The Data Warehouse ETL Toolkit
Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data

Ralph Kimball
Joe Caserta
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
## Chapter 2  ETL Data Structures

### How the Data Warehouse Is Changing

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
</tr>
</tbody>
</table>

### The Mission of the ETL Team

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
</tr>
</tbody>
</table>

### Resolving Architectural Conflict: A Hybrid Approach

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
</tr>
</tbody>
</table>

#### To Stage or Not to Stage

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
</tr>
</tbody>
</table>

#### Designing the Staging Area

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
</tr>
</tbody>
</table>

#### Data Structures in the ETL System

<table>
<thead>
<tr>
<th>Flat Files</th>
<th>XML Data Sets</th>
<th>Relational Tables</th>
<th>Independent DBMS Working Tables</th>
<th>Third Normal Form Entity/Relation Models</th>
<th>Nonrelational Data Sources</th>
<th>Dimensional Data Models: The Handoff from the Back Room to the Front Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>35</td>
<td>40</td>
<td>41</td>
<td>42</td>
<td>42</td>
<td>45</td>
</tr>
</tbody>
</table>

#### Fact Tables

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
</tr>
</tbody>
</table>

#### Dimension Tables

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
</tr>
</tbody>
</table>

#### Atomic and Aggregate Fact Tables

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
</tr>
</tbody>
</table>

#### Surrogate Key Mapping Tables

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
</tr>
</tbody>
</table>

#### Planning and Design Standards

<table>
<thead>
<tr>
<th>Impact Analysis</th>
<th>Metadata Capture</th>
<th>Naming Conventions</th>
<th>Auditing Data Transformation Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>49</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>

#### Summary

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
</tr>
</tbody>
</table>

## Part II  Data Flow

### Chapter 3  Extracting

#### Part 1: The Logical Data Map

<table>
<thead>
<tr>
<th>Designing Logical Before Physical</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
</tr>
</tbody>
</table>

#### Inside the Logical Data Map

<table>
<thead>
<tr>
<th>Components of the Logical Data Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
</tr>
</tbody>
</table>

#### Using Tools for the Logical Data Map

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
</tr>
</tbody>
</table>

#### Building the Logical Data Map

<table>
<thead>
<tr>
<th>Data Discovery Phase</th>
<th>Data Content Analysis</th>
<th>Collecting Business Rules in the ETL Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>71</td>
<td>73</td>
</tr>
</tbody>
</table>

#### Interreating Heterogeneous Data Sources

<table>
<thead>
<tr>
<th>Part 2: The Challenge of Extracting from Disparate Platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
</tr>
</tbody>
</table>

#### Connecting to Diverse Sources through ODBC

<table>
<thead>
<tr>
<th>Mainframe Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working with COBOL Copybooks</td>
</tr>
<tr>
<td>EBCDIC Character Set</td>
</tr>
<tr>
<td>Converting EBCDIC to ASCII</td>
</tr>
<tr>
<td>73</td>
</tr>
</tbody>
</table>
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

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 #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
Screens
Contents

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
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
Chapter 6 Delivering Fact Tables

The Basic Structure of a Fact Table
Guaranteeing Referential Integrity
Surrogate Key Pipeline
    Using the Dimension Instead of a Lookup Table
Fundamental Grains
    Transaction Grain Fact Tables
    Periodic Snapshot Fact Tables
    Accumulating Snapshot Fact Tables
Preparing for Loading Fact Tables
    Managing Indexes
    Managing Partitions
    Outwitting the Rollback Log
    Loading the Data
    Incremental Loading
    Inserting Facts
    Updating and Correcting Facts
    Negating Facts
    Updating Facts
    Deleting Facts
    Physically Deleting Facts
    Logically Deleting Facts
Factless Fact Tables
Augmenting a Type 1 Fact Table with Type 2 History
Graceful Modifications
Multiple Units of Measure in a Fact Table
Collecting Revenue in Multiple Currencies
Late Arriving Facts
Aggregations
    Design Requirement #1
    Design Requirement #2
    Design Requirement #3
    Design Requirement #4
    Administering Aggregations, Including Materialized Views
Delivering Dimensional Data to OLAP Cubes
    Cube Data Sources
    Processing Dimensions
    Changes in Dimension Data
    Processing Facts
    Integrating OLAP Processing into the ETL System
    OLAP Wrap-up
Summary
<table>
<thead>
<tr>
<th>Part III</th>
<th>Implementation and operations</th>
<th>255</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 7</td>
<td>Development</td>
<td>257</td>
</tr>
<tr>
<td></td>
<td>Current Marketplace ETL Tool Suite Offerings</td>
<td>258</td>
</tr>
<tr>
<td></td>
<td>Current Scripting Languages</td>
<td>260</td>
</tr>
<tr>
<td></td>
<td>Time Is of the Essence</td>
<td>260</td>
</tr>
<tr>
<td></td>
<td>Push Me or Pull Me</td>
<td>261</td>
</tr>
<tr>
<td></td>
<td>Ensuring Transfers with Sentinels</td>
<td>262</td>
</tr>
<tr>
<td></td>
<td>Sorting Data during Preload</td>
<td>263</td>
</tr>
<tr>
<td></td>
<td>Sorting on Mainframe Systems</td>
<td>264</td>
</tr>
<tr>
<td></td>
<td>Sorting on Unix and Windows Systems</td>
<td>266</td>
</tr>
<tr>
<td></td>
<td>Trimming the Fat (Filtering)</td>
<td>269</td>
</tr>
<tr>
<td></td>
<td>Extracting a Subset of the Source File Records on Mainframe Systems</td>
<td>269</td>
</tr>
<tr>
<td></td>
<td>Extracting a Subset of the Source File Fields</td>
<td>270</td>
</tr>
<tr>
<td></td>
<td>Extracting a Subset of the Source File Records on Unix and Windows Systems</td>
<td>271</td>
</tr>
<tr>
<td></td>
<td>Extracting a Subset of the Source File Fields</td>
<td>273</td>
</tr>
<tr>
<td></td>
<td>Creating Aggregated Extracts on Mainframe Systems</td>
<td>274</td>
</tr>
<tr>
<td></td>
<td>Creating Aggregated Extracts on UNIX and Windows Systems</td>
<td>274</td>
</tr>
<tr>
<td></td>
<td>Using Database Bulk Loader Utilities to Speed Inserts</td>
<td>276</td>
</tr>
<tr>
<td></td>
<td>Preparing for Bulk Load</td>
<td>278</td>
</tr>
<tr>
<td></td>
<td>Managing Database Features to Improve Performance</td>
<td>280</td>
</tr>
<tr>
<td></td>
<td>The Order of Things</td>
<td>282</td>
</tr>
<tr>
<td></td>
<td>The Effect of Aggregates and Group Bys on Performance</td>
<td>286</td>
</tr>
<tr>
<td></td>
<td>Performance Impact of Using Scalar Functions</td>
<td>287</td>
</tr>
<tr>
<td></td>
<td>Avoiding Triggers</td>
<td>287</td>
</tr>
<tr>
<td></td>
<td>Overcoming ODBC the Bottleneck</td>
<td>288</td>
</tr>
<tr>
<td></td>
<td>Benefiting from Parallel Processing</td>
<td>288</td>
</tr>
<tr>
<td></td>
<td>Troubleshooting Performance Problems</td>
<td>292</td>
</tr>
<tr>
<td></td>
<td>Increasing ETL Throughput</td>
<td>294</td>
</tr>
<tr>
<td></td>
<td>Reducing Input/Output Contention</td>
<td>296</td>
</tr>
<tr>
<td></td>
<td>Eliminating Database Reads/Writes</td>
<td>296</td>
</tr>
<tr>
<td></td>
<td>Filtering as Soon as Possible</td>
<td>297</td>
</tr>
<tr>
<td></td>
<td>Partitioning and Parallelizing</td>
<td>297</td>
</tr>
<tr>
<td></td>
<td>Updating Aggregates Incrementally</td>
<td>298</td>
</tr>
<tr>
<td></td>
<td>Taking Only What You Need</td>
<td>299</td>
</tr>
<tr>
<td></td>
<td>Bulk Loading/Eliminating Logging</td>
<td>299</td>
</tr>
<tr>
<td></td>
<td>Dropping Databases Constraints and Indexes</td>
<td>299</td>
</tr>
<tr>
<td></td>
<td>Eliminating Network Traffic</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>Letting the ETL Engine Do the Work</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>300</td>
</tr>
</tbody>
</table>

| Chapter 8 | Operations | 301 |
|  | Operations | 302 |
|  | Scheduling and Support | 302 |
|  | Reliability, Availability, Manageability Analysis for ETL | 302 |
|  | ETL Scheduling 101 | 303 |
## 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
Planning and Leadership
- 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
- 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 ETL Systems** | 419

**Chapter 11 Real-Time ETL 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
Reporting Only or Integration, Too?
Just the Facts or Dimension Changes, Too?
Alerts, Continuous Polling, or Nonevents?
Data Integration or Application Integration?
Point-to-Point versus Hub-and-Spoke
Customer Data Cleanup Considerations
Real-Time ETL Approaches
  Microbatch ETL
  Enterprise Application Integration
  Capture, Transform, and Flow
  Enterprise Information Integration
  The Real-Time Dimension Manager
  Microbatch Processing
  Choosing an Approach—A Decision Guide
Summary

**Chapter 12 Conclusions**
Deepening the Definition of ETL
The Future of Data Warehousing and ETL in Particular
  Ongoing Evolution of ETL Systems

**Index**