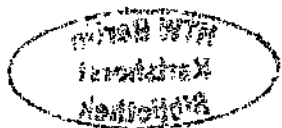


Data Warehouse Design

Modern Principles and Methodologies

Matteo Golfarelli
Stefano Rizzi



Translated by Claudio Pagliarani



New York Chicago San Francisco
Lisbon London Madrid Mexico City
Milan New Delhi San Juan
Seoul Singapore Sydney Toronto

Contents

Acknowledgments	xiii
Foreword	xv
Preface	xvii
1 Introduction to Data Warehousing	1
1.1 Decision Support Systems	2
1.2 Data Warehousing	4
1.3 Data Warehouse Architectures	7
1.3.1 Single-Layer Architecture	7
1.3.2 Two-Layer Architecture	8
1.3.3 Three-Layer Architecture	10
1.3.4 An Additional Architecture Classification	12
1.4 Data Staging and ETL	15
1.4.1 Extraction	15
1.4.2 Cleansing	16
1.4.3 Transformation	17
1.4.4 Loading	18
1.5 Multidimensional Model	18
1.5.1 Restriction	22
1.5.2 Aggregation	23
1.6 Meta-data	25
1.7 Accessing Data Warehouses	27
1.7.1 Reports	27
1.7.2 OLAP	29
1.7.3 Dashboards	36
1.8 ROLAP, MOLAP, and HOLAP	37
1.9 Additional Issues	39
1.9.1 Quality	39
1.9.2 Security	41
1.9.3 Evolution	41
2 Data Warehouse System Lifecycle	43
2.1 Risk Factors	43
2.2 Top-Down vs. Bottom-Up	44
2.2.1 Business Dimensional Lifecycle	46
2.2.2 Rapid Warehousing Methodology	48
2.3 Data Mart Design Phases	50
2.3.1 Analysis and Reconciliation of Data Sources	51
2.3.2 Requirement Analysis	52

2.3.3	Conceptual Design	52
2.3.4	Workload Refinement and Validation of Conceptual Schemata	53
2.3.5	Logical Design	53
2.3.6	Physical Design	53
2.3.7	Data-Staging Design	53
2.4	Methodological Framework	54
2.4.1	Scenario 1: Data-Driven Approach	55
2.4.2	Scenario 2: Requirement-Driven Approach	57
2.4.3	Scenario 3: Mixed Approach	58
2.5	Testing Data Marts	58
3	Analysis and Reconciliation of Data Sources	61
3.1	Inspecting and Normalizing Schemata	64
3.2	The Integration Problem	65
3.2.1	Different Perspectives	67
3.2.2	Equivalent Modeling Constructs	68
3.2.3	Incompatible Specifications	68
3.2.4	Common Concepts	69
3.2.5	Interrelated Concepts	70
3.3	Integration Phases	71
3.3.1	Preintegration	71
3.3.2	Schema Comparison	72
3.3.3	Schema Alignment	75
3.3.4	Merging and Restructuring Schemata	76
3.4	Defining Mappings	77
4	User Requirement Analysis	79
4.1	Interviews	80
4.2	Glossary-based Requirement Analysis	83
4.2.1	Facts	84
4.2.2	Preliminary Workload	87
4.3	Goal-oriented Requirement Analysis	89
4.3.1	Introduction to Tropos	90
4.3.2	Organizational Modeling	92
4.3.3	Decision-making Modeling	95
4.4	Additional Requirements	97
5	Conceptual Modeling	99
5.1	The Dimensional Fact Model: Basic Concepts	103
5.2	Advanced Modeling	108
5.2.1	Descriptive Attributes	109
5.2.2	Cross-Dimensional Attributes	111
5.2.3	Convergence	112
5.2.4	Shared Hierarchies	113
5.2.5	Multiple Arcs	114
5.2.6	Optional Arcs	115

5.2.7	Incomplete Hierarchies	116
5.2.8	Recursive Hierarchies	117
5.2.9	Additivity	118
5.3	Events and Aggregation	120
5.3.1	Aggregating Additive Measures	123
5.3.2	Aggregating Non-additive Measures	124
5.3.3	Aggregating with Convergence and Cross-dimensional Attributes	127
5.3.4	Aggregating with Optional or Multiple Arcs	128
5.3.5	Empty Fact Schema Aggregation	131
5.3.6	Aggregating with Functional Dependencies among Dimensions	133
5.3.7	Aggregating along Incomplete or Recursive Hierarchies	133
5.4	Time	137
5.4.1	Transactional vs. Snapshot Schemata	137
5.4.2	Late Updates	140
5.4.3	Dynamic Hierarchies	143
5.5	Overlapping Fact Schemata	145
5.6	Formalizing the Dimensional Fact Model	148
5.6.1	Metamodel	148
5.6.2	Intensional Properties	149
5.6.3	Extensional Properties	151
6	Conceptual Design	155
6.1	Entity-Relationship Schema-based Design	156
6.1.1	Defining Facts	157
6.1.2	Building Attribute Trees	159
6.1.3	Pruning and Grafting Attribute Trees	165
6.1.4	One-to-One Relationships	169
6.1.5	Defining Dimensions	169
6.1.6	Time Dimensions	172
6.1.7	Defining Measures	174
6.1.8	Generating Fact Schemata	174
6.2	Relational Schema-based Design	180
6.2.1	Defining Facts	180
6.2.2	Building Attribute Trees	181
6.2.3	Other Phases	185
6.3	XML Schema-based Design	187
6.3.1	Modeling XML Associations	187
6.3.2	Preliminary Phases	189
6.3.3	Selecting Facts and Building Attribute Trees	190
6.4	Mixed-approach Design	193
6.4.1	Mapping Requirements	194
6.4.2	Building Fact Schemata	194
6.4.3	Refining	196
6.5	Requirement-driven Approach Design	196

7	Workload and Data Volume	199
7.1	Workload	199
7.1.1	Dimensional Expressions and Queries on Fact Schemata	200
7.1.2	Drill-Across Queries	206
7.1.3	Composite Queries	207
7.1.4	Nested GPSJ Queries	209
7.1.5	Validating a Workload in a Conceptual Schema	209
7.1.6	Workload and Users	211
7.2	Data Volumes	213
8	Logical Modeling	217
8.1	MOLAP and HOLAP Systems	217
8.1.1	The Problem of Sparsity	219
8.2	ROLAP Systems	221
8.2.1	Star Schema	221
8.2.2	Snowflake Schema	224
8.3	Views	226
8.3.1	Relational Schemata with Aggregate Data	229
8.4	Temporal Scenarios	232
8.4.1	Dynamic Hierarchies: Type 1	233
8.4.2	Dynamic Hierarchies: Type 2	234
8.4.3	Dynamic Hierarchies: Type 3	236
8.4.4	Dynamic Hierarchies: Full Data Logging	237
8.4.5	Deleting Tuples	239
9	Logical Design	241
9.1	From Fact Schemata to Star Schemata	242
9.1.1	Descriptive Attributes	242
9.1.2	Cross-dimensional Attributes	242
9.1.3	Shared Hierarchies	243
9.1.4	Multiple Arcs	244
9.1.5	Optional Arcs	248
9.1.6	Incomplete Hierarchies	249
9.1.7	Recursive Hierarchies	251
9.1.8	Degenerate Dimensions	252
9.1.9	Additivity Issues	255
9.1.10	Using Snowflake Schemata	256
9.2	View Materialization	257
9.2.1	Using Views to Answer Queries	262
9.2.2	Problem Formalization	263
9.2.3	A Materialization Algorithm	266
9.3	View Fragmentation	268
9.3.1	Vertical View Fragmentation	269
9.3.2	Horizontal View Fragmentation	272

10	Data-staging Design	275
10.1	Populating Reconciled Databases	276
10.1.1	Extracting Data	277
10.1.2	Transforming Data	282
10.1.3	Loading Data	283
10.2	Cleansing Data	285
10.2.1	Dictionary-based Techniques	287
10.2.2	Approximate Merging	287
10.2.3	Ad-hoc Techniques	290
10.3	Populating Dimension Tables	290
10.3.1	Identifying the Data to Load	290
10.3.2	Replacing Keys	291
10.4	Populating Fact Tables	293
10.5	Populating Materialized Views	294
11	Indexes for the Data Warehouse	299
11.1	B ⁺ -Tree Indexes	299
11.2	Bitmap Indexes	302
11.2.1	Bitmap Indexes vs. B ⁺ -Trees	304
11.2.2	Advanced Bitmap Indexes	306
11.3	Projection Indexes	309
11.4	Join and Star Indexes	311
11.4.1	Multi-join Indexes	313
11.5	Spatial Indexes	317
11.6	Join Algorithms	320
11.6.1	Nested Loop	320
11.6.2	Sort-merge	321
11.6.3	Hash Join	322
12	Physical Design	325
12.1	Optimizers	325
12.1.1	Rule-based Optimizers	330
12.1.2	Cost-based Optimizers	335
12.1.3	Histograms	337
12.2	Index Selection	340
12.2.1	Indexing Dimension Tables	341
12.2.2	Indexing Fact Tables	342
12.3	Additional Physical Design Elements	343
12.3.1	Splitting a Database Into Tablespaces	343
12.3.2	Allocating Data Files	345
12.3.3	Disk Block Size	348
13	Data Warehouse Project Documentation	351
13.1	Data Warehouse Level	352
13.1.1	Data Warehouse Schemata	352
13.1.2	Deployment Schema	354

13.2	Data Mart Level	357
13.2.1	Bus and Overlapping Matrices	357
13.2.2	Operational Schema	358
13.2.3	Data-Staging Schema	360
13.2.4	Domain Glossary	365
13.2.5	Workload and Users	366
13.2.6	Logical Schema and Physical Schema	368
13.2.7	Testing Documents	370
13.3	Fact Level	371
13.3.1	Fact Schemata	371
13.3.2	Attribute and Measure Glossaries	372
13.4	Methodological Guidelines	373
14	A Case Study	375
14.1	Application Domain	375
14.2	Planning the TranSport Data Warehouse	375
14.3	The Sales Data Mart	376
14.3.1	Data Source Analysis and Reconciliation	376
14.3.2	User Requirement Analysis	389
14.3.3	Conceptual Design	390
14.3.4	Logical Design	395
14.3.5	Data-Staging Design	398
14.3.6	Physical Design	400
14.4	The Marketing Data Mart	400
15	Business Intelligence: Beyond the Data Warehouse	403
15.1	Introduction to Business Intelligence	403
15.2	Data Mining	406
15.2.1	Association Rules	408
15.2.2	Clustering	409
15.2.3	Classifiers and Decision Trees	410
15.2.4	Time Series	411
15.3	What-If Analysis	412
15.3.1	Inductive Techniques	413
15.3.2	Deductive Techniques	414
15.3.3	Methodological Notes	415
15.4	Business Performance Management	417
	Glossary	423
	Bibliography	429
	Index	445