# Table of Contents

**Preface** ..... xvi

**Section I**
**Conceptual Model and Development**

**Chapter I**
Development of Data Warehouse Conceptual Models: Method Engineering Approach ..... 1
*Laila Niedrite, University of Latvia, Latvia*
*Maris Treimanis, University of Latvia, Latvia*
*Darja Solodovnikova, University of Latvia, Latvia*
*Liga Grundmane, University of Latvia, Latvia*

**Chapter II**
Conceptual Modeling Solutions for the Data Warehouse ..... 24
*Stefano Rizzi, DEIS-University of Bologna, Italy*

**Chapter III**
A Machine Learning Approach to Data Cleaning in Databases and Data Warehouses ..... 43
*Hamid Haidarian Shahri, University of Maryland, USA*

**Chapter IV**
Interactive Quality-Oriented Data Warehouse Development ..... 59
*Maurizio Pighin, IS & SE- Lab, University of Udine, Italy*
*Lucio Ieronutti, IS & SE- Lab, University of Udine, Italy*

**Chapter V**
Integrated Business and Production Process Data Warehousing ..... 88
*Dirk Draheim, University of Lunsbruck, Austria*
*Oscar Mangisengi, BWIN Interactive Entertainment, AG & SMS Data System, GmbH, Austria*

## Section II
## OLAP and Pattern

**Chapter VI**
Selecting and Allocating Cubes in Multi-Node OLAP Systems: An Evolutionary Approach ..... 99
*Jorge Loureiro, Instituto Politécnico de Viseu, Portugal*
*Orlando Belo, Universidade do Minho, Portugal*

**Chapter VII**
Swarm Quant' Intelligence for Optimizing Multi-Node OLAP Systems ..... 132
*Jorge Loureiro, Instituto Politécnico de Viseu, Portugal*
*Orlando Belo, Universidade do Minho, Portugal*

**Chapter VIII**
Multidimensional Anlaysis of XML Document Contents with OLAP Dimensions ..... 155
*Franck Ravat, IRIT, Universite Toulouse, France*
*Olivier Teste, IRIT, Universite Toulouse, France*
*Ronan Tournier, IRIT, Universite Toulouse, France*

**Chapter IX**
A Multidimensional Pattern Based Approach for the Design of Data Marts ..... 172
*Hanene Ben-Abdallah, University of Sfax, Tunisia*
*Jamel Feki, University of Sfax, Tunisia*
*Mounira Ben Abdallah, University of Sfax, Tunisia*

## Section III
## Spatio-Temporal Data Warehousing

**Chapter X**
A Multidimensional Methodology with Support for Spatio-Temporal Multigranularity in the Conceptual and Logical Phases ..... 194
*Concepción M. Gascueña, Polytechnic of Madrid University, Spain*
*Rafael Guadalupe, Polytechnic of Madrid University, Spain*

**Chapter XI**
Methodology for Improving Data Warehouse Design using Data Sources Temporal Metadata ..... 231
*Francisco Araque, University of Granada, Spain*
*Alberto Salguero, University of Granada, Spain*
*Cecilia Delgado, University of Granada, Spain*

**Chapter XII**
Using Active Rules to Maintain Data Consistency in Data Warehouse Systems ..... 252
*Shi-Ming Huang, National Chung Cheng University Taiwan*
*John Tait, Information Retrieval Faculty, Austria*
*Chun-Hao Su, National Chung Cheng University, Taiwan*
*Chih-Fong Tsai, National Central University, Taiwan*

**Chapter XIII**
Distributed Approach to Continuous Queries with kNN Join Processing in Spatial Telemetric Data Warehouse ..... 273
*Marcin Gorawski, Silesian Technical University, Poland*
*Wojciech Gębczyk, Silesian Technical University, Poland*

**Chapter XIV**
Spatial Data Warehouse Modelling ..... 282
*Maria Luisa Damiani, Università di Milano, Italy & Ecole Polytechnique Fédérale, Switzerland*
*Stefano Spaccapietra, Ecole Polytechnique Fédérale de Lausanne, Switzerland*

**Section IV**
**Benchmarking and Evaluation**

**Chapter XV**
Data Warehouse Benchmarking with DWEB ..... 302
*Jérôme Darmont, University of Lyon (ERIC Lyon 2), France*

**Chapter XVI**
Analyses and Evaluation of Responses to Slowly Changing Dimensions in Data Warehouses ..... 324
*Lars Frank, Copenhagen Business School, Denmark*
*Christian Frank, Copenhagen Business School, Denmark*

**Compilation of References** ..... 338

**About the Contributors** ..... 361

**Index** ..... 367

# Detailed Table of Contents

**Preface** ..... xvi

**Section I**
**Conceptual Model and Development**

**Chapter I**
Development of Data Warehouse Conceptual Models: Method Engineering Approach ..... 1
*Laila Niedrite, University of Latvia, Latvia*
*Maris Treimanis, University of Latvia, Latvia*
*Darja Solodovnikova, University of Latvia, Latvia*
*Liga Grundmane, University of Latvia, Latvia*

There are many methods in the area of data warehousing to define requirements for the development of the most appropriate conceptual model of a data warehouse. There is no universal consensus about the best method, nor are there accepted standards for the conceptual modeling of data warehouses. Only few conceptual models have formally described methods how to get these models. Therefore, problems arise when in a particular data warehousing project, an appropriate development approach, and a corresponding method for the requirements elicitation, should be chosen and applied. Sometimes it is also necessary not only to use the existing methods, but also to provide new methods that are usable in particular development situations. It is necessary to represent these new methods formally, to ensure the appropriate usage of these methods in similar situations in the future. It is also necessary to define the contingency factors, which describe the situation where the method is usable.This chapter represents the usage of method engineering approach for the development of conceptual models of data warehouses. A set of contingency factors that determine the choice between the usage of an existing method and the necessity to develop a new one is defined. Three case studies are presented. Three new methods: user-driven, data-driven, and goal-driven are developed according to the situation in the particular projects and using the method engineering approach.

**Chapter II**
Conceptual Modeling Solutions for the Data Warehouse ..... 24
*Stefano Rizzi, DEIS-University of Bologna, Italy*

In the context of data warehouse design, a basic role is played by conceptual modeling, that provides a higher level of abstraction in describing the warehousing process and architecture in all its aspects, aimed at achieving independence of implementation issues. This chapter focuses on a conceptual model called the DFM that suits the variety of modeling situations that may be encountered in real projects of small to large complexity. The aim of the chapter is to propose a comprehensive set of solutions for conceptual modeling according to the DFM and to give the designer a practical guide for applying them in the context of a design methodology. Besides the basic concepts of multidimensional modeling, the other issues discussed are descriptive and cross-dimension attributes; convergences; shared, incomplete, recursive, and dynamic hierarchies; multiple and optional arcs; and additivity.

**Chapter III**

A Machine Learning Approach to Data Cleaning in Databases and Data Warehouses ..... 43
*Hamid Haidarian Shahri, University of Maryland, USA*

Entity resolution (also known as duplicate elimination) is an important part of the data cleaning process, especially in data integration and warehousing, where data are gathered from distributed and inconsistent sources. Learnable string similarity measures are an active area of research in the entity resolution problem. Our proposed framework builds upon our earlier work on entity resolution, in which fuzzy rules and membership functions are defined by the user. Here, we exploit neuro-fuzzy modeling for the first time to produce a unique adaptive framework for entity resolution, which automatically learns and adapts to the specific notion of similarity at a meta-level. This framework encompasses many of the previous work on trainable and domain-specific similarity measures. Employing fuzzy inference, it removes the repetitive task of hard-coding a program based on a schema, which is usually required in previous approaches. In addition, our extensible framework is very flexible for the end user. Hence, it can be utilized in the production of an intelligent tool to increase the quality and accuracy of data.

**Chapter IV**

Interactive Quality-Oriented Data Warehouse Development ..... 59
*Maurizio Pighin, IS & SE- Lab, University of Udine, Italy*
*Lucio Ieronutti, IS & SE- Lab, University of Udine, Italy*

Data Warehouses are increasingly used by commercial organizations to extract, from a huge amount of transactional data, concise information useful for supporting decision processes. However, the task of designing a data warehouse and evaluating its effectiveness is not trivial, especially in the case of large databases and in presence of redundant information. The meaning and the quality of selected attributes heavily influence the data warehouse's effectiveness and the quality of derived decisions. Our research is focused on interactive methodologies and techniques targeted at supporting the data warehouse design and evaluation by taking into account the quality of initial data. In this chapter we propose an approach for supporting the data warehouses development and refinement, providing practical examples and demonstrating the effectiveness of our solution. Our approach is mainly based on two phases: the first one is targeted at interactively guiding the attributes selection by providing quantitative information measuring different statistical and syntactical aspects of data, while the second phase, based on a set of 3D visualizations, gives the opportunity of run-time refining taken design choices according to data examination and analysis. For experimenting proposed solutions on real data, we have developed

a tool, called ELDA (EvaLuation DAta warehouse quality), that has been used for supporting the data warehouse design and evaluation.

**Chapter V**
Integrated Business and Production Process Data Warehousing ..... 88
*Dirk Draheim, University of Lunsbruck, Austria*
*Oscar Mangisengi, BWIN Interactive Entertainment, AG & SMS Data System, GmbH, Austria*

Nowadays tracking data from activity checkpoints of unit transactions within an organization's business processes becomes an important data resource for business analysts and decision-makers to provide essential strategic and tactical business information. In the context of business process-oriented solutions, business-activity monitoring (BAM) architecture has been predicted as a major issue in the near future of the business-intelligence area. On the other hand, there is a huge potential for optimization of processes in today's industrial manufacturing. Important targets of improvement are production efficiency and product quality. Optimization is a complex task. A plethora of data that stems from numerical control and monitoring systems must be accessed, correlations in the information must be recognized, and rules that lead to improvement must be identified. In this chapter we envision the vertical integration of technical processes and control data with business processes and enterprise resource data. As concrete steps, we derive an activity warehouse model based on BAM requirements. We analyze different perspectives based on the requirements, such as business process management, key performance indication, process and state based-workflow management, and macro- and micro-level data. As a concrete outcome we define a meta-model for business processes with respect to monitoring. The implementation shows that data stored in an activity warehouse is able to efficiently monitor business processes in real-time and provides a better real-time visibility of business processes.

## Section II
## OLAP and Pattern

**Chapter VI**
Selecting and Allocating Cubes in Multi-Node OLAP Systems: An Evolutionary Approach ..... 99
*Jorge Loureiro, Instituto Politécnico de Viseu, Portugal*
*Orlando Belo, Universidade do Minho, Portugal*

OLAP queries are characterized by short answering times. Materialized cube views, a pre-aggregation and storage of group-by values, are one of the possible answers to that condition. However, if all possible views were computed and stored, the amount of necessary materializing time and storage space would be huge. Selecting the most beneficial set, based on the profile of the queries and observing some constraints as materializing space and maintenance time, a problem denoted as cube views selection problem, is the condition for an effective OLAP system, with a variety of solutions for centralized approaches. When a distributed OLAP architecture is considered, the problem gets bigger, as we must deal with another dimension—space. Besides the problem of the selection of multidimensional structures, there's now a node allocation one; both are a condition for performance. This chapter focuses on distributed OLAP systems, recently introduced, proposing evolutionary algorithms for the selection and allocation of the

distributed OLAP Cube, using a distributed linear cost model. This model uses an extended aggregation lattice as framework to capture the distributed semantics, and introduces processing nodes' power and real communication costs parameters, allowing the estimation of query and maintenance costs in time units. Moreover, as we have an OLAP environment, whit several nodes, we will have parallel processing and then, the evaluation of the fitness of evolutionary solutions is based on cost estimation algorithms that simulate the execution of parallel tasks, using time units as cost metric.

**Chapter VII**

Swarm Quant' Intelligence for Optimizing Multi-Node OLAP Systems ..... 132

*Jorge Loureiro, Instituto Politécnico de Viseu, Portugal*

*Orlando Belo, Universidade do Minho, Portugal*

Globalization and market deregulation has increased business competition, which imposed OLAP data and technologies as one of the great enterprise's assets. Its growing use and size stressed underlying servers and forced new solutions. The distribution of multidimensional data through a number of servers allows the increasing of storage and processing power without an exponential increase of financial costs. However, this solution adds another dimension to the problem: space. Even in centralized OLAP, cube selection efficiency is complex, but now, we must also know where to materialize subcubes. We have to select and also allocate the most beneficial subcubes, attending an expected (changing) user profile and constraints. We now have to deal with materializing space, processing power distribution, and communication costs. This chapter proposes new distributed cube selection algorithms based on discrete particle swarm optimizers; algorithms that solve the distributed OLAP selection problem considering a query profile under space constraints, using discrete particle swarm optimization in its normal(Di-PSO), cooperative (Di-CPSO), multi-phase (Di-MPSO), and applying hybrid genetic operators.

**Chapter VIII**

Multidimensional Anlaysis of XML Document Contents with OLAP Dimensions ..... 155

*Franck Ravat, IRIT, Universite Toulouse, France*

*Olivier Teste, IRIT, Universite Toulouse, France*

*Ronan Tournier, IRIT, Universite Toulouse, France*

With the emergence of Semi-structured data format (such as XML), the storage of documents in centralised facilities appeared as a natural adaptation of data warehousing technology. Nowadays, OLAP (On-Line Analytical Processing) systems face growing non-numeric data. This chapter presents a framework for the multidimensional analysis of textual data in an OLAP sense. Document structure, metadata, and contents are converted into subjects of analysis (facts) and analysis axes (dimensions) within an adapted conceptual multidimensional schema. This schema represents the concepts that a decision maker will be able to manipulate in order to express his analyses. This allows greater multidimensional analysis possibilities as a user may gain insight within a collection of documents.

**Chapter IX**

A Multidimensional Pattern Based Approach for the Design of Data Marts ..... 172

*Hanene Ben-Abdallah, University of Sfax, Tunisia*

*Jamel Feki, University of Sfax, Tunisia*

*Mounira Ben Abdallah, University of Sfax, Tunisia*

Despite their strategic importance, the wide-spread usage of decision support systems remains limited by both the complexity of their design and the lack of commercial design tools. This chapter addresses the design complexity of these systems. It proposes an approach for data mart design that is practical and that endorses the decision maker involvement in the design process. This approach adapts a development technique well established in the design of various complex systems for the design of data marts (DM): Pattern-based design. In the case of DM, a multidimensional pattern (MP) is a generic specification of analytical requirements within one domain. It is constructed and documented with standard, real-world entities (RWE) that describe information artifacts used or produced by the operational information systems (IS) of several enterprises. This documentation assists a decision maker in understanding the generic analytical solution; in addition, it guides the DM developer during the implementation phase. After over viewing our notion of MP and their construction method, this chapter details a reuse method composed of two adaptation levels: one logical and one physical. The logical level, which is independent of any data source model, allows a decision maker to adapt a given MP to their analytical requirements and to the RWE of their particular enterprise; this produces a DM schema. The physical specific level projects the RWE of the DM over the data source model. That is, the projection identifies the data source elements necessary to define the ETL procedures. We illustrate our approaches of construction and reuse of MP with examples in the medical domain.

## Section III
## Spatio-Temporal Data Warehousing

**Chapter X**
A Multidimensional Methodology with Support for Spatio-Temporal Multigranularity in the Conceptual and Logical Phases ..... 194
*Concepción M. Gascueña, Polytechnic of Madrid University, Spain*
*Rafael Guadalupe, Polytechnic of Madrid University, Spain*

The Multidimensional Databases (MDB) are used in the Decision Support Systems (DSS) and in Geographic Information Systems (GIS); the latter locates spatial data on the Earth's surface and studies its evolution through time. This work presents part of a methodology to design MDB, where it considers the Conceptual and Logical phases, and with related support for multiple spatio-temporal granularities. This will allow us to have multiple representations of the same spatial data, interacting with other, spatial and thematic data. In the Conceptual phase, the conceptual multidimensional model—FactEntity (FE)—is used. In the Logical phase, the rules of transformations are defined, from the FE model, to the Relational and Object Relational logical models, maintaining multidimensional semantics, and under the perspective of multiple spatial, temporal, and thematic granularities. The FE model shows constructors and hierarchical structures to deal with the multidimensional semantics on the one hand, carrying out a study on how to structure "a fact and its associated dimensions." Thus making up the Basic factEnty, and in addition, showing rules to generate all the possible Virtual factEntities. On the other hand, with the spatial semantics, highlighting the Semantic and Geometric spatial granularities.

**Chapter XI**
Methodology for Improving Data Warehouse Design using Data Sources Temporal Metadata ..... 231
*Francisco Araque, University of Granada, Spain*
*Alberto Salguero, University of Granada, Spain*
*Cecilia Delgado, University of Granada, Spain*

One of the most complex issues of the integration and transformation interface is the case where there are multiple sources for a single data element in the enterprise Data Warehouse (DW). There are many facets due to the number of variables that are needed in the integration phase. This chapter presents our DW architecture for temporal integration on the basis of the temporal properties of the data and temporal characteristics of the data sources. If we use the data arrival properties of such underlying information sources, the Data Warehouse Administrator (DWA) can derive more appropriate rules and check the consistency of user requirements more accurately. The problem now facing the user is not the fact that the information being sought is unavailable, but rather that it is difficult to extract exactly what is needed from what is available. It would therefore be extremely useful to have an approach which determines whether it would be possible to integrate data from two data sources (with their respective data extraction methods associated). In order to make this decision, we use the temporal properties of the data, the temporal characteristics of the data sources, and their extraction methods. In this chapter, a solution to this problem is proposed.

**Chapter XII**
Using Active Rules to Maintain Data Consistency in Data Warehouse Systems ..... 252
*Shi-Ming Huang, National Chung Cheng University Taiwan*
*John Tait, Information Retrieval Faculty, Austria*
*Chun-Hao Su, National Chung Cheng University, Taiwan*
*Chih-Fong Tsai, National Central University, Taiwan*

Data warehousing is a popular technology, which aims at improving decision-making ability. As the result of an increasingly competitive environment, many companies are adopting a "bottom-up" approach to construct a data warehouse, since it is more likely to be on time and within budget. However, multiple independent data marts/cubes can easily cause problematic data inconsistency for anomalous update transactions, which leads to biased decision-making. This research focuses on solving the data inconsistency problem and proposing a temporal-based data consistency mechanism (TDCM) to maintain data consistency. From a relative time perspective, we use an active rule (standard ECA rule) to monitor the user query event and use a metadata approach to record related information. This both builds relationships between the different data cubes, and allows a user to define a VIT (valid interval temporal) threshold to identify the validity of interval that is a threshold to maintain data consistency. Moreover, we propose a consistency update method to update inconsistent data cubes, which can ensure all pieces of information are temporally consistent.

**Chapter XIII**
Distributed Approach to Continuous Queries with kNN Join Processing in Spatial Telemetric Data Warehouse ..... 273
*Marcin Gorawski, Silesian Technical University, Poland*
*Wojciech Gębczyk, Silesian Technical University, Poland*

This chapter describes realization of distributed approach to continuous queries with kNN join processing in the spatial telemetric data warehouse. Due to dispersion of the developed system, new structural members were distinguished: the mobile object simulator, the kNN join processing service, and the query manager. Distributed tasks communicate using JAVA RMI methods. The kNN queries (k Nearest Neighbour) joins every point from one dataset with its k nearest neighbours in the other dataset. In our approach we use the Gorder method, which is a block nested loop join algorithm that exploits sorting, join scheduling, and distance computation filtering to reduce CPU and I/O usage

**Chapter XIV**
Spatial Data Warehouse Modelling ..... 282
*Maria Luisa Damiani, Università di Milano, Italy & Ecole Polytechnique Fédérale, Switzerland*
*Stefano Spaccapietra, Ecole Polytechnique Fédérale de Lausanne, Switzerland*

This chapter is concerned with multidimensional data models for spatial data warehouses. Over the last few years different approaches have been proposed in the literature for modelling multidimensional data with geometric extent. Nevertheless, the definition of a comprehensive and formal data model is still a major research issue. The main contributions of the chapter are twofold: First, it draws a picture of the research area; second it introduces a novel spatial multidimensional data model for spatial objects with geometry (MuSD – multigranular spatial data warehouse). MuSD complies with current standards for spatial data modelling, augmented by data warehousing concepts such as spatial fact, spatial dimension and spatial measure. The novelty of the model is the representation of spatial measures at multiple levels of geometric granularity. Besides the representation concepts, the model includes a set of OLAP operators supporting the navigation across dimension and measure levels.

**Section IV**
**Benchmarking and Evaluation**

**Chapter XV**
Data Warehouse Benchmarking with DWEB ..... 302
*Jérôme Darmont, University of Lyon (ERIC Lyon 2), France*

Performance evaluation is a key issue for designers and users of Database Management Systems (DBMSs). Performance is generally assessed with software benchmarks that help, for example test architectural choices, compare different technologies, or tune a system. In the particular context of data warehousing and On-Line Analytical Processing (OLAP), although the Transaction Processing Performance Council (TPC) aims at issuing standard decision-support benchmarks, few benchmarks do actually exist. We present in this chapter the Data Warehouse Engineering Benchmark (DWEB), which allows generating various ad-hoc synthetic data warehouses and workloads. DWEB is fully parameterized to fulfill various data warehouse design needs. However, two levels of parameterization keep it relatively easy to tune. We also expand on our previous work on DWEB by presenting its new Extract, Transform, and Load (ETL) feature, as well as its new execution protocol. A Java implementation of DWEB is freely available online, which can be interfaced with most existing relational DMBSs. To the best of our knowledge, DWEB is the only easily available, up-to-date benchmark for data warehouses.

**Chapter XVI**
Analyses and Evaluation of Responses to Slowly Changing Dimensions in Data Warehouses ..... 324
*Lars Frank, Copenhagen Business School, Denmark*
*Christian Frank, Copenhagen Business School, Denmark*

A *Star Schema Data Warehouse* looks like a star with a central, so-called *fact table*, in the middle, surrounded by so-called *dimension tables* with one-to-many relationships to the central fact table. Dimensions are defined as *dynamic* **or** *slowly changing* if the attributes or relationships of a dimension can be updated. Aggregations of fact data to the level of the related dynamic dimensions might be misleading if the fact data are aggregated without considering the changes of the dimensions. In this chapter, we will first prove that the problems of SCD (Slowly Changing Dimensions) in a datawarehouse may be viewed as a special case of the read skew anomaly that may occur when different transactions access and update records without concurrency control. That is, we prove that aggregating fact data to the levels of a dynamic dimension should not make sense. On the other hand, we will also illustrate, by examples, that in some situations it does make sense that fact data is aggregated to the levels of a dynamic dimension. That is, it is the semantics of the data that determine whether historical dimension data should be preserved or destroyed. Even worse, we also illustrate that for some applications, we need a history preserving response, while for other applications at the same time need a history destroying response. Kimball et al., (2002), have described three classic solutions/responses to handling the aggregation problems caused by slowly changing dimensions. In this chapter, we will describe and evaluate four more responses of which one are new. This is important because all the responses have very different properties, and it is not possible to select a best solution without knowing the semantics of the data.

**Compilation of References** ..... 338

**About the Contributors** ..... 361

**Index** ..... 367