



# Contents

<b>Acknowledgments</b>	<b>xvii</b>	
<b>About the Authors</b>	<b>xix</b>	
<b>Introduction</b>	<b>xxi</b>	
<b>Part I</b>	<b>Requirements, Realities, and Architecture</b>	<b>1</b>
<b>Chapter 1</b>	<b>Surrounding the Requirements</b>	<b>3</b>
	Requirements	4
	Business Needs	4
	Compliance Requirements	4
	Data Profiling	5
	Security Requirements	6
	Data Integration	7
	Data Latency	7
	Archiving and Lineage	8
	End User Delivery Interfaces	8
	Available Skills	9
	Legacy Licenses	9
	Architecture	9
	ETL Tool versus Hand Coding (Buy a Tool Suite or Roll Your Own?)	10
	The Back Room – Preparing the Data	16
	The Front Room – Data Access	20
	The Mission of the Data Warehouse	22
	What the Data Warehouse Is	22
	What the Data Warehouse Is Not	23
	Industry Terms Not Used Consistently	25

Resolving Architectural Conflict: A Hybrid Approach	27
How the Data Warehouse Is Changing	27
The Mission of the ETL Team	28
<b>2 ETL Data Structures</b>	<b>29</b>
To Stage or Not to Stage	29
Designing the Staging Area	31
Data Structures in the ETL System	35
Flat Files	35
XML Data Sets	38
Relational Tables	40
Independent DBMS Working Tables	41
Third Normal Form Entity/Relation Models	42
Nonrelational Data Sources	42
Dimensional Data Models: The Handoff from the Back Room to the Front Room	45
Fact Tables	45
Dimension Tables	46
Atomic and Aggregate Fact Tables	47
Surrogate Key Mapping Tables	48
Planning and Design Standards	48
Impact Analysis	49
Metadata Capture	49
Naming Conventions	51
Auditing Data Transformation Steps	51
Summary	52
<b>Data Flow</b>	<b>53</b>
<b>3 Extracting</b>	<b>55</b>
Part 1: The Logical Data Map	56
Designing Logical Before Physical	56
Inside the Logical Data Map	58
Components of the Logical Data Map	58
Using Tools for the Logical Data Map	62
Building the Logical Data Map	62
Data Discovery Phase	63
Data Content Analysis	71
Collecting Business Rules in the ETL Process	73
Integrating Heterogeneous Data Sources	73
Part 2: The Challenge of Extracting from Disparate Platforms	76
Connecting to Diverse Sources through ODBC	76
Mainframe Sources	78
Working with COBOL Copybooks	78
EBCDIC Character Set	79
Converting EBCDIC to ASCII	80

Transferring Data between Platforms	80
Handling Mainframe Numeric Data	81
Using PICTures	81
Unpacking Packed Decimals	83
Working with Redefined Fields	84
Multiple OCCURS	85
Managing Multiple Mainframe Record Type Files	87
Handling Mainframe Variable Record Lengths	89
Flat Files	90
Processing Fixed Length Flat Files	91
Processing Delimited Flat Files	93
XML Sources	93
Character Sets	94
XML Meta Data	94
Web Log Sources	97
W3C Common and Extended Formats	98
Name Value Pairs in Web Logs	100
ERP System Sources	102
Part 3: Extracting Changed Data	105
Detecting Changes	106
Extraction Tips	109
Detecting Deleted or Overwritten Fact Records at the Source	111
Summary	111
<b>Chapter 4 Cleaning and Conforming</b>	<b>113</b>
Defining Data Quality	115
Assumptions	116
Part 1: Design Objectives	117
Understand Your Key Constituencies	117
Competing Factors	119
Balancing Conflicting Priorities	120
Formulate a Policy	122
Part 2: Cleaning Deliverables	124
Data Profiling Deliverable	125
Cleaning Deliverable #1: Error Event Table	125
Cleaning Deliverable #2: Audit Dimension	128
Audit Dimension Fine Points	130
Part 3: Screens and Their Measurements	131
Anomaly Detection Phase	131
Types of Enforcement	134
Column Property Enforcement	134
Structure Enforcement	135
Data and Value Rule Enforcement	135
Measurements Driving Screen Design	136
Overall Process Flow	136
The Show Must Go On—Usually	138
Screens	139

Known Table Row Counts	140
Column Nullity	140
Column Numeric and Date Ranges	141
Column Length Restriction	143
Column Explicit Valid Values	143
Column Explicit Invalid Values	144
Checking Table Row Count Reasonability	144
Checking Column Distribution Reasonability	146
General Data and Value Rule Reasonability	147
Part 4: Conforming Deliverables	148
Conformed Dimensions	148
Designing the Conformed Dimensions	150
Taking the Pledge	150
Permissible Variations of Conformed Dimensions	150
Conformed Facts	151
The Fact Table Provider	152
The Dimension Manager: Publishing Conformed Dimensions to Affected Fact Tables	152
Detailed Delivery Steps for Conformed Dimensions	153
Implementing the Conforming Modules	155
Matching Drives Deduplication	156
Surviving: Final Step of Conforming	158
Delivering	159
Summary	160
<b>er 5 Delivering Dimension Tables</b>	<b>161</b>
The Basic Structure of a Dimension	162
The Grain of a Dimension	165
The Basic Load Plan for a Dimension	166
Flat Dimensions and Snowflaked Dimensions	167
Date and Time Dimensions	170
Big Dimensions	174
Small Dimensions	176
One Dimension or Two	176
Dimensional Roles	178
Dimensions as Subdimensions of Another Dimension	180
Degenerate Dimensions	182
Slowly Changing Dimensions	183
Type 1 Slowly Changing Dimension (Overwrite)	183
Type 2 Slowly Changing Dimension (Partitioning History)	185
Precise Time Stamping of a Type 2 Slowly Changing Dimension	190
Type 3 Slowly Changing Dimension (Alternate Realities)	192
Hybrid Slowly Changing Dimensions	193
Late-Arriving Dimension Records and Correcting Bad Data	194
Multivalued Dimensions and Bridge Tables	196
Ragged Hierarchies and Bridge Tables	199
Technical Note: POPULATING HIERARCHY BRIDGE TABLES	201

Using Positional Attributes in a Dimension to Represent Text Facts	204
Summary	207
<b>Chapter 6 Delivering Fact Tables</b>	<b>209</b>
The Basic Structure of a Fact Table	210
Guaranteeing Referential Integrity	212
Surrogate Key Pipeline	214
Using the Dimension Instead of a Lookup Table	217
Fundamental Grains	217
Transaction Grain Fact Tables	218
Periodic Snapshot Fact Tables	220
Accumulating Snapshot Fact Tables	222
Preparing for Loading Fact Tables	224
Managing Indexes	224
Managing Partitions	224
Outwitting the Rollback Log	226
Loading the Data	226
Incremental Loading	228
Inserting Facts	228
Updating and Correcting Facts	228
Negating Facts	229
Updating Facts	230
Deleting Facts	230
Physically Deleting Facts	230
Logically Deleting Facts	232
Factless Fact Tables	232
Augmenting a Type 1 Fact Table with Type 2 History	234
Graceful Modifications	235
Multiple Units of Measure in a Fact Table	237
Collecting Revenue in Multiple Currencies	238
Late Arriving Facts	239
Aggregations	241
Design Requirement #1	243
Design Requirement #2	244
Design Requirement #3	245
Design Requirement #4	246
Administering Aggregations, Including Materialized Views	246
Delivering Dimensional Data to OLAP Cubes	247
Cube Data Sources	248
Processing Dimensions	248
Changes in Dimension Data	249
Processing Facts	250
Integrating OLAP Processing into the ETL System	252
OLAP Wrap-up	253
Summary	253

<b>Implementation and operations</b>	<b>255</b>
<b>Development</b>	<b>257</b>
Current Marketplace ETL Tool Suite Offerings	258
Current Scripting Languages	260
Time Is of the Essence	260
Push Me or Pull Me	261
Ensuring Transfers with Sentinels	262
Sorting Data during Preload	263
Sorting on Mainframe Systems	264
Sorting on Unix and Windows Systems	266
Trimming the Fat (Filtering)	269
Extracting a Subset of the Source File Records on Mainframe Systems	269
Extracting a Subset of the Source File Fields	270
Extracting a Subset of the Source File Records on Unix and Windows Systems	271
Extracting a Subset of the Source File Fields	273
Creating Aggregated Extracts on Mainframe Systems	274
Creating Aggregated Extracts on UNIX and Windows Systems	274
Using Database Bulk Loader Utilities to Speed Inserts	276
Preparing for Bulk Load	278
Managing Database Features to Improve Performance	280
The Order of Things	282
The Effect of Aggregates and Group Bys on Performance	286
Performance Impact of Using Scalar Functions	287
Avoiding Triggers	287
Overcoming ODBC the Bottleneck	288
Benefiting from Parallel Processing	288
Troubleshooting Performance Problems	292
Increasing ETL Throughput	294
Reducing Input/Output Contention	296
Eliminating Database Reads/Writes	296
Filtering as Soon as Possible	297
Partitioning and Parallelizing	297
Updating Aggregates Incrementally	298
Taking Only What You Need	299
Bulk Loading/Eliminating Logging	299
Dropping Databases Constraints and Indexes	299
Eliminating Network Traffic	300
Letting the ETL Engine Do the Work	300
Summary	300
<b>Operations</b>	<b>301</b>
Scheduling and Support	302
Reliability, Availability, Manageability Analysis for ETL	302
ETL Scheduling 101	303

Scheduling Tools	304
Load Dependencies	314
Metadata	314
Migrating to Production	315
Operational Support for the Data Warehouse	316
Bundling Version Releases	316
Supporting the ETL System in Production	319
Achieving Optimal ETL Performance	320
Estimating Load Time	321
Vulnerabilities of Long-Running ETL processes	324
Minimizing the Risk of Load Failures	330
Purging Historic Data	330
Monitoring the ETL System	331
Measuring ETL Specific Performance Indicators	331
Measuring Infrastructure Performance Indicators	332
Measuring Data Warehouse Usage to Help Manage ETL Processes	337
Tuning ETL Processes	339
Explaining Database Overhead	340
ETL System Security	343
Securing the Development Environment	344
Securing the Production Environment	344
Short-Term Archiving and Recovery	345
Long-Term Archiving and Recovery	346
Media, Formats, Software, and Hardware	347
Obsolete Formats and Archaic Formats	347
Hard Copy, Standards, and Museums	348
Refreshing, Migrating, Emulating, and Encapsulating	349
Summary	350
<b>Chapter 9 Metadata</b>	<b>351</b>
Defining Metadata	352
Metadata—What Is It?	352
Source System Metadata	353
Data-Staging Metadata	354
DBMS Metadata	355
Front Room Metadata	356
Business Metadata	359
Business Definitions	360
Source System Information	361
Data Warehouse Data Dictionary	362
Logical Data Maps	363
Technical Metadata	363
System Inventory	364
Data Models	365
Data Definitions	365
Business Rules	366
ETL-Generated Metadata	367

ETL Job Metadata	368
Transformation Metadata	370
Batch Metadata	373
Data Quality Error Event Metadata	374
Process Execution Metadata	375
Metadata Standards and Practices	377
Establishing Rudimentary Standards	378
Naming Conventions	379
Impact Analysis	380
Summary	380
<b>Responsibilities</b>	<b>383</b>
Planning and Leadership	383
Having Dedicated Leadership	384
Planning Large, Building Small	385
Hiring Qualified Developers	387
Building Teams with Database Expertise	387
Don't Try to Save the World	388
Enforcing Standardization	388
Monitoring, Auditing, and Publishing Statistics	389
Maintaining Documentation	389
Providing and Utilizing Metadata	390
Keeping It Simple	390
Optimizing Throughput	390
Managing the Project	391
Responsibility of the ETL Team	391
Defining the Project	392
Planning the Project	393
Determining the Tool Set	393
Staffing Your Project	394
Project Plan Guidelines	401
Managing Scope	412
Summary	416
<b>Real Time Streaming ETL Systems</b>	<b>419</b>
<b>Real-Time ETL Systems</b>	<b>421</b>
Why Real-Time ETL?	422
Defining Real-Time ETL	424
Challenges and Opportunities of Real-Time Data	
Warehousing	424
Real-Time Data Warehousing Review	425
Generation 1—The Operational Data Store	425
Generation 2—The Real-Time Partition	426
Recent CRM Trends	428
The Strategic Role of the Dimension Manager	429
Categorizing the Requirement	430

Data Freshness and Historical Needs	430
Reporting Only or Integration, Too?	432
Just the Facts or Dimension Changes, Too?	432
Alerts, Continuous Polling, or Nonevents?	433
Data Integration or Application Integration?	434
Point-to-Point versus Hub-and-Spoke	434
Customer Data Cleanup Considerations	436
Real-Time ETL Approaches	437
Microbatch ETL	437
Enterprise Application Integration	441
Capture, Transform, and Flow	444
Enterprise Information Integration	446
The Real-Time Dimension Manager	447
Microbatch Processing	452
Choosing an Approach—A Decision Guide	456
Summary	459
<b>Chapter 12 Conclusions</b>	<b>461</b>
Deepening the Definition of ETL	461
The Future of Data Warehousing and ETL in Particular	463
Ongoing Evolution of ETL Systems	464
<b>Index</b>	<b>467</b>