation shows that scanning the whole column with 40 cores takes $\approx$230 ms. Despite that this scan speed is unthinkable for a row-oriented database, the speed might not be sufficient for all applications.

18.3 Inverted Index

Now, let us investigate a more complicated, but more efficient algorithm presented in detail in [FSKP12]. We consider an inverted index for the attribute *city*. An inverted index maps each distinct value to a position list, which contains all positions where the distinct value can be found in the column. The index for the dictionary-encoded column consists of the following two parts (Fig. 18.3):

- Index offsets (IO) : this vector stores for each dictionary entry (or in other words, for each unique value of the attribute vector) the offset for the position list in the positions vector. This means that the offset vector stores references to the first occurrence of the particular dictionary code in the positions vector.
- Index positions (IP) : the index position vector contains a position list of all distinct values of the attribute vector sorted by the integer valueID. In contrast, the attribute vector stores valueIDs by position.

Let us see how much data the CPU has to read using an index. We continue with the query shown in Listing 18.1. The following steps need to be executed to determine the position of *berlin* in the attribute vector.

1. We need to perform a binary search on the dictionary to determine the dictionary position related to *berlin*. As depicted in Fig. 18.4, *berlin* is at position 1.
2. The dictionary position of *berlin* corresponds directly to the position of the index offset vector shown in Fig. 18.5. In this example, the dictionary position of *berlin* is 1, so the corresponding index offset vector position is 1.
3. Since the attribute of the respective search criterion is not necessarily a primary key, it is possible that the same value is used by many records. Consequently, more than one attribute vector entry can be filled with that value. As explained in the beginning of this chapter, the index position vector represents a sorted list of the values in the attribute vector. To determine the range of values to read

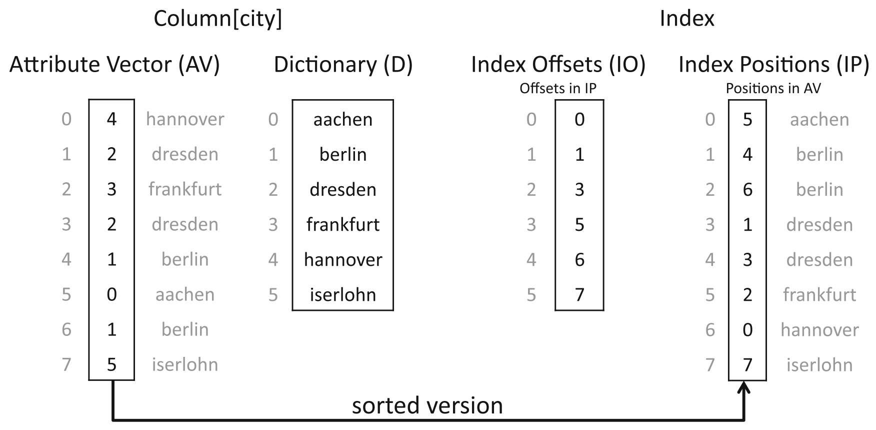


Fig. 18.3 Index offset and index positions

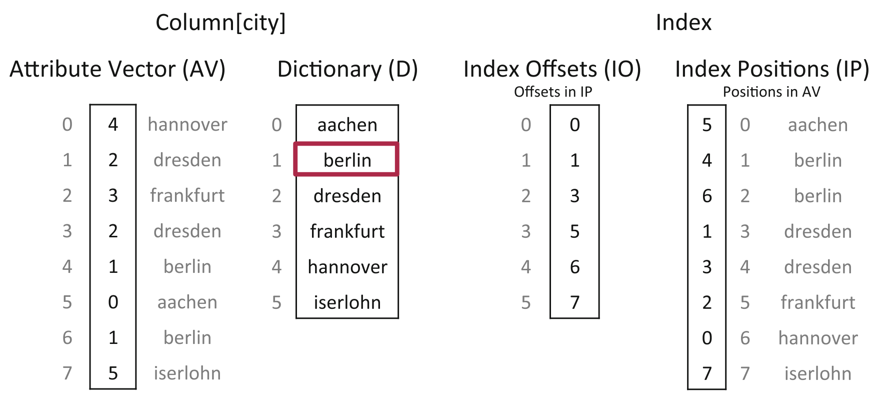


Fig. 18.4 Query processing using indices: Step 1

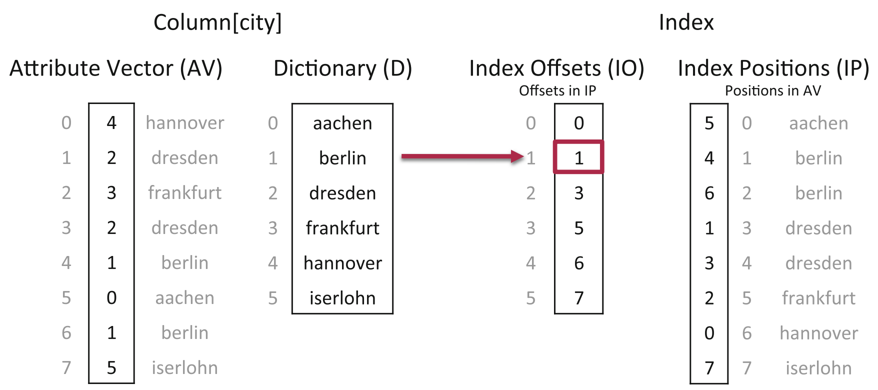


Fig. 18.5 Query processing using indices: Step 2

from the index position vector, we simply read the value of the index offset vector at the position, we determined in Step 1, and the value of the next higher position (see Fig. 18.6).

4. Since 3 is already the offset of the next value in the index positions vector, we only need to read positions 1 and 2. As shown in Fig. 18.7, IP vector position 1 contains the value 4 and position 2 contains the value 6, which are the exact positions of the dictionary code for *berlin* in the attribute vector. By retrieving the offsets of *berlin* and *dresden* from the IO vector, we are able to determine the exact range of all values we need to read in order to resolve the respective attribute vector positions of *berlin*.
5. With the positions resolved in Step 4, we are able to jump directly to the respective attribute vector positions of all other columns of that table in order to materialize the complete records of all the people that live in Berlin (Fig. 18.8).

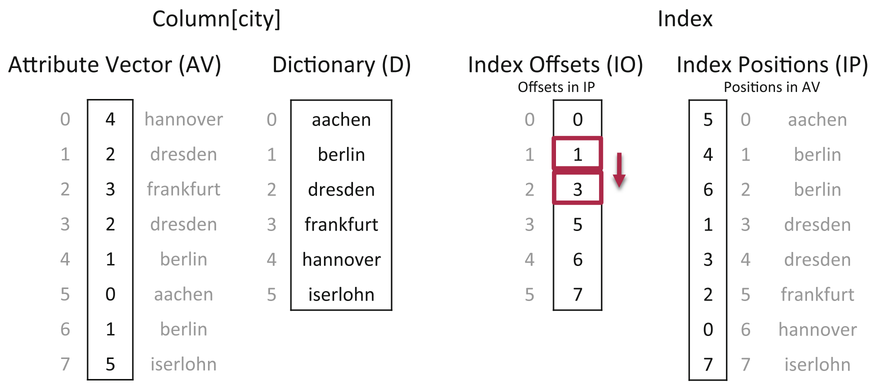


Fig. 18.6 Query processing using indices: Step 3

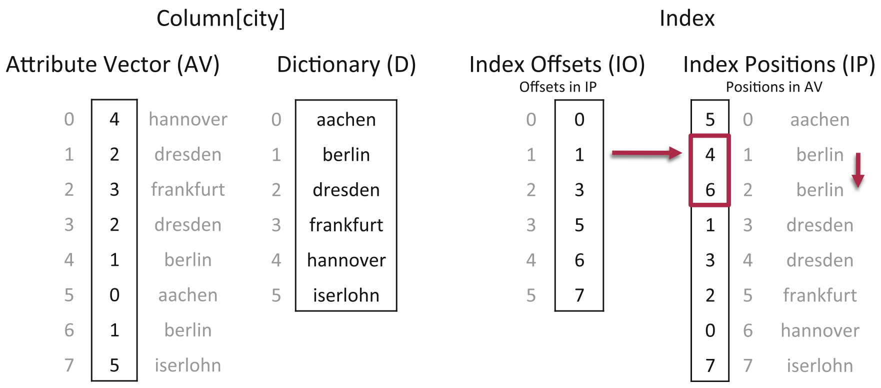


Fig. 18.7 Query processing using indices: Step 4

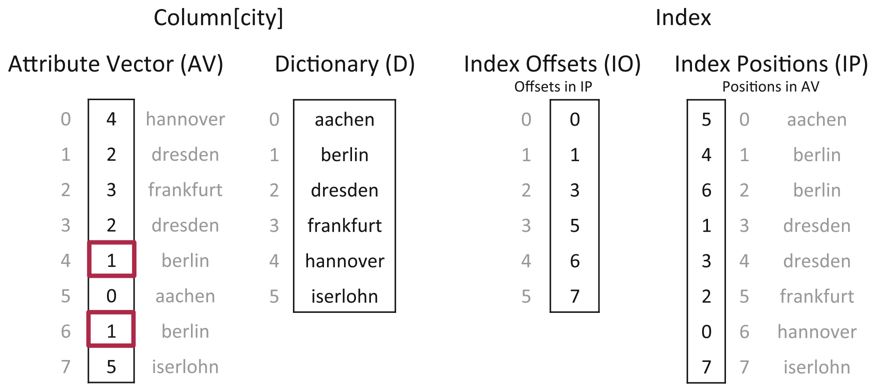


Fig. 18.8 Query processing using indices: Step 5

Using this approach, we reduce the data volume read by a CPU from the main memory by providing a data structure that does not require the scan of the entire attribute vector. Investigations regarding the influence of using indices on memory traffic and performance are shown in Sect. 18.4.

18.4 Discussion

In the previous section, we explained the idea of using an inverted index on a dictionary-encoded column to increase response time for lookup requests. An index increases the memory consumption per column. In this section we compare data lookup using full table scan against an index, regarding memory consumption and lookup performance. We first introduce the following symbols that we use.

18.4.1 Memory Consumption

In the beginning of this chapter, we explained that an index consists of an IO vector and an IP vector. To determine the overall size of the index, we need to calculate the size of these two structures.

$$I_m = IO_m + IP_m$$

The allocated memory of a vector can simply be calculated by multiplying its length (number of entries) with its width (size of a single entry).

$$IO_m = IO_l \cdot IO_w$$
$$IP_m = IP_l \cdot IP_w$$

The length of IP directly corresponds to the length of the attribute vector AV_l, since it is basically a sorted version of the corresponding attribute vector. The width of IP is determined by the bit-encoded length of the attribute vector, since it contains direct positions to the values in the attribute vector.

$$IP_l = AV_l$$
$$IP_w = \lceil log_2(AV_l) \rceil \text{ bits}$$

The length of IO directly corresponds to the length of the dictionary D_l, which in turn is determined by the number of distinct values in the respective column. The width of IO is derived from the biggest offset into IP, because IO contains the bit-encoded offsets used to determine the position ranges in IP. As the maximum offset stored in IO can be the length of IP, the resulting width of IO is $\lceil log_2(IP_l) \rceil$.

$$IO_l = D_l$$
$$IO_w = \lceil log_2(IP_l) \rceil \text{ bits}$$

Summarizing, we combine the above formulas to a single equation for calculating the size of an index structure.

$$I_m = D_l \cdot \lceil log_2(IP_l) \rceil + AV_l \cdot \lceil log_2(AV_l) \rceil \text{bits}$$
$$I_m = (D_l + AV_l) \cdot (\lceil log_2(AV_l) \rceil) \text{bits}$$

Let us now calculate the actual size of an index for the *city* column of our *world_population* table from Fig. 18.1. We need to determine D_l, IP_l, and AV_l. Based on the assumption that there are about 1 million cities around the world and that the world population is 8 billion, we just need to insert these numbers into our formula.

$$I_m = 10^6 \cdot \lceil log_2(8 \cdot 10^9) \rceil + 8 \cdot 10^9 \cdot \lceil log_2(8 \cdot 10^9) \rceil \text{bits}$$

So from this formula, we get an index size of about 31 GB for the *city* column.

18.4.2 Lookup Performance

Independent of using an index or not, we need to perform a binary search on the dictionary to determine the encoded value for the respective search term. Let us assume that we need to read $log_2(D_l)$ entries to perform the binary search. Since the binary search on the dictionary has to be done for both access methods we can ignore it, when we compare them.

Description	Unit	Symbol
Memory consumption of the index	bits	I_m
Length of the index offset vector	–	IO_l
Width of the index offset vector	bits	IO_w
Memory consumption of index offset vector	bits	IO_m
Length of the index positions vector	–	IP_l
Width of the index positions vector	bits	IP_w
Memory consumption of index positions vector	bits	IP_m
Length of dictionary (number of distinct values in column)	–	D_l
Length of attribute vector	–	AV_l
Width of attribute vector	bits	AV_w

In case of a full column scan, we need to traverse the attribute vector sequentially, by reading AV_l entries, each with a size of $\lceil log_2(D_l) \rceil$ bits. Again, assuming 8 billion rows in the table and 1 million cities, we need to read

$$8 \cdot 10^9 \cdot \lceil log_2(10^6) \rceil \text{bits} = 160.000.000.000 \, \text{bits}$$

for a full attribute vector scan.

Now, when using an index the situation is different. After the dictionary lookup, we directly read the upper and lower limit from the index offset vector (see Fig. 18.6). We neglect this step for our performance consideration since the number of bytes read does not impact the overall scan performance. Having determined the upper and lower limit for the index positions vector, we need to traverse through it (see Fig. 18.7). The number of entries to read from IP depends on the distribution of values in the column. Reading an attribute value that is used more frequently, we need to read more entries, on less frequently used values we need to read less. Assuming a uniform distribution of values we need to read $AV_l \div D_l$ entries. The width of the entries is $\lceil log_2(AV_l) \rceil$. Combining both equations, we get

$$IndexPositions = \frac{AV_l \cdot \lceil log_2(AV_l) \rceil}{D_l}$$

for the number of bits to read from the index positions vector. Taking our world population example, where we look for all people living in Berlin, we come up with

$$\frac{8 \cdot 10^9 \cdot \lceil log_2(8 \cdot 10^9) \rceil \text{bits}}{10^6} = 264.000\,\text{bits}$$

to read using an index. Assuming a CPU performance of 2MB/ms/core, a single core needs about 9 s to scan the complete attribute vector. Accessing the column using an index, the CPU needs 0.0157 ms to read the attribute vector positions of the people living in Berlin. Thus, in this example, we improve performance by a factor of ≈573,248 with the index, compared to a sequential attribute vector scan.

We compare the theoretical memory traffic for the attribute vector scan and the position read in Fig. 18.9 for different dictionary sizes on a column with 30 million entries. With a uniform distribution, the index leads to less memory traffic, if at least 8 distinct values are present.

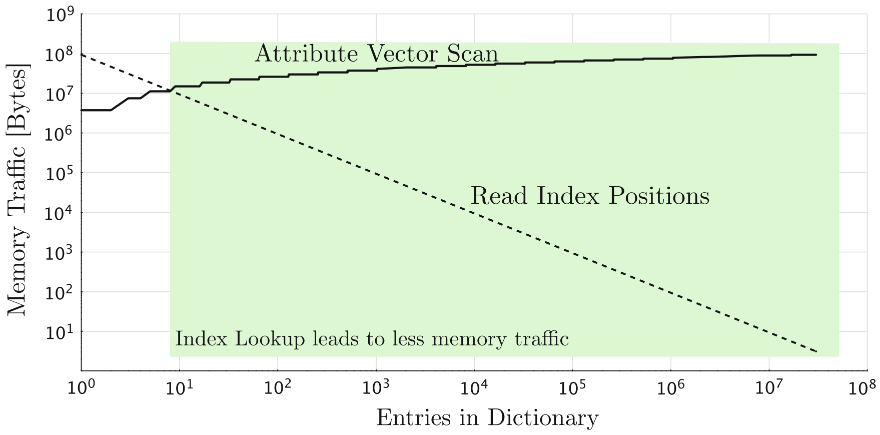


Fig. 18.9 Attribute vector scan versus index position list read for a column with 30 million entries (note the log-log scale)

18.5 Self Test Questions

1. **Index Characteristics**
   Introducing an index...
   (a) decreases memory consumption
   (b) increases memory consumption
   (c) speeds up inserts
   (d) slows down look-ups.
2. **Inverted Index**
   What is an inverted index?
   (a) A structure that contains the distinct values of the dictionary in reverse order
   (b) A list of text entries that have to be decrypted, it is used for enhanced security
   (c) A structure that contains the delta of each entry in comparison to the largest value
   (d) A structure that maps each distinct value to a position list, which contains all positions where the value can be found in the column.

Reference

[FSKP12] M. Faust, D. Schwalb, J. Krueger, H. Plattner, Fast lookups for in-memory column stores: group-key indices, lookup and maintenance. in *ADMS '12: Proceedings of the 3rd International Workshop on Accelerating Data Management Systems Using Modern Processor and Storage Architectures at VLDB'12*, 2012

Chapter 19
Join

This chapter is about joins and their execution in in-memory database systems. In general, joins are a way to combine tuples of two or more tables. There are two general categories of joins which can be than further specialized:

- *Inner-Joins* creates a result table that combines the tuples from the two input tables based on a join-predicate. It therefore combines each tuple from the first table with each tuple of the second table to check for the join-predicate.
- *Outer-joins* are used to fetch information that might not be there. An example where an outer join can be used are sales in a certain region and period. If the objective is to display the sales for all regions over the whole year and there is a region with no sales in a certain period then that information would be lost with a regular equi-join. In contrast, the outer-join inserts *Null* as a value if there is no matching tuple in the second relation. This allows obtaining the information from the result directly (without guessing at missing regions in the example).

Further specialization of the join types are:

- *equi-joins* are the most common and most important join type. They allow the selection of tuples from both relations which satisfy a given equality predicate.
- *semi-joins* return only one half of the intermediate join result. The other half is discarded.

The most prominent use case for joins are normalized database schemas where joins are executed based on foreign keys. Another use case of joins are applications where tables come from different sources. They use the same attribute, e.g., a customer number or a material number, but might have different names. The applications is able to tell the database that those two attributes are the same and based on this information it is possible to join the tables.

Different types of relations between two tables exist. These are one-to-one, one-to-many, and many-to-many relations. A one-to-one relation connects one tuple of the first table with one tuple from the second table. This means joining the tables' results in a maximum of one join partner for each tuple. In a one-to-many relationship, each tuple of the first table is joined with multiple tuples of the second table. An example for a one-to-many relation would be a normalized world population table where a foreign key in the world population table represents the country and the names of the countries are located in a separate country table that contains exactly one tuple per country. The resulting relation between county and the world population table connects each country from the country table with all its citizens in the world population table. In a many-to-many relationship, each tuple from the first relation may be joined to multiple tuples from the second one and each tuple from the second relation may be joined to multiple tuples from the first one. An example for a many-to-many relation is the books-and-authors relation. A book can have many authors and an author can have multiple books. A many-to-many relation can be implemented as an additional table containing pairs of foreign keys that lists the matching tuples of the two tables with the many-to-many relation.

19.1 Join Execution in Main Memory

Looking at the properties of main memory shows that sequential scans are many times faster than random access. In turn, the target of join algorithms in main memory systems is to leverage sequential scans as much as possible. As a consequence, random lookups are avoided as much as possible. Another target of these algorithms is to avoid materialization of data as long as possible in order to work with the much smaller position information, e.g., to fit more data into one cache line (see Chap. 16).

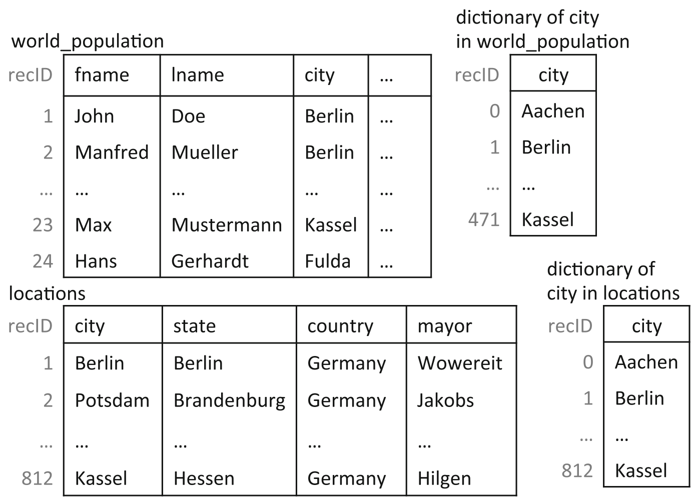
The next two sections present the *hash-join* and the *sort-merge join* as two join algorithms that leverage the features of in-memory databases. First, a hash-join is based on a hash function. The function maps input values to fixed length output values. It has the advantage that the access is very cheap. The sort-merge join uses the properties of a distinctly sortable data type. After sorting the two join columns, the columns are merged.

To present the algorithms, we use the world population table and an extra locations table that provides more details about specific locations (both are simplified for the examples to keep the IDs small). See Fig. 19.1 for an overview of the example tables.

We furthermore use the statement in Listing 19.1. It retrieves the name of each city in the state ‘Hessen’, the mayor of that city and the number of people living there. For simplicity, we assume that city names are distinct per state. As it is an inner join, only cities with at least one join partner in the world population table are shown in the result set.

```
SELECT count(*) As population, wp.city, locations.mayor
FROM world_population AS wp
INNER JOIN locations ON world_population.city = locations.city
WHERE locations.state = 'Hessen'
GROUP BY wp.city, locations.mayor;
```

Listing 19.1: SQL statement including a join


world_population

recID	fname	lname	city	...
1	John	Doe	Berlin	...
2	Manfred	Mueller	Berlin	...
...	...	...	...	...
23	Max	Mustermann	Kassel	...
24	Hans	Gerhardt	Fulda	...

dictionary of city in world_population

recID	city
0	Aachen
1	Berlin
...	...
471	Kassel

locations

recID	city	state	country	mayor
1	Berlin	Berlin	Germany	Wowereit
2	Potsdam	Brandenburg	Germany	Jakobs
...	...	...	...	...
812	Kassel	Hessen	Germany	Hilgen

dictionary of city in locations

recID	city
0	Aachen
1	Berlin
...	...
812	Kassel

Fig. 19.1 The example join tables

19.2 Hash-Join

The hash-join is based on a hash function, which allows access in constant time. A second feature of this function is that it maps to fixed length keys, even though the input of the function might have a variable length.

The hash-join algorithm itself consists of two phases: a hash phase and a hash-join phase. During the hash phase, the smaller relation (lower table cardinality) is sequentially scanned on the predicate column and the hash key of each attribute value is calculated. The key is inserted into a hash map together with the value's position in the table. To keep the hash map small and its creation fast, the smaller one of the two joined tables is used for hash map creation.

During the hash-join phase, the bigger relation is sequentially scanned. Each value is probed into the hash map by calculating the hash key and looking it up in the hash map. If the value exists, the position of the currently probed tuple and the tuple's position in the hash map are returned as a matching pair. A row is skipped if its key does not exist in the hash map.

19.2.1 Example Hash-Join

This section provides a deeper insight into the hash-join algorithm by analyzing the example query's execution as displayed in Listing 19.1. The first step of the query execution is to find all predicates that can be evaluated before the actual join execution starts. This allows reducing the size of intermediate results and in turn saves memory and other computing resources. Regarding the example, this means the filter

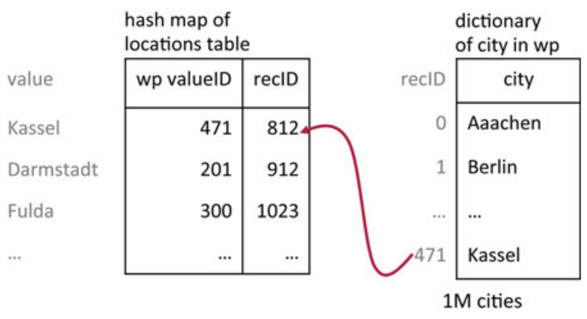


Fig. 19.2 Hash map creation

operation of the *WHERE* clause is executed first. The result is a reduced amount of cities that are relevant for the join. In our example, the result is a list of all positions of cities of the locations table in ‘Hessen’, which are <812,912,1023,4581,…>.

At this point the actual join starts with the hash phase. The cities in ‘Hessen’ are the smaller input relation, which means they are sequentially scanned to build the hash map as shown in Fig. 19.2. The hash map contains the valueID of the city column of the world population table as the hash-encoded key. That valueID translation is done using the city dictionaries of both tables. As an example, take a look at the city ‘Kassel’ with the recordID 812. To get the valueID of ‘Kassel’ in the world population table, it is looked up in its city dictionary. The hash-encoded key of the hash map is generated from this valueID, which is 471 for ‘Kassel’. The second part that is stored in the hash map is the position information of the locations table. In the case of ‘Kassel’ that is 812. Creating the hash map creates a mapping from the world population table to the locations table.

The next phase is the actual hash-join phase, which is also called *probe phase*. Figure 19.3 displays a part of the example’s probe phase. The city column of the world population table is scanned and each item’s city valueID is probed into the hash map to check if the hash map contains the key. This access is of constant complexity. When a key is found, the recordID of the currently probed row and the recordID that is stored together with the matching hash key are returned as a matching pair. Those pairs are the result set of the actual join operation.

In a last step, the rest of the query is executed. This includes executing the COUNT aggregation, based on position pairs, to calculate how many people are living in each city. Finally, the mayor and the name of the city can be fetched and the query’s materialized result is complete.

Hash-joins have one significant problem, which is the hash function. The reason for this is that a hash function for longer byte values like strings is difficult to find. If the function does not absolutely fit, the algorithm has to deal with collisions. Collisions occur if multiple distinct values are hashed to the same key. Those collisions make the hash-join algorithm more complex as it has to deal with them. An alternative to the hash-join is the sort-merge join, which will be discussed in the next section.

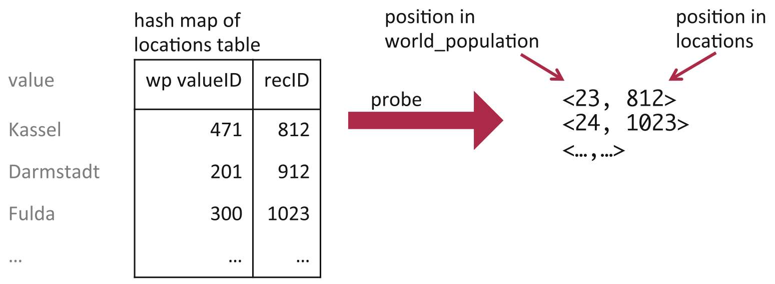


Fig. 19.3 Hash-join phase

19.3 Sort-Merge Join

The sort-merge join consists of multiple phases that can be implemented as a semi-join. First, the predicate column of the smaller relation is scanned to build a list of unique valueIDs. This list is used afterwards to create a translation table, which contains the unique values of each relation. Removing all rows without a join partner from the mapping table completes the first part of the semi-join.

Now, the bigger relation is sequentially scanned and all matching positions are retrieved using the translation table. The result is a list of all matching positions. Based on the position lists of both tables, the results from both parts of the semi-join are combined and the matching tuples are returned.

19.3.1 Example Sort-Merge Join

This section shows the execution of a sort-merge join with the help of an example. The query from Listing 19.1 is used again but this time a different join algorithm is employed. An overview of both join tables and their join column dictionaries is shown in Fig. 19.1. ‘Kassel’, the city of the last example, has position 812 in the locations table. The valueID of ‘Kassel’ is 471 in the world population city dictionary and 812 in the locations city dictionary.

The first phase of the sort-merge join execution is very similar to the hash-join execution. First, the filter operation is executed to reduce the number of input tuples to the actual join operation.

In the join execution’s second part, a unique list of valueIDs is created from the smaller relation. This is a list of all valueIDs of the cities in ‘Hessen’, i.e. 812, 231, and 510. It is the base for the translation table. The example translation table is shown in Fig. 19.4. It provides a mapping from a city’s valueID in the world population dictionary to the valueID in the locations table dictionary. Each valueID from the dictionary of the locations table is used to retrieve the actual

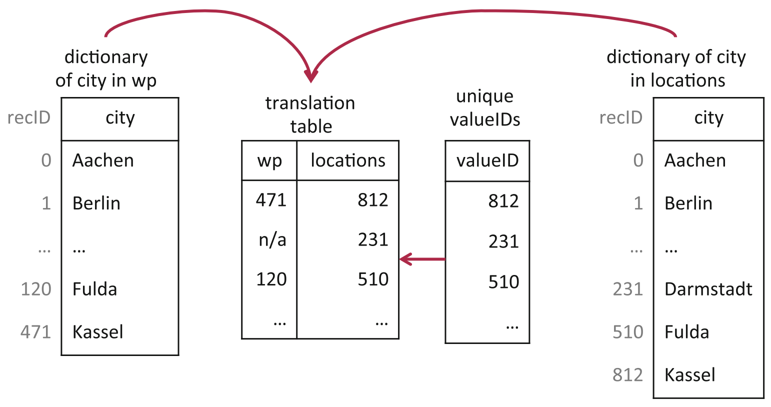


Fig. 19.4 Building the translation table

value, which in turn is used to perform a lookup on the dictionary of the world population table.

The next step of the sort-merge join is to remove all rows from the mapping table, where one part of the mapping is missing. In the example, this is the case for the row where 231 is the valueID of the location table and a world population's valueID is not available. Removing those rows completes the first part of the semi-join. So far, the algorithm mainly worked on dictionaries instead of actual data.

The second part of the semi-join makes use of the translation table to find all matching positions in the world population table. Afterwards, the list of matching positions in the world population and the list of positions in the locations table are available and both list are sorted. The city column of locations is already sorted because the values come from the dictionary of the column "city" of the table "locations", which is alphabetically sorted. That means it is now possible to combine the two semi-joins and create a list of intermediate matching position pairs. This step is illustrated in Fig. 19.5. This is the result of the join itself.

The last step is again very similar to the last step of the hash-join. The aggregation, i.e., the number of inhabitants of all cities in 'Hessen', is calculated and afterwards the actual names of the city and the mayor are retrieved from the dictionary.

19.4 Choosing a Join Algorithm

Besides hash-join and sort-merge join, the *nested-loop join* is a third option to perform a join. Basically, it scans one relation and for each scanned tuple, the whole other relation is scanned for matching tuples. This results in a complexity of $\mathcal{O}(n \cdot m)$ with the first table having n tuples and the second having m tuples.

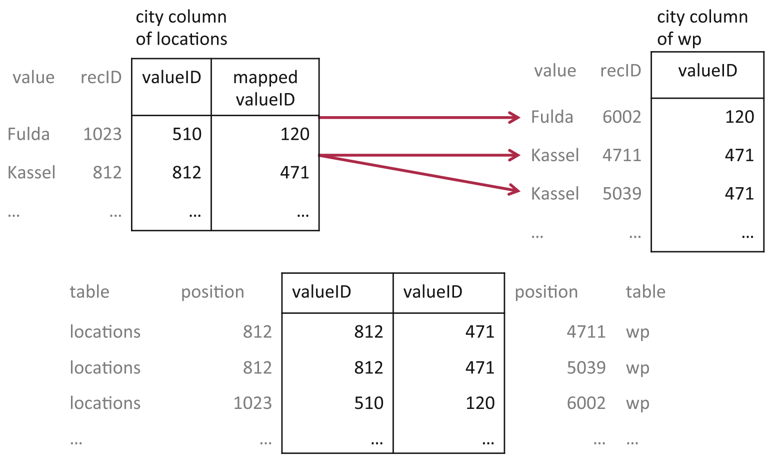


Fig. 19.5 Matching pairs from both position lists

The algorithm of the hash-join has a complexity of $\mathcal{O}(n+m)$ as the first relation's join column is scanned once to build the hash map and afterwards the second relation's join column is scanned once to probe the values into the hash map.

The sort-merge join complexity is $\mathcal{O}(n \cdot log(n) + m \cdot log(m))$. In general, it is worse than the hash-join's complexity because it requires sorting before the merge can be done.

To sum this up, the hash-join performs best in general. It performs better if there is an index on the attribute, which improves building the hash map. The limitation of the algorithm is the hash map. If the map becomes too large or of it is too complicated to build at all, there might be better alternatives. A reliable fall back algorithm is the sort-merge join, because it does not require an index on the attribute. The nested-loop algorithm is suitable for very small data sets, where creating the data structures of the other algorithms would create too much overhead.

19.5 Self Test Questions

1. **Hash-Join Complexity**
   What is the complexity of the Hash-Join?
   (a) O(n+m)
   (b) O(n^2/m^2)
   (c) O(n · m)
   (d) O(n · log(n)+m+log(m))

2. **Sort-Merge Join Complexity**
   What is the complexity of the Sort-Merge Join?
   (a) O(n+m)
   (b) $O(n^2/m^2)$
   (c) $O(n \cdot m)$
   (d) $O(n \cdot \log(n)+m \cdot \log(m))$
3. **Join Algorithm Small Data Set**
   Given is an extremely small data set. Which join algorithm would you choose in order to get the best performance?
   (a) All join algorithms have the same performance
   (b) Nested-Loop Join
   (c) Sort-Merge Join
   (d) Hash-Join
4. **Join Algorithm Large Data Set**
   Imagine a large data set with an index. Which join algorithm would you choose in order to get the best performance?
   (a) Nested-Loop Join
   (b) Sort-Merge Join
   (c) All join algorithms have the same performance
   (d) Hash-Join
5. **Equi-Join**
   What is the Equi-Join?
   (a) If you select tuples from both relations, you use only one half of the join relations and the other half of the table is discarded
   (b) If you select tuples from both relations, you will always select those tuples, that qualify according to a given equality predicate
   (c) It is a join algorithm that ensures that the result consists of equal amounts from both joined relations
   (d) It is a join algorithm to fetch information, that is probably not there. So if you select a tuple from one relation and this tuple has no matching tuple on the other relation, you would insert their NULL values there.
6. **One-to-One-Relation**
   What is a one-to-one relation?
   (a) A one-to-one relation between two objects means that for each object on the left side, there are one or more objects on the right side of the joined table and each object of the right side has exactly one join partner on the left
   (b) A one-to-one relation between two objects means that for exactly one object on the left side of the join exists exactly one object on the right side and vice versa

(c) A one-to-one relation between two objects means that each object on the left side is joined to one or more objects on the right side of the table and vice versa each object on the right side has one or more join partners on the left side of the table
(d) Each query which has exactly one join between exactly two tables is called a one-to-one relation, because one table is joined to exactly one other table.

Chapter 20
Aggregate Functions

This chapter discusses aggregate functions. It outlines what aggregate functions are, how they work, and how they can be executed in an in-memory database system.

Aggregate functions are a specific set of functions that take multiple rows as an input to create an output. This means, they work on data sets instead of single values. Grouping the input data based on specified group attributes creates the sets. Basic aggregate functions are, e.g., *COUNT*, *SUM*, *AVERAGE*, *MEDIAN*, *MAXIMUM* and *MINIMUM*. Furthermore, it is possible to create additional functions for special purposes, e.g., OLAP functions that extend basic functions.

The basic SQL syntax to use an aggregate function can be seen in Listing 20.1.

```
SELECT AggregateFunction(attribute1), attribute2, attribute3
FROM table_name
WHERE attribute2 = some_value
GROUP BY attribute2, attribute3
HAVING AggregateFunction(attribute1) > 5;
```

Listing 20.1: SQL aggregate function syntax

The *GROUP BY* clause specifies the attributes by which the input relation is grouped. All selected attributes that are not used in the *GROUP BY* clause should specify an aggregate function in the select clause, otherwise their value might be undefined. The *WHERE* and the *HAVING* clauses are optional.

20.1 Aggregation Example Using the COUNT Function

Let us consider an example for the use of the *COUNT* aggregate function. Assume an input table containing the complete world population as shown in Fig. 20.1.

The goal is to count all citizens per country. Using the *COUNT* aggregate function, an SQL query to achieve this is depicted in Listing 20.2.

recID	fname	lname	gender	city	country	birthday
0	John	Smith	m	Chicago	USA	12.03.1964
1	Mary	Brown	f	London	UK	12.05.1964
2	Jane	Doe	f	Palo Alto	USA	23.04.1976
3	John	Doe	m	Palo Alto	USA	17.06.1952
4	Peter	Schmidt	m	Potsdam	GER	11.11.1975
...	...	...	...		...	...

Fig. 20.1 Input relation containing the world population

```
SELECT country, COUNT(*) AS citizens
FROM world_population
GROUP BY country;
```

Listing 20.2: Example SQL query using the COUNT aggregate function

Figure 20.2 shows how such a query would be processed. First, the system runs through the attribute vector for the *country* column. For each new encountered country valueID, an entry with initial value "1" is added to the result map. If the encountered valueID has been added before, the respective entry in the result map is increased by one. That way, a result map is created which contains the valueIDs of each country and its number of occurrence. Second, the countries are fetched from the respective dictionary using the valuesIDs and the final result of the *COUNT* function is created. The result contains pairs of country names and the countries' number of occurrences in the source table, which corresponds to the number of citizens.

Other aggregate functions show a similar pattern. For *SUM*, the number of occurrences for each valueID is counted in an auxiliary data structure and the sum is calculated in a final step by summing up the number of occurrences multiplied

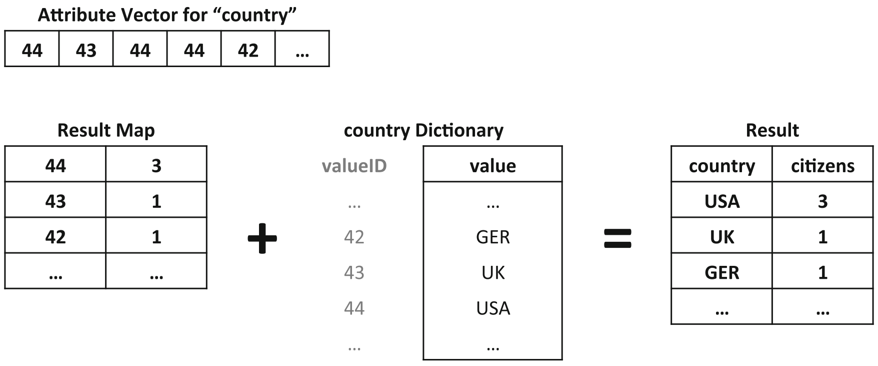


Fig. 20.2 Count example

with the respective value of each valueId. *AVERAGE* can be calculated by dividing *SUM* by *COUNT*. To retrieve the median, the complete relation has to be sorted and the middle value is returned. *MAXIMUM* and *MINIMUM* compare a temporary extreme value with the next value from the relation and replace it if the new value is higher (for *MAXIMUM*) or lower (for *MINIMUM*), respectively.

20.2 Self Test Questions

1. **Aggregate Function Definition**
   What are aggregate functions?
   (a) A set of functions that transform data types from one to another data
   (b) A set of indexes that speed up processing a specific report
   (c) A set of tuples that are grouped together according to specific requirements
   (d) A specific set of functions that summarize multiple rows from an input data set.
2. **Aggregate Functions**
   Which of the following is an aggregate function?
   (a) HAVING
   (b) MINIMUM
   (c) SORT
   (d) GROUP BY

Chapter 21
Parallel Select

Chapter 18 introduced the concept of an inverted index to prevent the database system from performing a full column scan every time a query searches for a predicate in the column. However, maintaining an index is expensive and consumes additional memory. So the decision to use an index should be made carefully, balancing all the pros and cons an index would bring in the particular situation. This chapter discusses how to speed up a full column scan despite adding an index to it. Chapter 17 introduced parallelism as a means to parallelize the execution of database operations. In this chapter, we present a detailed description of how parallelism can be used to speed up the execution of a SELECT operation.

21.1 Parallelization

The purpose of a SELECT operation is to find all positions in a column, that correspond to a certain predicate, meaning that we need to scan the attribute vector to find the position of all codes that match the dictionary entry of the predicate. Splitting the vector into chunks of data can parallelize a sequential scan over the attribute vector of a column. Each chunk of data can be processed independently and the results will be combined. Let us look at an example. We want to find the names of all men from Italy. The corresponding query to that question is shown in Listing 21.1. Please note that this query is just for demonstration purposes. There exist better implementations of that query, which we knowingly neglect to construct a simplified, yet less optimal example.

```
SELECT fname, lname
FROM world_population
WHERE country = 'Italy' AND gender = 'm';
```

Listing 21.1: Query to select all men from Italy

Figure 21.1 shows the resulting query plan, when the columns are split into four chunks. In this case, the attribute vector scans can be performed in parallel by eight independent threads.

As can be seen in Listing 21.1, we need two parallel *UNION* operations to combine the results from the different threads. The *positional AND* operation is executed sequentially again. Fortunately, the *positional AND* operation can be parallelized, too. If the two columns are equally partitioned, meaning that the number of partitions is equal and the position ranges in each corresponding partition are the same, the *positional AND* can be executed in parallel as well. Figure 21.2 shows such as scenario based on our example. The two attribute vectors of the columns are partitioned into four chunks. The position ranges in the respective chunks are the same for both columns (0. . .2, 3. . .5, 6. . .8, and 9. . .11). Parallel scans in the individual chunks will result in the selection shown in Fig. 21.3. After the parallel scans determined the individual positions that

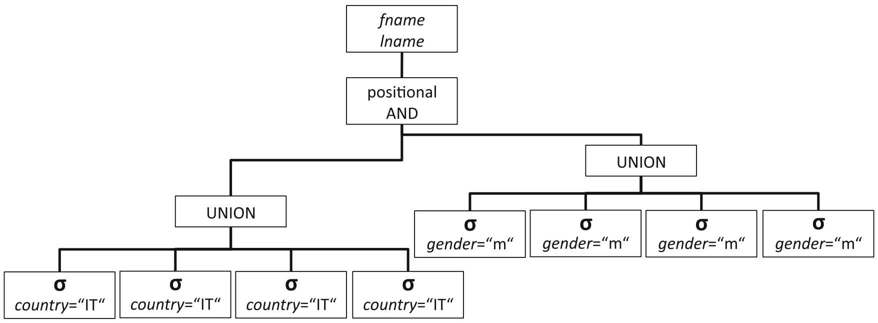
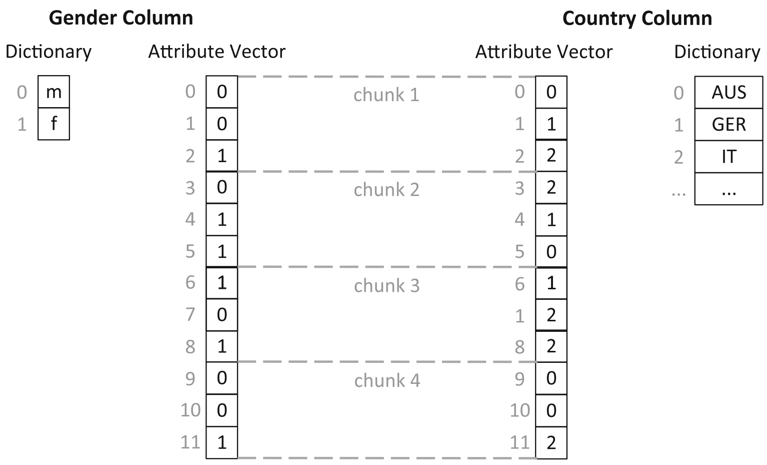


Fig. 21.1 Parallel scan, partitioning each column into 4 chunks



Gender position	Gender value	Chunk	Country position	Country value
0	0	chunk 1	0	0
1	0		1	1
2	1		2	2
3	0	chunk 2	3	2
4	1		4	1
5	1		5	0
6	1	chunk 3	6	1
7	0		1	2
8	1		8	2
9	0	chunk 4	9	0
10	0		10	0
11	1		11	2

Fig. 21.2 Equally partitioned columns

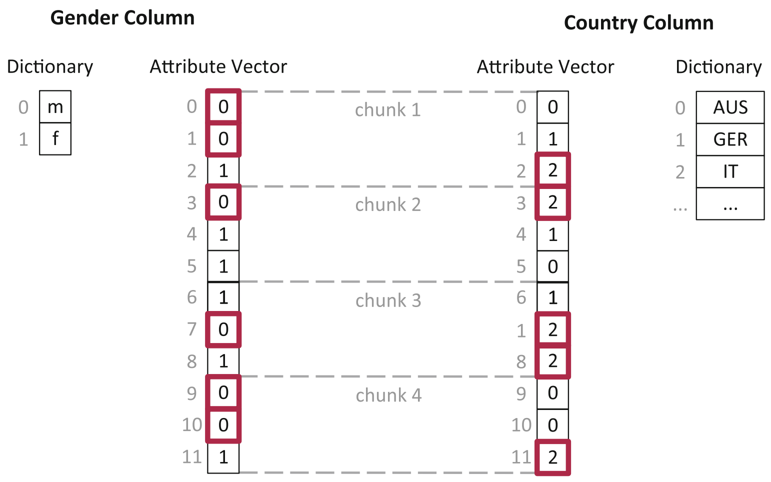


Fig. 21.3 Result of parallel scans

correspond to the search predicate in the respective column, we need to compare the positions within each chunk. If the same position has been marked in both columns, we have found a record that fulfills our search predicates. Now, this operation can be performed in parallel for each individual chunk. Figure 21.4 shows the result of the parallel comparison.

The last operation in the query is the *UNION*, which collects the results from the different *positional AND* operations. Once, we parallelize the *positional AND*, the executed query plan changes. The new query plan is depicted in Fig. 21.5.

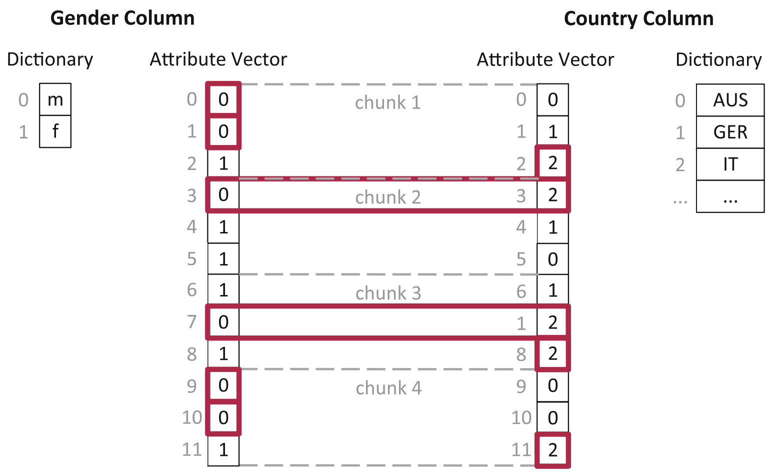


Fig. 21.4 Result of *Positional AND*

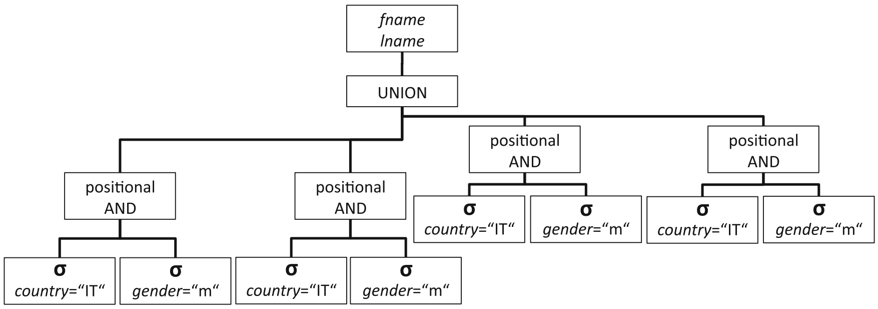


Fig. 21.5 Parallel scans with parallel *Positional AND*

21.2 Self Test Questions

1. **Amdahl's Law**
   Amdahl's Law states that...
   (a) the number of CPUs doubles every year
   (b) the level of parallelization can be no higher than the number of available CPUs
   (c) the speedup of parallelization is limited by the time needed for the sequential fractions of the program
   (d) the amount of available memory doubles every year.
2. **Query Execution Plans in Parallelizing SELECTS**
   When a SELECT statement is executed in parallel...
   (a) all other SELECT statements are paused
   (b) its query execution plan becomes much simpler compared to sequential execution
   (c) its query execution plan is adapted accordingly
   (d) its query execution plan is not changed at all.

Chapter 22
Workload Management and Scheduling

22.1 The Power of Speed

One of the most important factors that determine the usability for an application is response time. Psychological studies show that the acceptable maximum application response time for a human is about three seconds. After three seconds, the user might loose concentration and does something else. Once the application has finished its processing and is ready for the next human input, the user has to focus on the application again. These context switches are extremely expensive as they constantly interrupt the user and lead to being unproductive.

Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) are very similar from a technical perspective, but completely different from the user interaction perspective. OLAP transactions are long lasting, whereas OLTP transactions are short, but are being executed frequently and by many users in parallel.

A table scan necessary for an OLAP query can easily be parallelized to a high degree by a query optimizer. If the query optimizer decides that the query can be executed in parallel, it separates the query into multiple sub-queries, executes every sub-query in parallel and combines their intermediate results. To mix a highly parallelized analytical workload with a transactional workload, two queues are required. One queue is used for transactional queries and one for parallelizable analytical queries. If analytical queries are so small that they do not need to be parallelized they are treated like normal transactional queries, so they go in the same queue as these. Consequently, the query optimizer can execute transactional queries right away and tackle analytical queries with otherwise unused resources.

The additional challenge is to allow multiple users to explore the data simultaneously whenever they want. With the new architecture of SanssouciDB, it is possible to give users the freedom to analyze data in a very fast and unrestricted manner. This leads from a user who is not frustrated by using IT systems to a user, who actually enjoys it. As a result, the user works more with the available data and explores it beyond minimal duty.

In the corporate world, users are often separated into different groups with varying execution priorities. It is usual that top-managers get the highest priority and clerks the lower one. This is disadvantageous, because top-managers use the system extremely seldom, whereas clerks have to work with the system on a daily basis. Consequently, it is desirable to empower also the clerks with the full processing capacity of the IT landscape as they often detect uncommon patterns early and know how to navigate through the data. The here presented architecture provides full processing speed for all hierarchies throughout the company.

22.2 Scheduling

The scheduler has to assign the computer's resources such as processing power, memory, or network bandwidth to the running queries. Each query can be broken into tasks, which are eligible for being distributed across the available resources. During query execution, the original query is decomposed into a query plan that is executed to generate the expected result. To optimize the scheduling decision, statistical information about queries and their execution is used as well.

In general, scheduling goes along with several kinds of complications: interdependencies between queries, different resource utilization for the queries (e.g. some are CPU-bound or memory-bandwidth bound etc.), restricted resources, locality of operations (e.g. a filter task should be executed at the node where the data is stored) and so on.

These possible complications make finding an optimal execution plan a very complex task which is NP-hard in most cases. Therefore, a good solution is often sufficient: such a solution can be found by applying heuristics and by making assumptions about some relevant parameters. This enables a near-optimal scheduling decision in the shortest possible period [Pla11].

22.3 Mixed Workload Management

Typically, optimizing resource usage and scheduling is a workload-dependent problem and becomes even more complicated when two different types of workloads are mixed. As said before, a transactional workload is characterized by short running queries that must be executed within tight time constraints. In contrast, an analytical workload consists of more complex and computationally heavier queries. Running mixed workloads on a single database instance leads to potentially conflicting optimization goals. The response time for transactional queries must be guaranteed. At the same time, the response time for analytical queries should be as short as possible. This can only be achieved by using the aforementioned scheduler, which makes sure that the available resources are fully utilized.

22.4 Self Test Questions

1. **Resource Conflicts**
   Which three hardware resources are usually taken into account by the scheduler in a distributed in-memory database setup?

   (a) CPU processing power, main memory, network bandwidth
   (b) Main memory, disk, tape drive
   (c) CPU processing power, graphics card, monitor
   (d) Network bandwidth, power supply unit, main memory.

2. **Workload Management Scheduling Strategy**
   Why does a complex workload scheduling strategy might have disadvantages in comparison to a simple resource allocation based on heuristics or a uniform distribution, e.g. Round Robin?

   (a) The execution of a scheduling strategy itself consumes more resources than a simplistic scheduling approach. A strategy is usually optimized for a certain workload—if this workload changes abruptly, the scheduling strategy might perform worse than a uniform distribution
   (b) Heuristics are always better than complex scheduling strategies
   (c) A scheduling strategy is based on general workloads and thus might not reach the best performance for specific workloads compared to heuristics or a uniform distribution, while its application is cheap
   (d) Round-Robin is usually the best scheduling strategy.

3. **Analytical Queries in Workload Management**
   Analytical queries typically are ...

   (a) long running with soft time constraints
   (b) short running with soft time constraints
   (c) short running with strict time constraints
   (d) long running with strict time constraints.

4. **Query Response Times**
   Query response times ...

   (a) can be increased so the user can do as many tasks as possible in parallel because context switches are cheap
   (b) have to be as short as possible, so the user stays focused at the task at hand
   (c) should never be decreased as users are unfamiliar with such system behavior and can become frustrated
   (d) have no impact on a users work behavior.

Reference

[Pla11] H. Plattner, SanssouciDB: an in-memory database for mixed-workload processing, in *BTW* (Springer, Berlin, 2011), pp. 2–21

Chapter 23
Parallel Join

Join procedures are cost intensive tasks even for in-memory databases. As today's systems introduce more and more parallelism, intra-operator parallelization moves into focus. This chapter discusses possible schemes to parallelize a hash-join algorithm, as described in Chap. 19. Hash-join algorithms usually consist of two phases:

1. In the hashing phase, a hash map for the smaller join relation is created,
2. In the probing phase, a sequential scan over the larger join relation is performed while probing into the hash map.

To optimize the performance of hash-join algorithms, sequential memory access should be favored while random access should be avoided. Moreover, early materialization should be avoided and one should work with valueIDs as long as possible. However, this chapter will not concentrate on performance optimizations but deals with parallelization schemes for hash-join algorithms.

Various methods to parallelize join algorithms exist. We will first outline a simple, only partially parallelized hash-join algorithm in Sect. 23.1, followed by a more complex and fully parallelized version in Sect. 23.2.

23.1 Partially Parallelized Hash-Join

A simple way to parallelize a hash-join is to keep the hashing phase sequential and to only parallelize the probing phase.

In the sequential hashing phase, a hash table for the smaller join relation is created. This is done by sequentially scanning the join column, computing the respective hash value of each element and storing the position in the elements hash bucket.

In the probing phase, the larger join relation is partitioned horizontally across the available threads and the probing is performed in parallel. Each thread works with a local copy of the hash table and stores its results in a local result table.

DOI: 10.1007/978-3-642-36524-9_23, © Springer-Verlag Berlin Heidelberg 2013

When all probe threads have finished, the local result tables are merged into a unified result table.

Although the join is not fully parallelized, this approach works well in practice, as the probing phase on the larger join relation dominates the join costs. However, one disadvantage is the distribution of the hash table across all cores. As it will be randomly accessed during the probe phase and it is likely that the hash table is too large for the first level cache, accesses are likely to induce cache misses.

23.2 Parallel Hash-Join

Another parallelization approach executes both, the hash phase and the probe phase of the join, in parallel. In the first phase, the hash table for the smaller input relation A is calculated in parallel, as outlined in Fig. 23.1. In the second phase, the larger input relation B is probed in parallel against the hash table of A, as described before in Sect. 23.1.

In the first phase, the smaller input relation A is prepared by calculating its hash table. The smaller relation is chosen, in order to keep the resulting hash map as small as possible. The hash table stores a list of positions for each value in the relation. Multiple *hash threads* work independently on the relation A, which is sliced up into multiple parts as shown in step (a) of Fig. 23.1. Each hash thread scans its part of the input table, hashes the values and writes its results into a small local hash table, as outlined by step (b). When the local hash table of a thread reaches a predefined size limit and does not fit into the cache anymore, it is written back to a buffer in main memory and a new local hash table is created by the

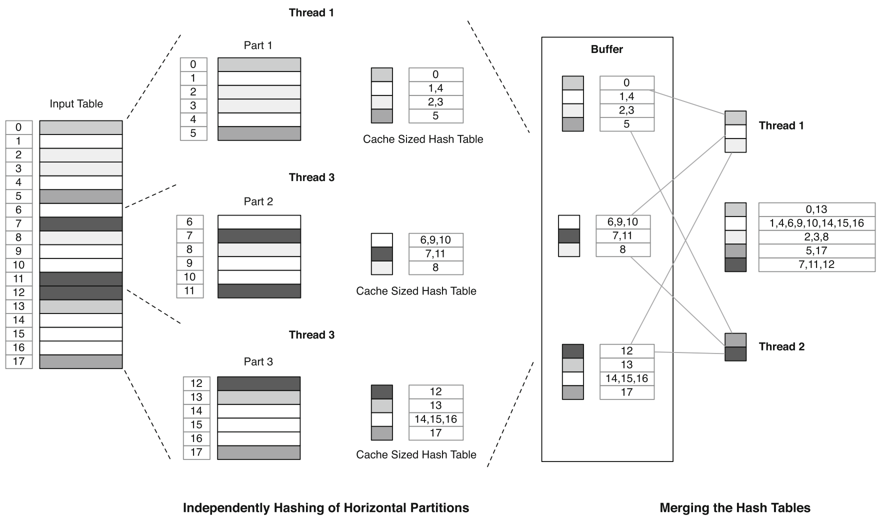


Fig. 23.1 Parallelized hashing phase of a join algorithm

thread. Each full local hash table is added to a queue, symbolized by the *buffer* in step (c). Multiple *merge threads* process the added tables in the buffer, merging them to a unified hash table for relation A, depicted in step (d). Each merge thread only processes its exclusively assigned values, so that the write synchronization on the unified hash table can be reduced to a minimum.

In the second phase, the probing is parallelized over the available threads as outlined before in Sect. 23.1. This means, the larger join relation is partitioned horizontally and each thread stores its results in a local result table. When all threads have finished, the local result tables of the probe threads are concatenated and materialized. For a more detailed discussion of parallel join algorithms, the interested reader is referred to [KKL+09, MBK82].

23.3 Self Test Questions

1. **Parallelizing Hash-Join Phases**
   What is the disadvantage when the probing phase of a join algorithm is parallelized and the hashing phase is performed sequentially?
   (a) Sequentially performing the hashing phase introduces inconsistencies in the produced hash values
   (b) The algorithm still has a large sequential part that limits its potential to scale
   (c) The sequential hashing phase will run slower due to the large resource utilization of the parallel probing phase
   (d) The table has to be split into smaller parts, so that every core, which performs the probing, can finish.

References

[KKL+09] C. Kim, T. Kaldewey, V.W. Lee, E. Sedlar, A.D. Nguyen, N. Satish, J. Chhugani, A. Di Blas, P. Dubey, Sort vs. Hash revisited: fast join implementation on modern multicore CPUs, in *VLDB*, 2009

[MBK82] S. Manegold, P. Boncz, M. Kersten, Optimizing main-memory join on modern hardware. IEEE Trans. Knowl. Data Eng. **19**, 412–426 (2008)

Chapter 24
Parallel Aggregation

Similar to the parallel join mechanism described in Chap. 23, aggregation operations can also be accelerated using parallelism and hash-based algorithms. In this chapter, we discuss how parallel aggregation is implemented in SanssouciDB. Note that multiple other ways to implement parallel aggregation are also conceivable. However, we focus on our parallel implementation using hashing and thread-local storage.

24.1 Aggregate Functions Revisited

The concept of aggregate functions has already been discussed in Chap. 20. Aggregate functions operate on single columns but usually take a large number of rows into account. Examples for aggregate functions are *COUNT, SUM, AVERAGE, MIN, MAX* or *STDDEV*. Which aggregates are returned is typically specified using *GROUP BY* and/or *HAVING* clauses in SQL. The *GROUP BY* clause is used to express that the aggregate function shall be computed for every distinct value of the specified attribute(s). For example, the following query would incur a *COUNT* operation with two different counters, namely one for female and another one for male entries:

```
SELECT COUNT(*)
FROM world_population
GROUP BY gender;
```

Listing 24.1: Simple SQL query using the COUNT aggregate function

The *HAVING* clause behaves similar to the SQL *WHERE* clause, the difference being that the filter criterion is specified with respect to the aggregate function.

24.2 Parallel Aggregation Using Hashing

Let us revisit the following example from Chap. 20:

```
SELECT country, COUNT(*) AS citizens
FROM world_population
GROUP BY country;
```

Listing 24.2: Another SQL example using the COUNT aggregate function

The result of this query lists the number of all citizens for every distinct value of the “country” column. Table 24.1 shows what the result might look like:

In the following, we describe how this result is computed using the parallel aggregation algorithm in SanssouciDB. Figure 24.1 visualizes how the algorithm works. It consists of two phases, the hashing phase and the aggregation phase.

In the hashing phase, shown in the upper part of Fig. 24.1, our world population table is horizontally partitioned into n parts (cf. Sect. 9.3). The chosen n determines how many threads will be used for the parallel aggregation. In a lightly loaded system, n might be as high as the number of CPU cores available on the machine (cf. Chap. 22). Note that the horizontal partitioning occurs logically and dynamically, i.e. the way the table is stored physically remains unaltered. Each of the n threads is now assigned a partition and creates an empty thread-local hash table. The hash table is used to store

- the hash value of a country and
- the number of occurrences of that country in the world population table.

The threads then begin to scan the “country” column. For each row in the partition of a thread, the thread checks whether the current country is already contained in its thread-local hash table. If yes, the number stored in correspondence with the hash value of the country is increased by one. If not, the hash value of the country is stored along with a “1”, since the thread has just found the first inhabitant of the country in the current partition. Note that creating a new entry in the hash table might result in a situation where the size of the hash table exceeds the size of the CPU cache. Whenever this would be the case, the hash table is stored and a new

Table 24.1 Possible result for query in listing 24.2

Country	Citizens
China	1,347,350,000
France	65,350,000
Germany	81,844,000
Italy	59,464,000
India	1,210,193,000
United Kingdom	62,262,000
United States	314,390,000
Japan	127,570,000

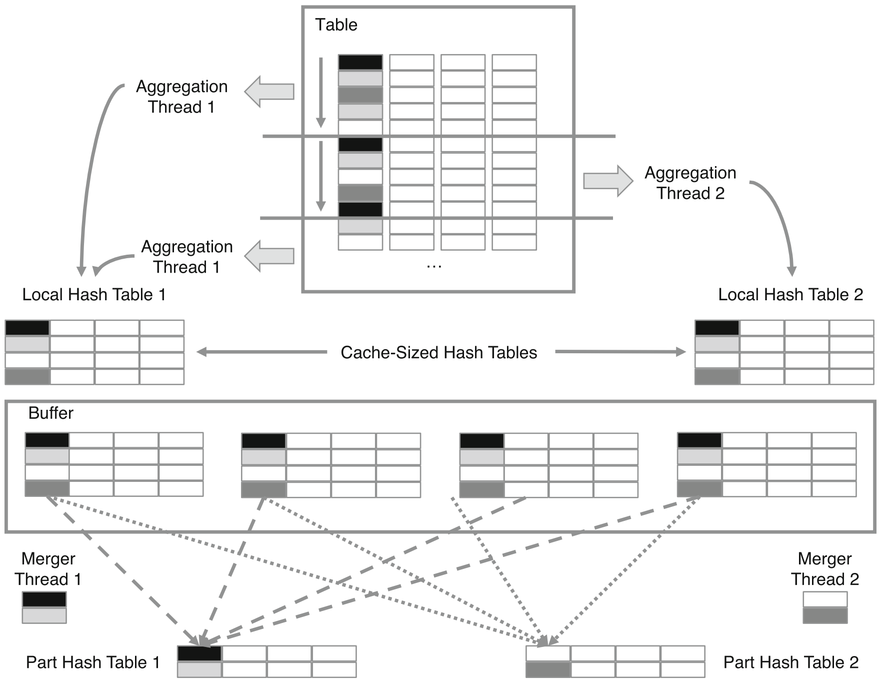


Fig. 24.1 Parallel aggregation in SanssouciDB

hash table is created. The new key/value pair is inserted into the just created empty hash table. Ensuring that the hash table never exceeds the size of the CPU cache is crucial for facilitating fast lookups in the hash table. Since a lookup occurs for every row, which is being scanned, the lookup and alteration of the hash table should not incur a cache miss. When all threads have finished scanning their assigned partition, the aggregation phase begins.

In the aggregation phase, the buffered hash tables are merged. This is accomplished by so-called merger threads. In this phase the results from the part hash-tables are further aggregated. The buffered tables are merged using range partitioning. Each merger thread is responsible for a certain range, indicated by a different color in Fig. 24.1. The partitioning criterion is defined on the keys of the local hash tables.

Considering the example in Listing 24.1, the hash values for the "gender" attribute could be partitioned as follows: All hash keys whose binary representation starts with "0" are assigned to one merger thread, and all keys whose binary representation starts with "1" are assigned to another merger thread. In our example given in Listing 24.2, the hash values for the "country" attribute might be partitioned into eight regions. Thus, assuming that there are roughly 200 countries in the world, each merger thread would be responsible for $\sim \frac{200}{8}$ countries if the hash function ensures a mostly uniform value distribution.

Each merger thread accesses all buffered hash tables and looks for hash values in the range that is has been assigned. Since access to the buffered hash tables is read-only, there are no contention issues due to synchronization. Similar to the hashing phase, each merger thread has a private hash table where it maintains the total number of citizens per country as it goes through all buffered hash tables from the previous phase. The result is obtained by simply concatenating the private hash tables of the merger threads.

A more detailed description of the parallel aggregation algorithm in SanssouciDB can be found in [Pla11].

24.3 Self Test Questions

1. **Aggregation—GROUP BY**
   Assume a query that returns the number of citizens of a country, e.g.: SELECT country, COUNT(·)
   FROM world_population
   GROUP BY country;

   The world_population table contains the names and countries of all citizens of the world.

   The GROUP BY clause is used to express ...

   (a) the graphical format of the results for display
   (b) an additional filter criteria based on an aggregate function
   (c) that the aggregate function shall be computed for every distinct value of country
   (d) the sort order of countries in the result set.

2. **Number of Threads**
   How many threads will be used during the second phase of the described parallel aggregation algorithm when the table is split into 20 chunks and the GROUP BY attribute has 6 distinct values?

   (a) exactly 20 threads
   (b) at most 6 threads
   (c) at least 10 threads
   (d) at most 20 threads.

Reference

[Pla11] H. Plattner, D.B. Sanssouci, An in-memory database for processing enterprise workloads, in *BTW, LNI*, ed. by T. Härder, W. Lehner, B. Mitschang, H. Schöning, H. Schwarz, vol. **180** (GI, 2011), pp. 2–21

Part IV
Advanced Database Storage Techniques

Chapter 25
Differential Buffer

25.1 The Concept

The database architecture discussed so far was optimized for read operations. In the previously described approach an insert of a single tuple can force a restructuring of the whole table if a new attribute value occurs and the dictionary has to be resorted. To overcome this problem, we will introduce the differential buffer in this chapter.

The concept of the differential buffer (sometimes also called "delta buffer" or "delta store") divides the database into a main store and the differential buffer. All inserts, updates, and delete operations are performed on the differential buffer. Thus, data modifications happen in the differential buffer only. The read-optimized main store is not touched by any data modifying operation. The overall current state of the data is the conjunction of the differential buffer and the main store, thus every read operation has to be performed on the main store and the differential buffer, too. Since the differential buffer is orders of magnitudes smaller than the main store, this has only a small impact on the reading performance.

During query execution, a query is logically split into a query on the compressed main partition and the differential buffer. After the results of both subsets are retrieved, the intermediate representation must be combined to build the full valid result representing the current state of the database (Fig. 25.1).

25.2 The Implementation

In the differential buffer, we keep the concept of a column-oriented layout and the use of dictionaries. However, to improve write performance, the dictionaries are not sorted but the values are stored in insertion order. Thus, resorting of the differential buffer will not occur. To speed up accesses to values in the unsorted dictionary, we use CSB+ trees [RR00]. The CSB+ tree is mainly used as an index to look up valueIDs for a given value. While this overall approach is optimized for write performance the biggest drawback of this data storage format is that we cannot execute the queries in the exact same way as we do it in the main partition.

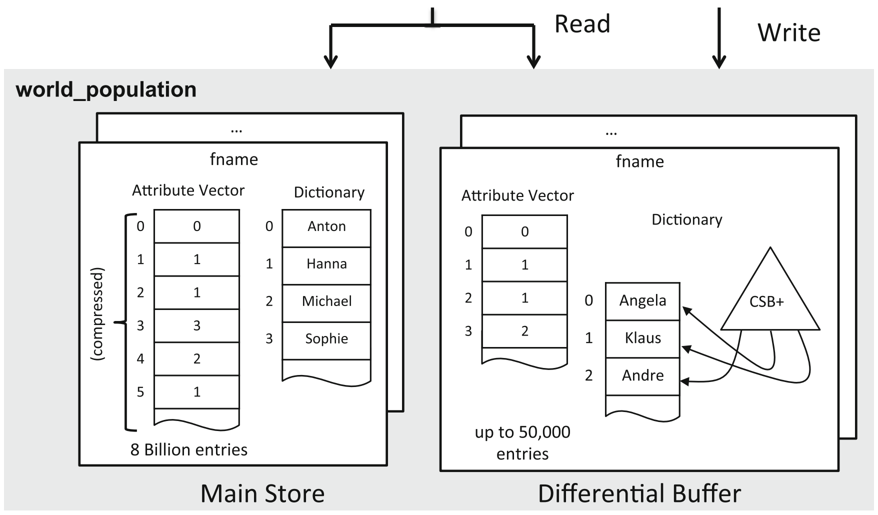


Fig. 25.1 The differential buffer concept

One example are range queries. In the compressed main partition we can guarantee based on the explicit order of the valueIDs that we can represent a range of valueIDs by its boundaries. Due to the insertion order of the tuples in the differential buffer we have to explicitly identify and carry each qualifying valueID as for comparison. For simple single point queries that only access one tuple this is not of importance, but as stated, might cause a problem for the mentioned range queries [HBK+11].

The general implementation of the differential buffer is thereby as follows. First, we keep a list of all occurring values and a CSB+ tree to allow for logarithmic search in all unique values. While the unique values are not stored in a specific order as it is done in the compressed main partition, the CSB+ tree allows to define an ordering criterion on an attribute to perform fast searches on that attribute. The disadvantage of this approach is the space overhead of the tree structure. While the tree is not required to explicitly store the actual value, which is stored in a separate list in insertion order, it still needs to build the required tree structure and thus the space overhead can be approximated by a factor of two depending on the actual fill level of the leaf nodes inside the tree.

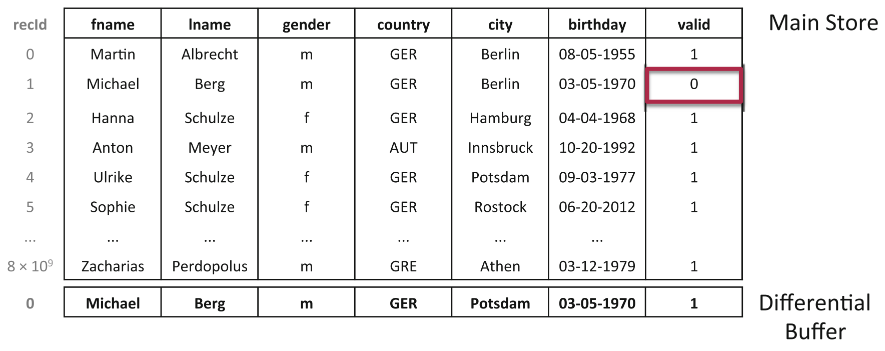
Since read performance is critical to our mixed workload enterprise applications, we need to make sure that the differential buffer size is always kept as small as possible. To achieve this we use an online reorganization process that merges the changes that are stored in the differential buffer with the compressed main partition to build a new compressed main partition. The detailed description of the merge process will follow in Chap. 27.

25.3 Tuple Lifetime

Because the compressed main partition of a table cannot be modified, we need a new way to identify tuple lifetime for the records stored there. If we have to update a record in the compressed main partition the first idea is that we will now perform an additional insert with the same values in the delta partition and do nothing in the main partition. The problem that arises with this implementation is that we then can no longer distinguish between the result of the compressed main partition and the delta partition. This is especially important if there are multiple modifications for a single record. To overcome this limitation we need to add a special system bit vector to the table that manages the validity of a tuple in the compressed main partition and the differential buffer. For each record, this validity vector stores a single bit that indicates if the record at this position is valid or not. To ensure fast read and write access to this vector, it stays uncompressed.

During query execution the lookup of the valid tuple is handled as follows: First the query is executed normally as it would be without the validity vector. In parallel, we execute the same query in the differential buffer. Afterwards, when the result for the compressed main partition is available, the result positions are verified with the validity vector to remove all positions from the intermediate result that are not valid. This strategy is illustrated in Fig. 25.2. In this example Michael Berg moves from Berlin to Potsdam. Since we cannot modify the main structure directly we have to execute two operations. First, we invalidate the old tuple in the main partition by unsetting the valid bit, and second, we insert the complete tuple with the new location in the differential buffer so we have all information about Michael Berg available again. Both operations have to be able to be executed atomically as one single operation so that no information will be lost at any time.

The drawback of this approach is that during query execution a small additional overhead is added. However, the benefits greatly outweigh the disadvantages, especially because using specialized SIMD instructions enables us to check multiple positions at once for validity.



recId	fname	lname	gender	country	city	birthday	valid	
0	Martin	Albrecht	m	GER	Berlin	08-05-1955	1	Main Store
1	Michael	Berg	m	GER	Berlin	03-05-1970	$\boxed{0}$	
2	Hanna	Schulze	f	GER	Hamburg	04-04-1968	1	
3	Anton	Meyer	m	AUT	Innsbruck	10-20-1992	1	
4	Ulrike	Schulze	f	GER	Potsdam	09-03-1977	1	
5	Sophie	Schulze	f	GER	Rostock	06-20-2012	1	
...	...	...	...	...	...	...		
8×10^9	Zacharias	Perdopolus	m	GRE	Athen	03-12-1979	1	
0	**Michael**	**Berg**	**m**	**GER**	**Potsdam**	**03-05-1970**	**1**	Differential Buffer

Fig. 25.2 Michael Berg moves from Berlin to Potsdam

25.4 Self Test Questions

1. **The Differential Buffer**
   What is the differential buffer?
   (a) A buffer where exception and error messages are stored
   (b) A buffer where different results for one and the same query are stored for later usage
   (c) A dedicated storage area in the database where inserts, updates and deletions are buffered
   (d) A buffer where queries are buffered until there is an idle CPU that takes one new task over.
2. **Performance of the Differential Buffer**
   Why might the performance of read queries decrease, if a differential buffer is used?
   (a) Because only one query at a time can be answered by using the differential buffer
   (b) Because read queries have to go against the main store and the differential buffer, which is write-optimized
   (c) Because inserts collected in the differential buffer have to be merged into the main store every time a read query comes in
   (d) Because the CPU cannot perform the query before the differential buffer is full.
3. **Querying the Differential Buffer**
   If we use a differential buffer, we have the problem that several tuples belonging to one real world entry might be present in the main store as well as in the differential buffer. How did we solve this problem?
   (a) This statement is completely wrong because multiple tuples for one real world entry must never exist
   (b) All attributes of every doubled occurrence are set to NULL in the compressed main store.
   (c) We introduced a validity bit
   (d) We use a specialized garbage collector that just keeps the most recent entry.

References

[HBK+11] F. Hübner, J.-H. Böse, J. Krüger, C. Tosun, A. Zeier, H. Plattner, A cost-aware strategy for merging differential stores in column-oriented in-memory DBMS. in *BIRTE* (2011), pp. 38–52

[RR00] J. Rao, K.A. Ross, Making b+- trees cache conscious in main memory. in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD '00*, (ACM, New York, 2000), pp. 475–486

Chapter 26
Insert-Only

26.1 Definition of the Insert-Only Approach

Data stored in database tables changes over time and these changes should be traceable, as enterprises need to access their historical data. Additionally, the possibility to access any data that has been stored in the database in the past and keeping historical data is mandatory e.g. for financial audits.

In this chapter, we discuss the insert-only approach. This approach means that applications do not perform updates and deletions on physically stored data tuples, but add tuples and manage their validity instead. A small glimpse how this can be done was already shown with the validity column in Chap. 25. By using insert-only, all data changes are recorded in the same logical database table, we abstract from the introduced separation between the main store and the differential buffer here. In other words, the insert-only approach can be formulated as following: outdated data is not overwritten or deleted, but invalidated. Invalidation can be done by means of additional attributes which indicate the current revision of a tuple. That makes accessing the previous revisions of data very simple: just by using the key of the tuple and the revision attribute results in the retrieval of the requested tuple in its desired version. This already provides the kind of traceability that is legally required for financial applications in many countries. In addition, there are some business related benefits and some technical reasons for insert-only, such as:

- So called time travel queries are easily possible. Time travel queries allow users to see the data like it was in certain point in the past. Simple access to historical data helps a company's management to efficiently analyze the history and development of the enterprise, which can be helpful for strategic decisions;
- This approach can simplify implementation of parallelization mechanisms, e.g. multiversion concurrency control (will be discussed below);
- In the context of in-memory column-store databases, where the encoded data is stored and appropriate dictionaries for the attributes are used, an insert-only approach simplifies working with dictionaries, as no dictionary cleaning is necessary in this case.

DOI: 10.1007/978-3-642-36524-9_26, © Springer-Verlag Berlin Heidelberg 2013

So, how do we differentiate between the actual tuples and outdated ones? We consider the two following implementation possibilities:

- Point representation: to determine the validity of tuples, one field is used. In the field "valid from" the insertion date is stored.
- Interval representation: to determine the validity of tuples, two fields are used. The fields contain information about the valid time interval: "valid from" and "valid to" dates.

Let us consider in more detail the corresponding implementation, use cases, advantages, and disadvantages of both implementation possibilities. To explain the approaches, we take the example table "world_population". In the following, we explain the concept with dates for simplicity. In reality, timestamps with a precision of one microsecond are used.

26.2 Point Representation

When using point representation, the "valid from" date is stored with every tuple in the database table. The field contains the date from the moment when the tuple was created. The obvious advantage of this method is a fast write of new tuples. On any update, only the tuple with the new values and the current "valid from" date has to be entered, the other tuples do not need to be changed. Consider the following initial state of the table (Table 26.1). Please note that this time the ids are stored explicitly and will be used to reference tuples (in reality, any primary key for the tuples will do, it does not have to be separate ids).

Now we want to update the tuple with id 1, because Michael moves from Berlin to Hamburg. The update in the database table is done on 07-02-2012. In case of an insert-only approach and point representation, the update result will look as depicted in Table 26.2.

So, the old data is not deleted, but the new record for the tuple with the same key and the different "valid from" date is inserted. To do the update, the user or the client application issues the following SQL statement:

```
UPDATE world_population
SET city = 'Hamburg'
WHERE id = 1
```

Listing 26.1: Update request from user or application

Table 26.1 Initial state of example table using point representation

Id	Fname	Lname	Gender	Country	City	Birthday	Valid from
0	Martin	Albrecht	m	Germany	Berlin	08-05-1955	10-11-2011
1	Michael	Berg	m	Germany	Berlin	03-05-1970	10-11-2011
2	Hanna	Schulze	f	Germany	Hamburg	04-04-1968	10-11-2011

Table 26.2 Example table using point representation after updating the tuple with $id = 1$

Id	Fname	Lname	Gender	Country	City	Birthday	Valid from
0	Martin	Albrecht	m	Germany	Berlin	08-05-1955	10-11-2011
1	Michael	Berg	m	Germany	Berlin	03-05-1970	10-11-2011
2	Hanna	Schulze	f	Germany	Hamburg	04-04-1968	10-11-2011
1	Michael	Berg	m	Germany	Hamburg	03-05-1970	07-02-2012

The update statement can be semantically regarded as the following insert operation (Listing 26.2), if the overwriting behavior of the update is discarded, which is the case for insert only. All attributes that are not directly mentioned in the update statement are retrieved from the most recent entry of the tuple. So internally, the following insert statement is executed:

```
INSERT INTO world_population
VALUES (1, 'Michael', 'Berg', 'm', 'Germany',
        'Hamburg', '03-05-1970', '07-02-2012')
```

Listing 26.2: Generated insert statement from update request

At the same time, a point representation approach causes the following disadvantage: it can be less efficient for read operations when the user only needs the most recent tuples. Every time when searching for the recent tuple, the other tuples of that entry (so here these with the same id) have to be checked, to make sure that the most recent one has been found. In other words, for identifying which tuple is the most recent one, we need to fetch all tuples and sort them by their `valid_from` timestamp. After that we get the most recent tuple.

Taking the updated example table, let us assume we want to select the newest record with the $id = 1$. The following operation has to be performed in this case (Listing 26.3):

```
SELECT * FROM world_population
WHERE id = 1
ORDER BY validFrom DESC LIMIT 1
```

Listing 26.3: Point representation: retrieve most recent entry

The mentioned properties make the point representation approach efficient for OLTP operations where write operations are required more often than read operations.

26.3 Interval Representation

When using interval representation, both "valid from" and "valid to" dates are stored with every tuple in the database table. The fields contain the creation date of a tuple and the point in time when it was invalidated.

Table 26.3 Initial state of example table using interval representation

Id	Fname	Lname	Gender	Country	City	Birthday	Valid from	Valid to
0	Martin	Albrecht	m	Germany	Berlin	08-05-1955	10-11-2011	
1	Michael	Berg	m	Germany	Berlin	03-05-1970	10-11-2011	
2	Hanna	Schulze	f	Germany	Hamburg	04-04-1968	10-11-2011	

Table 26.4 Example table using interval representation after updating the tuple with *id* = 1

Id	Fname	Lname	Gender	Country	City	Birthday	Valid from	Valid to
0	Martin	Albrecht	m	Germany	Berlin	08-05-1955	10-11-2011	
1	Michael	Berg	m	Germany	Berlin	03-05-1970	10-11-2011	07-02-2012
2	Hanna	Schulze	f	Germany	Hamburg	04-04-1968	10-11-2011	
1	Michael	Berg	m	Germany	Hamburg	03-05-1970	07-02-2012	

As in the point representation, in interval representation the complete tuple is stored on attribute change with the "valid from" date. Additionally, the "valid to" date field is updated in the tuple that is substituted by a newer one. This "valid to" date will be equal to the "valid from" date in the new tuple. Obviously, a write operation is more complex in this case. Consider the same table as in the point representation example, with the initial state shown in Table 26.3.

Again, we want to update the tuple with the *id* = 1 again such that the city will be changed to 'Hamburg'. In case of interval representation, the result of the update will look as depicted in Table 26.4.

Not only the new tuple with the updated field values is inserted, but also the old tuple is updated. Two operations have to be done in this case, and that makes writes for the interval representation less efficient (see Listing 26.4).

```
(I)     UPDATE world_population SET validTo = '07-02-2012'
        WHERE id = 1 AND validTo IS NULL

(II)    INSERT INTO world_population
        VALUES (1, 'Michael', 'Berg', 'm', 'Germany',
                'Hamburg', '03-05-1970', '07-02-2012')
```

Listing 26.4: Operations for update in insert only

On the other hand, an additional "valid to" field simplifies read operations in comparison to point representation. In the case of an interval representation, there is no need to fetch and sort the tuples to get the most recent ones; only the tuples with the appropriate key and an empty "valid to" date have to be selected. To select the most recent tuple for Michael, the following operation is executed (Listing 26.5):

```
SELECT *
FROM world_population
WHERE id = 1 AND validTo IS NULL
```

Listing 26.5: Interval representation: retrieve most recent entry

The mentioned properties make the interval representation approach especially efficient for OLAP operations, where read operations are required more often than write operations.

26.4 Concurrency Control: Snapshot Isolation

Taking the multi-core architecture of modern CPUs and the possibility to parallelize queries into account, different ways of parallelization and concurrency control have to be investigated. As mentioned above, an insert-only approach not only helps to match business requirements, but also simplifies technical aspects of an in-memory column store. Let us discover in more detail how the insert-only approach can help to simplify snapshot isolation.

The following approaches for concurrency control are commonly used:

- Locking: in this case, a transaction that locks resources works with them exclusively. An operation can be started only if all required resources are available.
- Optimistic concurrency control: in this case, data for an operation is stored isolated, in a so called virtual snapshot.

In the secondly mentioned approach, all manipulations are done on the data that is valid from the time the transaction was started which is the so-called virtual snapshot. So, when using multiversion concurrency control, transactions that need to update data actually insert new versions into the database, but concurrent transactions will still see a consistent state based on previous versions, because they all work on their own "virtual copy" of the data. Obviously, that can cause conflicts.

Now, let us see an example how the insert-only approach can simplify multiversion concurrency control. We consider interval representation in this example. Let us take the following simple table about the salary of employees, with the following state at the time point T1 (Table 26.5):

This table will be the source data for two concurrently executing transactions. T_1 is the time when the first transaction starts. It works on the data for the employee with the id 0. At the point of time T_1, it reads the salary value "10000". At the point of time T_3 (which is in our example again '07-07-2012' and a timestamp exact to microseconds in reality), it updates the record with id 0 as following (Table 26.6):

Table 26.5 Example table using interval representation to show concurrent updates

EmplId	Salary	Valid from	Valid to
0	10000	10-11-2011	
1	20000	10-11-2011	
2	15000	10-11-2011	

Table 26.6 Example table using interval representation to show concurrent updates after first update

EmplId	Salary	Valid from	Valid to
0	10000	10-11-2011	07-07-2012
1	20000	10-11-2011	
2	15000	10-11-2011	
0	12000	07-07-2012	

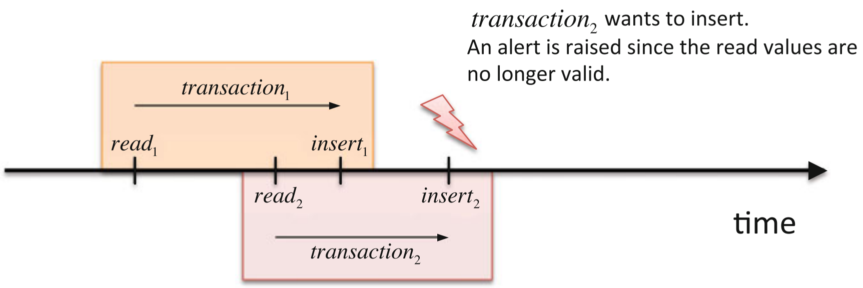

$transaction_2$ wants to insert.
An alert is raised since the read values are no longer valid.
$transaction_1$
$read_1$
$insert_1$
$read_2$
$insert_2$
$transaction_2$
time

Fig. 26.1 Snapshot isolation

The second concurrent transaction starts at the point of time T_2, so that T_2 lies in between T_1 and T_3. It also works with the data with $id = 0$. But at the moment T_2, it has no access to the updated version by the first transaction, and cannot see the updates done by the first transaction at the point of time T_3. The concurrency control starts to work when the second transaction tries to update the record with id = 0 at the point in time T_4, which is later than T_3 (Fig. 26.1).

26.5 Insert-Only: Advantages and Challenges

As described above, we never delete data from a table. This raises the question "How will the insert-only approach influence memory consumption?". On any change of the tuple, for each row update, an additional tuple with the additional timestamp will be inserted into the database and the stored data volume will increase. This fact seems to increase memory consumption considerably. But is it really so? To answer this question, let us see what types of updates are usually performed in business applications:

- Aggregate updates
- Status updates
- Value updates

Taking the advantages of an in-memory column-based database into account, aggregates can be efficiently calculated on the fly, so we can avoid the updates of the first type altogether. Concerning the remaining update types, a study was conducted at HPI [KKG+11]. The study showed that a typical SAP financials application is not update-intensive: only an average of about 5 % of all operations are updates (see Chap. 3). That already lessens the problem much. Adding the fact that most of these updates are status updates, another simple trick helps to reduce the remaining impact on memory consumption. Since most status fields only consist of one bit of information, which means that a status is now set, they can directly be replaced with the timestamp of the change. When doing the status update this way, all information is encapsulated in one field and the update can be done in place, so no additional record has to be written. These characteristics and improvements lead to the result that the memory consumption only increases moderately when insert-only is being used.

26.6 Self Test Questions

1. **Statements Concerning Insert-Only**
   Considering an insert-only approach, which of the following statements is true?
   (a) When given a differential buffer, historical data can be used to further speed up the insert performance
   (b) Old data items are deleted as they are not necessary any longer
   (c) Historical data has to be stored in a separate database to reduce the overall database size
   (d) Data is not deleted, but invalidated instead.

2. **Benefits of Historic Data**
   Which of the following is NOT a reason why historical data is kept by an enterprise?
   (a) Historic data can be used to analyze the development of the company
   (b) It is legally required in many countries to store historical data
   (c) Historical data can provide snapshots of the database at certain points in time
   (d) Historical data can be analyzed to boost query performance.

3. **Accesses for Point Representation**
   Considering point representation and a table with one tuple, that was invalidated five times, how many tuples have to be checked to find the most recent tuple?
   (a) Five
   (b) Two, the most recent one and the one before that
   (c) Only one, that is, the first which was inserted
   (d) Six.

4. **Physical Delete instead of Insert-Only**
   What would be necessary if physical deletion of tuples was implemented in SanssouciDB?
   (a) Dictionary cleaning which would cause rewriting of the attribute vector
   (b) The latest snapshot has to be reloaded after a deletion to maintain data integrity
   (c) Deletion of tuples is part of SanssouciDB
   (d) Assurance of compatibility to other DBMS.
5. **Statement concerning Insert-Only**
   Which of the following statements concerning insert-only is true?
   (a) Point representation allows faster read operations than interval representation due to its lower impact on tuple size
   (b) In interval representation, four operations have to be executed to invalidate a tuple
   (c) Interval representation allows more efficient write operations than point representation
   (d) Point representation allows more efficient write operations than interval representation.

Reference

[KKG+11] J. Krueger, C. Kim, M. Grund, N. Satish, D. Schwalb, J. Chhugani, H. Plattner, P. Dubey, A. Zeier, Fast updates on read-optimized databases using multi-core CPUs, in *PVLDB*, 2011

Chapter 27
The Merge Process

Using a differential buffer as an additional data structure to improve write performance requires to cyclically combine this data with the compressed main partition. This process is called "merge".

The reasons for merging are two-fold. On the one hand, merging the data of the differential buffer into the compressed main partition decreases the memory consumption due to better compression. On the other hand, merging both data structures improves the read performance due to the sorting of the dictionary of the read-optimized main store. In an enterprise context, the requirements of the merge process are that it

- can be executed asynchronously,
- has as little impact as possible on all other operations, and
- does not block any OLTP or OLAP transactions.

To achieve this goal, the merge process creates a new empty differential buffer and a copy of the main store upfront the actual merging phase to avoid looking during the merge. Incoming data modifications are passed to the new differential buffer. Using this approach, we are able to reduce the time when we have to explicitly lock the compressed main and delta partitions. Using the copy of the main store and the new differential buffer, the merged table is only locked for the short time period when the pointer to the main store of a table is set to the new main store after the merge. In this online merge concept, the table is available for reads and the second delta for write operations during the merge process.

Now, we will describe the merge algorithm and acquired locks in more detail. During the merge, the process requires additional system resources (CPU and main memory) that have to be considered during system sizing and scheduling.

For the differential buffer concept, it is important to mention that all update, insert, and delete operations are captured as technical inserts into the differential buffer while a dedicated valid tuple vector per table and store ensures consistency. When using a differential buffer, the update, insert, and delete performance of the database is limited by two factors:

- the insert rate into the write-optimized structure and
- the performance with which the system can merge the accumulated modifications into the read-optimized main store.

By introducing a differential buffer, the read performance is decreased depending on the number of tuples in the differential buffer. Join operations are especially slowed down since results have to be materialized. This means that intermediate values from the differential buffer cannot be directly compared to those from the compressed main partition. This can force the execution engine to switch to an execution based on early materialization which can have a severe performance impact (see Chap. 16). Consequently, the merge process has to be executed if the performance impact becomes too large. It is triggered by one of the following events:

- The number of tuples in the differential buffer for a table exceeds a defined threshold.
- The memory consumption of the differential buffer exceeds a specified limit.
- The differential buffer log for a columnar table exceeds the defined limit.
- The merge process is triggered explicitly by a specific SQL command.

27.1 The Asynchronous Online Merge

To enable the execution of queries during a running merge operation, we introduce the concept of an asynchronous merge. The overall requirement for this process is that it will not block any concurrent modifying transactions. Figure 27.1 illustrates this concept.

By introducing a second differential buffer during the merge phase, data changes on the table can still be applied, even during the merge process. Consequently, read operations have to access both differential buffers to query the

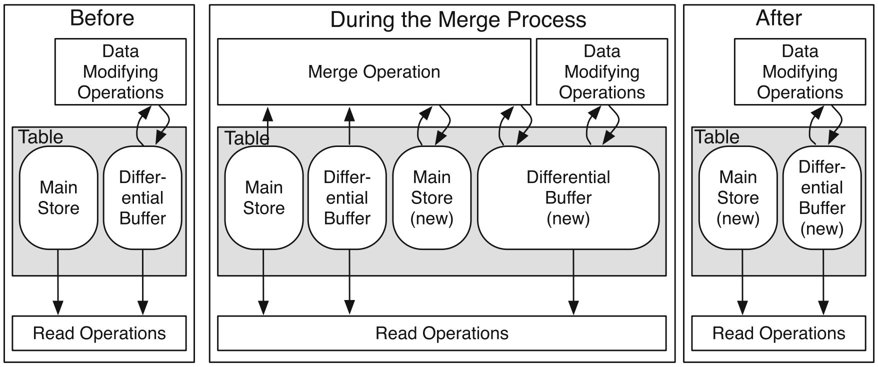


Fig. 27.1 The concept of the online merge

current state during the merge. To maintain consistency, the merge process requires a lock at the beginning and at the end of the process of switching the stores and applying necessary data modifications—such as valid tuple modifications—which have occurred during the merge process. Open transactions are not affected by the merge, since their changes are copied from the old into the new differential buffer and can be processed in parallel to the merge process. To finish the merge operation, the old main store is replaced with the new one. Within this last step of the merge process, a snapshot of the new main store is persisted, which defines a new starting point for log replay in case of failures (see Chap. 28).

The merge process consists of three phases: (i) prepare merge, (ii) attribute merge, and (iii) commit merge. Phase (ii) is hereby carried out for each attribute of the table.

27.1.1 Prepare Merge Phase

The prepare merge phase locks both the differential buffer as well as the main store, and creates a new empty differential buffer for all new inserts, updates, and deletes that occur during the merge process. Additionally, the current validity vectors of the old differential buffer and the main store are copied because these may be changed by concurrent updates or deletes applied during the merge, which may affect tuples involved in this process.

27.1.2 Attribute Merge Phase

The attribute merge phase as outlined in Fig. 27.2 consists of two steps. In the first step of the attribute merge operation, the differential buffer and main store dictionaries are combined into one sorted result dictionary. In addition, a value mapping is created as an auxiliary mapping structure to map the positions from the old dictionary to the new dictionary for the differential buffer and the main store. These auxiliary structures are actually not necessary for the algorithm, but avoid expensive lookups in the old dictionaries and improve cache utilization.

The input dictionaries to be consolidated are the main store's dictionary and the sorted dictionary extracted from the differential buffer's CSB+ tree. Having the sorted dictionaries, both are merged and form the resulting dictionary containing the main store's and differential buffer's distinct values.

In the second step of the attribute merge, the values from the two attribute vectors are copied into a new combined attribute vector. Therefor the auxiliary structure is used. To ensure that the sizing of the new attribute vector is correct, we calculate the required value width based on the size of the new dictionary.

An exemplary run of the attribute merge phase is shown in Sect. 27.2.

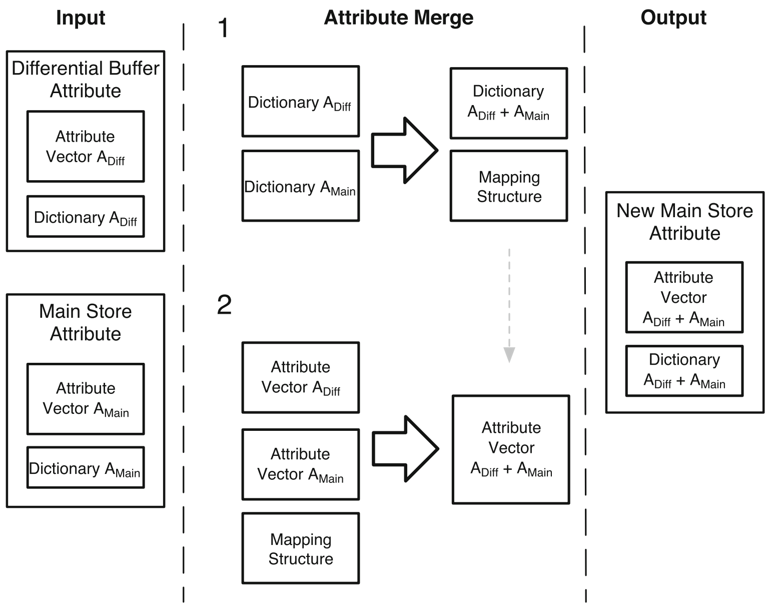

Input
Attribute Merge
Output
Differential Buffer Attribute
Attribute Vector A_{Diff}
Dictionary A_{Diff}
Main Store Attribute
Attribute Vector A_{Main}
Dictionary A_{Main}
1
Dictionary A_{Diff}
Dictionary A_{Main}
Dictionary $A_{Diff} + A_{Main}$
Mapping Structure
2
Attribute Vector A_{Diff}
Attribute Vector A_{Main}
Mapping Structure
Attribute Vector $A_{Diff} + A_{Main}$
New Main Store Attribute
Attribute Vector $A_{Diff} + A_{Main}$
Dictionary $A_{Diff} + A_{Main}$

Fig. 27.2 The attribute merge

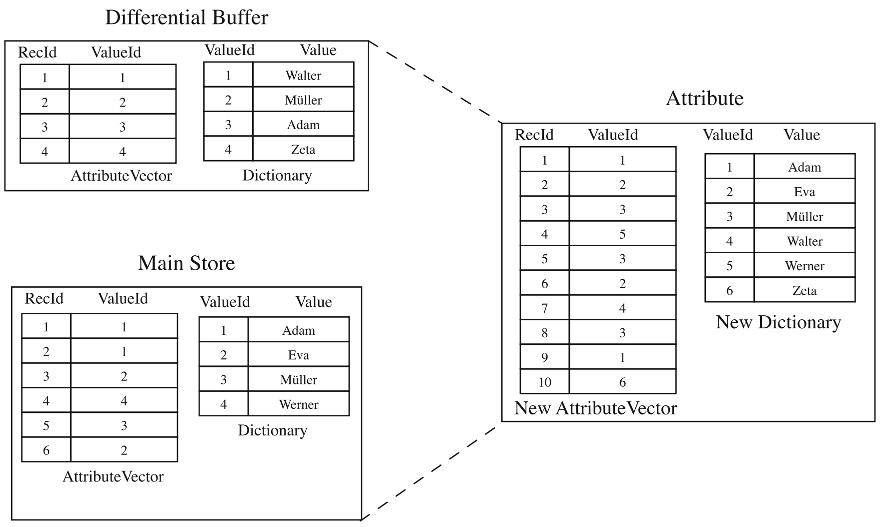
27.1.3 Commit Merge Phase

The commit merge phase starts by acquiring a write lock of the table. This ensures that all running queries are finished before the switch to the new main store including the updated valueIDs takes place. Then, the valid tuple vector that was copied in the prepare phase is applied to the actual vector to mark invalidated rows. As the last step, the new main store replaces the original differential buffer as well as the old main store and the memory allocations of both are freed.

The result of the merge process for a simple example is shown in Fig. 27.3. The new attribute vector holds all tuples of the original main store, as well as the ones from the differential buffer. Note that the new dictionary includes all values from the main store and the differential buffer and it is sorted to allow for binary search and range queries that incorporate materialization query execution strategies.

27.2 Exemplary Attribute Merge of a Column

The Attribute Merge described in Sect. 27.1 will be explained with a simplified example, showing the merge of a single column. Please note that this does include the optimizations of the single column merge concept described in Sect. 27.3. As also shown in Fig. 27.4, the overall process consists of two distinct steps.


Differential Buffer

AttributeVector

RecId	ValueId
1	1
2	2
3	3
4	4

Dictionary

ValueId	Value
1	Walter
2	Müller
3	Adam
4	Zeta

Main Store

AttributeVector

RecId	ValueId
1	1
2	1
3	2
4	4
5	3
6	2

Dictionary

ValueId	Value
1	Adam
2	Eva
3	Müller
4	Werner

Attribute

New AttributeVector

RecId	ValueId
1	1
2	2
3	3
4	5
5	3
6	2
7	4
8	3
9	1
10	6

New Dictionary

ValueId	Value
1	Adam
2	Eva
3	Müller
4	Walter
5	Werner
6	Zeta

Fig. 27.3 The attribute in the main store after the merge process

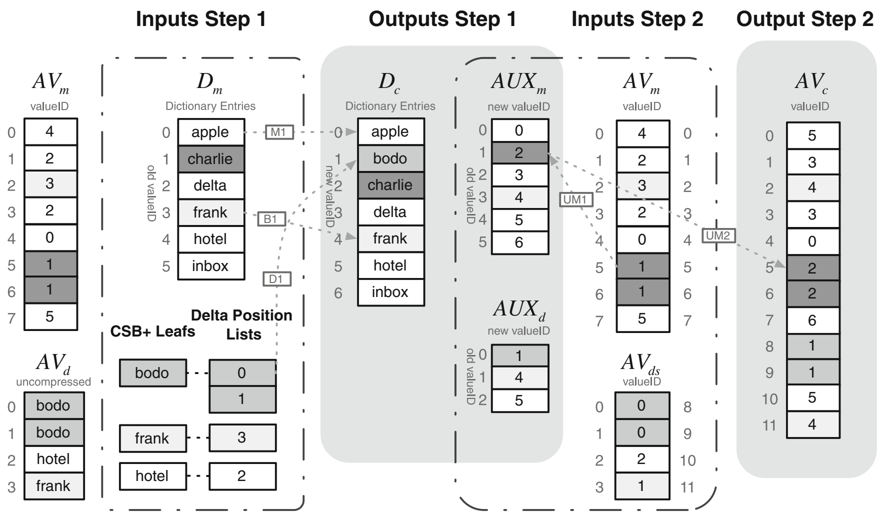


Fig. 27.4 The merge process for a column (adapted from [FSKP12])

The first step takes the dictionary of the main store's attribute vector (here denoted as D_m) of the main store column (AV_m) and its counterpart, the sorted list of values (AV_d), which is extracted from the differential buffer's CSB+ tree, as input and produces the combined sorted dictionary D_c as well as the auxiliary structures AUX_m and AUX_d as output. To merge both dictionaries, first the pointer on both lists is set to the first element. The values of the pointers of both dictionaries are then compared in each iteration, the smaller value is added to the result, and the corresponding pointer is incremented. In the event that both values are equal, the value is added once and both pointers are incremented. To be able to later update the attribute vector to the newly created dictionary, every time a value is added to the merged dictionary the mapping information from the corresponding old dictionary to the merged one is added to the corresponding auxiliary structure. If one of the two pointers reaches the end of its input dictionary, the remaining items of the other dictionary are copied directly to the result vector, since both dictionaries are sorted.

That means, that at the beginning of this step, the pointer on D_m marks "apple", and the pointer on AV_d marks "bodo". As "apple" is lexically located before "bodo", it is added to D_c first and the pointer on D_m is advanced to "charlie". Afterwards, "bodo" is added to D_c and the pointer on the CSB+ tree is advanced to "frank". As can be seen, entries (like "apple", arrow M1) only present in the old main store dictionary, entries only present in the CSB+ tree of the differential buffer ("bodo", arrow D1) and entries present in both structures ("frank", arrow B1) get transferred to the new combined dictionary D_c. While constructing the combined dictionary, the auxiliary structures AUX_m and AUX_d are filled with the resulting new valueIDs which will most probably differ from the old valueIDs. This is necessary for both structures, not just the differential buffer: for the main store, the new valueIDs might be increased in comparison to the old ones due to the addition of new entries from the differential buffer.

The second step builds up the combined attribute vector AV_c using the attribute vector AV_m of the old main store, the leafs of the CSB+ tree (AV_{ds}) which represent the attribute vector of the differential buffer in a sorted order encoded by the CSB+ tree and the just created auxiliary structures AUX_m and AUX_d. The valueIDs of the outdated attribute vectors are sequentially scanned. Each value is updated with the help of the appropriate auxiliary structure and then just added to AV_c. The arrows UM1 and UM2 show an example for the entry "charlie", which was represented by the valueID 1, but is now represented by the valueID 2 due to the addition of the value "bodo" to the combined dictionary. All in all, the resulting combined attribute vector is the concatenation of the existing attribute vectors with updated valueIDs.

27.3 Merge Optimizations

In addition to the described asynchronous online merge, this section presents further optimizations.

27.3.1 Using the Main Store's Dictionary

The first optimization is the usage of the main store's dictionary in the differential buffer. One of the major advantages of writing to a differential buffer is that adding new elements can be done without the eventual penalty of re-sorting the dictionary or re-encoding the attribute vector. A disadvantage is that an additional dictionary is created which has to be incorporated into the main store's dictionary during the merge. But in several cases, using the main store's dictionary in the differential buffer as well can improve the merge performance. This is the case when the main store's dictionary is already saturated. Typical examples are columns storing years, country codes, or postal codes. These columns are saturated early and new elements are rather rare.

In these cases, all elements in the differential buffer use the final dictionary positions of the main store. This can severely reduce the memory consumption of the differential buffer. Consequently, the merge of an attribute that reuses the main dictionary is a simple concatenation of the differential buffer tuples to the main store's attribute vector. In cases where the probability of data modifications introducing new values to the dictionary is high, using the main store's dictionary in the differential buffer is rather expensive. This is the case for attributes such as the time of day, entity identifier, or similar.

27.3.2 Single Column Merge

Another possible optimization concerns memory consumption. During the merge phase, the complete new main store is kept in main memory. At the point of highest memory consumption, more than twice the size of the original main store plus the size of the differential buffer is required to be stored in main memory to execute the proposed merge process. Tables in enterprise applications often consist of millions of tuples while having hundreds of attributes. As a consequence, requiring full table copies can lead to a huge overhead since at least twice the size of the largest table has to be available in memory to allow the merge process to run. For example, the financial accounting table of a large consumer products company contains about 250 million line items with 300 attributes. The uncompressed size with variable length fields of the table is about 250 GB and can be compressed with bit compressed dictionary encoding to 20 GB [KGZP10]. However, to run the merge process, at least 40 GB of main memory are necessary.

To avoid storing a complete table in memory twice, Krueger et al. present the so-called Single Column Merge that merges a table column-wise [KGW+11]. Consequently, not the whole table needs to be kept in memory twice, but only a single column. Thus, if all columns are merged sequentially, the required amount of memory is reduced to the size of the differential buffer and the compressed table, plus the size of the largest resulting column. A drawback of this approach is, that querying as well as transaction management on a partially merged table becomes more complex.

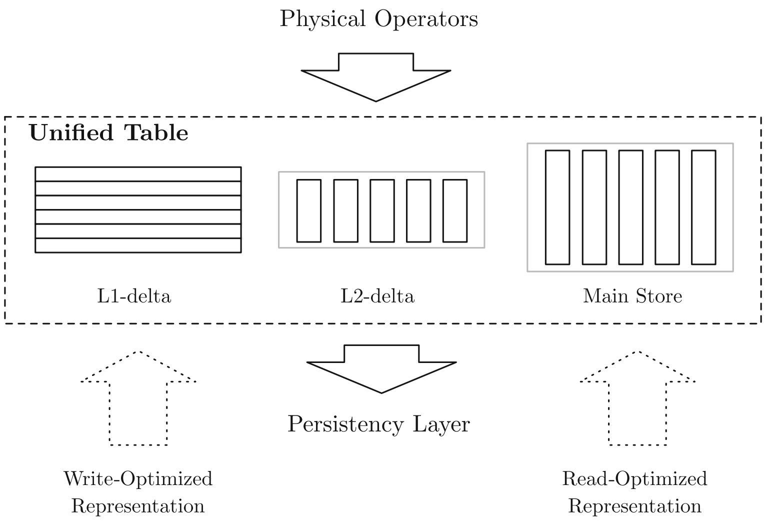


Fig. 27.5 Unified table concept (adapted from [SFL+12])

27.3.3 *Unified Table Concept*

To further improve the transactional capabilities of column stores, [SFL+12] present a modified differential buffer concept, called Unified Table Concept. Hereby, an additional data structure in form of an in-memory row store—called L1-delta—is used (see Fig. 27.5), while the L2-delta (i.e. the differential buffer in SanssouciDB) and the main store have similar structures as in SanssouciDB.

In this concept, each data modification is first written to the L1-delta. This structure stores approximately 10,000 to 100,000 rows. The L1-delta is merged with the L2-delta at regular intervals or when a certain row-limit is reached. The L2-delta is suited to store up to 10 million rows and is merged with the main store. Consequently, this approach introduces bulk-loading improvements to the differential buffer and uses a highly write-optimized data structure (i.e., row store) for incoming data modifications.

27.4 Self Test Questions

1. **What is the Merge?**
   The merge process...

   (a) incorporates the data of the write-optimized differential buffer into the read-optimized main store
   (b) combines the main store and the differential buffer to increase the parallelism

(c) merges the columns of a table into a row-oriented format
(d) optimizes the write-performance.

2. **When to Merge?**
When is the merge process triggered?

(a) When the number of tuples within the differential buffer exceeds a specified threshold
(b) When the space on disk runs low and the main store needs to be further compressed
(c) Before each SELECT operation
(d) After each INSERT operation.

References

[FSKP12] M. Faust, D. Schwalb, J. Krueger, H. Plattner, Fast lookups for in-memory column stores: group-key indices, lookup and maintenance, in *ADMS '12: Proceedings of the 3rd International Workshop on Accelerating Data Management Systems Using Modern Processor and Storage Architectures at VLDB'12*, 2012

[KGW+11] J. Krueger, M. Grund, J. Wust, A. Zeier, H. Plattner, Merging differential updates in in-memory column store, in *DBKDA*, 2011

[KGZP10] J. Krueger, M. Grund, A. Zeier, H. Plattner, Enterprise application-specific data management, in *EDOC*, pp. 131–140, 2010

[SFL+12] V. Sikka, F. Färber, W. Lehner, S.K. Cha, T. Peh, C. Bornhövd, Efficient transaction processing in SAP HANA database: the end of a column store myth, in *SIGMOD Conference*, pp. 731–742, 2012

Chapter 28
Logging

Databases need to provide durability guarantees (as part of ACID[1]) to be used in productive enterprise applications. To provide these guarantees, fault-tolerance and high availability have to be ensured. However, since hardware failures or power outages cannot be avoided or foreseen, measures have to be taken which allow the system to recover from failures.

The standard procedure to enable durable recovery is logging. With the help of logging and recovery protocols, databases can be brought back to the last consistent state before the failure. This is achieved by check pointing the current system and logging subsequent data modifications. Data is written into log files, which are stored on persistent memory such as hard disk drives (HDD) or solid-state drives (SSD).

Please note that these requirements are true for any database, regardless of being an in-memory database or not.

28.1 Logging Infrastructure

A key consideration when talking about logging is performance, both for writing the logs as well as for reading logs back into memory when recovering. As discussed in Sect. 4.6, the performance gap between disk and CPU is steadily increasing. Consequently, logging has to be primarily optimized with respect to minimizing I/O operations.

Figure 28.1 outlines the logging infrastructure of SanssouciDB. The logging data, which is written to disk, consist of three parts:

- Snapshot of the main store
- Value logs
- Dictionary logs.

[1] ACID stands for Atomicity, Consistency, Isolation, Durability. These properties guarantee reliability for database transactions and are considered as the foundation for reliable enterprise computing.

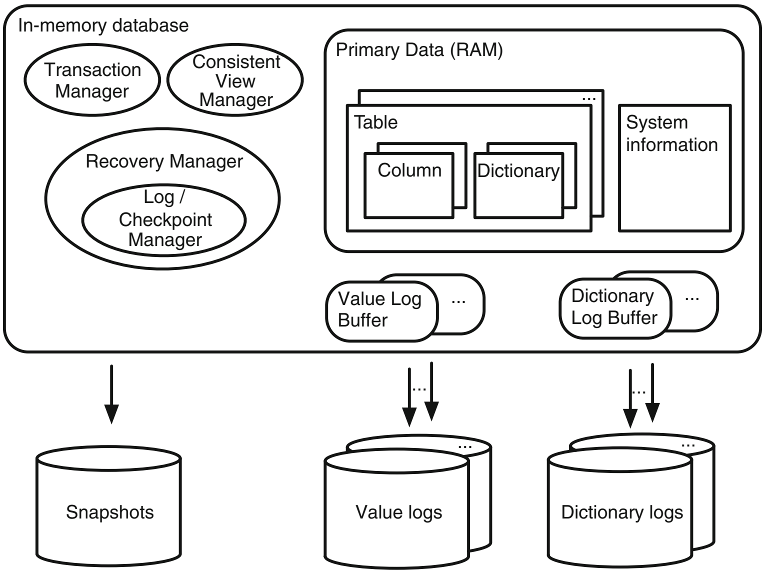


Fig. 28.1 Logging infrastructure

Check pointing [Bor84] is used to create a snapshot of the database at a certain point in time, at which the data is in a consistent state. According to [HR83], a database is in a consistent state "if and only if it contains the results of all committed transactions". The snapshot is a direct copy of the read-optimized main store and it is written to disk periodically. The purpose of check pointing is to speed up the recovery process, since only log entries after the snapshot have to be replayed, while the main store can be loaded from the snapshot directly. To log the data of the differential store, which is not part of the snapshot, value logs as well as dictionary logs are used to track committed changes.

SanssouciDB's logging infrastructure differs from most traditional databases. SanssouciDB adapted the infrastructure to leverage the columnar data structures and to reduce I/O performance penalties. Amongst these optimizations are:

- **Snapshot format**: at each checkpoint, a snapshot of the main store is written to disk using a binary file format. This means that an exact copy of the main store in memory is written to disk, which can later be directly restored without any loading overhead in case of a recovery.
- **Checkpoint timing**: the ideal timing for check pointing is when the differential buffer is relatively small compared to the main store. That is right after the merge.
- **Storing meta data**: to speed up the recovery process, additional meta data is written to disk. With the help of this meta data, the required memory can be allocated before loading. Thus, expensive re-allocations and data movements can be avoided. Data hereby written is, e.g., the number of tuples in the main store and the number of bits used in each dictionary.

- **Separation into value and dictionary logs**: the two major performance optimizations for logging in SanssouciDB are the reduction of the log size and the parallelization of logging. This is achieved by using dictionary-encoded logging, which is discussed in detail in the next section.

28.2 Logical Versus Dictionary-Encoded Logging

The obvious way to log data modifications is logical logging. As depicted in Fig. 28.2, logical logging simply writes the SQL statement and its parameters (recordID and attribute values) to disk.

Logical logging has two major shortcomings. First, logging and recovery cannot be parallelized since the order of the log has to be preserved during replay to recover the dictionary and the corresponding attribute vector elements. Second, logical logging writes values directly to disk and, therefore, does not leverage compression as used in SanssouciDB. Consequently, logical logging writes comparatively large data volumes to disk.

To avoid these shortcomings, SanssouciDB uses a logging schema, which separates the dictionary-encoded data (and their corresponding dictionary inserts) from the transactional context, the so-called dictionary-encoded logging [WBR+12]. This approach allows recovering the attribute vectors and dictionaries in parallel and permits to replay the log entries in an arbitrary order. Furthermore, dictionary-encoded logging reduces the log size due to the usage of dictionary-compression, which speeds up the recovery process significantly.

In which cases dictionary-encoded logging is advantageous over logical logging depends on the data characteristics. In enterprise applications, the same data characteristics that favor dictionary-compressed column-stores also apply to dictionary-encoded logging. Amongst these characteristics are, e.g., the low number of distinct values, which leads to fewer dictionary log entries, and the distribution of values.

Value distribution can be described using the Zipf distribution. Intuitively, the Zipf distribution describes—depending on the variable *alpha*—how heavily the distribution is drawn to one value. In the case of $alpha = 0$, the distribution equals a uniform distribution and every value occurs equally often. As alpha increases, less values are more frequently occurring (see exemplary distributions for varying alpha values in Fig. 28.3).

The authors in [HBK+11] state that the majority of columns analyzed from financial, sales and distribution modules of an enterprise resource planning (ERP)

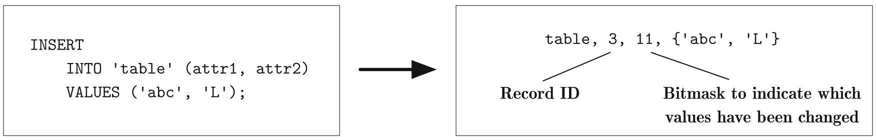

INSERT
INTO 'table' (attr1, attr2)
VALUES ('abc', 'L');
table, 3, 11, {'abc', 'L'}
Record ID
Bitmask to indicate which values have been changed

Fig. 28.2 Logical logging

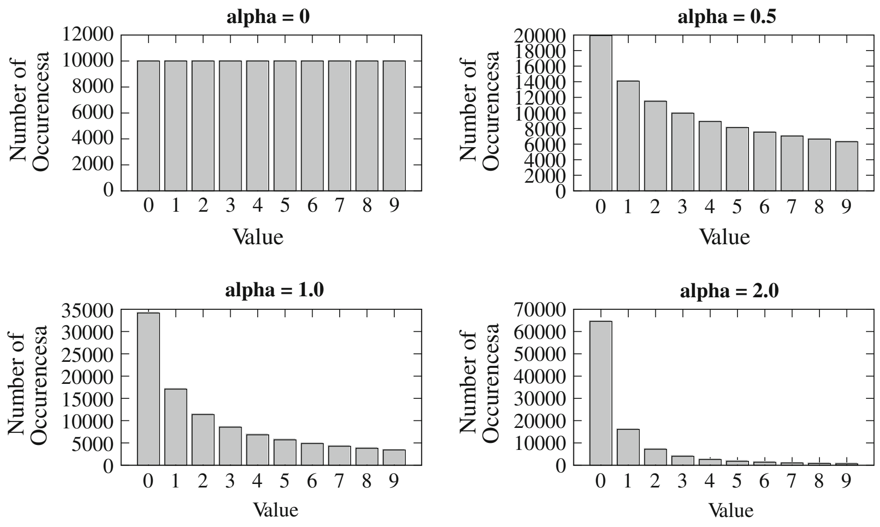


Fig. 28.3 Exemplary Zipf distributions for varying alpha values

system were following a power-law distribution—a small set of values that occur frequently, while the majority of values is rare. Furthermore, [HBK+11] identified an average alpha value of 1.581 in enterprise systems.

Figure 28.4 shows the results of an experiment that measures the cumulated log size per query for a varying value distribution. In this experiment, one million INSERT queries of single zipf distributed values with a total of 1,000 distinct values have been simulated. With an alpha value of 1.581, the dictionary is already saturated after ≈30,000 queries. Afterwards, queries rarely add entries to the dictionary log.

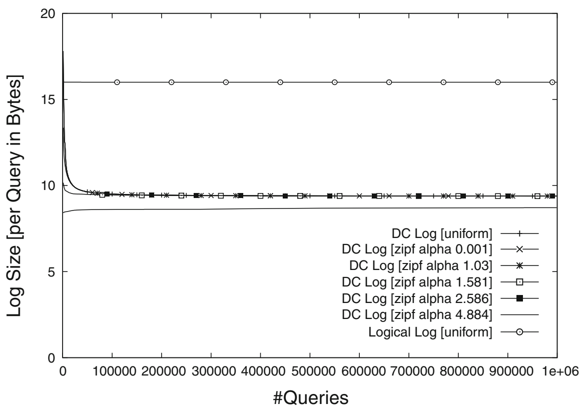


Fig. 28.4 Cumulated average log size per query for varying value distributions (*DC* = *Dictionary-Compressed*)

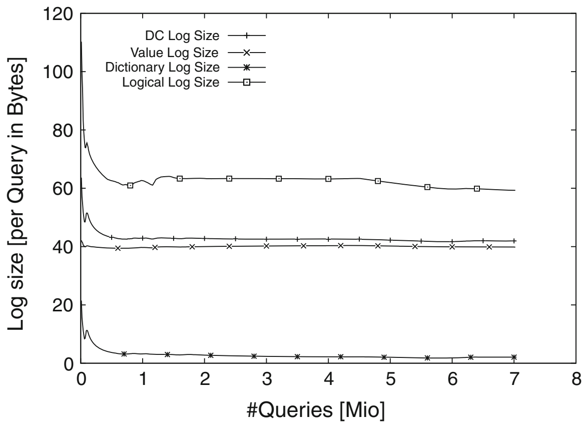


Fig. 28.5 Log size comparison of logical logging and dictionary-encoded logging

As shown for an alpha value of 4.884 (see Fig. 28.4), the more heavily the distribution is drawn to one value, the smaller the accumulated log size will be. Please note that logging is only needed for queries that modify the data set to save the changes.

A comparison of the log sizes for logical and dictionary-encoded logging is shown in Fig. 28.5. These values have been measured on a productive enterprise system with seven million write operations on the sales item table. Due to the high compression of recurring values, the dictionary-encoded logging reduces the log size by 29 %. As a consequence, dictionary-encoded logging is in favor of logical logging, since it exploits typical data distributions in enterprise systems. For more details about enterprise data characteristics see Chap. 3.

28.3 Example

Figure 28.6 shows an example of dictionary-encoded logging. Here, three SQL queries (`insert`, `update`, and `delete`) of three different transactions are logged.

The first query (`INSERT INTO T1 (Attr1, Attr2) VALUES ('abc', 'L');`) inserts a new row into table `T1`. This query has the transaction ID 9 (TID = 9), which is stored in the following format:

$$L_t = \{``t", TID\}$$

Since both values ('abc' and 'L') are not yet stored in the corresponding dictionaries, new entries will be added. Dictionary logs `Ld` are created each time a transaction adds new values into a dictionary. Therefore, the table identifier `t`, the column index c_i, the added value `v`, and the corresponding valueID `VID` are logged. The letter "d" at the first position of the dictionary log entry marks that this is a dictionary log entry, similar to the letter "t" marking the transaction log entry above.

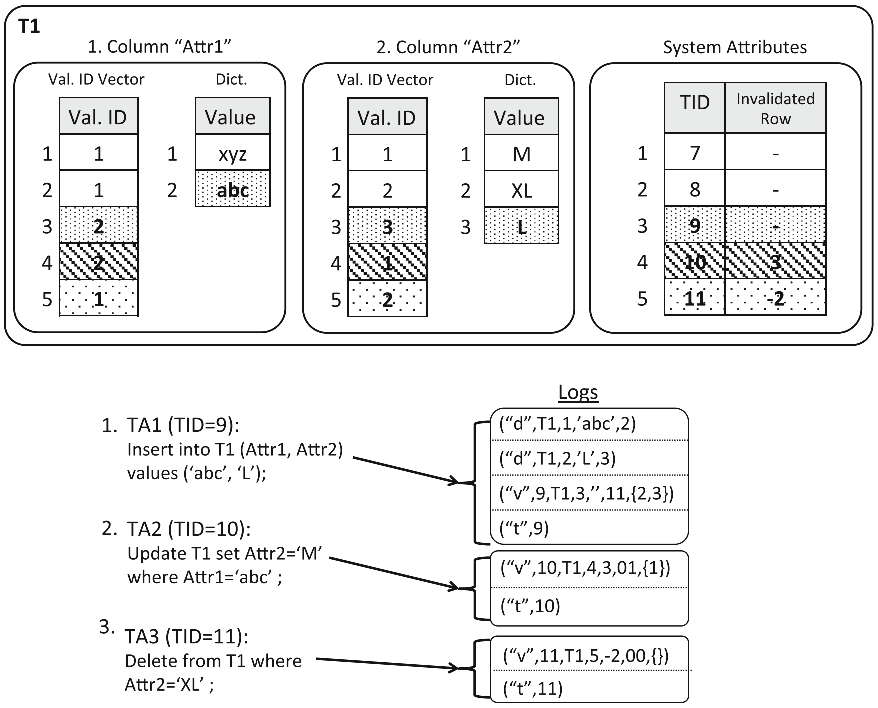


Fig. 28.6 Example: logging for dictionary-encoded columns

$$L_d = \{``d'', t, c_i, v, VID\}$$

Consequently, for transaction 9, two dictionary logs are stored holding the IDs of the modified dictionaries for table T1 (i.e. column ID), the corresponding positions in the dictionaries, and the newly inserted values:

$$(``d'', T1, 1, `abc', 2) \; and \; (``d'', T1, 2, `L', 3)$$

Value logs `Lv` store the actual values, which are appended to the attribute vectors. Value log entries store more than just plain data structure changes as done in the dictionary logs, since they have to be linked to the corresponding transactions.

$$L_v = \{``v'', TID, t, RID, IRID, bm_n, (VID_1, \ldots, VID_n)\}$$

Each value log `Lv` hereby stores the transaction identifier `TID`, the table identifier `t`, and the rowID `RID` in the attribute vector. The letter "v" at the first position marks this log entry as a value log entry. The values are stored in a vector of `VIDs`, whereby the bit mask bm_n stores the corresponding columns (`n` is the number of attributes in table `t`). If a row is invalidated by the new row (e.g. due to an update or a delete), the ID of the invalidated row is stored in `IRID`.

The second query (`UPDATE T1 SET Attr2 = 'M' WHERE Attr1 = 'abc';`) in Fig. 28.6 alters a row, without introducing new values. Thus only one transactional log entry and one value log entry, storing the new dictionary position for attribute "Attr1", are stored.

The delete query (`DELETE FROM T1 WHERE Attr2 = 'XL';`) invalidates a row. But in this case, the result might not be obvious. When a row is deleted, a new line is added to the table. In this example, transaction 11 marks row 2 as invalid. This is achieved via hidden system attributes (i.e. columns storing the TID and the corresponding IDs of invalidated rows). While the TID field of a certain row always persists the transaction ID that inserted this row, the invalidated row field is only written for updates and deletes. To mark, that a line has not just been updated but entirely deleted, the invalidated row field is prefixed (see entry "-2" in Fig. 28.6). Both, for updates and deletes, the unchanged fields of the inserted tuple are copied from the invalidated row instead of being left empty. The reason is twofold: first, with fixed length attribute vectors, empty fields provide no advantages in terms of performance or memory consumption. Second, copying the row avoids additional lookups to get the values of the invalidated row. This is especially advantageous for long-running transactions and queries, which potentially need to include outdated rows.

It is furthermore important to understand when and in which order the stored logs are written to disk. Once a transaction is committed, first the dictionary buffers have to be written to disk. This has to be ensured to avoid value logs referencing valueIDs that cannot be recovered. Afterwards, the value logs are written to disk. Finally, if both logs were written to disk successfully, the committed transaction log is written to disk. Both, the value log entries and the transaction log entries are collected in the same log buffer (the value log buffer in Fig. 28.1).

28.4 Self Test Questions

1. **Snapshot Statements**
   Which statement about snapshots is wrong?
   (a) The recovery process is faster when using a snapshot because only log files after the snapshot need to be replayed
   (b) The snapshot contains the current read-optimized store
   (c) A snapshot is an exact image of a consistent state of the database to a given time
   (d) A snapshot is ideally taken after each insert statement.

2. **Recovery Characteristics**
   Which of the following choices is a desirable characteristic of any recovery mechanism?
   (a) Recovery of only the latest data

   (b) Returning the results in the right sorting order
   (c) Maximal utilization of system resources
   (d) Fast recovery without any data loss.

3. **Situations for Dictionary-Encoded Logging**
   When is dictionary-encoded logging superior?

   (a) If large values are inserted only one time
   (b) If the number of distinct values is high
   (c) If all values are different
   (d) If large values are inserted multiple times.

4. **Small Log Size**
   Which logging method results in the smallest log size?

   (a) Common logging
   (b) Log sizes never differ
   (c) Dictionary-encoded logging
   (d) Logical logging.

5. **Dictionary-Encoded Log Size**
   Why has dictionary-encoded logging the smaller log size in comparison to logical logging?

   (a) Because of interpolation
   (b) Because it stores only the differences of predicted values and real values
   (c) Because of the reduction of recurring values
   (d) Actual log sizes are equal, the smaller size is only a conversion error when calculating the log sizes.

References

[Bor84] A.J. Borr, Robustness to crash in a distributed database: a non shared-memory multi-processor approach, in *VLDB*, eds. by U. Dayal, G. Schlageter, L.H. Seng (Morgan Kaufmann, 1984), pp. 445–453

[HBK+11] F. Hübner, J.-H. Böse, J. Krüger, C. Tosun, A. Zeier, H. Plattner, A cost-aware strategy for merging differential stores in column-oriented in-memory DBMS, in *BIRTE* (2011), pp. 38–52

[HR83] T. Härder, A. Reuter, Principles of transaction-oriented database recovery. ACM Comput. Surv. **15**(4), 287–317 (1983)

[WBR+12] J. Wust, J.-H. Boese, F. Renkes, S. Blessing, J. Krueger, H. Plattner, Efficient logging for enterprise workloads on column-oriented in-memory databases, in *CIKM* (ACM, 2012)

Chapter 29
Recovery

To handle steadily growing volumes of data and intensifying workloads, modern enterprise systems have to scale out, using multiple servers within the enterprise system landscape. With the growing number of servers—and consequently growing number of racks and CPUs—the probability of hardware-induced failures is rising.

Productive enterprise systems are expected to never fail or to securely fail-over once a defect is detected. When a server fails, it may have to be rebooted and restored, or another server has to take over the workload of the failed server. In either way, to restore the previous state of the server before its failure, data stored on persistent memory has to be loaded back into the in-memory database. This process is called “recovery”. Using snapshots and log data—as presented in the previous Chap. 28—a database can be rebuild to the latest consistent state.

The recovery process, which is presented in this section, relies on dictionary-encoded logging [WBR+12]. It is executed in two subsequent tasks (I) read meta data and prepare data structures, (II) read logging data and recover database.

29.1 Reading Meta Data

In addition to logging committed transactions, SanssouciDB logs meta data to speed up the recovery process. With additional knowledge about the data structures, which have to be recovered, expensive data movements and re-allocations can be avoided. Examples for stored meta data are, e.g., the location of the latest snapshot, the number of rows in the main store, or the bits required for the dictionary encoding of columns.

As an example, take the replaying of the dictionary logs. Without knowing in advance how many elements have been persisted before the system failure, the allocated space for the dictionary would probably have to be resized several times. A resize usually implies moving the whole data set to a new allocation. If the number of elements and the number of required bits is known in advance, the allocated space can be sized accordingly without the need of re-allocations or data movements. The number of dictionary elements is stored, since scanning the dictionary log for the latest log to receive the dictionary size is not efficient even when scanning in reverse order. The reason is that transactions do not have to be logged in their incoming order. Thus, finding the maximum dictionary position in the dictionary log might cause reading large parts of the dictionary log file.

29.2 Recovering the Database

After the data structure allocation has taken place, the recovery process continues to replay the database logs. As part of this process, the snapshot of a table's main store is reloaded into memory. At the same time the dictionary log files (containing the dictionary log entries) and the value log files (containing the value and transaction log entries) are replayed.

Due to the dictionary-encoded logging described in Sect. 28.2, the files can be processed in parallel. The import of the dictionary logs and the main store is rather straightforward, while reading the value and transaction log entries from the value log file is a bit more complex. To avoid replaying not committed transactions, the value log file is read in reverse order. This way it is ensured that only value log entries are replayed, whose transactions have been successfully committed. Remember, value and transaction log entries are written to the same file with the strict order of writing the transaction log entry after all value and dictionary log entries have been successfully written.

After the import of the value log file, a second run over the imported tuples is performed. This is caused by the dictionary-encoded logging, which only logs changed attributes of tuples, thus reducing I/O operations. Consequently, the imported tuples have to be checked for empty attributes and they have to be completed if necessary. This is done by iterating over all versions of the tuple, using the validation flag.

29.3 Self Test Questions

1. **Recovery**
   What is recovery?
   (a) It is the process of recording all data during the run time of a system
   (b) It is the process of restoring a server to the last consistent state before its crash
   (c) It is the process of improving the physical layout of database tables to speed up queries
   (d) It is the process of cleaning up main memory, that is "recovering" space.

2. **Server Failure**
   What happens in the situation of a server failure?
   (a) The system has to be rebooted and restored if possible, while another server takes over the workload
   (b) The power supply is switched to backup power supply so the data within the main memory of the server is not lost
   (c) The failure of a server has no impact whatsoever on the workload
   (d) All data is saved to persistent storage in the last moment before the server shuts down.

Reference

[WBR+12] J. Wust, J.-H. Boese, F. Renkes, S. Blessing, J. Krueger, H. Plattner. Efficient logging for enterprise workloads on column-oriented in-memory databases, in *CIKM 2012* (ACM, 2012)

Chapter 30
On-the-Fly Database Reorganization

In typical enterprise applications, schema and data layout have to be changed from time to time. The main cases for such changes are software upgrades, software customization, or workload changes. Therefore, the option of database reorganization, such as adding an attribute to a table or changing attribute properties, is required.

In row-oriented databases, database reorganization is typically time-consuming as well as cost-intensive. That is why most row-oriented database management systems usually do not allow data definition operations while the database is online [AGJ+08]. Consequently, downtime of the database server has to be taken into account. In contrast, modifications within a column store database, such as SanssouciDB, can be done dynamically without any downtime. The following sections explain what database reorganization looks like in row stores and column stores.

30.1 Reorganization in a Row Store

In row stores, database reorganization is expensive. As mentioned in Sect. 8.2 all attributes of a tuple are stored sequentially in the same memory block, where each block contains multiple rows. The left side of Fig. 30.1 shows a table that includes a unique identifier, the first name, and last name of citizens.

If an additional attribute, for example, state is added and no space is available in the block, adding a new attribute requires a reorganization of the storage for the entire table. The same issue occurs when the size of an attribute is increased. The right side of Fig. 30.1 shows the table's storage after the attribute *state* is added. Each row is extended by that attribute and all following rows are moved within the block (and the following blocks if necessary).

To be able to dynamically change the data layout, a common approach for row stores is to create a logical schema on top of the physical data layout [AGJ+08].

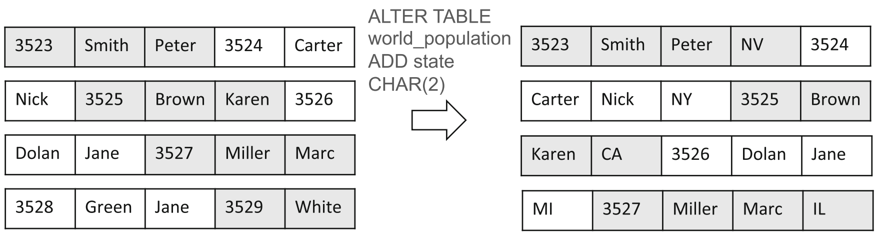


Fig. 30.1 Example memory layout for a row store

This method allows changing the logical schema without modifying the physical representation of the database but also decreases the performance because of the overhead of accessing the meta data and data of the logical tables. Another approach is using schema versioning for database systems [Rod95]. These advanced approaches will not be discussed any further as part of this learning material.

30.2 On-the-Fly Reorganization in a Column Store

In column-oriented databases, each column is stored independently from the other columns in a separate block, see the example in Fig. 30.2. New attributes can be added very easily, because they will be created in a new memory area. Locking for changing the data layout is only required for a very short period, during which solely the meta data of the table is adapted.

As mentioned in Sect. 16.4, in SanssouciDB new columns will not be materialized until the first value is added. The dictionary and the attribute vector of new columns remain non-existent as long as the column does not contain any values. The addition of a column has no impact whatsoever on existing applications if they solely request their required attributes from the database (meaning they do not use *SELECT* * statements).

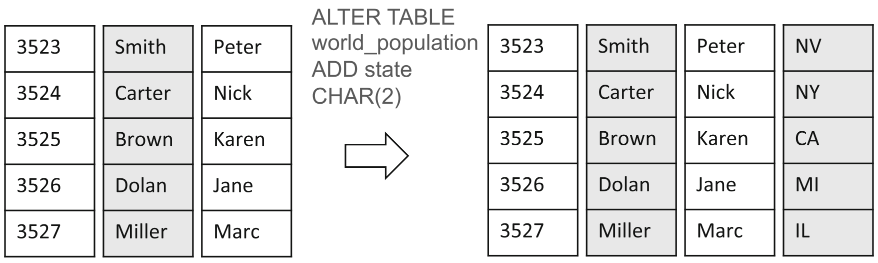


Fig. 30.2 Example memory layout for a column store

30.3 Excursion: Multi-Tenancy Requires Online Reorganization

This sections gives a typical use case where online reorganization is required in a database system.

In a *single-tenant* system, each customer (tenant) has its own database instance on a physically separated server machine. In this case, maintenance costs for the service provider will be very high and in addition, tenants do not even use their system permanently with the complete utilization level. In contrast, on *multi-tenant* systems different customers share the same resources on the same machine. By providing a single administration framework for the whole system, multi-tenancy can improve the efficiency of system management [JA07] and increases the utilization of systems. The Software-as-a-Service provider Salesforce.com[1] first employed this technique on a large scale.

Multi-tenancy can be implemented in three different ways with different levels of granularity: shared machine, shared database instance, and shared table.

In the *shared machine* implementation (see Fig. 30.3a) each customer has its own database process and these processes are executed on the same server. The advantages of this approach are a good isolation among other tenants and easy customer migrations from one machine to another. Major limitations are that this approach does not support memory pooling and each database needs its own connection pool. Moreover, administrative operations cannot be applied on all database instances simultaneously (in bulk).

In the *shared database instance* implementation (see Fig. 30.3b) each customer has its own tables, but shares the database instance with other customers. In this case, connection pools can be shared between customers and pooling of memory is better compared with the previous approach. On the other hand, isolation between customers is reduced. This approach allows simultaneous execution of many administrative operations on all database instances.

In the *shared table* approach (see Fig. 30.3c) many tenants share the common database and each customer has its own rows, which are marked by an additional attribute, e.g., *tenantID*. With this approach, resource pooling performs best and sharing of connection pools between customers is possible. Administrative operations can be carried out in bulk by running queries over the column containing the *tenantID*.

Multi-tenant systems using the shared table approach are a typical environment where on-the-fly database reorganization is necessary. These systems aim at maintaining the ability of individual tenants to make custom changes to their database tables, while not affecting other tenants using the same resources. In row stores, the entire database or table would be completely locked to process data definition operations. In the column store, the table is locked only for the amount

[1] http://www.salesforce.com

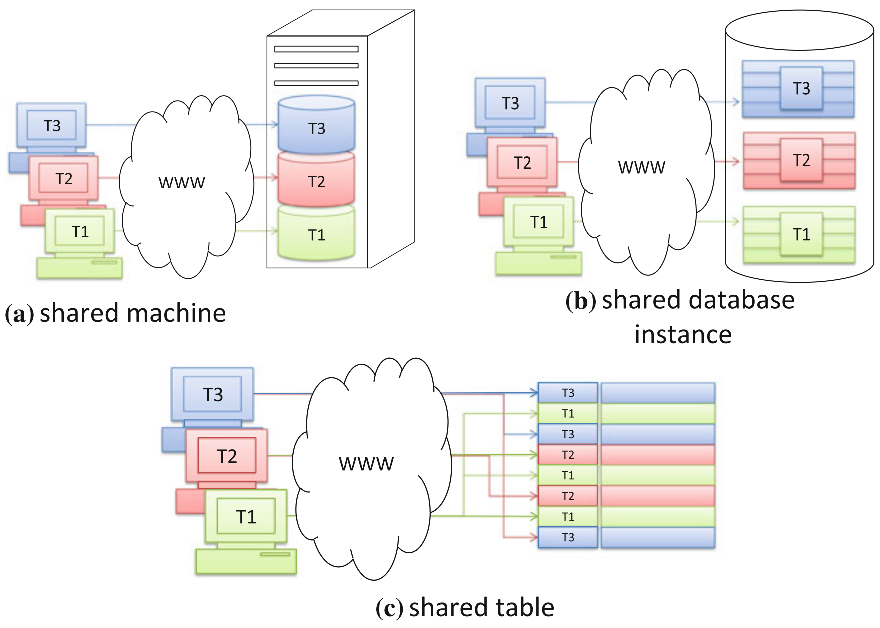


Fig. 30.3 Multi-tenancy granularity levels

of time that is needed to complete the data definition operation and the lock is only restricted to the meta data of the table.

30.4 Hot and Cold Data

In addition to ever increasing growth of data, intern (i.e. controlling) and extern (i.e. tax law) requirements demand that data should be kept at hand for many years. The separation of data into *cold data* and *hot data* is an approach to handle the increased, but still limited capacity of main memory efficiently, while keeping all data of a company available for reporting needs.

Business objects in enterprise applications have their own life cycle. In Fig. 30.4, an example life cycle for a sales opportunity is shown. The life cycle of a business object can be separated into active (hot) and passive (cold) states. The sequence of events a business object has passed through determines its state. A business object becomes passive and can be moved to cold data if it will not be changed any longer and, thus, the object will be accessed less often. Passive objects will still be used for reporting. In our example of the sales opportunity, it becomes passive if the opportunity is won and, thus, it is turned into a sales order by the sales representative, or if it is lost, e.g., canceled by the customer.

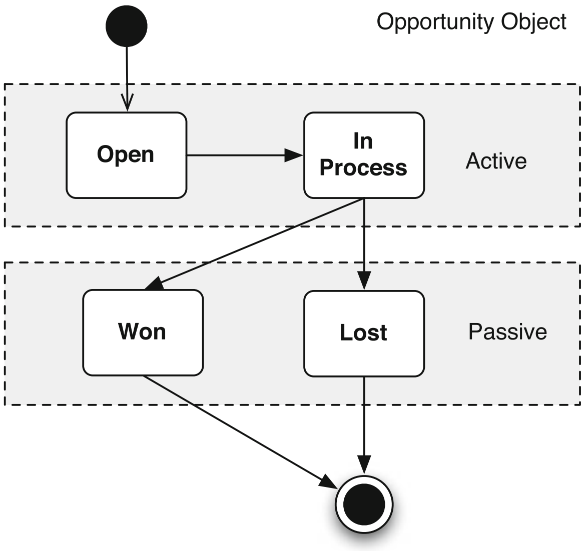


Fig. 30.4 The life cycle of a sales order

Probably more than 90 % off all queries are going against hot data. Hot and cold data can be treated differently as access patterns differ, e.g., read-only access in cold data versus read and write access in hot data. Different data partitioning (due to differing access patterns), other storage media (DRAM for hot data, SSD for cold data), and different materialization strategies can further be used dependent on the data state.

30.5 Self Test Questions

1. **Separation of Hot and Cold Data**
   How should the data separation into hot and cold take place?
   (a) Randomly, to ensure efficient utilization of storage areas
   (b) Round-robin, to ensure uniform distribution of data among hot and cold stores
   (c) Manually, upon the end of the life cycle of an object
   (d) Automatically, depending on the state of the object in its life cycle.
2. **Data Reorganization in Row Stores**
   The addition of a new attribute within a table that is stored in row-oriented format ...
   (a) is not possible
   (b) is an expensive operation as the complete table has to be reconstructed to make place for the additional attribute in each row

   (c) is possible on-the-fly, without any restrictions of queries running concurrently that use the table
   (d) is very cheap, as only meta data has to be adapted.

3. **Cold Data**
   What is cold data?

   (a) Data, which is not modified any longer and that is accessed less frequently
   (b) The rest of the data within the database, which does not belong to the result of the current query
   (c) Data that is used in a majority of queries
   (d) Data, which is still accessed frequently and on which updates are still expected.

4. **Data Reorganization**
   The addition of an attribute in the column store ...

   (a) slows down the response time of applications that only request the attributes they need from the database
   (b) speeds up the response time of applications that always request all possible attributes from the database
   (c) has no impact on existing applications if they only request the attributes they need from the database
   (d) has no impact on applications that always request all possible attributes from the table.

5. **Single-Tenancy**
   In a single-tenant system ...

   (a) all customers are placed on one single shared server and they also share one single database instance
   (b) each tenant has its own database instance on a shared server
   (c) power consumption per customer is best and therefore it should be favored
   (d) each tenant has its own database instance on a physically separated server.

6. **Shared Machine**
   In the shared machine implementation of multi-tenancy ...

   (a) each tenant has an own exclusive machine, but these share their resources (CPU, RAM) and their data via a network
   (b) all tenants share one server machine, but have own database processes
   (c) each tenant has an own exclusive machine, but these share their resources (CPU, RAM) but not their data via a network
   (d) all tenants share the same physical machine, but the CPU cores are exclusively assigned to the tenants.

7. **Shared Database Instance**
   In the shared database instance implementation of multi-tenancy ...
   (a) the risk of failures is minimized because more technical staff (from different tenants) will have a look at the shared database
   (b) all tenants share one server machine and one main database process, tables are also shared
   (c) each tenant has its own server, but the database instance is shared between the tenants via an InfiniBand network
   (d) all tenants share one server machine and one main database process, tables are tenant exclusive, access control is managed within the database.

References

[AGJ+08] S. Aulbach, T. Grust, D. Jacobs, A. Kemper, J. Rittinger, Multi-tenant databases for software as a service: schema-mapping techniques, in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08* (ACM, New York, NY, USA, 2008), pp. 1195–1206

[JA07] D. Jacobs, S. Aulbach, Ruminations on multi-tenant databases, in *BTW, LNI*, ed. by A. Kemper, H. Schöning, T. Rose, M. Jarke, T. Seidl, C. Quix, C. Brochhaus, vol. 103 (GI, 2007), pp. 514–521

[Rod95] J.F. Roddick, A survey of schema versioning issues for database systems. Inf. Softw. Technol. **37**, 383–393 (1995)

Part V
Foundations for a New Enterprise Application Development Era

Chapter 31
Implications on Application Development

In the previous chapters, we introduced the ideas behind our new database architecture and their technical details. In addition, we showed that the in-memory approach can significantly improve the performance of existing database applications.

In this chapter, we discuss how the existing applications should be redesigned and how new applications should be designed to take full advantage of the new database technology. Our research and the prototypes we built show that in-memory technology greatly influences the design and development of enterprise applications. The main driver for these changes is the drastically reduced response time for database queries. Now, even more complex analytical queries can be executed directly on the transactional data in less than one second. With this performance, we are able to develop new applications and enhance currently existing applications in a way that was not possible before. Modern applications can especially benefit from the database performance when it comes to better granularity and actuality of the processed data.

The most important approach to achieve this performance is to move application logic closer to the database. While traditional approaches try to encapsulate complex logic in the application server, with the advent of in-memory computing it becomes crucial to move data intensive logic as close as possible to the database. An additional advantage of moving data-intensive application logic closer to the database is that the amount of data that has to be transferred between the application server and the database system is significantly reduced when most of the data intensive operations are executed directly in the database system.

31.1 Optimizing Application Development for In-Memory Databases

A typical enterprise application contains three main architectural layers (see Fig. 31.1).

DOI: 10.1007/978-3-642-36524-9_31, © Springer-Verlag Berlin Heidelberg 2013

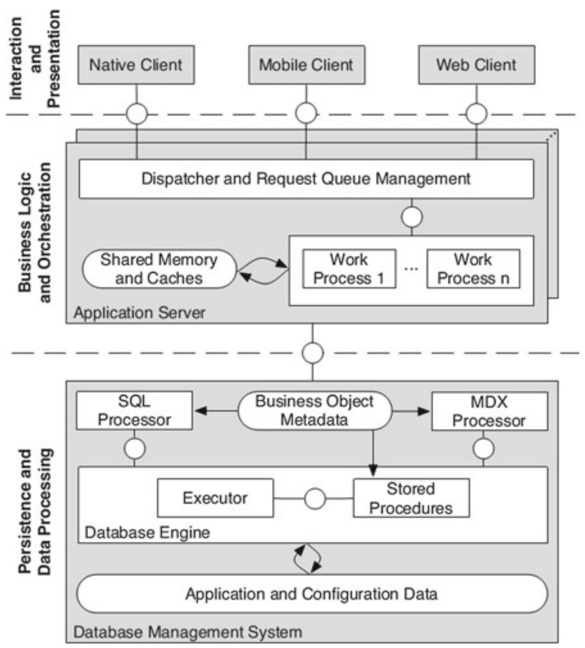


Fig. 31.1 Three tier enterprise application

These three main layers are usually distributed over three independent physical systems, which leads to a three tier setup. To ensure a common understanding of the terms layer and tier, the words are shortly explained: A *layer* separates program code and its responsibility on the logical level, but it does not state how the deployment of the code looks like. The word *tier* describes the physical architecture of a system, so it gives details about the hardware setup used to run the program code.

The interaction and presentation layer is responsible for providing a user interface. This includes the creation of views, which summarize required information to be shown in a clear and understandable fashion. Moreover, the presentation layer gets user information requests and forwards them to the other layers. The complete user interface may consist of many different independent presentation layers for different devices or platforms.

The business logic and orchestration layer acts as a mediator between the presentation and the persistence layer. It handles user requests obtained from the presentation layer. This can either be the direct execution of data operations (by using the application's cache) or delegation of calls to the persistence layer.

Data persistence and processing provides interfaces for requesting data with the help of declarative query languages (such as SQL or Multidimensional Expression (MDX)) and prepares data for further processing in the upper layers.

31.1.1 Moving Business Logic into the Database

As mentioned before, in traditional applications, the application logic is mainly stored in the orchestration layer to allow easier scaling of the complete application. To leverage the full performance, we have to identify which application logic should be moved closer to the persistence layer. The ultimate goal is to leave only such logic in the orchestration layer, that provides functionality that is orthogonal to what can be handled within the user interaction request. This reduced layer would than mostly translate user requests into SQL and MDX queries, or calls to stored procedures on the database system.

To illustrate the impact, we will explain the changes and the effects using an example that performs an analytical operation directly on the transactional data. In the following, two different implementations of the same user request will be compared. The request identifies all due invoices per customer and aggregates their amount (which is usually referred to as dunning). Dunning is one of the most important applications for consumer companies. It is typically a very time-consuming task, because it involves read operations on large amounts of transactional data.

Listing 31.1 implements business logic directly in the application layer. It depends on given object structures and encodes the algorithms in terms of the used programming language.

Using this approach, all customer data is required to be loaded from the database and an object instance for each customer will be created. To create the object, all attributes will be loaded, although only one attribute is needed. After that, for all invoices of each customer, it will be determined whether it is considered paid or not. For that, it is checked whether the due date at which the invoice should be paid, has already passed. Finally, the total unpaid amount for each customer is aggregated.

For each iteration of the inner loop, a query is executed in the database to retrieve all attributes of the customer invoice.

```
for customer in allCustomers() do
  for invoice in customer.unpaidInvoices() do
    if invoice.dueDate < Date.today()
      dueInvoiceVolume[customer.id] +=
       invoice.totalAmount
    end
  end
end
```

Listing 31.1: Imperative implementation of a simple report

The second approach, presented in Listing 31.2, uses a single SQL query to retrieve the same result set. All calculations, filtering and aggregations are handled close to the data. Therefore, the efficient operators implementation introduced in previous chapters can be used. The other advantage is that only the required result set is returned to the application layer. Consequently, the network traffic is reduced.

```
SELECT invoices.customerId,
   SUM(invoices.totalAmount) AS dueInvoiceVolume
FROM invoices
WHERE invoices.isPaid IS FALSE AND
   invoices.dueDate < CURDATE()
GROUP BY invoices.customerId
```

Listing 31.2: Declarative implementation of a simple report

When using small amounts of data, the performance differences are barely noticeable. However, once the system is in a production and is filled with realistic amounts of data, using the imperative approach results in much slower response times. Accordingly, it is very important to test performance of different algorithms with realistic customer data sets that represent realistic sizing settings and value distributions.

The ability to express application logic using SQL can be a huge advantage because expensive calculations are done inside the database. That way, calculations as well as comparisons can work directly on the compressed data. Only as the last step, when returning the results, the compressed values are converted to the original values to present them in human readable format.

31.1.2 Stored Procedures

An additional possibility to move application logic into the database are *stored procedures*, which allow to reuse data-intensive application logic. The main benefits of using stored procedures are:

- Business logic centralization and reuse
- Reduction of application code and simplifying of change management
- Reduction of network traffic
- Pre-compilation of queries increases the performance for repeated execution.

Stored procedures are typically written in a special mixed imperative-declarative programming language (see Listing 31.3). Such programming languages support both declarative database queries (such SQL) and imperative controlling sequences (loops, conditions) and concepts (e.g. variables, parameters). Once a stored procedure is defined, it can be used (and reused) by several applications. Applicability across different applications is usually established via individual invocation parameters (our tiny example does not contain such parameters, but we could alter it so that we pass a country to the procedure which is used as a selection criterion and only customers of this country would be part of the dunning run).

```
// Definition
CREATE PROCEDURE dueInvoiceVolumePerCustomer()
BEGIN
  SELECT invoices.customerId,
    SUM(invoices.totalAmount) AS dueInvoiceVolume
  FROM invoices
  WHERE invoices.isPaid IS FALSE AND
    invoices.dueDate < CURDATE()
  GROUP BY invoices.customerId;
END;

// Invocation
CALL dueInvoiceVolumePerCustomer();
```

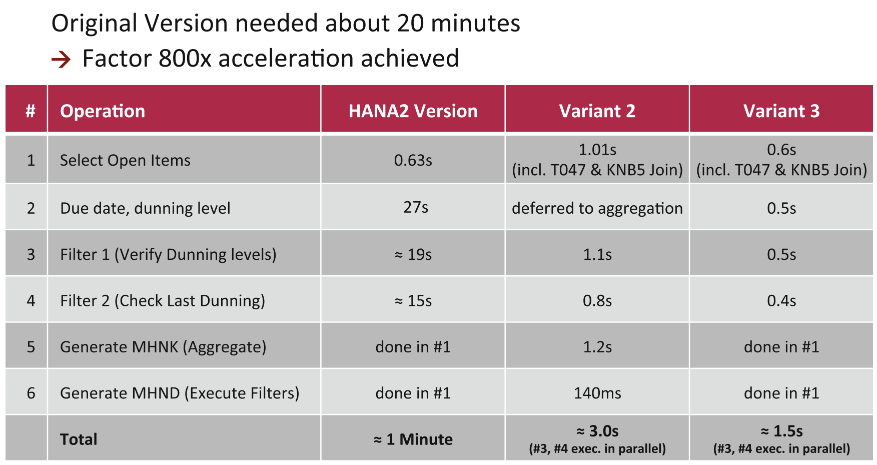
Listing 31.3: Creation of a stored procedure

31.1.3 Example Application

One prominent example, where we were able to achieve an astonishing performance increase over a traditional implementation was in the area of financial applications. Here, we analyzed the dunning run, meaning extraction of all overdue accounting entries from the accounting tables. The traditional picture of the dunning run showed that the original application was implemented as follows: First select all accounts to be dunned and transfer this list to the application server. Now, for each account, all open account items where selected and the due date for each calculated. Now for all items that will be dunned additional configuration logic was loaded and the materialized result set is written to a dedicated dunning table. From the discussion in the previous sections we see that this implementation is clearly disadvantageous since it executes a lot of individual SQL statements and transfers intermediate results from the database system to the application server and back. In addition, the implementation looks like a manual join implementation connecting accounts with account items.

In several iterations on the dunning implementation, we were able to reduce the overall runtime of the dunning implementation from initially 1200 to 1.5 s. Figure 31.2 shows the summary comparison of these implementations. The main difference between the versions is that the fastest implementation tries to push as much selection already down to the first filter predicates and executes as much as possible in parallel. Thus, we were able to achieve a speedup of factor 800.

To summarize, in our new implementation of the dunning run, we followed the principles that were presented earlier in this section. The most important of these principles is to move data-intensive application logic as close as possible to the database.



Original Version needed about 20 minutes
→ Factor 800x acceleration achieved

#	Operation	HANA2 Version	Variant 2	Variant 3
1	Select Open Items	0.63s	1.01s (incl. T047 & KNB5 Join)	0.6s (incl. T047 & KNB5 Join)
2	Due date, dunning level	27s	deferred to aggregation	0.5s
3	Filter 1 (Verify Dunning levels)	≈ 19s	1.1s	0.5s
4	Filter 2 (Check Last Dunning)	≈ 15s	0.8s	0.4s
5	Generate MHNK (Aggregate)	done in #1	1.2s	done in #1
6	Generate MHND (Execute Filters)	done in #1	140ms	done in #1
	Total	**≈ 1 Minute**	**≈ 3.0s (#3, #4 exec. in parallel)**	**≈ 1.5s (#3, #4 exec. in parallel)**

Fig. 31.2 Comparison of different dunning implementations

31.2 Best Practices

In the following section, the discussion of the chapter will be summarized by outlining the most important rules, which should be fulfilled by working with enterprise applications.

- *The right place for data processing*: This is an important decision, which developers have to make during implementation. The more data is processed during a single operation, the closer it should be executed to the database. Aggregations should be executed in the database while single record operations should be part of the application layer.
- *Avoid SELECT**: Only really required attributes for the application should be loaded. Developers often tend to load more data than is actually needed, because this apparently allows easier adoption to unforeseen use cases. The downside is, that this leads to intensive data transfer between the application and database servers which causes significant performance penalties. Furthermore, tuple reconstruction in a column-oriented data format is slightly more complex than in a row-oriented data format (see Chap. 13).
- *Use real data for application development*: Only real data can show possible bottlenecks of the application architecture and identify patterns that may have a negative impact on the application performance. Another benefit is that user feedback during development tends to be much more productive and precise, if real data is used.
- *Work in multi-disciplinary teams*: We believe that only joint, multidisciplinary efforts of user interface designers, application programmers, database specialists, and domain experts will lead to the creation of new, innovative applications. Each

of them has its own point of view and is able to optimize one aspect of a possible solution, but only if they jointly try to solve problems, the others will benefit from their knowledge.

31.3 Self Test Questions

1. **Architecture of a Banking Solution**
   Current financials solutions contain base tables, change history, materialized aggregates, reporting cubes, indices, and materialized views. The target financials solutions contains...

   (a) only base tables, reporting cubes, and the change history.
   (b) only base tables, algorithms, and some indexes.
   (c) only base tables, materialized aggregates, and materialized views.
   (d) only indexes, change history, and materialized aggregates.

2. **Criterion for Dunning**
   What is the criterion to send out dunning letters?

   (a) Bad stock-market price of the own company
   (b) Bad information about the customer is received from consumer reporting agencies
   (c) When the responsible accounting clerk has to achieve his rate of dunning letters
   (d) A customer payment is overdue.

3. **In-Memory Database for Financials**
   Why is it beneficial to use in-memory databases for financials systems?

   (a) Financial systems are usually running on mainframes. No speed up is needed. All long-running operations are conducted as batch jobs.
   (b) Operations like dunning can be performed in much shorter time.
   (c) Because of the high reliability of data in main memory, less maintenance work is necessary and labor costs could be reduced.
   (d) Easier algorithms are used within the applications, so shorter algorithm run time leads to more work for the end user. Business efficiency is improved.

4. **Connection between Object Fields and Columns**
   Assume that “overdue” is expressed in an enterprise system business object by four fields. How many columns play a role to store that information?

   (a) all columns of the table
   (b) two columns
   (c) four columns
   (d) one column.

5. **Languages for Stored Procedures**
   Languages for stored procedures are...
   (a) designed primarily to be human readable. They follow the spoken english grammar as close as possible.
   (b) strongly imperative, the database is forced to exactly fulfill the orders expressed via the procedure.
   (c) usually a mixture of declarative and imperative concepts.
   (d) strongly declarative, they just describe how the result set should look like. All aggregations and join predicates are automatically retrieved from the database, which has the information "stored" for that.

Chapter 32
Database Views

32.1 Advantages of Views

Database views define a transformation rule that is processed when the underlying data item is accessed [PZ11]. Thus, views describe a structured subset of the complete data available in the database.

The concept of database views come with major two advantages.

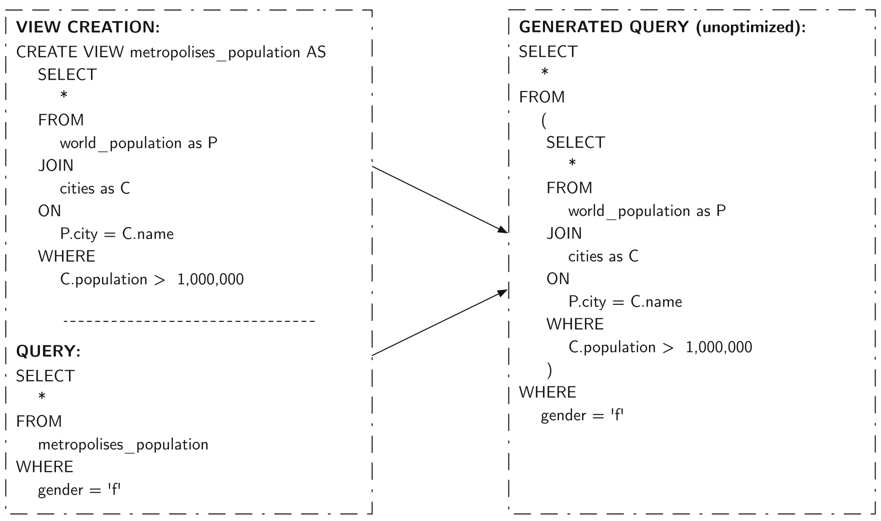
Firstly, they can be used to reduce the complexity of queries, i.e. special transformations and joins are hidden by the view while the transformed data is returned as result. Multiple views can be cascaded. Thus, complex queries can be orchestrated and easily maintained.

Secondly, views replace long-running transformations from the ETL processes by instant transformations. For example, if a data transformation is required ETL requires to transform all data before importing them. With the help of views, the transformation is performed just when a certain data item is accessed. This is advantageous, if only a small subset of all imported data are accessed.

Figure 32.1 shows the view metropolises_population, which returns only citizens from cities with more than one million inhabitants. With the help of this view, writing queries that select citizens of big cities is straight forward while the readability of the code is improved.

Another advantage of database views is that they can be used to create virtual data schemas that build stable interfaces for application development. Two important aspects concerning software quality, software maintenance and reusability, are enhanced by the decoupling of the application code from the actual data schema. A prominent example are data cubes, such as those used in many data warehouses [GBLP96].

Instead of materializing data redundantly as a cube schema, virtual cubes can be created on the fly with the help of database views. In contrast to traditional cubes, virtual cubes work directly on the raw data, i.e. they access always the latest data without any latency. As a result, the integration of third party software applications, such as Business Intelligence dashboards, Microsoft Excel, or web applications, is simplified, storage demands are reduced due to the elimination of redundant data,



```
VIEW CREATION:
CREATE VIEW metropolises_population AS
   SELECT
      *
   FROM
      world_population as P
   JOIN
      cities as C
   ON
      P.city = C.name
   WHERE
      C.population >  1,000,000

      ------------------------------

QUERY:
SELECT
   *
FROM
   metropolises_population
WHERE
   gender = 'f'
```

```
GENERATED QUERY (unoptimized):
SELECT
   *
FROM
   (
    SELECT
       *
    FROM
       world_population as P
    JOIN
       cities as C
    ON
       P.city = C.name
    WHERE
       C.population >  1,000,000
   )
WHERE
   gender = 'f'
```

Fig. 32.1 Using views to simplify join-queries

and the forecasts become more accurately since data used by virtual cubes is always the latest available data within the database.

32.2 Layered Views Concept

Figure 32.2 depicts the *layered view concept*. This concept describes the assembly of views in layers. Hereby, the data sources for a view can be either tables or other views. Views can built on column-oriented and row-oriented database tables equally. The layered view concept allows the integration of external data sources, such as further databases, to join them into one virtual table. Thus, the layered view concepts simplifies the development of queries and the combination of data.

32.3 Development Tools for Views

Graphical tools can be used for view creation. These tools are able to create complex join-views by interactively dragging one database table onto another, whereby join attribute(s) are automatically determined. Furthermore, view development tools provide performance analysis, e.g. of a joined table, and point out possible processing improvements by rearranging data or joins.

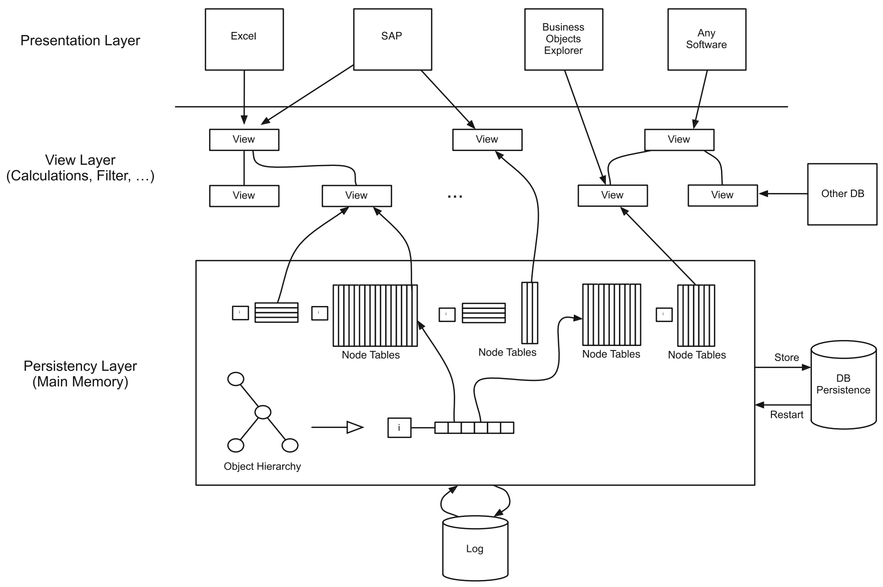


Fig. 32.2 The view layer concept

32.4 Self Test Questions

1. **View Locations**
   Where should a logical view be built to get the best performance?
   (a) in the GPU
   (b) in a third system
   (c) close to the data in the database
   (d) close to the user in the analytical application.
2. **Data Representation**
   What is the traditional representation of business data to the end user?
   (a) bits
   (b) videos
   (c) music
   (d) lists and tables.
3. **Views and Software Quality**
   Which aspects concerning software quality are improved by the introduction of database views?
   (a) Accessibility and availability
   (b) Testability and security

(c) Reliability and usability
(d) Reusability and maintainability.

References

[GBLP96] J. Gray, A. Bosworth, A. Layman, H. Pirahesh, Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-total, in *ICDE*, ed. by S.Y.W. Su (IEEE Computer Society, 1996), pp. 152–159
[PZ11] H. Plattner, A. Zeier, *In-Memory Data Management* (Springer, Heidelberg, 2011)

Chapter 33
Handling Business Objects

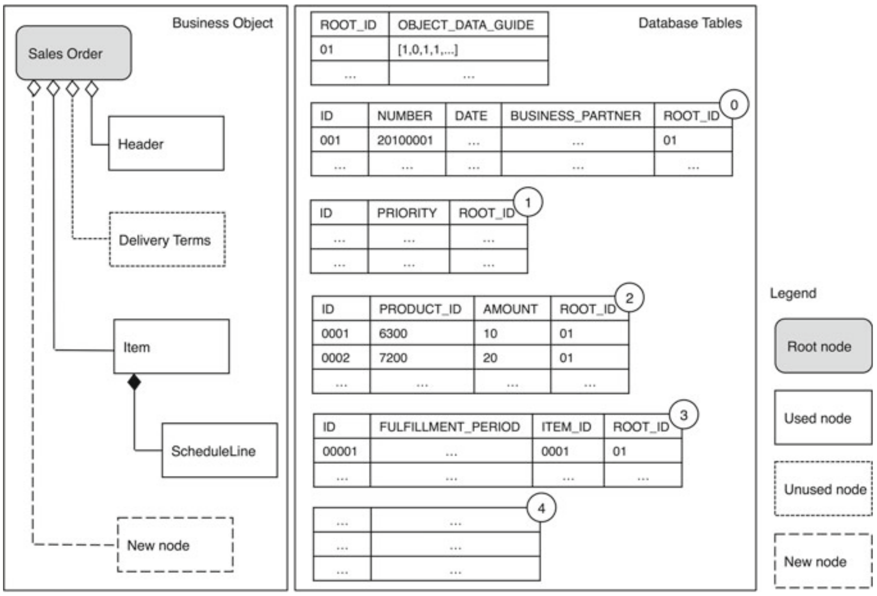
Enterprise applications are typically developed in an object-oriented fashion: Real world objects, such as production facilities or warehouses as well as artifacts like sales orders, are mapped to so-called business objects. A business object is an entity capable of storing information and state. It typically has a tree like structure with leaves holding information about the object or connections to other business objects.

As an example, the left hand side of Fig. 33.1 shows a business object representing a sales order. It consists of a *header* leaf with general information like the order number, the order date and the customer (business partner). As described in Chap. 3, typically only a small number of attributes of the provided ones are really used in productive systems. For the case at hand, the leaf storing the *delivery terms* is not used—this may be the case, when delivery terms have not been entered in the system, yet, or if the company running the system does not maintain this information in its enterprise application. Each sales order consists of a number of *items* and they each have an associated *schedule line* with information about their delivery.

33.1 Persisting Business Objects

The challenge now is, how to persist a business object in the relational database model, so that it still can be queried efficiently. Let us assume that the database only stores the sales order numbers in a *sales order* table and there is an additional table for each of the leaves. When extracting the sales order, there is no way of knowing, which leaves are really used and therefore one SELECT statement needs to be executed for every leaf. In case the sales order consists of 50 leaves, of which only 5 are used, a large number of wasted SELECT statements needs to be executed.

To avoid those unnecessary SELECTs, a *business object data guide* structure can help to store the information which leaves of a business object are populated with data. In our example, the root object stores a bit mask that contains the



ROOT_ID	OBJECT_DATA_GUIDE
01	[1,0,1,1,...]
...	...

ID	NUMBER	DATE	BUSINESS_PARTNER	ROOT_ID
001	20100001	...	...	01
...	...	...	...	...

ID	PRIORITY	ROOT_ID
...	...	...
...	...	...

ID	PRODUCT_ID	AMOUNT	ROOT_ID
0001	6300	10	01
0002	7200	20	01
...	...	...	...

ID	FULFILLMENT_PERIOD	ITEM_ID	ROOT_ID
00001	...	0001	01
...	...	...	...

...	...
...	...
...	...

Fig. 33.1 Sales order business object with object data guide representation

information which leaves are really populated. In the example of Fig. 33.1, the zero at the second position of the *object data guide* indicates that the *delivery terms* leaf is not filled with data and a SELECT statement on that table can be omitted.

33.2 Object-Relational Mapping

Another field of research is object-relational mapping (ORM) integrated in the database. Object-relational mapping is used to map objects—as used in most high-level programming languages—to their relational representations as used in relational databases. ORM inside the database is especially interesting for handling business objects on columnar databases.

One reason is the vast number of applications and systems, that are interacting with the business data. In contrast to most web applications, business applications are highly diverse. To deploy the same view on business objects throughout all applications a business objects repository should be used. Such a repository is a central place for business objects definitions inside the database, which are regularly pulled from applications and systems relying on the business objects. This way of modifying business objects or integrating new business processes (e.g. implemented as stored procedures directly on the database side) does not require to modify each application's ORM framework.

Another advantage for business object handling inside the database is the proximity to the actual data. Object-relational mappers aim at reducing the usage of "SELECT *" queries, as they often occur in applications. Since "SELECT *" should be avoided when possible on columnar stores, object-relational mapping inside the database can prevent such queries and enforce an efficient business object handling. One possibility to prevent such queries is to track regularly requested business objects and only query attributes, which are likely to be used. Any additional attributes, which are unexpectedly requested, would incur additional queries.

Furthermore, having business objects on the database side allows to implement business processes using stored procedures and thus reducing client application code. This way complex business processes can be implemented using business objects instead of raw relational data.

33.3 Self Test Questions

1. **Business Object Mapping**
   What is business object mapping?
   (a) Putting together a diagram of all used business objects. It is similar to a sitemap on webpages
   (b) Allocate an index to every business object and save it in the associated memory area
   (c) Representing every element of a business object in a table
   (d) Create a hash code of the business object and save this hash code instead of the whole object.

Chapter 34
Bypass Solution

As illustrated throughout the course, in-memory data management can enable significant advantages to data processing within enterprises. However, the transition for enterprise applications to an in-memory database will require radical changes to data organization and processing, resulting in major adaptations throughout the entire stack of enterprise applications. By considering conservative upgrade policies used by many ERP system customers, the adoption of in-memory technology is often delayed, because such radical changes do not align well with the evolutionary modification schemes of business-critical customer systems. Consequently, a risk-free approach is required to help enterprises to immediately leverage in-memory data management technology without disruption of their existing enterprise systems.

We propose a transition process that allows customers to benefit from in-memory technology without changing their running systems. This is a step by step, non-disruptive process that helps to transform traditionally separated operational and analytical systems into what we believe is the future for enterprise applications: transactional and analytical workloads handled by a single, in-memory database.

Within the first step of the transition, an in-memory database will run in parallel to the traditional database and the data will be stored in both systems. Secondly, new side-by-side applications using transactional data of the ERP system can be developed. In the next step, a new data warehouse (DW) solution can be introduced that is able to answer flexible, ad-hoc queries and operate without materialized views, aggregating all necessary information on the fly from the transactional data. Complex ETL processes are not required any longer. Finally, the traditional disk-based database of the ERP system can be replaced by a column-oriented dictionary-encoded in-memory database. This transition is described in more detail in the next section.

34.1 Transition Steps in Detail

First of all, we start with a commonly found initial architecture of an existing enterprise solution as illustrated in Fig. 34.1.

Typically, it consists of multiple OLTP and OLAP systems, each of them running on separate databases. The OLAP system consolidates data from multiple OLTP systems and external data sources. A costly and time-consuming ETL process between the OLTP and OLAP systems is used to pre-aggregate data for the OLAP system. Based on this architecture, the non-disruptive transition plan that we call "bypass solution" has been developed.

In the first step of this approach (see Fig. 34.2), the IMDB is installed and connected to the traditional database.

The only difference will be in data representation: data will be stored in columns. An initial load to the in-memory database (IMDB) creates a copy of the existing system state with all business objects in the IMDB. In spite of the huge volume of data to be reproduced, first experiments with massively parallel bulk

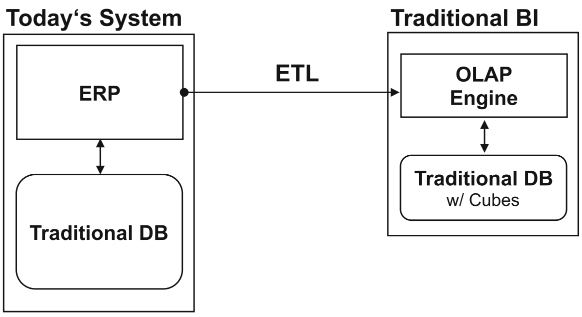


Fig. 34.1 Initial architecture

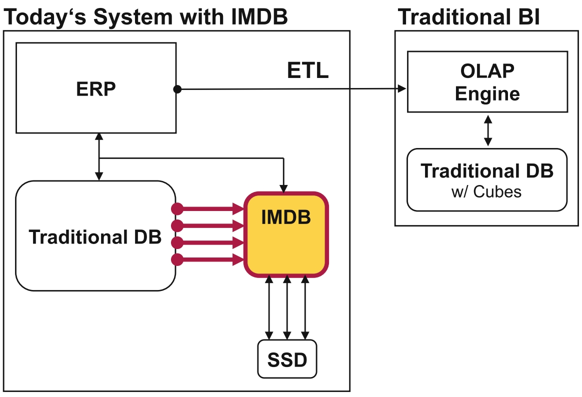


Fig. 34.2 Run IMDB in parallel

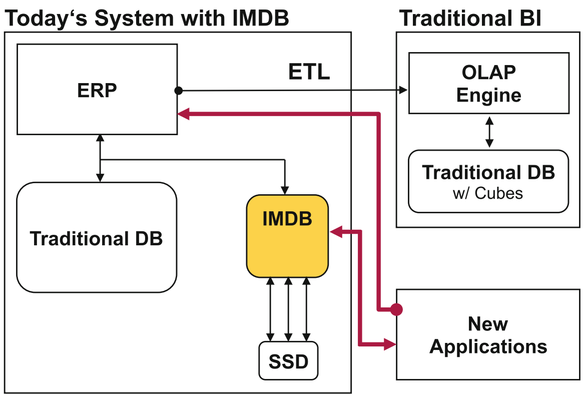


Fig. 34.3 Deploy new applications

loads of customer data have shown that even for the largest companies this one-time initialization can be done in only a few hours.

After the initial load, the two storages will be maintained in parallel, every document and change in a business object is stored in both databases. For this, established database replication technologies are used. The high compression rate in a column store helps to decrease the amount of main memory required for such parallel use of two databases, and hence it does not lead to a significant waste of resources. When calculating with an average data compression factor of 10, a 1 TB database system can be compressed so that only 100 GB of in-memory storage are needed. At the same time, using the parallel installation of the IMDB, we can estimate performance and memory consumption benefits of this architecture for concrete business cases and prove the need of moving the system to the new data storage.

In a second step, new applications can be developed leveraging the potentials of the new technology (see Fig. 34.3). These applications only read the replicated ERP system data. If they want to write data, they either use the traditional interfaces of the ERP system or, if they want to store additional data, they use a

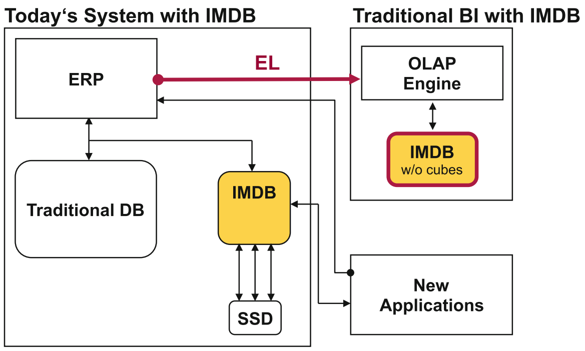


Fig. 34.4 Traditional data warehouse on IMDB

separate segment in the in-memory database. That way, business value by new revolutionary applications can be generated from the first day on.

In a third step, which can be done in parallel with the previous ones, the data warehouse system is ported to the IMDB as illustrated in Fig. 34.4. This will help to achieve another gain in reporting performance in comparison to disk-based OLAP systems.

The difference to the traditional system is that all materialized data cubes and aggregates will be removed. Instead, aggregates are computed on the fly and all data-intensive operations are pushed to the database level. In comparison with storing all materialized aggregates and indices, this reduces the amount of main memory, which is necessary for the OLAP system. It immediately leads to the following benefits: the data cubes in a traditional BI system are usually updated on a weekly basis or even less frequently. However, executives, management, and all other decision makers often demand up-to-date information. By running OLTP and OLAP systems on the same IMDB platform, this information can be delivered in real time. Another advantage is that the ETL process is radically simplified as the complex calculation of aggregates is omitted. The ETL process can partly be replaced by the same replication mechanism used between the traditional ERP database and the parallel installation of the in-memory database. Therefore, we called the replication in Fig. 34.4 simply EL, since no transformation takes place any longer, just extraction and load remain. Furthermore, the BI system will be more flexible as complex cube management and maintenance operations are abandoned and complexity is reduced.

In most cases, the migration to the in-memory database BI system can be conducted automatically. This is relatively simple as existing materialized views are replaced by non-materialized views. Analytical queries can be rewritten by

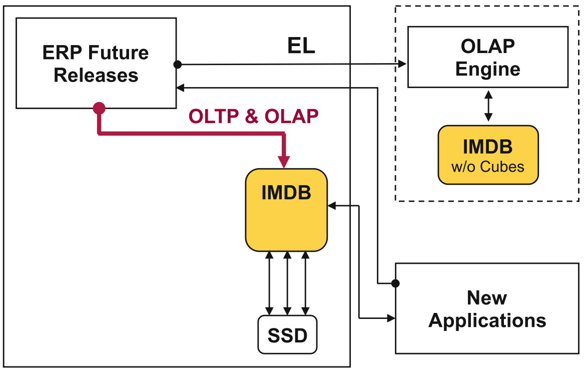


Fig. 34.5 Run OLTP and OLAP on IMDB

generators. However, in complex migrations scenarios, factors such as different SQL dialects, data structures, or software parts that rely on additional data in proprietary formats may pose an obstacle and might require manual interventions.

The final step of the suggested solution can be executed when the customer is comfortable with the parallel in-memory solution. In this step, the traditional OLTP database can be switched off. After that, the customer works only with the consolidated in-memory enterprise system that is used for both transactional and analytical queries (see Fig. 34.5).

The data warehouse system could theoretically be replaced in a setup with only one ERP system. Reality shows that many OLTP systems exist and external data has to be integrated into the analytical system. To this end, the traditional BW system is used as a data platform consolidating data from these OLTP systems and external sources.

Additionally, during system evolution, new extensions to the data model are possible. Adding new tables and new attributes to existing tables in the column store is done on the fly and this speeds up release cycles significantly.

34.2 Bypass Solution: Conclusion

As discussed above, the suggested bypass solution introduces a risk-free and non-disruptive transition to the in-memory database technology. The importance of this transition can be backed-up by selected customer examples: For a large financial service provider, the analysis of 33 million customer records could be reduced from 45 min on the traditional DBMS to 5 s on an IMDB. This increase in speed fundamentally changes the company's opportunities for customer relationship management, promotion planning, and cross selling.

In a similar use case, a large vendor in the construction industry is using an IMDB to analyze its nine million customer records and to create contact listings for specific regions, sales organizations, and branches. Customer contact listing is currently an IT process that may take two to three days to complete. A request must be sent to the IT department who must plan a background job that may take 30 min and the results have to be manually returned to the requester. With an IMDB, sales people can directly query the live system and create customer listings in any format they wish, in less than 10 s.

Concluding, the use of in memory technology can lead to a qualitative change in the business processes of an enterprise. The transaction that took days using the traditional process, can now be performed in the foreground on the fly. That will change the way of thinking and optimize many business processes in enterprises.

34.3 Self Test Questions

1. **Transition to IMDBs**
   What does the transition to in-memory database technology mean for enterprise applications?

   (a) Data organization and processing will change radically and enterprise applications need to be adapted
   (b) The data organization will not change at all, but the source code of the applications has to be adapted
   (c) There will be no impact on enterprise applications
   (d) All enterprise applications are significantly sped up without incurring any adaptions.

Self Test Solutions

Introduction

1. **Rely on Disks**
   Does an in-memory database still rely on disks?

   Possible Answers:

   (a) Yes, because disk is faster than main memory when doing complex calculations
   (b) No, data is kept in main memory only
   (c) Yes, because some operations can only be performed on disk
   (d) Yes, for archiving, backup, and recovery

   Correct Answer: (d)

   Explanation: Logs for archiving have to be stored on a persistent storage medium that preserves the content longer timespans. Main memory looses all information if the system is unpowered, therefore an other, persistent storage medium such as hard drives or SSDs has to be used for recovery and archiving.

New Requirements for Enterprise Computing

1. **Compression Factor**
   What is the average compression factor for accounting data in an in-memory column-oriented database?

   Possible Answers:

   (a) 100x
   (b) 10x
   (c) 50x
   (d) 5x

Correct Answer: (b)

Explanation: As mentioned in the text, the average compression factor that was measured in a financials application was about 10. Other measurements in different industries showed average compression factors in the same range.

2. **Data explosion**
   Consider the formula 1 race car tracking example, with each race car having 512 sensors, each sensor records 32 events per second whereby each event is 64 byte in size.
   How much data is produced by a F1 team, if a team has two cars in the race and the race takes two hours?
   For easier calculation, assume 1,000 byte = 1 kB, 1,000 kB = 1 MB, 1,000 MB = 1 GB.

   Possible Answers:

   (a) 14 GB
   (b) 15.1 GB
   (c) 32 GB
   (d) 7.7 GB

   Correct Answer: (b)

   Explanation: Total time: $2\,\text{h} = 2 \cdot 60 \cdot 60\,\text{s} = 7,200\,\text{s}$
   $\text{Total events per car} : 7,200\,\text{s} \cdot 512\ \text{sensors} \cdot 32\,\text{events/second/sensor}$
   $= 117,964,800\ \text{events}$
   $\text{Total events per team} : (2 \cdot \text{total events per car}) = 235,929,600\ \text{events}$
   $\text{Total amount of data per team} : 64\ \text{byte/event} \cdot 235,929,600\ \text{events}$
   $= 15,099,494,400\,\text{byte} \approx 15.1\,\text{GB}$

Enterprise Application Characteristics

1. **OLTP OLAP Separation Reasons**
   Why was OLAP separated from OLTP?

   Possible Answers:

   (a) Due to performance problems
   (b) For archiving reasons; OLAP is more suitable for tape-archiving
   (c) Out of security concerns
   (d) Because some customers only wanted either OLTP or OLAP and did not want to pay for both

Correct Answer: (a)

Explanation: The runtimes of analytical queries are significantly higher than these of transactional ones. Based on this characteristic, analytical processing negatively affected the day-to-day business i.e. in terms of delayed sales processing. The separation of analytical and transactional queries to different machines was the inevitable consequence of the hardware and database prerequisites of these times.

Changes in Hardware

1. **Speed per Core**
What is the speed of a single core when processing a simple scan operation (under optimal conditions)?

Possible Answers:

(a) 2 GB/ms/core
(b) 2 MB/ms/core
(c) 2 MB/s/core
(d) 200 MB/s/core

Correct Answer: (b)

Explanation: Today's CPUs access more than 95 % of the data to run programs out of their caches without incurring a cache miss. We assume that all data we want to deal with is in the Level 1 cache and ignore further delays from the fetching the data into the Level 1 cache. From Level 1 cache we need roughly 0.5 ns to load one byte, so in 1 millisecond, we could load about 2,000,000 byte, which is 2 MB (1,000,000 ns/0.5 ns per byte = 2,000,000 byte).

2. **Latency of Hard Disk and Main Memory**
Which statement concerning latency is wrong?

Possible Answers:

(a) The latency of main memory is about 100 ns
(b) A disk seek takes an average of 0.5 ms
(c) Accessing main memory is about 100,000 times faster than a disk seek
(d) 10 ms is a good estimation for a disk seek

Correct Answer: (b)

Explanation: Please have a look at Table 4.1 on page 48.

A Blueprint for SanssouciDB

1. **New Bottleneck**
   What is the new bottleneck of SanssouciDB that data access has to be optimized for?

   Possible Answers:

   (a) Disk
   (b) The ETL process
   (c) Main memory
   (d) CPU

   Correct Answer: (c)

   Explanation: Main memory access is the new bottleneck, since the CPU busses limit overall data transfer. CPU speed increased according to Moore's Law (in terms of parallelism) and outperforms the bus speed. Disks are only used for backup and archiving reasons in SanssouciDB and are therefore not of interest for actual production usage. ETL processes are not of concern either, since all queries run on transactional data in one system for online transaction processing and online analytical processing.

2. **Indexes**
   Can indexes still be used in SanssouciDB?

   Possible Answers:

   (a) No, because every column can be used as an index
   (b) Yes, they can still be used to increase performance
   (c) Yes, but only because data is compressed
   (d) No, they are not even possible in columnar databases

   Correct Answer: (b)

   Explanation: Indices are a valid optimization of the performance in SanssouciDB. The index concept does not rely on compression.

Dictionary Encoding

1. **Lossless Compression**
   For a column with few distinct values, how can dictionary encoding significantly reduce the required amount of memory without any loss of information?

Possible Answers:

(a) By mapping values to integers using the smallest number of bits possible to represent the given number of distinct values
(b) By converting everything into full text values. This allows for better compression techniques, because all values share the same data format.
(c) By saving only every second value
(d) By saving consecutive occurrences of the same value only once

Correct Answer: (a)

Explanation: The correct answer describes the main principle of dictionary encoding, which automatically results in a lossless compression if values appear more often than once. Saving only every second value is clearly lossy. The same applies for saving consecutive occurrences of the same value only once, if the quantity of occurrences is not saved as well. Additionally, this does not describe dictionary encoding, but Run-Length Encoding. Transforming numbers and other values into text values increases the data size, since each character value is at least 1 byte in order to allow the representation of the full alphabet. Number representations are usually limited to certain upper limits and achieve much compacter data sizes.

2. **Compression Factor on Whole Table**
Given a population table (50 millions rows) with the following columns:

- name (49 bytes, 20,000 distinct values)
- surname (49 bytes, 100,000 distinct values)
- age (1 byte, 128 distinct values)
- gender (1 byte, 2 distinct values)

What is the compression factor (uncompressed size/compressed size) when applying dictionary encoding?

Possible Answers:

(a) ≈ 20
(b) ≈ 90
(c) ≈ 10
(d) ≈ 5

Correct Answer: (a)

Explanation: Calculation without dictionary encoding:
Total size per row: $49 + 49 + 1 + 1$ (byte) $= 100$ byte
Total size: 100 byte $\cdot$ 50 million rows $= 5,000\,\text{MB}$

Calculation with dictionary encoding:
Number of bit needed for the attributes:

- names: $\log_2(20{,}000) < 15$
- surnames: $\log_2(100{,}000) < 17$
- ages: $\log_2(128) <= 7$
- genders: $\log_2(2) <= 1$

Size of the attribute vectors:

- 50 million rows $\cdot (15 + 17 + 7 + 1)$ bit $=$ 2,000 million bit $= 250\,\text{MB}$

Size of the dictionaries:

- names: $20{,}000 \cdot 49$ byte $= 980\,\text{KB}$
- surnames: $100{,}000 \cdot 49$ byte $= 4,9\,\text{MB}$
- ages: $128 \cdot 7$ byte $=$ 896 byte
- genders: $2 \cdot 1$ byte $=$ 2 byte

Total dictionary size: $4,9\,\text{MB} + 980\,\text{KB} + 896$ byte $+ 2$ byte $\approx 5\,\text{MB}$
Overall size: size of attribute vectors + size of dictionaries $= 250$ MB $+ 5$ MB
$= 255$ MB
Compression rate:
$5000\,\text{MB}/255\,\text{MB} = 19,6 \approx 20$

3. **Information in the Dictionary**
What information is saved in a dictionary in the context of dictionary encoding?

Possible Answers:

(a) Cardinality of a value
(b) All distinct values
(c) Hash of a value of all distinct values
(d) Size of a value in bytes

Correct Answer: (b)

Explanation: The dictionary is used for encoding the values of a column. Therefore it consists of a list of all distinct values to be encoded and the resulting encoded values (in most cases ascending numbers). The distinct values are used to encode the attributes in user queries during look-ups and to decode the retrieved numbers from query results back to meaningful, human-readable values.

4. Advantages through Dictionary Encoding
What is an advantage of dictionary encoding?

Possible Answers:

(a) Sequentially writing data to the database is sped up
(b) Aggregate functions are sped up
(c) Raw data transfer speed between application and database server is increased
(d) INSERT operations are simplified

Correct Answer: (b)

Explanation: Aggregate functions are sped up when using dictionary encoding. because less data has to be transferred from main memory to CPU. The raw data transfer speed between application and database server is not increased, this is a determined by the physical hardware and exploited to the maximum. Insert operations suffer from dictionary encoding, because new values that are not yet present in the dictionary, have to be added to the dictionary and might require a re-sorting of the related attribute vector. Due to that, sequentially writing data to the database is not sped up, either.

5. Entropy
What is entropy?

Possible Answers:

(a) Entropy limits the amount of entries that can be inserted into a database. System specifications greatly affect this key indicator.
(b) Entropy represents the amount of information in a given dataset. It can be calculated as the number of distinct values in a column (column cardinality) divided by the number of rows of the table (table cardinality).
(c) Entropy determines tuple lifetime. It is calculated as the number of duplicates divided by the number of distinct values in a column (column cardinality).
(d) Entropy limits the attribute sizes. It is calculated as the size of a value in bits divided by number of distinct values in a column the number of distinct values in a column (column cardinality).

Correct Answer: (b)

Explanation: As in information theory, in this context, entropy determines the amount of information content gained from the evaluation of a certain message, in this case the dataset.

Compression

1. Sorting Compressed Tables
Which of the following statements is correct?

Possible Answers:

(a) If you sort a table by the amount of data for a row, you achieve faster read access
(b) Sorting has no effect on possible compression algorithms
(c) You can sort a table by multiple columns at the same time
(d) You can sort a table only by one column

Correct Answer: (d)

Explanation: Some compression techniques achieve a better compression rate when they are applied to a sorted table, like Indirect Encoding. Furthermore, a table cannot be sorted by multiple columns at the same time, so the right answer is that a table can only be sorted by one column. A potential enhancement is that one can sort a table cascading (i.e. first sort by country, then sort the resulting groups by city,...), which improves accesses to the column used as secondary sorting attribute to some extent.

2. Compression and OLAP / OLTP
What do you have to keep in mind if you want to bring OLAP and OLTP together?

Possible Answers:

(a) You should not use any compression techniques because they increase CPU load
(b) You should not use compression techniques with direct access, because they cause major security concerns
(c) Legal issues may prohibit to bring certain OLTP and OLAP datasets together, so all entries have to be reviewed
(d) You should use compression techniques that give you direct positional access, since indirect access is too slow

Correct Answer: (d)

Explanation: Direct positional access is always favorable. It does not cause any difference to data security. Also, legal issues will not interfere with OLTP and OLAP datasets, since all OLAP data is generated out of OLTP data. The increased CPU load which occurs when using compression is tolerated because the compression leads to smaller data sizes which usually results in better cache usage and faster response times.

3. **Compression Techniques for Dictionaries**
   Which of the following compression techniques can be used to decrease the size of a sorted dictionary?

   Possible Answers:

   (a) Cluster Encoding
   (b) Prefix Encoding
   (c) Run-Length Encoding
   (d) Delta Encoding

   Correct Answer: (d)

   Explanation: Delta Encoding for Dictionaries is explained in detail in Sect. 7.5. Cluster Encoding, Run-Length Encoding, and Prefix Encoding can not be used on dictionaries because each dictionary entry is unique.

4. **Indirect Access Compression Techniques**
   Which of the explained compression techniques does not support direct access?

   Possible Answers:

   (a) Run-Length Encoding
   (b) Prefix Encoding
   (c) Cluster Encoding
   (d) Indirect Encoding

   Correct Answer: (c)

   Explanation: Cluster Encoding does not support direct access. The position of a record has to be computed via the bit vector.

5. **Compression Example Prefix Encoding**
   Suppose there is a table where all 80 million inhabitants of Germany are assigned to their cities. Germany consists of about 12,200 cities, so the valueID is represented in the dictionary via 14 bit. The outcome of this is that the attribute vector for the cities has a size of 140 MB. We compress this attribute vector with Prefix Encoding and use Berlin, which has nearly 4 million inhabitants, as the prefix value. What is the size of the compressed attribute vector?
   Assume that the needed space to store the amount of prefix values and the prefix value itself is neglectable, because the prefix value only consumes 22 bit to represent the number of citizens in Berlin and additional 14 bit to store the key for Berlin once. Further assume the following conversions: 1 MB = 1000 kB, 1 kB = 1000 B

Possible Answers:

(a) 0.1 MB
(b) 133 MB
(c) 63 MB
(d) 90 MB

Correct Answer: (b)

Explanation: Because we use Prefix Encoding for this attribute vector, we do not save the valueID for Berlin 4 million times to represent its inhabitants in the city column. Instead, we resort the table so that all people who live in Berlin are on the top. In the attribute vector we save the valueID for Berlin and the number of occurrences for that valueID. Then the valueIDs for the remaining 76 million people in Germany follow. So the new size of the attribute vector is made up of the size of the valueID for Berlin (14 bit), the size needed to save the number of occurrences for Berlin (22 bit) and the size of the remaining entries. The missing numbers can be calculated the following way: From 80 million people in Germany remain 76 million to store. Each entry needs 14 bit for the valueID of its city. So the size of the remaining entries is 76 million $\cdot$ 14 bit $=$ 1064000000 bit. Thus, the size of the attribute vector is 14 bit $+$ 22 bit $+$ 1,064,000,000 bit $=$ 1,064,000,036 bit, which is about 133 Mbyte (8 bit $=$ 1 byte).

6. **Compression Example Run-Length Encoding Germany**
Suppose there is a table where all 80 million inhabitants of Germany are assigned to their cities. The table is sorted by city. Germany consists of about 12,200 cities (represented by 14 bit). Using Run-Length Encoding with a start position vector, what is the size of the compressed city vector? Always use the minimal number of bits required for any of the values you have to choose. Further assume the following conversions: 1 MB $=$ 1,000 kB, 1 kB $=$ 1,000 B

Possible Answers:

(a) 1.2 MB
(b) 127 MB
(c) 5.2 KB
(d) 62.5 kB

Correct Answer: (d)

Explanation: We have to compute the size of (a) the value array and (b) the size of the start position array. The size of (a) is the distinct number of cities (12,200) times the size of each field of the value array (log_2(12,200)). The size of (b) is the number of entries in the dictionary (12,200) times the number of bit required to encode the highest possible number of inhabitants (log_2(80,000,000)). The result is thus 14 bit times 12,200 (170,800) plus 27 bit times 12,200 (329,400), summing up to 500,200 bit (or 62.5 kB) in total.

7. Compression Example Cluster Encoding

Assume the world population table with 8 billion entries. This table is sorted by countries. There are about 200 countries in the world. What is the size of the attribute vector for countries if you use Cluster Encoding with 1,024 elements per block assuming one block per country can not be compressed? Use the minimum required count of bits for the values. Further assume the following conversions: 1 MB = 1,000 KB, 1 kB = 1,000 B

Possible Answers:

(a) $\approx$ 9 MB
(b) $\approx$ 4 MB
(c) $\approx$ 0.5 MB
(d) $\approx$ 110 MB

Correct Answer: (a)

Explanation: To represent the 200 cities, 8 bit are needed for the valueID, because $log_2(200)$ is 8. With a cluster size of 1024 elements the number of blocks is 7,812,500 (8 billion entries/1,024 elements per block). Each country has one incompressible block, so there are 200 of them. The size of one incompressible block is the number of elements per block (1,024) times the size of one valueID (8 bit). The result is 8,192 bit and consequently the required size for the 200 blocks is 200 $\cdot$ 8,192 bit = 1,638,400 bit. For the remaining 7,812,300 compressible blocks, it is only necessary to store one valueID for each block. Hence the resulting size of the compressible blocks is 62,498,400 bit (7,812,300 $\cdot$ 8 bit). Finally there is the bit vector which indicates compressible and incompressible blocks. It requires 1 bit per block, so it has a size of 7,812,500 bit. The size of the whole compressed attribute vector is the sum of the size of the compressed and uncompressed blocks and the bit vector, which is 1,638,400 bit + 62,498,400 bit + 7,812,500 bit = 71,949,300 bit. That is about 9 MB, which is the correct answer.

8. Best Compression Technique for Example Table

Find the best compression technique for the name column in the following table. The table lists the names of all inhabitants of Germany and their cities, i.e. there are two columns: first_name and city. Germany has about 80 million inhabitants and 12,200 cities. The table is sorted by the city column. Assume that any subset of 1,024 citizens contains at most 200 different first names.

Possible Answers:

(a) Run-Length Encoding
(b) Indirect Encoding
(c) Prefix Encoding
(d) Cluster Encoding

Correct Answer: (b)

Explanation: In order to use Prefix Encoding or Run-Length Encoding to compress a column, a table has to be sorted by this specific column. In this example, we want to compress the "first_name" column, but the table is sorted by the column city. Therefore we can not use these two compression techniques. Cluster Encoding is possible in general and we could achieve high compression rates, but unfortunately Cluster Encoding does not support direct access. So choosing Cluster Encoding would prohibit direct access and consequently we would loose much performance. As a conclusion, Indirect Encoding is the best compression technique for this column, because it works with a good compression rate while keeping the possibility of direct access.

Data Layout in Main Memory

1. When DRAM can be accessed randomly with the same costs, why are consecutive accesses usually faster than stride accesses?

   Possible Answers:

   (a) With consecutive memory locations, the probability that the next requested location has already been loaded in the cache line is higher than with randomized/strided access. Furthermore is the memory page for consecutive accesses probably already in the TLB.
   (b) The bigger the size of the stride, the higher the probability, that two values are both in one cache line.
   (c) Loading consecutive locations is not faster, since the CPU performs better on prefetching random locations, than prefetching consecutive locations.
   (d) With modern CPU technologies like TLBs, caches and prefetching, all three access methods expose the same performance.

   Correct Answer: (a)

   Explanation: Having always the same distance between accessed addresses enables the prefetcher to predict the correct locations to load. For randomly accessed addresses this is obviously not possible. Furthermore, strides of zero, which is the case for consecutive attribute accesses using columnar layouts, are highly cash efficient, since solely the needed locations are loaded. To summarize, random memory accesses might have the same costs, but since the CPU loads more data than requested into the caches and the prefetcher even fetches unrequested data, data with a high information density (entropy) can be processed faster.

Partitioning

1. **Partitioning Types**
   Which partitioning types do really exist and are mentioned in the course?

   Possible Answers:

   (a) Selective Partitioning
   (b) Syntactic Partitioning
   (c) Range Partitioning
   (d) Block Partitioning

   Correct Answer: (c)

   Explanation: Range Partitioning is the only answer that really exists. It is a subtype of horizontal partitioning and separates tables into partitions by a predefined partitioning key, which determines how individual data rows are distributed to different partitions.

2. **Partitioning Type for given Query**
   Which partitioning type fits best for the column 'birthday' in the world population table, when we assume that the main workload is caused by queries like 'SELECT first_name, last_name FROM population WHERE birthday $> 01.01.1990$ AND birthday $< 31.12.2010$ AND country = 'England' Assume a non-parallel setting, so we can not scan partitions in parallel. The only parameter that is changed in the query is the country.

   Possible Answers:

   (a) Round Robin Partitioning
   (b) All partitioning types will show the same performance
   (c) Range Partitioning
   (d) Hash Partitioning

   Correct Answer: (c)

   Explanation: Range Partitioning separates tables into partitions by a predefined key. In the example that would lead to a distribution where all required tuples are in the same partition (or in the minimal number of partitions to cover the queried range) and our query only needs to access this (or these). Round Robin Partitioning would not be a good partitioning type in this example, because it assigns tuples turn by turn to each partition, so the data is separated across many different partitions which have to be accessed. Hash Partitioning uses a hash function to specify the partition assignment for each row, so the data is probably separated across many different partitions, too.

3. **Partitioning Strategy for Load Balancing**
   Which partitioning type is suited best to achieve fair load-balancing if the values of the column are non-uniformly distributed?

   Possible Answers:

   (a) Partitioning based on the number of attributes used modulo the number of systems
   (b) Range Partitioning
   (c) Round Robin Partitioning
   (d) All partitioning types will show the same performance

   Correct Answer: (c)

   Explanation: Round Robin Partitioning distributes tuples turn by turn to each partition, so all partitions have nearly the same number of tuples. In contrast, Range Partitioning assigns entries to the table by a predefined partitioning key. Because the values used as partitioning keys are normally distributed non-uniformly, it is difficult or maybe even impossible to find a key that segments the table into parts of the same size. Consequently, Round Robin Partitioning is the best strategy for fair load-balancing if the values of the column are non-uniformly distributed.

Delete

1. **Delete Implementations**
   Which two possible delete implementations are mentioned in the course?

   Possible Answers:

   (a) White box and black box delete
   (b) Physical and logical delete
   (c) Shifted and liquid delete
   (d) Column and row deletes

   Correct Answer: (b)

   Explanation: A physical delete erases the tuple content from memory, so that it is no longer there. A logical delete only marks the tuple as invalid, but it may still be queried for history traces.

2. Arrays to Scan for Specific Query with Dictionary Encoding

When applying a delete with two predicates, e.g. firstname=‘John’ AND lastname=‘Smith’ how many logical blocks in the IMDB are being looked at during determination which tuples to delete (all columns are dictionary encoded)?

Possible Answers:

(a) 1
(b) 2
(c) 4
(d) 8

Correct Answer: (c)

Explanation: First the two dictionaries for firstname and lastname to get the corresponding valueIDs and then the two attribute vectors to get the positions (recordIDs).

3. Fast Delete Execution

Assume a physical delete implementation and the following two SQL statements on our world population table:
(A) DELETE FROM world_population WHERE country=‘China’;
(B) DELETE FROM world_population WHERE country=‘Ireland’;
Which query will execute faster? Please only consider the concepts learned so far.

Possible Answers:

(a) Equal execution time
(b) A
(c) Depends on the ordering of the dictionary
(d) B

Correct Answer: (d)

Explanation: Based on the actual locations of used logical blocks in a physical delete implementation, the largest part of the time will be moving memory blocks. Therefore the number of deletes is essential for the runtime. Since China has a much larger population than Ireland, query B will be faster.

Insert

1. **Access Order of Structures during Insert**
   When doing an insert, what entity is accessed first?

   Possible Answers:

   (a) The attribute vector
   (b) The dictionary
   (c) No access of either entity is needed for an insert
   (d) Both are accessed in parallel in order to speed up the process

   Correct Answer: (b)

   Explanation: First the dictionary is scanned in order to figure out, whether the value to be inserted is already part of the dictionary or has to be added.

2. **New Value in Dictionary**
   Given the following entities:
   Old dictionary: ape, dog, elephant, giraffe
   Old attribute vector: 0, 3, 0, 1, 2, 3, 3
   Value to be inserted: lamb
   What value is the lamb mapped to in the new attribute vector?

   Possible Answers:

   (a) 1
   (b) 2
   (c) 3
   (d) 4

   Correct Answer: (d)

   Explanation: "lamb" starts with the letter "l" and therefore does belong after the entry "giraffe". "giraffe" was the last entry with the logical number 3 (fourth entry) in the old dictionary, so "lamb" gets the number 4 in the new dictionary.

3. **Insert Performance Variation over Time**
   Why might real world productive column stores experience faster insert performance over time?

Possible Answers:

(a) Because the dictionary reaches a state of saturation and, thus, rewrites of the attribute vector become less likely.
(b) Because the hardware will run faster after some run-in time.
(c) Because the column is already loaded into main-memory and does not have to be loaded from disk.
(d) An increase in insert performance should not be expected.

Correct Answer: (a)

Explanation: Consider for instance a database for the world population. Most first names probably did appear after writing a third of the world population. Future inserts can be done a little faster since less steps are required if the values are already present in the corresponding dictionary.

4. Resorting Dictionaries of Columns
Consider a dictionary encoded column store (without a differential buffer) and the following SQL statements on an initially empty table:
INSERT INTO students VALUES('Daniel', 'Bones', 'USA');
INSERT INTO students VALUES('Brad', 'Davis', 'USA');
INSERT INTO students VALUES('Hans', 'Pohlmann', 'GER');
INSERT INTO students VALUES('Martin', 'Moore', 'USA');
How many complete attribute vector rewrites are necessary?

Possible Answers:

(a) 2
(b) 3
(c) 4
(d) 5

Correct Answer: (b)

Explanation: Each column needs to looked at seperately. An attribute vector always gets rewritten, if the dictionary was resorted.

- First name: 'Daniel'; gets inserted, its the first dictionary entry. When 'Brad' gets added to the dictionary, it needs to be resorted and therefore the attribute vector needs to be rewritten. 'Hans' and 'Martin' are simply appended to the end of the dictionary each time. Thats a total of one rewrite for the first name.

For the other attributes, the process is equal, the actions are described in short:

- Last name: Bones, Davis, Pohlmann, Moore $\rightarrow$ rewrite
- Country: USA, USA $\rightarrow$ already present, GER $\rightarrow$ rewrite, USA $\rightarrow$ already present

In total, three rewrites are necessary.

5. Insert Performance
Which of the following use cases will have the worst insert performance when all values will be dictionary encoded?

Possible Answers:

(a) A city resident database, that store all the names of all the people from that city
(b) A database for vehicle maintenance data which stores failures, error codes and conducted repairs
(c) A password database that stores the password hashes
(d) An inventory database of a company storing the furnature for each room

Correct Answer: (c)

Explanation: Inserts take especially long when new unique dictionary entries are inserted. This will be most likely the case for time stamps and password hashes.

Update

1. Status Update Realization
How do we want to realize status updates for binary status variables?

Possible Answers:

(a) Single status field: “false” means state 1, “true” means state 2
(b) Two status fields: “true/false” means state 1, “false/true” means state 2
(c) Single status field: “null” means state 1, a timestamp signifies transition to state 2
(d) Single status field: timestamp 1 means state 1, timestamp 2 means state 2

Correct Answer: (c)

Explanation: By using “null” for state 1 and a timestamp for state 2, the maximum density of needed information is achieved. Given a binary status, the creation time of the initial status is available in the creation timestamp of the complete tuple, it does not have to be stored again. If the binary information is flipped, the “update” timestamp conserves all necessary information in the described manner. Just saving “true” or “false” would discard this information.

2. Value Updates

What is a “value update”?

Possible Answers:

(a) Changing the value of an attribute
(b) Changing the value of a materialized aggregate
(c) The addition of a new column
(d) Changing the value of a status variable

Correct Answer: (a)

Explanation: In typical enterprise applications, three different types of updates can be found. Aggregate updates change a value of a materialized aggregate, status updates change the value of a status variable and finally value updates change the value of an attribute. Adding a new, empty column is not regarded as an update of a tuple at all, because it manipulates the whole relation via the database schema.

3. Attribute Vector Rewriting after Updates

Consider the world population table (first name, last name) that includes all people in the world: Angela Mueller marries Friedrich Schulze and becomes Angela Schulze. Should the complete attribute vector for the last name column be rewritten?

Possible Answers:

(a) No, because ‘Schulze’ is already in the dictionary and only the valueID in the respective row will be replaced
(b) Yes, because ‘Schulze’ is moved to a different position in the dictionary
(c) It depends on the position: All values after the updated row need to be rewritten
(d) Yes, because after each update, all attribute vectors affected by the update are rewritten

Correct Answer: (a)

Explanation: Because the entry ‘Schulze’ is already in the dictionary, it implicitly has the correct position concerning the sort order and does not need to be moved. Furthermore, each attribute vector entry has a fixed size, so that every dictionary entry can be referenced without changing the position of the adjacent entries in the memory area. Based on these two facts, the answer is that the attribute vector does not need to be rewritten.

Tuple Reconstruction

1. Tuple Reconstruction on the Row Layout: Performance
Given a table with the following characteristics:

- Physical storage in rows
- The size of each field is 34 byte
- The number of attributes is 9
- A cache line has 64 byte
- The CPU processes 2 MB per millisecond.

Calculate the time required for reconstructing a full row. Please assume the following conversions: 1 MB = 1,000 kB, 1 kB = 1,000 B

Possible Answers:

(a) ≈ 0.1 µs
(b) ≈ 0.275 µs
(c) ≈ 0.16 µs
(d) ≈ 0.416 µs

Correct Answer: (c)

Explanation: For 9 attributes of each 34 byte rows, in total 306 byte have to be fetched. Given a cache line size of 64 byte, 5 cache lines have to be filled: $5 \cdot 64$ byte = 320 byte;
Total time needed: size of data to be read/processing speed = 320 byte/ 2,000,000 byte/ms/core = 0.16 µs

2. Tuple Reconstruction on the Column Layout: Performance
Given a table with the following characteristics:

- Physical storage in columns
- The size of each field is 34 byte
- The number of attributes is 9
- A cache line has 64 byte
- The CPU processes 2 MB per millisecond

Calculate the time required for reconstructing a full row. Please assume the following conversions: 1 MB = 1,000 KB, 1 kB = 1,000 B

Possible Answers:

(a) ≈ 0.16 µs
(b) ≈ 0.145 µs
(c) ≈ 0.288 µs
(d) ≈ 0.225 µs

Correct Answer: (c)

Explanation: Size of data to be read: number of attributes $\cdot$ cache line size $=$ $9 \cdot 64 = 576$ byte
Total time needed: size of data to be read/processing speed = 576 byte/2,000,000 byte/ms/core = 0.288 μs

3. Tuple Reconstruction in Hybrid Layout
A table containing product stock information has the following attributes: Warehouse (4 byte); Product Id (4 byte); Product Name Short (20 byte); Product Name Long (40 byte); Self Production (1 byte); Production Plant (4 byte); Product Group (4 byte); Sector (4 byte); Stock Volume (8 byte); Unit of Measure (3 byte); Price (8 byte); Currency (3 byte); Total Stock Value (8 byte); Stock Currency (3 byte)
The size of a full tuple is 114 byte.
The size of a cache-line is 64 byte.
The table is stored in main memory using a hybrid layout. The following fields are stored together:

- Stock Volume and Unit of Measure;
- Price and Currency;
- Total Stock Value and Stock Currency;

All other fields are stored column-wise.
Calculate and select from the list below the time required for reconstructing a full tuple using a single CPU core. Please assume the following conversions: 1 MB = 1,000 kB, 1 kB = 1,000 B

Possible Answers:

(a) ≈ 0.352 μs
(b) ≈ 0.020 μs
(c) ≈ 0.061 μs
(d) ≈ 0.427 μs

Correct Answer: (a)

Explanation: The correct answer is calculated as follows: first, the number of cache lines to be accessed is determined. Attributes that are stored together, can be read in one access if the total bit size to be retrieved does not exceed the size of a cache line.
Stock Volume and Unit of Measure: 8 byte + 3 byte < 64 byte $\rightarrow$ 1 cache line
Price and Currency: 8 byte + 3 byte < 64 byte $\rightarrow$ 1 cache line
Total Stock Value and Stock Currency: 8 byte + 3 byte < 64 byte $\rightarrow$ 1 cache line
All other 8 attributes are stored column wise, so one cache access per attribute is required, resulting in additional cache accesses.

The total amount of data to be read is therefore: 11 (cachelines) $\cdot$ 64 byte = 704 byte
704 byte/(2, 000, 000 byte/ms/core) $\cdot$ 1 core = 0.352 µs

4. **Comparison of Performance of the Tuple Reconstruction on Different Layouts**

A table containing product stock information has the following attributes: Warehouse (4 byte); Product Id (4 byte); Product Name Short (20 byte); Product Name Long (40 byte); Self Production (1 byte); Production Plant (4 byte); Product group (4 byte); Sector (4 byte); Stock Volume (8 byte); Unit of Measure (3 byte); Price (8 byte); Currency (3 byte); Total Stock Value (8 byte); Stock Currency (3 byte)
The size of a full tuple is 114 byte.
The size of a cache-line is 64 byte.
Which of the following statements are true?

Possible Answers:

(a) If the table is physically stored in column layout, the reconstruction of a single full tuple consumes ≈0.192 µs using a single CPU core.
(b) If the table is physically stored in row layout, the reconstruction of a single full tuple consumes ≈128 ns using a single CPU core.
(c) If the table is physically stored in column layout, the reconstruction of a single full tuple consumes ≈448 ns using a single CPU core.
(d) If the table is physically stored in row layout, the reconstruction of a single full tuple consumes ≈0.64 µs using a single CPU core.

Correct Answer: (c)

Explanation: To reconstruct a full tuple from row layout, we first need to calculate the count of cache accesses. Considering a size of 114 byte, we will need 2 cache accesses (114 byte/64 byte per cache line = 1.78 → 2) to read a whole tuple from main memory. With two cache accesses, each loading 64 byte, we read 128 byte from main memory. We assume, like in the questions before, that the reading speed of our system is 2 megabyte per millisecond per core. Now we divide 128 byte by 2 megabyte per millisecond per core and get a result of 0.000064 ms (0.064 µs). So the answers ≈0.64 µs and ≈128 ns with one core, if the table is stored in row layout, are both false.
In a columnar layout, we need to read every value individually from main memory, because they are not stored in a row (consecutive memory area) but in different attribute vectors. We have 14 attributes in this example, so we need 14 cache accesses, each reading 64 bytes. Thus the CPU has to read 14 $\cdot$ 64 byte = 896 byte from main memory. Like before, we assume that the reading speed is 2 MB per millisecond per core. By dividing the 896 byte by 2 MB/ms/core, we get the time one CPU core needs to reconstruct a full tuple. The result is 0.000448 ms (448 nanoseconds). So ≈448 ns with one core and columnar layout is the correct answer.

Scan Performance

1. **Loading Dictionary-Encoded Row-Oriented Tuples**
   Consider the example in Sect. 14.2 with dictionary-encoded tuples. In this example, each tuple has a size of 32 byte. What is the time that a single core processor needs to scan the whole world_population table if all data is stored in a dictionary-encoded row layout?

   Possible Answers:

   (a) 128 s
   (b) 256 s
   (c) 64 s
   (d) 96 s

   Correct Answer: (a)

   Explanation: The accessed data volume is 8 billion tuples · 32 byte each ≈256 GB. Therefore the expected response time is 256 GB/(2 MB/ms/core · 1 core) = 128 s

Select

1. **Table Size**
   What is the table size if it has 8 billion tuples and each tuple has a total size of 200 byte?

   Possible Answers:

   (a) ≈ 12.8 TB
   (b) ≈ 12.8 GB
   (c) ≈ 2 TB
   (d) ≈ 1.6 TB

   Correct Answer: (d)

   Explanation: The total size of the table is 8,000,000,000 · 200 byte = 1,600,000,000,000 bytes = 1.6 TB.

2. Optimizing SELECT

How could the performance of SELECT statements be improved?

Possible Answers:

(a) Reduce the number of indices
(b) By using the FAST SELECT keyword
(c) Order multiple sequential select statements from low selectivity to high selectivity
(d) Optimizers try to keep intermediate result sets large for maximum flexibility during query processing

Correct Answer: (c)

Explanation: During the execution of the SELECT statement, we have to search through the whole data of the database for entries with the demanded selection attributes. By ordering the selection criteria from low/strong (many rows are filtered out) to high/weak selectivity (few rows are filtered out), we reduce the amount of data we have to walk through in subsequent steps, which results in a shorter execution time for the overall SELECT statement.

3. Selection Execution Order

Given is a query that selects the names of all German women born after January 1, 1990 from the world_population table (contains data about all people in the world). In which order should the query optimizer execute the selections? Assume a sequential query execution plan.

Possible Answers:

(a) country first, birthday second, gender last
(b) country first, gender second, birthday last
(c) gender first, country second, birthday last
(d) birthday first, gender second, country last

Correct Answer: (a)

Explanation: To optimize the speed of sequential selections the most restrictive ones have to be executed first. While the gender restriction would reduce the amount of data (8 billion) to the half (4 billion), the birthday restriction would return about 1.6 billion tuples and the country restriction (Germany) about 80 million tuples. So the query optimizer should execute the country restriction first, followed by the birthday restriction, which filters additional 80 % of the 80 million Germans. The gender restriction then filters the last $\approx$50 % of entries.

4. **Selectivity Calculation**
   Given is the query to select the names from German men born after January 1, 1990 and before December 31, 2010 from the world population table (8 billion people). Calculate the selectivity.
   Selectivity = number of tuples selected / number of tuples in the table
   Assumptions:

   - there are about 80 million Germans in the table
   - males and females are equally distributed in each country
   - there is an equal distribution between all generations from 1910 until 2010

   Possible Answers:

   (a) 0.001
   (b) 0.005
   (c) 0.1
   (d) 1

   Correct Answer: (a)

   Explanation: Calculation:

   - 80 million Germans
   - 80 million (Germans) $\cdot$ 50 % = 40 million German males
   - 40 million (German males) $\cdot$ 20 % = 8 million German males between 1990 and2010
   - 8 million (German males between 1990 and 2010)/8 billion = 0.001

   The first selection is based on the assumption, that the distribution of males and females is equal. The second selection is based on the assumption that the population is equally distributed over the generations, so selecting a timespan of 20 years of 100 years is effectively 1/5th or 20 %. The selectivity is then calculated via: number of selected tuples/number of all tuples = 8 million/8 billion = 0.001

5. **Execution Plans**
   For any one SELECT statement...

   Possible Answers:

   (a) there always exist exactly two execution plans, which mirror each other
   (b) exactly one execution plan exists
   (c) several execution plans with the same result set, but differing performance may exist
   (d) several executions plans may exist that deliver differing result sets

Correct Answer: (c)

Explanation: For any SELECT statement, several execution plans with the same result set, but different runtimes may exist. As an example, we want to query all men living in Italy from world population table; the database offers three different execution plans. We could query for the gender 'male' first and then for the country 'Italy' in the result set or we start with the selection for 'Italy' and then we narrow the result to only males, or we might perform the two selections on 'male' and 'Italy' in parallel queries, both running on the full dataset and then create the intersection. All three execution plans create the same result set, but require different runtimes. For example its faster to query first for 'Italy'and then for 'male', because in this case first 8 billion entries (all entries) and then further select on the resulting 60 million entries (all Italiens), if you start with 'male' and then query for 'Italy' you have to scan through 8 billion (all Italiens) and then through 4 billion entries (all males).

Materialization Strategies

1. **Which Strategy is Faster?**
   Which materialization strategy—late or early materialization—provides the better performance?

   Possible Answers:

   (a) Early materialization
   (b) Late materialization
   (c) Depends on the characteristics of the executed query
   (d) Late and early materialization always provide the same performance

   Correct Answer: (c)

   Explanation: The question of which materialization strategy to use is dependent on several facts. Amongst them are e.g. the selectivity of queries and the execution strategy (pipelined or parallel). In general, late materialization is superior, if the query has a low selectivity and the queried table uses compression.

2. **Disadvantages of Early Materialization**
   Which of the following statements is true?

   Possible Answers:

   (a) The execution of an early materialized query plan can not be parallelized

(b) Whether late or early materialization is used is determined by the system clock
(c) Early materialization requires lookups into the dictionary, which can be very expensive and are not required when using late materialization
(d) Depending on the persisted value types of a column, using positional information instead of actual values can be advantageous (e.g. in terms of cache usage or SIMD execution)

Correct Answer: (d)

Explanation: Working with intermediate results provides advantages in terms of cache usage and parallel execution, since positional information usually has a smaller size than the actual values. Consequently, more items fit into a cache line, which additionally are of fixed length enabling parallel SIMD operations. The question of locking and parallelization is in general independent from the materialization strategy.

Parallel Data Processing

1. Shared Memory

What limits the use of shared memory?

Possible Answers:

(a) The number of workers, which share the same resources and the limited memory itself.
(b) The caches of each CPU
(c) The operation frequency of the processor
(d) The usage of SSE instructions.

Correct Answer: (a)

Explanation: By default, main memory is assigned to exactly one process, all other process can not access the reserved memory area. If a memory segment is shared between processes and this segment is full, additional shared memory has to be requested. The size of the memory area is therefore a limiting factor. But more important is the fact that several workers (this can be threads, processes, etc.) share the same resource. To avoid inconsistencies, measures have to be taken, e.g. locking or MVCC (multiversion concurrency control). While this does usually not affect low numbers of workers, any locking will automatically get a problem as soon as the number of workers reaches a certain limit. At this level, too many workers cannot work since they have to wait for locked resources.

Indices

1. Index Characteristics
Introducing an index...

Possible Answers:

(a) decreases memory consumption
(b) increases memory consumption
(c) speeds up inserts
(d) slows down look-ups

Correct Answer: (b)

Explanation: The index is an additional data structure and therefore consumes memory. Its purpose is to increase performance of scan operations.

2. Inverted Index
What is an inverted index?

Possible Answers:

(a) A structure that contains the distinct values of the dictionary in reverse order
(b) A list of text entries that have to be decrypted, it is used for enhanced security
(c) A structure that contains the delta of each entry in comparison to the largest value
(d) A structure that maps each distinct value to a position list, which contains all positions where the value can be found in the column

Correct Answer: (d)

Explanation: The inverted index consists of the index offset vector and the index position vector. The index offset vector stores a reference to the sequence of positions of each dictionary entry in the index position vector. The index position vector contains the positions (i.e. all occurances) for each value in the attribute vector. Thus, the inverted index is a structure that maps each distinct value to a position list, which contains all positions where the value can be found in the column.

Join

1. **Hash-Join Complexity**
   What is the complexity of the Hash-Join?

   Possible Answers:

   (a) O(n+m)
   (b) O(n^2/m^2)
   (c) O(n · m)
   (d) O(n · log(n)+m+log(m))

   Correct Answer: (a)

   Explanation: Let m and n be the cardinality of the input relations M and N with m < = n. The hash function used in the first phase of a Hash-Join, maps a variable-length value to a fixed length value in constant time and is applied to the smaller input relation. Thus, it takes m operations. In the second phase, the attribute vector of the larger relation is probed against the hash table which was generated in the first step. Again, the hash function is used and therefore n constant time operations take place. In short, the Hash-Join has a complexity of O(m+n).

2. **Sort-Merge Join Complexity**
   What is the complexity of the Sort-Merge Join?

   Possible Answers:

   (a) O(n+m)
   (b) O(n^2/m^2)
   (c) O(n · m)
   (d) O(n · log(n)+m · log(m))

   Correct Answer: (d)

   Explanation: Let m and n be the cardinality of the input relations M and N with m < = n. The runtime of the Sort-Merge Join is determined by the task to sort both input relations. As sorting algorithm, merge sort is used, that has runtime a complexity of O(n · log(n)) for input relation N. This complexity is based on the fact, that the merge join works recursive and divides the input into two parts, which are sorted and afterwards combined. The actual number of steps is determined by the number of recursion levels. The resulting equation for n elements can be assessed with the term n · log(n) according to the master theorem. Therefore, the Sort-Merge Join has a complexity of O(m · log(m)+n · log(n)).

3. Join Algorithm Small Data Set
Given is an extremely small data set. Which join algorithm would you choose in order to get the best performance?

Possible Answers:

(a) All join algorithms have the same performance
(b) Nested-Loop Join
(c) Sort-Merge Join
(d) Hash-Join

Correct Answer: (b)

Explanation: Even though the Nested-Loop Join has a much worse complexity than the other join algorithms, it manages to join the input relations without additional data structures. Therefore, no initialization is needed. In case of very small relations, this is a huge saving.

4. Join Algorithm Large Data Set
Imagine a large data set with an index. Which join algorithm would you choose in order to get the best performance?

Possible Answers:

(a) Nested-Loop Join
(b) Sort-Merge Join
(c) All join algorithms have the same performance
(d) Hash-Join

Correct Answer: (d)

Explanation: The Nested-Loop Join is not suitable for a large data set because its complexity is much worse than the complexity of the two other algorithms. The Sort-Merge Join has a worse runtime complexity than the Hash-Join because it requires sorting before the actual merge can take place. However, it does not require to build an additional hash structure, which can be complicated and time consuming depending on the circumstances. In this case the Hash-Join can use the existing index to speed up the building of the hash map, so the only possible obstacle is not of concern here. Hence, the Hash-Join is the most suitable algorithm for large data sets with an index to maximize the performance.

5. Equi-Join
What is the Equi-Join?

Possible Answers:

(a) If you select tuples from both relations, you use only one half of the join relations and the other half of the table is discarded
(b) If you select tuples from both relations, you will always select those tuples, that qualify according to a given equality predicate
(c) It is a join algorithm that ensures that the result consists of equal amounts from both joined relations
(d) It is a join algorithm to fetch information, that is probably not there. So if you select a tuple from one relation and this tuple has no matching tuple on the other relation, you would insert their NULL values there.

Correct Answer: (b)

Explanation: There are two general categories of joins: inner joins and outer joins. Inner joins create a result table that combines tuples from both input tables only if both regarded tuples meet the specified join condition. Based on a join predicate each tuple from the first table is combined with each tuple of the second table, the resulting cross-set is filtered (similar to a SELECT statement) on the join condition. Outer joins, in contrast, have more relaxed conditions on which tuples to include. If a tuple has no matching tuple in the other relation, the outer join inserts NULL values for the missing attributes in the result and includes the resulting combination. Further specializations of the two join types are for example the Semi-Join, which returns only the attributes of the left join partner if the join predicate is matched. Another specialization is the Equi-Join. It allows the retrieval of tuples that satisfy a given equality predicate for both sides of the tables to be joined. The combined result comprises the equal attribute twice, once from the left relation and once from the right relation. The so called Natural-Join is similar to the Equi-Join, except that it strips away the redundant column that is kept in the Equi-Join.

6. One-to-One-Relation
What is a one-to-one relation?

Possible Answers:

(a) A one-to-one relation between two objects means that for each object on the left side, there are one or more objects on the right side of the joined table and each object of the right side has exactly one join partner on the left
(b) A one-to-one relation between two objects means that for exactly one object on the left side of the join exists exactly one object on the right side and vice versa

(c) A one-to-one relation between two objects means that each object on the left side is joined to one or more objects on the right side of the table and vice versa each object on the right side has one or more join partners on the left side of the table
(d) Each query which has exactly one join between exactly two tables is called a one-to-one relation, because one table is joined to exactly one other table.

Correct Answer: (b)

Explanation: Three different types of relations between two tables exist. These are one-to-one, one-to-many and many-to-many relations. In a many-to-many relationship, each tuple from the first relation may be related to multiple tuples from the second one and vice versa. In a one-to-many relationship, each tuple of the first table might be related to multiple tuples of the second table, but each tuple of the second relation refers to exactly one tuple of the first relation. Consequently a one-to-one relation connects each tuple of the first relation with exactly one tuple from the second relation, additionally no tuple of the second relation stays unconnected, so the amount of entries in both relations is equal.

Aggregate Functions

1. **Aggregate Function Definition**
What are aggregate functions?

Possible Answers:

(a) A set of functions that transform data types from one to another data
(b) A set of indexes that speed up processing a specific report
(c) A set of tuples that are grouped together according to specific requirements
(d) A specific set of functions that summarize multiple rows from an input data set

Correct Answer: (d)

Explanation: Sometimes it is necessary to get a summary of a data set, like the average, the minimum, or the number of entries. Databases provide special functions for these tasks, called aggregate functions, that take multiple rows as an input and create an aggregated output. Instead of processing single values, these functions work on data sets, which are created by grouping the input data on specified grouping attributes.

2. Aggregate Functions
Which of the following is an aggregate function?

Possible Answers:

(a) HAVING
(b) MINIMUM
(c) SORT
(d) GROUP BY

Correct Answer: (b)

Explanation: MINIMUM (often expressed as MIN) is the only aggregate function listed here. HAVING is used in SQL Queries to add additional requirements for the resulting aggregate values in order to be accepted. GROUP BY is used to specify the columns on which the aggregations should take place (all tuples with equal values in this columns are grouped together, if multiple columns are specified, one result per unique attribute combination is computed). SORT is not a valid SQL expression at all.

Parallel Select

1. Amdahl's Law
Amdahl's Law states that ...

Possible Answers:

(a) the number of CPUs doubles every year
(b) the level of parallelization can be no higher than the number of available CPUs
(c) the speedup of parallelization is limited by the time needed for the sequential fractions of the program
(d) the amount of available memory doubles every year

Correct Answer: (c)

Explanation: While the execution time of the parallelizable code segments can be shortened by multiple cores, the runtime of the sequential fraction can not be decreased because it has to be executed by one CPU core only. As this time is constant, the execution time of the complete program code is at least as long as the sequential code segment, regardless how many cores are used. This main principle was first described by Amdahl and named after him. The increase of chip density by a factor of 2 about every 18–24 month is called "Moore's Law", the other two possible answers are incorrect at all.

2. Query Execution Plans in Parallelizing SELECTS
When a SELECT statement is executed in parallel ...

Possible Answers:

(a) all other SELECT statements are paused
(b) its query execution plan becomes much simpler compared to sequential execution
(c) its query execution plan is adapted accordingly
(d) its query execution plan is not changed at all

Correct Answer: (c)

Explanation: If columns are split into chunks, attribute vector scans can be executed in parallel with one thread per chunk. All results for chunks of the same column have to be combined by a UNION operation afterwards. If more than one column is scanned, a positional AND operation has to be performed additionally. In order to execute the AND operation in parallel too, the affected columns have to be partitioned equally. So parallelizing the AND operation changes the execution plan, therefore the answer that the execution plan has to be adapted, is correct.

Workload Management and Scheduling

1. Resource Conflicts
Which three hardware resources are usually taken into account by the scheduler in a distributed in-memory database setup?

Possible Answers:

(a) CPU processing power, main memory, network bandwidth
(b) Main memory, disk, tape drive
(c) CPU processing power, graphics card, monitor
(d) Network bandwidth, power supply unit, main memory

Correct Answer: (a)

Explanation: When scheduling queries in an in-memory database, storage outside of main memory is of no importance. Furthermore, graphics hardware and peripherals are no performance indicator for database systems.

2. **Workload Management Scheduling Strategy**
   Why does a complex workload scheduling strategy might have disadvantages in comparison to a simple resource allocation based on heuristics or a uniform distribution, e.g. Round Robin?

   Possible Answers:

   (a) The execution of a scheduling strategy itself consumes more resources than a simplistic scheduling approach. A strategy is usually optimized for a certain workload—if this workload changes abruptly, the scheduling strategy might perform worse than a uniform distribution
   (b) Heuristics are always better than complex scheduling strategies
   (c) A scheduling strategy is based on general workloads and thus might not reach the best performance for specific workloads compared to heuristics or a uniform distribution, while its application is cheap
   (d) Round-Robin is usually the best scheduling strategy.

   Correct Answer: (a)

   Explanation: Scheduling strategies reach a point where every further optimization is based on a specific workload. If one can predict future workloads based on the past ones, the scheduler can distribute queries for maximum performance regarding this scenario. However, if the workload changes unpredictably, there is no gain from specialized strategies. Under these circumstances, specialized optimizations are rather an unnecessary overhead, as they may require additional scheduling time, data structures, or increased resources

3. **Analytical Queries in Workload Management**
   Analytical queries typically are ...

   Possible Answers:

   (a) long running with soft time constraints
   (b) short running with soft time constraints
   (c) short running with strict time constraints
   (d) long running with strict time constraints

   Correct Answer: (a)

   Explanation: Analytical workloads consist of complex and computationally heavy queries. Hence analytical queries have a long execution time. Whereas the response time must be guaranteed for transactional queries, this is not the case for analytical ones. While they should be as short as possible, business processes will not abort if an analytical query takes 3 instead of 2 s. Therefore, analytical queries have soft time constraints in comparison to transactional ones.

4. Query Response Times
Query response times ...

Possible Answers:

(a) can be increased so the user can do as many tasks as possible in parallel because context switches are cheap
(b) have to be as short as possible, so the user stays focused at the task at hand
(c) should never be decreased as users are unfamiliar with such system behavior and can become frustrated
(d) have no impact on a users work behavior

Correct Answer: (b)

Explanation: Query response times have a huge impact on the user. If an operation takes too long, the mind tends to wander to other topics than the original task. The longer it takes, the further this process goes on. Refocusing on the task is in fact very exhausting and thus, waiting periods should be avoided to guaranty a convenient experience and reduce human errors.

Parallel Join

1. Parallelizing Hash-Join Phases
What is the disadvantage when the probing phase of a join algorithm is parallelized and the hashing phase is performed sequentially?

Possible Answers:

(a) Sequentially performing the hashing phase introduces inconsistencies in the produced hash values
(b) The algorithm still has a large sequential part that limits its potential to scale
(c) The sequential hashing phase will run slower due to the large resource utilization of the parallel probing phase
(d) The table has to be split into smaller parts, so that every core, which performs the probing, can finish

Correct Answer: (b)

Explanation: With Amdahls' Law in mind, the Hash-Join can only be as fast as the sum of all sequential parts of the algorithm. Parallelizing the probing phase shortens the time needed, but has no impact on the hashing phase, which still has to be done sequentially. Thus, there is always a huge part of the algorithm which cannot be parallelized.

Parallel Aggregation

1. **Aggregation—GROUP BY**
   Assume a query that returns the number of citizens of a country, e.g.:
   SELECT country, COUNT(·)
   FROM world_population
   GROUP BY country;
   The world_population table contains the names and countries of all citizens of the world.
   The GROUP BY clause is used to express ...

   Possible Answers:

   (a) the graphical format of the results for display
   (b) an additional filter criteria based on an aggregate function
   (c) that the aggregate function shall be computed for every distinct value of country
   (d) the sort order of countries in the result set

   Correct Answer: (c)

   Explanation: In general, the GROUP BY clause is used to express that all used aggregate function shall be computed for every distinct value (or value combinations) of the specified attributes. In this case, only one attribute is specified, so only sets having distinct values for country are aggregated. The sort order is specified in the ORDER BY clause, additional filter criteria can be added on the aggregated values with the HAVING clause.

2. **Number of Threads**
   How many threads will be used during the second phase of the described parallel aggregation algorithm when the table is split into 20 chunks and the GROUP BY attribute has 6 distinct values?

   Possible Answers:

   (a) exactly 20 threads
   (b) at most 6 threads
   (c) at least 10 threads
   (d) at most 20 threads

   Correct Answer: (b)

   Explanation: In the aggregation phase, the so called merger threads merge the buffered hash tables. Each thread is responsible for a certain range of the GROUP BY attribute. So if the GROUP BY attribute has 6 distinct values, the maximum number of threads is 6. If there are more than 6 threads, the surplus threads will get no pending values and are therefore not used.

Differential Buffer

1. **Differential Buffer**
   What is the differential buffer?

   Possible Answers:

   (a) A buffer where exception and error messages are stored
   (b) A buffer where different results for one and the same query are stored for later usage
   (c) A dedicated storage area in the database where inserts, updates and deletions are buffered
   (d) A buffer where queries are buffered until there is an idle CPU that takes one new task over

   Correct Answer: (c)

   Explanation: The main store is optimized for read operations. An insert of a tuple is likely to force restructuring of the whole table. To avoid this, we introduced the differential buffer, an additional storage area where all the data modifications like inserts, updates and delete operations are performed and buffered until they are integrated into the main store. As a result, we have a read optimized main store and a write optimized differential buffer. In combination, update operations as well as read operations are supported by optimal storage structures, which results in an increased overall performance.

2. **Performance of the Differential Buffer**
   Why might the performance of read queries decrease, if a differential buffer is used?

   Possible Answers:

   (a) Because only one query at a time can be answered by using the differential buffer
   (b) Because read queries have to go against the main store and the differential buffer, which is write-optimized
   (c) Because inserts collected in the differential buffer have to be merged into the main store every time a read query comes in
   (d) Because the CPU cannot perform the query before the differential buffer is full

   Correct Answer: (b)

   Explanation: New tuples are inserted into the differential buffer first, before being merged into the main store eventually. To speed up inserts as much as possible, the differential buffer is optimized for writing rather than reading tuples. Read queries against all data have to go against the differential buffer as well, what might cause a slowdown. To prevent noticeable effects for the user, the amount of values in the differential buffer is kept small in comparison to the main store. By exploiting the fact that the main store and the differential buffer can be scanned in parallel, noticeable speed losses are avoided.

3. Querying the Differential Buffer
If we use a differential buffer, we have the problem that several tuples belonging to one real world entry might be present in the main store as well as in the differential buffer. How did we solve this problem?

Possible Answers:

(a) This statement is completely wrong because multiple tuples for one real world entry must never exist
(b) All attributes of every doubled occurrence are set to NULL in the compressed main store
(c) We introduced a validity bit
(d) We use a specialized garbage collector that just keeps the most recent entry

Correct Answer: (c)

Explanation: Following the insert-only approach, we do not delete or change existing attributes or whole tuples. If we want to change or add attributes of an existing tuple in spite of the insert-only approach, we add an updated tuple with the changed as well as the unchanged values to the differential buffer. In order to solve the problem of multiple tuples belonging to one real world entry, we introduced a validity vector that indicates whether a tuple is the most current one and therefore is valid or not.

Insert Only

1. Statements Concerning Insert-Only
Considering an insert-only approach, which of the following statements is true?

Possible Answers:

(a) When given a differential buffer, historical data can be used to further speed up the insert performance
(b) Old data items are deleted as they are not necessary any longer
(c) Historical data has to be stored in a separate database to reduce the overall database size
(d) Data is not deleted, but invalidated instead

Correct Answer: (d)

Explanation: With an insert-only approach, no data is ever deleted or sourced out to another, separated database. Furthermore, when using a differential buffer, the insert performance is not dependent on the data already stored in the database. Without the differential buffer, a huge amount of data already in the table might indeed speed up inserts into sorted columns, because a quite saturated and therefore stable dictionary would reduce the resorting overhead.

2. **Benefits of Historic Data**
   Which of the following is NOT a reason why historical data is kept by an enterprise?

   Possible Answers:

   (a) Historic data can be used to analyze the development of the company
   (b) It is legally required in many countries to store historical data
   (c) Historical data can provide snapshots of the database at certain points in time
   (d) Historical data can be analyzed to boost query performance

   Correct Answer: (d)

   Explanation: With historical data, time-travel queries are possible. They allow users to see the data exactly like it was at any point in the past. This simple access to historical data helps a company's management to efficiently analyze the history and the development of the enterprise. Additionally, i.e. in Germany it is legally required to store particular commercial documents for tax audits. Historical data however will not improve the query performance, which is the correct answer consequently.

3. **Accesses for Point Representation**
   Considering point representation and a table with one tuple, that was invalidated five times, how many tuples have to be checked to find the most recent tuple?

   Possible Answers:

   (a) Five
   (b) Two, the most recent one and the one before that
   (c) Only one, that is, the first which was inserted
   (d) Six

   Correct Answer: (d)

Explanation: All tuples belonging to one real world entry have to be checked in order to determine the most recent tuple when using point representation. It is not sufficient to start from the end of the table and take the first entry that belongs to the desired real world entry, because in most cases the table is sorted by an attribute and not by the insertion order. In this case, six tuples have to be checked, all five invalidated ones and the current one.

4. **Physical Delete instead of Insert-Only**
   What would be necessary if physical deletion of tuples was implemented in SanssouciDB?

   Possible Answers:

   (a) Dictionary cleaning which would cause rewriting of the attribute vector
   (b) The latest snapshot has to be reloaded after a deletion to maintain data integrity
   (c) Deletion of tuples is part of SanssouciDB
   (d) Assurance of compatibility to other DBMS

   Correct Answer: (a)

   Explanation: Physical deletions would, in some cases, cause the need to clean the dictionary, because values that do not exist in the table any longer have to be eliminated.

5. **Statement concerning Insert-Only**
   Which of the following statements concerning insert-only is true?

   Possible Answers:

   (a) Point representation allows faster read operations than interval representation due to its lower impact on tuple size
   (b) In interval representation, four operations have to be executed to invalidate a tuple
   (c) Interval representation allows more efficient write operations than point representation
   (d) Point representation allows more efficient write operations than interval representation

   Correct Answer: (d)

   Explanation: Point representation will be less efficient for read operations, that only require the most recent tuple. Using point representation, all tuples of that entry have to be checked, to determine the most recent one. As a positive aspect, point representation allows more efficient write operations in comparison to interval representation, because on any update, only the tuple with the new values and the current ‘value from’ date has to be entered, the other tuples do not need to be changed. The insertion of the tuple with the new entries might however require the lookup of the former most recent tuple, to retrieve all unchanged values.

Merge

1. **What is the Merge?**
The merge process ...

Possible Answers:

(a) incorporates the data of the write-optimized differential buffer into the read-optimized main store
(b) combines the main store and the differential buffer to increase the parallelism
(c) merges the columns of a table into a row-oriented format
(d) optimizes the write-performance

Correct Answer: (a)

Explanation: If we are using a differential buffer as an additional data structure to improve the write performance of our database, we have to integrate this data into the compressed main partition periodically in order to uphold the benefits concerning the read performance. This process is called “merge”.

2. **When to Merge?**
When is the merge process triggered?

Possible Answers:

(a) When the number of tuples within the differential buffer exceeds a specified threshold
(b) When the space on disk runs low and the main store needs to be further compressed
(c) Before each SELECT operation
(d) After each INSERT operation

Correct Answer: (a)

Explanation: Holding too many tuples in the differential buffer, slows down read performance against all data. Therefore it is necessary to define a certain threshold at which the merge process is triggered. If it was too often (for example after every INSERT or even before every SELECT), the overhead of the merge would be larger than the possible penalty for querying both main store and differential buffer.

Logging

1. **Snapshot Statements**
   Which statement about snapshots is wrong?

   Possible Answers:

   (a) The recovery process is faster when using a snapshot because only log files after the snapshot need to be replayed
   (b) The snapshot contains the current read-optimized store
   (c) A snapshot is an exact image of a consistent state of the database to a given time
   (d) A snapshot is ideally taken after each insert statement

   Correct Answer: (d)

   Explanation: A snapshot is a direct copy of the main store. Because of that, all the data of the database has to be in the main store when a snapshot is created in order to represent the complete dataset. This is only the case after the merge process, not after each insert, because inserts are written into the differential buffer and otherwise would be omitted.

2. **Recovery Characteristics**
   Which of the following choices is a desirable characteristic of any recovery mechanism?

   Possible Answers:

   (a) Recovery of only the latest data
   (b) Returning the results in the right sorting order
   (c) Maximal utilization of system resources
   (d) Fast recovery without any data loss

   Correct Answer: (d)

   Explanation: It is of course preferable to recover all of the data in as few as possible time. A high utilization of resources might be a side effect, but is not a must.

3. **Situations for Dictionary-Encoded Logging**
   When is dictionary-encoded logging superior?

   Possible Answers:

   (a) If large values are inserted only one time

(b) If the number of distinct values is high
(c) If all values are different
(d) If large values are inserted multiple times

Correct Answer: (d)

Explanation: Dictionary Encoding replaces large values by a minimal number of bits to represent these values. As an additional structure, it requires some space, too. In return, huge savings are possible if values appear more than once, because the large values is only saved once and then referenced by the smaller key value whenever needed. If the number of distinct values is high or even maximal (when all values are different, which is also given if each value is inserted only once), the possible improvements can not be fully leveraged.

4. Small Log Size
Which logging method results in the smallest log size?

Possible Answers:

(a) Common logging
(a) Log sizes never differ
(a) Dictionary-encoded logging
(a) Logical logging

Correct Answer: (c)

Explanation: Common logging is wrong because it does not even exist. Logical logging writes all data uncompressed to disk. In contrast, dictionary-encoded logging saves the data using dictionaries, so the size of the data is implicitly smaller because of the compression of recurring values. Consequently, of the mentioned logging variants, dictionary-encoded logging has the smallest log size.

5. Dictionary-Encoded Log Size
Why has dictionary-encoded logging the smaller log size in comparison to logical logging?

Possible Answers:

(a) Because of interpolation
(b) Because it stores only the differences of predicted values and real values
(c) Because of the reduction of recurring values
(d) Actual log sizes are equal, the smaller size is only a conversion error when calculating the log sizes

Correct Answer: (c)

Explanation: Dictionary Encoding is a compression technique that encodes variable length values by smaller fixed-length encoded values using a mapping dictionary. As a consequences, it reduces the size of recurring values.

Recovery

1. Recovery
What is recovery?

Possible Answers:

(a) It is the process of recording all data during the run time of a system
(b) It is the process of restoring a server to the last consistent state before its crash
(c) It is the process of improving the physical layout of database tables to speed up queries
(d) It is the process of cleaning up main memory, that is "recovering" space

Correct Answer: (b)

Explanation: In case of a failure of the database server, the system has to be restarted and set to a consistent state. This is be done by loading the backup data stored on persistent storage back into the in-memory database. The overall process is called 'recovery'.

2. Server Failure
What happens in the situation of a server failure?

Possible Answers:

(a) The system has to be rebooted and restored if possible, while another server takes over the workload
(b) The power supply is switched to backup power supply so the data within the main memory of the server is not lost
(c) The failure of a server has no impact whatsoever on the workload
(d) All data is saved to persistent storage in the last moment before the server shuts down

Correct Answer: (a)

Explanation: A backup power supply is only a solution if there is a power outage but if a CPU or a mainboard causes an error, a backup power supply is useless. Saving data to persistent storage in the last moment before the server shuts down is not always possible, for example if there is a power outage and no backup power supply or if an electric component causes a short circuit and the server shuts down immediately to prevent further damage, there is no time to write the large amount of data to the slow disk. If a server has to shut down, incoming queries should be accepted and executed nonetheless. Therefore, the right answer is that the server has to be rebooted if possible and the data is restored as fast as possible. In the meantime the workload has to be distributed to other servers, if available.

On-the-Fly Database Reorganization

1. Separation of Hot and Cold Data
How should the data separation into hot and cold take place?

Possible Answers:

(a) Randomly, to ensure efficient utilization of storage areas
(b) Round-robin, to ensure uniform distribution of data among hot and cold stores
(c) Manually, upon the end of the life cycle of an object
(d) Automatically, depending on the state of the object in its life cycle

Correct Answer: (d)

Explanation: In enterprise applications, business objects have their own live cycles, which can be separated into active (hot) and passive (cold) states. The states are defined by a sequence of passed events. If a business object will not be changed any longer and is accessed less often, it becomes passive and can be moved to the storage area for cold data, which might be a slower and therefore cheaper hardware location than main memory.

2. Data Reorganization in Row Stores
The addition of a new attribute within a table that is stored in row-oriented format ...

Possible Answers:

(a) is not possible
(b) is an expensive operation as the complete table has to be reconstructed to make place for the additional attribute in each row
(c) is possible on-the-fly, without any restrictions of queries running concurrently that use the table
(d) is very cheap, as only meta data has to be adapted

Correct Answer: (b)

Explanation: In row-oriented tables all attributes of a tuple are stored consecutively. If an additional attribute is added, the storage for the entire table has to be reorganized, because each row has to be extended by the amount of space the newly added attribute requires. All following rows have to be moved to memory areas behind. Of course, the movement of tuples backwards can be parallelized if the size of the added attribute is known and constant, nonetheless is the piecewise relocation of the complete table relatively expensive.

3. **Cold Data**
What is cold data?

Possible Answers:

(a) Data, which is not modified any longer and that is accessed less frequently
(b) The rest of the data within the database, which does not belong to the result of the current query
(c) Data that is used in a majority of queries
(d) Data, which is still accessed frequently and on which updates are still expected

Correct Answer: (a)

Explanation: To reduce the amount of main memory needed to store the entire data set of an enterprise application, the data is separated into active (hot) and passive (cold) data. Active data is the data of business processes that are not yet completed and therefore stored in main memory for fast access. Passive data in contrast is data of business processes that are closed or completed and will not be changed any more and thus is moved to slower storage mediums like SSDs.

4. **Data Reorganization**
The addition of an attribute in the column store ...

Possible Answers:

(a) slows down the response time of applications that only request the attributes they need from the database
(b) speeds up the response time of applications that always request all possible attributes from the database
(c) has no impact on existing applications if they only request the attributes they need from the database
(d) has no impact on applications that always request all possible attributes from the table

Correct Answer: (c)

Explanation: In column-oriented tables each column is stored independently from the other columns in a separate block. Because a new attribute requires a new memory block, there is no impact on the existing columns and their layout in memory. Hence, there is also no impact on existing applications which access only the needed attributes that existed before.

5. **Single-Tenancy**
In a single-tenant system ...

Possible Answers:

(a) all customers are placed on one single shared server and they also share one single database instance
(b) each tenant has its own database instance on a shared server
(c) power consumption per customer is best and therefore it should be favored
(d) each tenant has its own database instance on a physically separated server

Correct Answer: (d)

Explanation: If each customer has its own database on a shared server, this is the 'shared machine' implementation of a multi-tenant system. If all customers share the same database on the same server, the implementation is called 'shared database'. A 'single-tenant' system provides each user an own database on a physically separated server. The power consumption per customer is not optimal in that implementation of multi-tenancy, because it is not possible to share hardware resources. Shared hardware resources allow to run several customers on one machine to utilize the resources in an optimal way and then shut down systems that are not required momentarily.

6. **Shared Machine**
In the shared machine implementation of multi-tenancy ...

Possible Answers:

(a) each tenant has an own exclusive machine, but these share their resources (CPU, RAM) and their data via a network
(b) all tenants share one server machine, but have own database processes
(c) each tenant has an own exclusive machine, but these share their resources (CPU, RAM) but not their data via a network
(d) all tenants share the same physical machine, but the CPU cores are exclusively assigned to the tenants

Correct Answer: (b)

Explanation: Please have a look at Sect. 30.3 on page 229.

7. **Shared Database Instance**
   In the shared database instance implementation of multi-tenancy ...

   Possible Answers:

   (a) the risk of failures is minimized because more technical staff (from different tenants) will have a look at the shared database
   (b) all tenants share one server machine and one main database process, tables are also shared
   (c) each tenant has its own server, but the database instance is shared between the tenants via an InfiniBand network
   (d) all tenants share one server machine and one main database process, tables are tenant exclusive, access control is managed within the database

   Correct Answer: (d)

   Explanation: Please have a look at Sect. 30.3 on page 229.

Implications

1. **Architecture of a Banking Solution**
   Current financials solutions contain base tables, change history, materialized aggregates, reporting cubes, indices, and materialized views. The target financials solutions contains ...

   Possible Answers:

   (a) only base tables, reporting cubes, and the change history
   (b) only base tables, algorithms, and some indexes
   (c) only base tables, materialized aggregates, and materialized views
   (d) only indexes, change history, and materialized aggregates

   Correct Answer: (b)

   Explanation: Because in-memory databases are considerably faster than their disk-focused counterparts, all views, aggregates and cubes can be computed on-the-fly. Also, the change history is dispensable when using the insert-only approach. Because main memory capacity is relatively expensive when compared to disk capacity, it is more important than ever to discard unneeded and redundant data. The base tables and the algorithms are still necessary, because they are the essential atomic parts of the database and indexes improve the performance while requiring only small amounts of memory.

2. **Criterion for Dunning**
What is the criterion to send out dunning letters?

Possible Answers:

(a) Bad stock-market price of the own company
(b) Bad information about the customer is received from consumer reporting agencies
(c) When the responsible accounting clerk has to achieve his rate of dunning letters
(d) A customer payment is overdue

Correct Answer: (d)

Explanation: A dunning letter is send to remind a customer to pay his outstanding invoices. If a customer doesn't pay within the term of payment, he is overdue. Then a company has to send a dunning letter to call on the customer to pay, before it could demand an interest rate from him.

3. **In-Memory Database for Financials**
Why is it beneficial to use in-memory databases for financials systems?

Possible Answers:

(a) Financial systems are usually running on mainframes. No speed up is needed. All long-running operations are conducted as batch jobs
(b) Operations like dunning can be performed in much shorter time
(c) Because of the high reliability of data in main memory, less maintenance work is necessary and labor costs could be reduced
(d) Easier algorithms are used within the applications, so shorter algorithm run time leads to more work for the end user. Business efficiency is improved

Correct Answer: (b)

Explanation: Operations like dunning are very time-consuming tasks, because they involve read operations on large amounts of transactional data. Column oriented in-memory databases reduce the duration of the dunning run because of their extremely high read performance.

4. **Connection between Object Fields and Columns**
Assume that "overdue" is expressed in an enterprise system business object by four fields. How many columns play a role to store that information?

Possible Answers:

(a) all columns of the table
(b) two columns

(c) four columns
(d) one column

Correct Answer: (c)

Explanation: Each field of a business object is an independent attribute and consequently stored in a separate column. Because each field is possibly important to store the information, all four columns play a role.

5. **Languages for Stored Procedures**
Languages for stored procedures are ...

Possible Answers:

(a) designed primarily to be human readable. They follow the spoken english grammar as close as possible
(b) strongly imperative, the database is forced to exactly fulfill the orders expressed via the procedure
(c) usually a mixture of declarative and imperative concepts
(d) strongly declarative, they just describe how the result set should look like. All aggregations and join predicates are automatically retrieved from the database, which has the information "stored" for that

Correct Answer: (c)

Explanation: Languages for stored procedures typically support declarative database queries expressed in SQL, imperative control sequences like loops and conditions and concepts like variables and parameters. Hence, languages for stored procedures usually support a mixture of declarative and imperative concepts and combine the best parts of the two programming paradigms to be very efficient.

Views

1. **View Locations**
Where should a logical view be built to get the best performance?

Possible Answers:

(a) in the GPU
(b) in a third system
(c) close to the data in the database
(d) close to the user in the analytical application

Correct Answer: (c)

Explanation: One chosen principle for SanssouciDB is that any data intensive operation should be executed in the database to speed it up. In consequence, views, which focus on a certain aspect of the data and often do some calculations to enrich the information with regard to the desired focus, should be placed close to the data. It this case, that means they should be located directly in the database.

2. **Data Representation**
   What is the traditional representation of business data to the end user?

   Possible Answers:

   (a) bits
   (b) videos
   (c) music
   (d) lists and tables

   Correct Answer: (d)

   Explanation: Bits are the internal representation of all data, but unconverted they are not suitable to be read for people. Humans are used to work rich media like sounds, videos and pictures. Business data can not be delivered via sounds, videos or pictures in a condensed form, so the best approximation and therefore optimal representation is to deliver the data in structured text formats, which are lists and tables. If available, these lists and tables should further be displayed in aggregated ways i.e. charts to help users to grasp the proportions reflected by the data.

3. **Views and Software Quality**
   Which aspects concerning software quality are improved by the introduction of database views?

   Possible Answers:

   (a) Accessibility and availability
   (b) Testability and security
   (c) Reliability and usability
   (d) Reusability and maintainability

   Correct Answer: (d)

   Explanation: The introduction of database views allows a decoupling of the application code from the actual data schema. This improves the reusability and maintainability, because changes on the application code are possible without requiring changes to the data schema and vice versa. Additionally, existing application code can be used for many different data schemes, if the required schemas can be mapped via views to the data schema used by the application code. Availability, reliability and testability are not improved by using views.

Handling Business Objects

1. **Business Object Mapping**
   What is business object mapping?

   Possible Answers:

   (a) Putting together a diagram of all used business objects. It is similar to a sitemap on webpages
   (b) Allocate an index to every business object and save it in the associated memory area
   (c) Representing every element of a business object in a table
   (d) Create a hash code of the business object and save this hash code instead of the whole object

   Correct Answer: (c)

   Explanation: A business object is an entity capable of storing information and state. It typically has a tree like structure with leaves holding information about the object or connections to other business objects. So if business object are stored in a database, every element of it has to be represented in a table. This is called business object mapping.

ByPass Solution

1. **Transition to IMDBs**
   What does the transition to in-memory database technology mean for enterprise applications?

   Possible Answers:

   (a) Data organization and processing will change radically and enterprise applications need to be adapted
   (b) The data organization will not change at all, but the source code of the applications has to be adapted
   (c) There will be no impact on enterprise applications
   (d) All enterprise applications are significantly sped up without incurring any adaptions

Correct Answer: (a)

Explanation: Traditional database systems separated the operational and the analytical part. With in-memory databases this is not necessary anymore, because they are fast enough to combine both parts. Furthermore, SanssouciDB stores the data in column-based format, while most traditional databases use a row-oriented format. Query speeds in enterprise applications will receive some improvements without any changes to the program code due to the fact that all data is in main memory and aggregations can be computed faster using column orientation. To fully leverage the potentials of the presented concepts, adaptions of the existing applications will be necessary so that the effects that the two mentioned major changes, the reunion of OLTP and OLAP and the column orientation, can be exploited in the program.

Glossary

ACID	Property of a database management system to always ensure atomicity, consistency, isolation, and durability of its transactions.
Active Data	Data of a business transaction that is not yet completed and is therefore always kept in main memory to ensure low latency access.
Aggregation	Operation on data that creates a summarized result, for example, a sum, maximum, average, and so on. Aggregation operations are common in enterprise applications.
Analytical Processing	Method to enable or support business decisions by giving fast and intuitive access to large amounts of enterprise data.
Application Programming Interface (API)	An interface for application programmers to access the functionality of a software system.
Atomicity	Database concept that demands that all actions of a transaction are executed or none of them.
Attribute	A characteristic of an entity describing a certain detail of it.
Availability	Characteristic of a system to continuously operate according to its specification, measured by the ratio between the accumulated time of correct operation and the overall interval.

Available-to-Promise (ATP)	Determining whether sufficient quantities of a requested product will be available in current and planned inventory levels at a required date in order to allow decision making about accepting orders for this product.
Batch Processing	Method of carrying out a larger number of operations without manual intervention.
Benchmark	A set of operations run on specified data in order to evaluate the performance of a system.
Blade	Server in a modular design to increase the density of available computing power.
Business Intelligence	Methods and processes using enterprise data for analytical and planning purposes, or to create reports required by management.
Business Logic	Representation of the actual business tasks of the problem domain in a software system.
Business Object	Representation of a real-life entity in the data model, for example, a purchasing order.
Cache	A fast but rather small memory that serves as buffer for larger but slower memory.
Cache Coherence	State of consistency between the versions of data stored in the local caches of a CPU cache.
Cache-Conscious Algorithm	An algorithm is cache conscious if program variables that are dependent on hardware configuration parameters (for example, cache size and cache-line length) need to be tuned to minimize the number of cache misses.
Cache Line	Smallest unit of memory that can be transferred between main memory and the processor's cache. It is of a fixed size, which depends on the respective processor type.
Cache Miss	A failed request for data from a cache because it did not contain the requested data.

Cache-Oblivious Algorithm	An algorithm is cache oblivious if no program variables that are dependent on hardware configuration parameters (for example, cache size and cache-line length) need to be tuned to minimize the number of cache misses.
Characteristic-Oriented Database System	A database system that is tailored towards the characteristics of special application areas. Examples are text mining, stream processing and data warehousing.
Cloud Computing	An IT provisioning model, which emphasizes the on-demand, elastic pay-per-use rendering of services or provisioning of resources over a network.
Column Store	Database storage engine that stores each column (attribute) of a table sequentially in a contiguous area of memory.
Compression	Encoding information in such a way that its representation consumes less space in memory.
Compression Rate	The ratio to what size the data on which compression is applied can be shrinked. A compression rate of 5 means that the compressed size is only 20 % of the original size.
Concurrency Control	Techniques that allow the simultaneous and independent execution of transactions in a database system without creating states of unwanted incorrectness.
Consistency	Database concept that demands that only correct database states are visible to the user despite the execution of transactions.
Consolidation	Placing the data of several customers on one server machine, database or table in a multi-tenant setup.
Cube	Specialized OLAP data structure that allows multi-dimensional analysis of data.
Customer Relationship Management (CRM)	Business processes and respective technology used by a company to organize its interaction with its customers.

Data Aging	The changeover from active data to passive data.
Data Center	Facility housing servers and associated ICT components.
Data Dictionary	Meta data repository.
Data Layout	The structure in which data is organized in the database; that is, the database's physical schema.
Data Mart	A database that maintains copies of data from a specific business area, for example, sales or production, for analytical processing purposes.
Data Warehouse	A database that maintains copies of data from operational databases for analytical processing purposes.
Database Management System (DBMS)	A set of administrative programs used to create, maintain and manage a database.
Database Schema	Formal description of the logical structure of a database.
Demand Planning	Estimating future sales by combining several sources of information.
Design Thinking	A methodology that combines an end-user focus with multidisciplinary collaboration and iterative improvement. It aims at creating desirable, user-friendly, and economically viable design solutions and innovative products and services.
Desirability	Design thinking term expressing the practicability of a system from a human-usability point of view.
Dictionary	In the context of this book, the compressed and sorted repository holding all distinct data values referenced by SanssouciDB's main store.
Dictionary Encoding	Light-weight compression technique that encodes variable length values by smaller fixed-length encoded values using a mapping dictionary.
Differential Buffer	A write-optimized buffer to increase write performance of the SanssouciDB column store. Sometimes also referred to as differential store or delta store.

Distributed System	A system consisting of a number of autonomous computers that communicate over a computer network.
Dunning	The process of scanning through open invoices and identifying overdue ones, in order to take appropriate steps according to the dunning level.
Durability	Database concept that demands that all changes made by a transaction become permanent after this transaction has been committed.
Enterprise Application	A software system that helps an organization to run its business. A key feature of an enterprise application is its ability to integrate and process up-to-the-minute data from different business areas providing a holistic, real-time view of the entire enterprise.
Enterprise Resource Planning (ERP)	Enterprise software to support the resource planning processes of an entire company.
Entropy	Average information containment of a sign system.
Extract-Transform-Load (ETL) Process	A process that extracts data required for analytical processing from various sources, then transforms it (into an appropriate format, removing duplicates, sorting, aggregating, etc.) such that it can be finally loaded into the target analytical system.
Fault Tolerance	Quality of a system to maintain operation according to its specification, even if failures occur.
Feasibility	Design thinking term expressing the practicability of a system from a technical point of view.
Front Side Bus (FSB)	Bus that connects the processor with main memory (and the rest of the computer).
Horizontal Partitioning	The splitting of tables with many rows, into several partitions each having fewer rows.
Hybrid Store	Database that allows mixing column- and row-wise storage.

In-Memory Database	A database system that always keeps its primary data completely in main memory.
Index	Data structure in a database used to optimize read operations.
Insert-Only	New and changed tuples are always appended; already existing changed and deleted tuples are then marked as invalid.
Inter-Operator Parallelism	Parallel execution of independent plan operators of one or multiple query plans.
Intra-Operator Parallelism	Parallel execution of a single plan operation independently of any other operation of the query plan.
Isolation	Database concept demanding that any two concurrently executed transactions have the illusion that they are executed alone. The effect of such an isolated execution must not differ from executing the respective transactions one after the other.
Join	Database operation that is logically the cross product of two or more tables followed by a selection.
Latency	The time that a storage device needs between receiving the request for a piece of data and transmitting it.
Locking	A method to achieve isolation by regulating the access to a shared resource.
Logging	Process of persisting change information to non-volatile storage.
Main Memory	Physical memory that can be directly accessed by the central processing unit (CPU).
Main Store	Read-optimized and compressed data tables of SanssouciDB that are completely stored in main memory and on which no direct inserts are allowed.
MapReduce	A programming model and software framework for developing applications that allows for parallel processing of vast amounts of data on a large number of servers.

Materialized View	Result set of a complex query, which is persisted in the database and updated automatically.
Memory Hierarchy	The hierarchy of data storage technologies characterized by increasing response time but decreasing cost.
Merge Process	Process in SanssouciDB that periodically moves data from the write-optimized differential store into the main store.
Meta Data	Data specifying the structure of tuples in database tables (and other objects) and relationships among them, in terms of physical storage.
Mixed Workload	Database workload consisting both of transactional and analytical queries.
Multi-Core Processor	A microprocessor that comprises more than one core (processor) in a single integrated circuit.
Multi-Tenancy	The consolidation of several customers onto the operational system of the same server machine.
Multithreading	Concurrently executing several threads on the same processor core.
Network Partitioning Fault	Fault that separates a network into two or more sub-networks that cannot reach each other anymore.
Node	Partial structure of a business object.
Normalization	Designing the structure of the tables of a database in such a way that anomalies cannot occur and data integrity is maintained.
Object Data Guide	A database operator and index structure introduced to allow queries on whole business objects.
Online Analytical Processing (OLAP)	see Analytical Processing.
Online Transaction Processing (OLTP)	see Transactional Processing.
Operational Data Store	Database used to integrate data from multiple operational sources and to then update data marts and/or data warehouses.

Object-Relational Mapping (ORM)	A technique that an object-oriented programm could use a relational database as if it is an object-oriented database.
Padding	Approach to modify memory structures so that they exhibit better memory access behavior but requiring the trade-off of having additional memory consumption.
Passive Data	Data of a business transaction that is closed/completed and will not be changed anymore. For SanssouciDB, it may therefore be moved to non-volatile storage.
Prefetching	A technique that asynchronously loads additional cache lines from main memory into the CPU cache to hide memory latency.
Query	Request sent to a DBMS in order to retrieve data, manipulate data, execute an operation, or change the database structure.
Query Plan	The set and order of individual database operations, derived by the query optimizer of the DBMS, to answer an SQL query.
Radio-Frequency Identification (RFID)	Wireless technology to support fast tracking and tracing of goods. The latter are equipped with tags containing a unique identifier that can be readout by reader devices.
Real Time	In the context of this book, defined as, within the timeliness constraints of the speed-of-thought concept.
Real-Time Analytics	Analytics that have all information at its disposal the moment they are called for (within the timeliness constraints of the speed-of-thought concept).
Recoverability	Quality of a DBMS to allow for recovery after a failure has occurred.
Recovery	Process of re-attaining a correct database state and operation according to the database's specification after a failure has occurred.

Relational Database	A database that organizes its data in relations (tables) as sets of tuples (rows) having the same attributes (columns) according to the relational model.
Response Time at the Speed of Thought	Response time of a system that is perceived as instantaneous by a human user because of his/her own mental processes. It normally lies between 550 and 750 ms.
Return on Investment (ROI)	Economic measure to evaluate the efficiency of an investment.
Row Store	Database storage engine that stores all tuples sequentially; that is, each memory block may contain several tuples.
Sales Analysis	Process that provides an overview of historical sales numbers.
Sales Order Processing	Process with the main purpose of capturing sales orders.
SanssouciDB	The in-memory database described in this book.
Scalability	Desired characteristic of a system to yield an efficient increase in service capacity by adding resources.
Scale-out	Capable of handling increasing workloads by adding new machines and using these multiple machines to provide the given service.
Scale up	Capable of handling increasing workloads by adding new resources to a given machine to provide the given service.
Scan	Database operation evaluating a simple predicate on a column.
Scheduling	Process of ordering the execution of all queries (and query plan operators) of the current workload in order to maintain a given optimality criterion.
Sequential Reading	Reading a given memory block by block.
Shared Database Instance	Multi-tenancy implementation scheme in which each customer has its own tables, and sharing takes place on the level of the database instances.

Shared Machine	Multi-tenancy implementation scheme in which each customer has its own database process, and these processes are executed on the same machine; that is, several customers share the same server.
Shared Table	Multi-tenancy implementation scheme in which sharing takes place on the level of database tables; that is, data from different customers is stored in one and the same table.
Shared Disk	All processors share one view to the non-volatile memory, but computation is handled individually and privately by each computing instance.
Shared Memory	All processors share direct access to a global main memory and a number of disks.
Shared Nothing	Each processor has its own memory and disk(s) and acts independently of the other processors in the system.
Single Instruction Multiple Data (SIMD)	A multiprocessor instruction that applies the same instructions to many data streams.
Smart Grid	An electricity network that can intelligently integrate the behavior and actions of all users connected to it - generators, consumers and those that do both in order to efficiently deliver sustainable, economic and secure electricity supplies.
Software-as-a-Service (SaaS)	Provisioning of applications as cloud services over the Inter- net.
Solid-State Drive (SSD)	Data storage device that uses microchips for non-volatile, high- speed storage of data and exposes itself via standard communication protocols.
Speedup	Measure for scalability defined as the ratio between the time consumed by a sequential system and the time consumed by a parallel system to carry out the same task.

Star Schema	Simplest form of a data warehouse schema with one fact table (containing the data of interest, for example, sales numbers) and several accompanying dimension tables (containing the specific references to view the data of interest, for example, state, country, month) forming a star-like structure.
Stored Procedure	Procedural programs that can be written in SQL or PL/SQL and that are stored and accessible within the DBMS.
Streaming SIMD Extensions (SSE)	An Intel SIMD instruction set extension for the x86 processor architecture.
Structured Data	Data that is described by a data model, for example, business data in a relational database.
Structured Query Language (SQL)	A standardized declarative language for defining, querying, and manipulating data.
Supply Chain Management (SCM)	Business processes and respective technology to manage the flow of inventory and goods along a company's supply chain.
Table	A set of tuples having the same attributes.
Tenant	(1) A set of tables or data belonging to one customer in a multi-tenant setup. (2) An organization with several users querying a set of tables belonging to this organization in a multi-tenant setup.
Thread	Smallest schedulable unit of execution of an operating system.
Three-tier Architecture	Architecture of a software system that is separated in a presentation, a business logic, and a data layer (tier).
Time Travel Query	Query returning only those tuples of a table that were valid at the specified point in time.
Total Cost of Ownership (TCO)	Accounting technique that tries to estimate the overall life- time costs of acquiring and operating equipment, for example, software or hardware assets.
Transaction	A set of actions on a database executed as a single unit according to the ACID concept.

Transactional Processing	Method to process every-day business operations as ACID transactions such that the database remains in a consistent state.
Translation Lookaside Buffer (TLB)	A cache that is part of a CPU's memory management unit and is employed for faster virtual address translation.
Trigger	A set of actions that are executed within a database when a certain event occurs; for example, a specific modification takes place.
Tuple	A real-world entity's representation as a set of attributes stored as element in a relation. In other words, a row in a table.
Unstructured Data	Data without data model or that a computer program cannot easily use (in the sense of understanding its content). Examples are word processing documents or electronic mail.
Vertical Partitioning	The splitting of the attribute set of a database table and distributing it across two (or more) tables.
Viability	Design thinking term expressing the practicability of a system from an economic point of view.
View	Virtual table in a relational database whose content is defined by a stored query.
Virtual Machine	A program mimicking an entire computer by acting like a physical machine.
Virtual Memory	Logical address space offered by the operating process for a programm which is independent of the amount of actual main memory.
Virtualization	Method to introduce a layer of abstraction in order to provide a common access to a set of diverse physical and thereby virtualized resources.

Index