

Contents

1	Data Warehouse Practice: An Overview	1
1.1	Data Warehouse Components.....	2
1.2	Designing the Data Warehouse.....	5
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.....	12
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 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
3.2.1.1	Preintegration.....	33
3.2.1.2	Schema Comparison	34
3.2.1.3	Schema Conforming	35
3.2.1.4	Schema Merging and Restructuring	36
3.2.2	Data Integration – Virtual.....	36
3.2.2.1	Carnot	36
3.2.2.2	SIMS.....	37
3.2.2.3	Information Manifold	37
3.2.2.4	TSIMMIS	38
3.2.3	Data Integration – Materialized	39
3.2.3.1	Squirrel	39
3.2.3.2	WHIPS.....	40
3.3	Towards Systematic Methodologies for Source Integration	40

3.3.1	Architecture for Source Integration	41
3.3.1.1	Conceptual Perspective.....	42
3.3.1.2	Logical Perspective.....	42
3.3.1.3	Mappings	43
3.3.2	Methodology for Source Integration	43
3.3.2.1	Source-Driven Integration	43
3.3.2.2	Client-Driven Integration.....	45
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
4.2.2.1	Snapshot Sources	58
4.2.2.2	Specific Sources	59
4.2.2.3	Logged Sources	59
4.2.2.4	Queryable Sources	59
4.2.2.5	Replicated Sources	60
4.2.2.6	Callback Sources	60
4.2.2.7	Internal Action Sources	61
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
4.3.4.1	User-Specified Rules	65
4.3.4.2	Automatically-derived Rules	66
4.3.5	Concluding Remarks on Data Cleaning.....	67
4.4	Update Propagation into Materialized Views	68
4.4.1	Notations and Definitions	68
4.4.2	View Maintenance: General Results	69
4.4.2.1	Characterizing (Self) Maintainable Views.....	69
4.4.2.2	Optimization	69
4.4.2.3	Joint Maintenance of a Set of Views	70
4.4.2.4	Evaluating View Maintenance Algorithms.....	71
4.4.3	View Maintenance in Data Warehouses – Specific Results	72
4.4.3.1	Consistency.....	72
4.4.3.2	Optimization	73
4.4.3.3	Temporal Data Warehouses.....	73
4.5	Toward a Quality-Oriented Refreshment Process.....	73
4.5.1	Quality Analysis for Refreshment	74
4.5.1.1	Quality Dimensions	74

4.5.1.2	Quality Factors	75
4.5.1.3	Design Choices	75
4.5.1.4	Links between Quality Factors and Design Choices	76
4.5.2	Implementing the Refreshment Process	77
4.5.2.1	Planning the Refreshment Process.....	77
4.5.3	Workflow Modeling with Rules	80
4.5.3.1	Main Features of the Toolkit	82
4.5.3.2	Functional Architecture of the Active Refreshment System	82
4.6	Concluding Remarks.....	84
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	96
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....	102
5.5.2	Which Formal Framework to Choose?.....	103
5.6	Conclusions.....	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 vs. Data Warehouse Queries	109
6.1.5	Canned Queries vs. Ad-hoc Queries.....	110
6.1.6	Multidimensional Queries	110
6.1.6.1	Querying the Dimensions	111
6.1.6.2	Querying Factual Information	111
6.1.7	Extensions of SQL.....	112
6.2	Query Processing Techniques	113
6.2.1	Data Access	113
6.2.1.1	Indexes.....	113
6.2.1.2	Aggregate Query Processing with Indexes	114
6.2.1.3	Join-Indexes for Stars	115
6.2.1.4	The Extended Datacube Model	115
6.2.2	Evaluation Strategies	116
6.2.2.1	Interleaving Group-By and Join	116
6.2.2.2	Optimization of Nested Subqueries	117
6.2.3	Exploitation of Redundancy	117
6.2.3.1	Which Views are Useful for Answering a Query?	118
6.2.3.2	What is the Expected Size of an Aggregate View?	121
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	Meta Data Interchange Specification (MDIS)	125
7.1.2	The Telos Language	125
7.1.3	Microsoft Repository.....	127
7.2	A Repository Model for the DWQ Framework	128
7.2.1	Conceptual Perspective.....	130
7.2.2	Logical Perspective	130
7.2.3	Physical Perspective	131
7.2.4	Applying the Architecture Model.....	131
7.3	Defining Data Warehouse Quality.....	136
7.3.1	Data Quality.....	136
7.3.2	Stakeholders and Goals in Data Warehouse Quality	137
7.3.3	State of Practice in Data Warehouse Quality.....	140
7.4	Managing Data Warehouse Quality.....	142
7.4.1	Quality Function Deployment	142
7.4.2	The Need for Richer Quality Models: An Example	143
7.4.3	The Goal-Question-Metric Approach.....	144
7.4.4	Repository Support for the GQM Approach.....	146
7.4.4.1	The Quality Meta Model	147
7.4.4.2	Implementation Support for the Quality Meta Model	149
7.4.4.3	Understanding, Controlling and Improving Quality with the Repository	151
7.5	Towards Quality-Driven Data Warehouse Design.....	152
7.5.1	Linking Quality Factors to Warehouse Components.....	152
7.5.2	An Example: Optimizing the Materialization of DW Views.....	153
7.6	Conclusions.....	157
	References.....	159
	Appendix A. ISO Standards Information Quality.....	179
	Appendix B. Glossary.....	183
B.1	Data Warehouse Systems.....	183
B.2	Data Quality.....	184
B.3	Source Integration.....	186
B.4	Multidimensional Aggregation	187
B.5	Query Optimization	189
	Index	191