The Data Warehouse ETL Toolkit

Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data

> Ralph Kimball Joe Caserta

> > WILEY Wiley Publishing, Inc.

Acknowledgments

About the Authors

Introduction

Part	Requirements,	Realities, and	Architecture
------	---------------	----------------	--------------

Chapter 1	Surrounding the Requirements
	Requirements
	Business Needs
	Compliance Requirements
	Data Profiling
	Security Requirements
	Data Integration
	Data Latency
	Archiving and Lineage
	End User Delivery Interfaces
	Available Skills
	Legacy Licenses
	Architecture
	ETL Tool versus Hand Coding (Buy a Tool Suite or Roll
	Your Own?)
	The Back Room - Preparing the Data
	The Front Room - Data Access
	The Mission of the Data Warehouse
	What the Data Warehouse Is
	What the Data Warehouse Is Not
	Industry Terms Not Used Consistently

	Resolving Architectural Conflict: A Hybrid Approach	27
	How the Data Warehouse Is Changing	27
	The Mission of the ETL Team	28
Chapter 2	EIL Data Structures	29
-	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	4.7
	Room to the Front Room	45
	Fact Tables	45
	Dimension Tables	40
	Atomic and Aggregate Fact Tables	47
	Suffogate Key Mapping Tables	48
	Impact A polygic	40
	Motadata Captura	49
	Naming Conventions	51
	Auditing Data Transformation Steps	51
	Summary	52
Part II	Data Flow	53
Chapter 3	Extracting	55
-	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
	Integratine Heteroeeneous Data Sources	13
	Plan 2: The Unahenge of Extracting from Disparate	
	rial Office Connecting to Diverse Sources through ODBC	
	Mainframe Sources	
	Working with COBOL Copybooks	
	FBCDIC Character Set	
	Converting EBCDIC to ASCII	
	Converting EBCDie to Tibeli	

	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
Chanter 4	Cleaning and Conforming	
Chapter 4	Cleaning and Conforming	
Chapter 4	Cleaning and Conforming Defining Data Quality	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverable	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverable Cleaning Deliverable #1: Error Event Table	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverables Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverables Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverables Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points Part 3: Screens and Their Measurements	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverables Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points Part 3: Screens and Their Measurements Anomaly Detection Phase	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverable Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points Part 3: Screens and Their Measurements Anomaly Detection Phase Types of Enforcement	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverables Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points Part 3: Screens and Their Measurements Anomaly Detection Phase Types of Enforcement	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverables Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points Part 3: Screens and Their Measurements Anomaly Detection Phase Types of Enforcement Column Property Enforcement	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverable Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points Part 3: Screens and Their Measurements Anomaly Detection Phase Types of Enforcement Column Property Enforcement Structure Enforcement	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverables Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points Part 3: Screens and Their Measurements Anomaly Detection Phase Types of Enforcement Column Property Enforcement Structure Enforcement Data and Value Rule Enforcement Measurements Driving Screen Design	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverables Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points Part 3: Screens and Their Measurements Anomaly Detection Phase Types of Enforcement Column Property Enforcement Structure Enforcement Data and Value Rule Enforcement Measurements Driving Screen Design Overall Process Flow	
Chapter 4	Cleaning and Conforming Defining Data Quality Assumptions Part 1: Design Objectives Understand Your Key Constituencies Competing Factors Balancing Conflicting Priorities Formulate a Policy Part 2: Cleaning Deliverables Data Profiling Deliverable Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #1: Error Event Table Cleaning Deliverable #2: Audit Dimension Audit Dimension Fine Points Part 3: Screens and Their Measurements Anomaly Detection Phase Types of Enforcement Column Property Enforcement Structure Enforcement Data and Value Rule Enforcement Measurements Driving Screen Design Overall Process Flow The Show Must Go On—Usually	

	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
Chapter 5	Delivering Dimension Tables	161
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension	161 162
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension	161 162 165
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension	161 162 165 166
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions	161 162 165 166 167
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions	161 162 165 166 167 170
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions	161 162 165 166 167 170 174
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions	161 162 165 166 167 170 174 176
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two	161 162 165 166 167 170 174 176 176
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles	161 162 165 166 167 170 174 176 176 178
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension	161 162 165 166 167 170 174 176 176 178 180
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions	161 162 165 166 167 170 174 176 176 178 180 182
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimensions	161 162 165 166 167 170 174 176 176 178 180 182 183
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimension (Overwrite)	161 162 165 166 167 170 174 176 176 178 180 182 183 183
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimension (Overwrite) Type 1 Slowly Changing Dimension (Partitioning History)	161 162 165 166 167 170 174 176 176 178 180 182 183 183 183
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimensions Type 1 Slowly Changing Dimension (Overwrite) Type 2 Slowly Changing Dimension (Partitioning History) Precise Time Stamping of a Type 2 Slowly Changing	161 162 165 166 167 170 174 176 176 176 178 180 182 183 183 183
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimension (Overwrite) Type 1 Slowly Changing Dimension (Partitioning History) Precise Time Stamping of a Type 2 Slowly Changing Dimension	161 162 165 166 167 170 174 176 176 178 180 182 183 183 185 190
Chapter 5	 Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimension (Overwrite) Type 1 Slowly Changing Dimension (Partitioning History) Precise Time Stamping of a Type 2 Slowly Changing Dimension Type 3 Slowly Changing Dimension (Alternate Realities) 	161 162 165 166 167 170 174 176 176 178 180 182 183 183 185 190 192
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimension (Overwrite) Type 1 Slowly Changing Dimension (Overwrite) Type 2 Slowly Changing Dimension (Partitioning History) Precise Time Stamping of a Type 2 Slowly Changing Dimension Type 3 Slowly Changing Dimension (Alternate Realities) Hybrid Slowly Changing Dimensions	161 162 165 166 167 170 174 176 176 178 180 182 183 183 183 185 190 192
Chapter 5	Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimension (Overwrite) Type 1 Slowly Changing Dimension (Overwrite) Type 2 Slowly Changing Dimension (Partitioning History) Precise Time Stamping of a Type 2 Slowly Changing Dimension Type 3 Slowly Changing Dimension (Alternate Realities) Hybrid Slowly Changing Dimensions Late-Arriving Dimension Records and Correcting Bad Data	161 162 165 166 167 170 174 176 176 178 180 182 183 183 183 185 190 192 193 194
Chapter 5	 Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimension (Overwrite) Type 1 Slowly Changing Dimension (Partitioning History) Precise Time Stamping of a Type 2 Slowly Changing Dimension Type 3 Slowly Changing Dimension (Alternate Realities) Hybrid Slowly Changing Dimensions Late-Arriving Dimension Records and Correcting Bad Data Multivalued Dimensions and Bridge Tables 	161 162 165 166 167 170 174 176 176 178 180 182 183 183 183 185 190 192 193 194 196
Chapter 5	 Delivering Dimension Tables The Basic Structure of a Dimension The Grain of a Dimension The Grain of a Dimension The Basic Load Plan for a Dimension Flat Dimensions and Snowflaked Dimensions Date and Time Dimensions Big Dimensions Small Dimensions One Dimension or Two Dimensional Roles Dimensions as Subdimensions of Another Dimension Degenerate Dimensions Slowly Changing Dimension (Overwrite) Type 1 Slowly Changing Dimension (Overwrite) Type 2 Slowly Changing Dimension (Alternate Realities) Hybrid Slowly Changing Dimensions Late-Arriving Dimension Records and Correcting Bad Data Multivalued Dimensions and Bridge Tables 	 161 162 165 166 167 170 174 176 176 178 180 182 183 183 185 190 192 193 194 196 199

	Using Positional Attributes in a Dimension to Represent	
	Text Facts	204
	Summary	207
Chapter 6	Delivering Fact Tables	209
	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	235
	Graceful Modifications	235
	Collecting Devenue in Multiple Communics	238
	Late Arriving Facto	239
	Late Arriving Facis	241
	Aggregations	243
	Design Requirement #1	244
	Design Requirement #2	245
	Design Requirement #4	246
	Administering Aggregations, Including Materialized	
	Viewe	246
	VIEWS Delivering Dimensional Data to OLAP Cubes	240
	Cube Deta Sources	247
	Cube Data Sources Processing Dimensions	240
	Changes in Dimension Data	240 240
	Changes in Dimension Data Drocessing Facts	249 250
	Integrating OLAP Processing into the FTL System	250
	OI AP Wran-un	252
	Summary	253
	Summury	200

Part III	Implementation and operations	255
Chapter 7	Development	257
	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	280
	Associations Associations	207
	Avoiding Inggers	201
	Benefiting from Parallel Processing	288
	Troubleshooting Performance Problems	200
	Increasing FTL Throughput	292
	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
Chapter 8	Operations	301
	Scheduling and Support	302
	Reliability, Availability, Manageability Analysis for ETL ETL Scheduling 101	302 303

	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 EIL 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	554
	Measuring Data Warehouse Usage to Help Manage EIL	
	Processes	
	Tuning EIL Processes	
	Explaining Database Overhead	
	EIL System Security	
	Securing the Development Environment	
	Securing the Production Environment	
	Short-Term Archiving and Recovery	
	Long-Term Archiving and Recovery	
	Media, Formats, Software, and Hardware	
	Obsolete Formats and Archaic Formats	
	Hard Copy, Standards, and Museums	
	Refreshing, Migrating, Emulating, and Encapsulating	
	Summary	
Chapter 9	Metadata	
•	Defining Metadata	
	Metadata—What Is It?	
	Source System Metadata	
	Data-Staging Metadata	
	DBMS Metadata	
	Front Room Metadata	
	Business Metadata	
	Business Definitions	
	Source System Information	
	Data Warehouse Data Dictionary	
	Logical Data Maps	
	Technical Metadata	
	System Inventory	
	Data Models	
	Data Definitions	
	Business Rules	
	ETL-Generated Metadata	

	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
Chapter 10	Responsibilities	383
1	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
Part IV	Real Time Streaming EIL Systems	419
Chapter 11	Real-Time EIL Systems	421
	Why Real-Time ETL?	422
	Defining Real-Time ETL	424
	Challenges and Opportunities of Real-Time Data	
	Warehousing	
	Real-Time Data Warehousing Review	
	Generation 1—The Operational Data Store	
	Generation 2—The Real-Time Partition	
	Recent CRM Trends	
	The Strategic Role of the Dimension Manager	
	Categorizing the Requirement	

	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
Chapter 12	Conclusions	461
•	Deepening the Definition of ETL	461
	The Future of Data Warehousing and ETL in Particular	463
	Ongoing Evolution of ETL Systems	464
Index		467