

Matthias Jarke • Maurizio Lenzerini
Yannis Vassiliou #Panos Vassiliadis

Fundamentals of Data Warehouses

Second, Revised and Extended Edition

With 59 Figures



Springer

Contents

1	Data Warehouse Practice: An Overview	1
1.1	Data Warehouse Components	2
1.2	Designing the Data Warehouse	4
1.3	Getting Heterogeneous Data into the Warehouse	5
1.4	Getting Multidimensional Data out of the Warehouse	6
1.5	Physical Structure of Data Warehouses	10
1.6	Metadata Management	13
1.7	Data Warehouse Project Management	13
2	Data Warehouse Research: Issues and Projects	15
2.1	Data Extraction and Reconciliation	15
2.2	Data Aggregation and Customization	15
2.3	Query Optimization	16
2.4	Update Propagation	17
2.5	Modeling and Measuring Data Warehouse Quality	17
2.6	Some Major Research Projects in Data Warehousing	19
2.7	Three Perspectives of Data Warehouse Metadata	21
3	Source Integration	27
3.1	The Practice of Source Integration	27
3.1.1	Tools for Data Warehouse Management	28
3.1.2	Tools for Data Integration	29
3.2	Research in Source Integration	30
3.2.1	Schema Integration	32
3.2.2	Data Integration - Virtual	36
3.2.3	Data Integration - Materialized	38
3.3	Towards Systematic Methodologies for Source Integration	40
3.3.1	Architecture for Source Integration	41
3.3.2	Methodology for Source Integration	43
3.4	Concluding Remarks	45
4	Data Warehouse Refreshment	47
4.1	What is Data Warehouse Refreshment?	47
4.1.1	Refreshment Process within the Data Warehouse Lifecycle	47
4.1.2	Requirements and Difficulties of Data Warehouse Refreshment	50
4.1.3	Data Warehouse Refreshment: Problem Statement	52

4.2	Incremental Data Extraction.....	54
4.2.1	Wrapper Functionality.....	55
4.2.2	Change Monitoring.....	56
4.3	Data Cleaning.....	62
4.3.1	Conversion and Normalization Functions.....	63
4.3.2	Special-Purpose Cleaning.....	64
4.3.3	Domain-Independent Cleaning.....	64
4.3.4	Rule-Based Cleaning.....	65
4.3.5	Concluding Remarks on Data Cleaning.....	67
4.4	Update Propagation into Materialized Views.....	67
4.4.1	Notations and Definitions.....	68
4.4.2	View Maintenance: General Results.....	68
4.4.3	View Maintenance in Data Warehouses - Specific Results.....	71
4.5	Toward a Quality-Oriented Refreshment Process.....	73
4.5.1	Quality Analysis for Refreshment.....	73
4.5.2	Implementing the Refreshment Process.....	77
4.5.3	Workflow Modeling with Rules.....	80
4.6	Implementation of the Approach.....	83
5	Multidimensional Data Models and Aggregation.....	87
5.1	Multidimensional View of Information.....	90
5.2	ROLAP Data Model.....	92
5.3	MOLAP Data Model.....	95
5.4	Logical Models for Multidimensional Information.....	97
5.5	Conceptual Models for Multidimensional Information.....	100
5.5.1	Inference Problems for Multidimensional Conceptual Modeling....	101
5.5.2	Which Formal Framework to Choose?.....	103
5.6	Conclusion.....	105
6	Query Processing and Optimization.....	107
6.1	Description and Requirements for Data Warehouse Queries.....	107
6.1.1	Queries at the Back End.....	108
6.1.2	Queries at the Front End.....	108
6.1.3	Queries in the Core.....	109
6.1.4	Transactional Versus Data Warehouse Queries.....	109
6.1.5	Canned Queries Versus Ad-hoc Queries.....	110
6.1.6	Multidimensional Queries.....	110
6.1.7	Extensions of SQL.....	112
6.2	Query Processing Techniques.....	113
6.2.1	Data Access.....	113
6.2.2	Evaluation Strategies.....	116
6.2.3	Exploitation of Redundancy.....	117
6.3	Conclusions and Research Directions.....	121

7	Metadata and Data Warehouse Quality	123
7.1	Metadata Management in Data Warehouse Practice.....	124
7.1.1	Metadata Interchange Specification (MDIS).....	125
7.1.2	The Telos Language.....	125
7.1.3	Microsoft Repository.....	127
7.1.4	OIM and CWM.....	128
7.2	A Repository Model for the DWQ Framework.....	129
7.2.1	Conceptual Perspective.....	131
7.2.2	Logical Perspective.....	132
7.2.3	Physical Perspective.....	132
7.2.4	Applying the Architecture Model.....	133
7.3	Defining Data Warehouse Quality.....	138
7.3.1	Data Quality.....	139
7.3.2	Stakeholders and Goals in Data Warehouse Quality.....	140
7.3.3	State of Practice in Data Warehouse Quality.....	143
7.4	Representing and Analyzing Data Warehouse Quality.....	144
7.4.1	Quality Function Deployment.....	145
7.4.2	The Need for Richer Quality Models: An Example.....	146
7.4.3	The Goal-Question-Metric Approach.....	147
7.4.4	Repository Support for the GQM Approach.....	148
7.5	A Detailed Example: Quality Analysis in Data Staging.....	154
7.5.1	Evaluation of the Quality of a DSA Schema.....	158
7.5.2	Analyzing the Quality of a View.....	160
8	Quality-Driven Data Warehouse Design	165
8.1	Interactions between Quality Factors and DW Tasks.....	165
8.2	The DWQ Data Warehouse Design Methodology.....	166
8.2.1	Source Integration.....	167
8.2.2	Multidimensional Aggregation and OLAP Query Generation.....	169
8.2.3	Design Optimization and Data Reconciliation.....	171
8.2.4	Operational Support.....	172
8.3	Optimizing the Materialization of DW Views.....	174
8.4	Summary and Outlook.....	178
	Bibliography	181
	Appendix A. ISO Standards Information Quality	203
	Appendix B. Glossary	207
	Index	215