Introduction xxxi

Part I  Getting Started  1

Chapter 1  ETL Primer  3
OLTP versus Data Warehousing  3
What Is ETL?  5
The Evolution of ETL Solutions  5
ETL Building Blocks  7
ETL, ELT, and EII  8
ELT  9
EII: Virtual Data Integration  10
Data Integration Challenges  11
Methodology: Agile BI  12
ETL Design  14
Data Acquisition  14
Beware of Spreadsheets  15
Design for Failure  15
Change Data Capture  16
Data Quality  16
Data Profiling  16
Data Validation  17
ETL Tool Requirements  17
Connectivity  17
Platform Independence  18
Scalability  18
Design Flexibility  19
Reuse  19
Extensibility  19
Chapter 2  Kettle Concepts  23

Design Principles  23
The Building Blocks of Kettle Design  25
  Transformations  25
    Steps  26
    Transformation Hops  26
    Parallelism  27
    Rows of Data  27
    Data Conversion  29
  Jobs  30
    Job Entries  31
    Job Hops  31
    Multiple Paths and Backtracking  32
    Parallel Execution  33
    Job Entry Results  34
  Transformation or Job Metadata  36
  Database Connections  37
    Special Options  38
    The Power of the Relational Database  39
    Connections and Transactions  39
    Database Clustering  40
  Tools and Utilities  41
  Repositories  41
  Virtual File Systems  42
Parameters and Variables  43
  Defining Variables  43
  Named Parameters  44
  Using Variables  44
Visual Programming  45
  Getting Started  46
  Creating New Steps  47
  Putting It All Together  49
Summary  51

Chapter 3  Installation and Configuration  53

Kettle Software Overview  53
  Integrated Development Environment: Spoon  55
  Command-Line Launchers: Kitchen and Pan  57
  Job Server: Carte  57
  Encr.bat and encr.sh  58
Installation  58
Java Environment 58
  Installing Java Manually 58
  Using Your Linux Package Management System 59
Installing Kettle 59
  Versions and Releases 59
  Archive Names and Formats 60
  Downloading and Uncompressing 60
  Running Kettle Programs 61
  Creating a Shortcut Icon or Launcher for Spoon 62
Configuration 63
  Configuration Files and the .kettle Directory 63
  The Kettle Shell Scripts 69
    General Structure of the Startup Scripts 70
    Adding an Entry to the Classpath 70
    Changing the Maximum Heap Size 71
  Managing JDBC Drivers 72
Summary 72

Chapter 4  An Example ETL Solution—Sakila 73
Sakila 73
  The Sakila Sample Database 74
    DVD Rental Business Process 74
    Sakila Database Schema Diagram 75
    Sakila Database Subject Areas 75
    General Design Considerations 77
    Installing the Sakila Sample Database 77
  The Rental Star Schema 78
    Rental Star Schema Diagram 78
    Rental Fact Table 79
    Dimension Tables 79
    Keys and Change Data Capture 80
    Installing the Rental Star Schema 81
Prerequisites and Some Basic Spoon Skills 81
  Setting Up the ETL Solution 82
    Creating Database Accounts 82
  Working with Spoon 82
    Opening Transformation and Job Files 82
    Opening the Step's Configuration Dialog 83
    Examining Streams 83
    Running Jobs and Transformations 83
The Sample ETL Solution 84
  Static, Generated Dimensions 84
    Loading the dim_date Dimension Table 84
    Loading the dim_time Dimension Table 86
Recurring Load 87
  The load_rentals Job 88
### Contents

- The load_dim_staff Transformation 91
- Database Connections 91
- The load_dim_customer Transformation 95
- The load_dim_store Transformation 98
- The fetch_address Subtransformation 99
- The load_dim_actor Transformation 101
- The load_dim_film Transformation 102
- The load_fact_rental Transformation 107

**Summary** 109

**Part II**  ETL 111

**Chapter 5**  ETL Subsystems 113

- Introduction to the 34 Subsystems 114
- Extraction 114
  - Subsystems 1–3: Data Profiling, Change Data Capture, and Extraction 115
- Cleaning and Conforming Data 116
  - Subsystem 4: Data Cleaning and Quality Screen Handler System 116
  - Subsystem 5: Error Event Handler 117
  - Subsystem 6: Audit Dimension Assembler 117
  - Subsystem 7: Deduplication System 117
  - Subsystem 8: Data Conformer 118
- Data Delivery 118
  - Subsystem 9: Slowly Changing Dimension Processor 118
  - Subsystem 10: Surrogate Key Creation System 119
  - Subsystem 11: Hierarchy Dimension Builder 119
  - Subsystem 12: Special Dimension Builder 120
  - Subsystem 13: Fact Table Loader 121
  - Subsystem 14: Surrogate Key Pipeline 121
  - Subsystem 15: Multi-Valued Dimension Bridge Table Builder 121
  - Subsystem 16: Late-Arriving Data Handler 122
  - Subsystem 17: Dimension Manager System 122
  - Subsystem 18: Fact Table Provider System 122
  - Subsystem 19: Aggregate Builder 123
  - Subsystem 20: Multidimensional (OLAP) Cube Builder 123
  - Subsystem 21: Data Integration Manager 123
- Managing the ETL Environment 123

**Summary** 126

**Chapter 6**  Data Extraction 127

- Kettle Data Extraction Overview 128
- File-Based Extraction 128
  - Working with Text Files 128
  - Working with XML files 133
- Special File Types 134
Contents

Database-Based Extraction 134
Web-Based Extraction 137
   Text-Based Web Extraction 137
   HTTP Client 137
   Using SOAP 138
Stream-Based and Real-Time Extraction 138
Working with ERP and CRM Systems 138
   ERP Challenges 139
   Kettle ERP Plugins 140
Working with SAP Data 140
ERP and CDC Issues 146
Data Profiling 146
   Using eobjects.org DataCleaner 147
      Adding Profile Tasks 149
      Adding Database Connections 149
      Doing an Initial Profile 151
   Working with Regular Expressions 151
   Profiling and Exploring Results 152
   Validating and Comparing Data 153
   Using a Dictionary for Column Dependency Checks 153
   Alternative Solutions 154
   Text Profiling with Kettle 154
CDC: Change Data Capture 154
   Source Data–Based CDC 155
   Trigger-Based CDC 157
   Snapshot-Based CDC 158
   Log-Based CDC 162
   Which CDC Alternative Should You Choose? 163
Delivering Data 164
Summary 164

Chapter 7 Cleansing and Conforming 167
Data Cleansing 168
   Data-Cleansing Steps 169
   Using Reference Tables 172
      Conforming Data Using Lookup Tables 172
      Conforming Data Using Reference Tables 175
   Data Validation 179
      Applying Validation Rules 180
      Validating Dependency Constraints 183
Error Handling 183
   Handling Process Errors 184
   Transformation Errors 186
   Handling Data (Validation) Errors 187
Auditing Data and Process Quality 191
Deduplicating Data 192
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handling Exact Duplicates</td>
<td>193</td>
</tr>
<tr>
<td>The Problem of Non-Exact Duplicates</td>
<td>194</td>
</tr>
<tr>
<td>Building Deduplication Transforms</td>
<td>195</td>
</tr>
<tr>
<td>Step 1: Fuzzy Match</td>
<td>197</td>
</tr>
<tr>
<td>Step 2: Select Suspects</td>
<td>198</td>
</tr>
<tr>
<td>Step 3: Lookup Validation Value</td>
<td>198</td>
</tr>
<tr>
<td>Step 4: Filter Duplicates</td>
<td>199</td>
</tr>
<tr>
<td>Scripting</td>
<td>200</td>
</tr>
<tr>
<td>Formula</td>
<td>201</td>
</tr>
<tr>
<td>JavaScript</td>
<td>202</td>
</tr>
<tr>
<td>User-Defined Java Expressions</td>
<td>202</td>
</tr>
<tr>
<td>Regular Expressions</td>
<td>203</td>
</tr>
<tr>
<td>Summary</td>
<td>205</td>
</tr>
<tr>
<td>Chapter 8 Handling Dimension Tables</td>
<td>207</td>
</tr>
<tr>
<td>Managing Keys</td>
<td>208</td>
</tr>
<tr>
<td>Managing Business Keys</td>
<td>209</td>
</tr>
<tr>
<td>Keys in the Source System</td>
<td>209</td>
</tr>
<tr>
<td>Keys in the Data Warehouse</td>
<td>209</td>
</tr>
<tr>
<td>Business Keys</td>
<td>209</td>
</tr>
<tr>
<td>Storing Business Keys</td>
<td>210</td>
</tr>
<tr>
<td>Looking Up Keys with Kettle</td>
<td>210</td>
</tr>
<tr>
<td>Generating Surrogate Keys</td>
<td>210</td>
</tr>
<tr>
<td>The “Add sequence” Step</td>
<td>211</td>
</tr>
<tr>
<td>Working with auto_increment or IDENTITY Columns</td>
<td>217</td>
</tr>
<tr>
<td>Keys for Slowly Changing Dimensions</td>
<td>217</td>
</tr>
<tr>
<td>Loading Dimension Tables</td>
<td>218</td>
</tr>
<tr>
<td>Snowflaked Dimension Tables</td>
<td>218</td>
</tr>
<tr>
<td>Top-Down Level-Wise Loading</td>
<td>219</td>
</tr>
<tr>
<td>Sakila Snowflake Example</td>
<td>219</td>
</tr>
<tr>
<td>Sample Transformation</td>
<td>221</td>
</tr>
<tr>
<td>Database Lookup Configuration</td>
<td>222</td>
</tr>
<tr>
<td>Sample Job</td>
<td>225</td>
</tr>
<tr>
<td>Star Schema Dimension Tables</td>
<td>226</td>
</tr>
<tr>
<td>Denormalization</td>
<td>226</td>
</tr>
<tr>
<td>Denormalizing to 1NF with the “Database lookup” Step</td>
<td>226</td>
</tr>
<tr>
<td>Change Data Capture</td>
<td>227</td>
</tr>
<tr>
<td>Slowly Changing Dimensions</td>
<td>228</td>
</tr>
<tr>
<td>Types of Slowly Changing Dimensions</td>
<td>228</td>
</tr>
<tr>
<td>Type 1 Slowly Changing Dimensions</td>
<td>229</td>
</tr>
<tr>
<td>The Insert / Update Step</td>
<td>229</td>
</tr>
<tr>
<td>Type 2 Slowly Changing Dimensions</td>
<td>232</td>
</tr>
<tr>
<td>The “Dimension lookup / update” Step</td>
<td>232</td>
</tr>
<tr>
<td>Other Types of Slowly Changing Dimensions</td>
<td>237</td>
</tr>
<tr>
<td>Type 3 Slowly Changing Dimensions</td>
<td>237</td>
</tr>
<tr>
<td>Hybrid Slowly Changing Dimensions</td>
<td>238</td>
</tr>
</tbody>
</table>
Part III Management and Deployment 293

Chapter 11 ETL Development Lifecycle 295
Solution Design 295
  Best and Bad Practices 296
  Data Mapping 297
  Naming and Commentary Conventions 298
  Common Pitfalls 299
ETL Flow Design 300
  Reusability and Maintainability 300
Agile Development 301
Testing and Debugging 306
  Test Activities 307
ETL Testing 308
  Test Data Requirements 308
  Testing for Completeness 309
  Testing Data Transformations 311
  Test Automation and Continuous Integration 311
Upgrade Tests 312
Debugging 312
Documenting the Solution 315
  Why Isn’t There Any Documentation? 316
    Myth 1: My Software Is Self-Explanatory 316
    Myth 2: Documentation Is Always Outdated 316
Kettle Documentation Features 317
  Generating Documentation 319
Summary 320

Chapter 12 Scheduling and Monitoring 321
Scheduling 321
  Operating System–Level Scheduling 322
    Executing Kettle Jobs and Transformations from the Command Line 322
    UNIX-Based Systems: cron 326
    Windows: The at utility and the Task Scheduler 327
Using Pentaho’s Built-in Scheduler 327
  Creating an Action Sequence to Run Kettle Jobs and Transformations 328
  Kettle Transformations in Action Sequences 329
Creating and Maintaining Schedules with the Administration Console 330
Attaching an Action Sequence to a Schedule 333
Monitoring 333
  Logging 333
    Inspecting the Log 333
Logging Levels 335
Writing Custom Messages to the Log 336
E-mail Notifications 336
Configuring the Mail Job Entry 337
Summary 340

**Chapter 13 Versioning and Migration** 341
Version Control Systems 341
File-Based Version Control Systems 342
Organization 342
Leading File-Based VCSs 343
Content Management Systems 344
Kettle Metadata 344
Kettle XML Metadata 345
Transformation XML 345
Job XML 346
Global Replace 347
Kettle Repository Metadata 348
The Kettle Database Repository Type 348
The Kettle File Repository Type 349
The Kettle Enterprise Repository Type 350
Managing Repositories 350
Exporting and Importing Repositories 350
Upgrading Your Repository 351
Version Migration System 352
Managing XML Files 352
Managing Repositories 352
Parameterizing Your Solution 353
Summary 356

**Chapter 14 Lineage and Auditing** 357
Batch-Level Lineage Extraction 358
Lineage 359
Lineage Information 359
Impact Analysis Information 361
Logging and Operational Metadata 363
Logging Basics 363
Logging Architecture 364
Setting a Maximum Buffer Size 365
Setting a Maximum Log Line Age 365
Log Channels 366
Log Text Capturing in a Job 366
Logging Tables 367
Transformation Logging Tables 367
Job Logging Tables 373
Summary 374
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part IV</td>
<td>375</td>
</tr>
<tr>
<td>Chapter 15</td>
<td>377</td>
</tr>
<tr>
<td>Performance Tuning</td>
<td>377</td>
</tr>
<tr>
<td>Transformation Performance: Finding the Weakest Link</td>
<td>377</td>
</tr>
<tr>
<td>Finding Bottlenecks by Simplifying</td>
<td>379</td>
</tr>
<tr>
<td>Finding Bottlenecks by Measuring</td>
<td>380</td>
</tr>
<tr>
<td>Copying Rows of Data</td>
<td>382</td>
</tr>
<tr>
<td>Improving Transformation Performance</td>
<td>384</td>
</tr>
<tr>
<td>Improving Performance in Reading Text Files</td>
<td>384</td>
</tr>
<tr>
<td>Using Lazy Conversion for Reading Text Files</td>
<td>385</td>
</tr>
<tr>
<td>Single-File Parallel Reading</td>
<td>385</td>
</tr>
<tr>
<td>Multi-File Parallel Reading</td>
<td>386</td>
</tr>
<tr>
<td>Configuring the NIO Block Size</td>
<td>386</td>
</tr>
<tr>
<td>Changing Disks and Reading Text Files</td>
<td>386</td>
</tr>
<tr>
<td>Improving Performance in Writing Text Files</td>
<td>387</td>
</tr>
<tr>
<td>Using Lazy Conversion for Writing Text Files</td>
<td>387</td>
</tr>
<tr>
<td>Parallel Files Writing</td>
<td>387</td>
</tr>
<tr>
<td>Changing Disks and Writing Text Files</td>
<td>387</td>
</tr>
<tr>
<td>Improving Database Performance</td>
<td>388</td>
</tr>
<tr>
<td>Avoiding Dynamic SQL</td>
<td>388</td>
</tr>
<tr>
<td>Handling Roundtrips</td>
<td>388</td>
</tr>
<tr>
<td>Handling Relational Databases</td>
<td>390</td>
</tr>
<tr>
<td>Sorting Data</td>
<td>392</td>
</tr>
<tr>
<td>Sorting on the Database</td>
<td>393</td>
</tr>
<tr>
<td>Sorting in Parallel</td>
<td>393</td>
</tr>
<tr>
<td>Reducing CPU Usage</td>
<td>394</td>
</tr>
<tr>
<td>Optimizing the Use of JavaScript</td>
<td>394</td>
</tr>
<tr>
<td>Launching Multiple Copies of a Step</td>
<td>396</td>
</tr>
<tr>
<td>Selecting and Removing Values</td>
<td>397</td>
</tr>
<tr>
<td>Managing Thread Priorities</td>
<td>397</td>
</tr>
<tr>
<td>Adding Static Data to Rows of Data</td>
<td>397</td>
</tr>
<tr>
<td>Limiting the Number of Step Copies</td>
<td>398</td>
</tr>
<tr>
<td>Avoiding Excessive Logging</td>
<td>398</td>
</tr>
<tr>
<td>Improving Job Performance</td>
<td>399</td>
</tr>
<tr>
<td>Loops in Jobs</td>
<td>399</td>
</tr>
<tr>
<td>Database Connection Pools</td>
<td>400</td>
</tr>
<tr>
<td>Summary</td>
<td>401</td>
</tr>
<tr>
<td>Chapter 16</td>
<td>403</td>
</tr>
<tr>
<td>Parallelization, Clustering, and Partitioning</td>
<td>403</td>
</tr>
<tr>
<td>Multi-Threading</td>
<td>403</td>
</tr>
<tr>
<td>Row Distribution</td>
<td>404</td>
</tr>
<tr>
<td>Row Merging</td>
<td>405</td>
</tr>
<tr>
<td>Row Redistribution</td>
<td>406</td>
</tr>
<tr>
<td>Data Pipelining</td>
<td>407</td>
</tr>
<tr>
<td>Consequences of Multi-Threading</td>
<td>408</td>
</tr>
<tr>
<td>Database Connections</td>
<td>408</td>
</tr>
</tbody>
</table>
Order of Execution 409
Parallel Execution in a Job 411
Using Carte as a Slave Server 411
The Configuration File 411
Defining Slave Servers 412
Remote Execution 413
Monitoring Slave Servers 413
Carte Security 414
Services 414
Clustering Transformations 417
Defining a Cluster Schema 417
Designing Clustered Transformations 418
Execution and Monitoring 420
Metadata Transformations 421
Rules 422
Data Pipelining 425
Partitioning 425
Defining a Partitioning Schema 425
Objectives of Partitioning 427
Implementing Partitioning 428
Internal Variables 428
Database Partitions 429
Partitioning in a Clustered Transformation 430
Summary 430

Chapter 17 Dynamic Clustering in the Cloud 433
Dynamic Clustering 433
Setting Up a Dynamic Cluster 434
Using the Dynamic Cluster 436
Cloud Computing 437
EC2 438
Getting Started with EC2 438
Costs 438
Customizing an AMI 439
Packaging a New AMI 442
Terminating an AMI 442
Running a Master 442
Running the Slaves 443
Using the EC2 Cluster 444
Monitoring 445
The Lightweight Principle and Persistence Options 446
Summary 447

Chapter 18 Real-Time Data Integration 449
Introduction to Real-Time ETL 449
Real-Time Challenges 450
Requirements 451
Transformation Streaming 452
A Practical Example of Transformation Streaming 454
Debugging 457
Third-Party Software and Real-Time Integration 458
Java Message Service 459
Creating a JMS Connection and Session 459
Consuming Messages 460
Producing Messages 460
Closing Shop 460
Summary 461

Part V Advanced Topics 463

Chapter 19 Data Vault Management 465
Introduction to Data Vault Modeling 466
Do You Need a Data Vault? 466
Data Vault Building Blocks 467
Hubs 467
Links 468
Satellites 469
Data Vault Characteristics 471
Building a Data Vault 471
Transforming Sakila to the Data Vault Model 472
Sakila Hubs 472
Sakila Links 473
Sakila Satellites 474
Loading the Data Vault: A Sample ETL Solution 477
Installing the Sakila Data Vault 477
Setting Up the ETL Solution 477
Creating a Database Account 477
The Sample ETL Data Vault Solution 478
Sample Hub: hub_actor 478
Sample Link: link_customer_store 480
Sample Satellite: sat_actor 483
Loading the Data Vault Tables 485
Updating a Data Mart from a Data Vault 486
The Sample ETL Solution 486
The dim_actor Transformation 486
The dim_customer Transformation 488
The dim_film Transformation 492
The dim_film_actor_bridge Transformation 492
The fact_rental Transformation 493
Loading the Star Schema Tables 495
Summary 495
Chapter 20  Handling Complex Data Formats  497
Non-Relational and Non-Tabular Data Formats  498
Non-Relational Tabular Formats  498
Handling Multi-Valued Attributes  498
Using the Split Field to Rows Step  499
Handling Repeating Groups  500
Using the Row Normaliser Step  500
Semi- and Unstructured Data  501
Kettle Regular Expression Example  503
Configuring the Regex Evaluation Step  504
Verifying the Match  507
Key/Value Pairs  508
Kettle Key/Value Pairs Example  509
Text File Input  509
Regex Evaluation  510
Grouping Lines into Records  511
Denormaliser: Turning Rows into Columns  512
Summary  513

Chapter 21  Web Services  515
Web Pages and Web Services  515
Kettle Web Features  516
General HTTP Steps  516
Simple Object Access Protocol  517
Really Simple Syndication  517
Apache Virtual File System Integration  517
Data Formats  517
XML  518
Kettle Steps for Working with XML  518
Kettle Job Entries for XML  519
HTML  520
JavaScript Object Notation  520
Syntax  521
JSON, Kettle, and ETL/DI  522
XML Examples  523
Example XML Document  523
XML Document Structure  523
Mapping to the Sakila Sample Database  524
Extracting Data from XML  525
Overall Design: The import_xml_into_db Transformation  526
Using the XSD Validator Step  528
Using the “Get Data from XML” Step  530
Generating XML Documents  537
Overall Design: The export_xml_from_db Transformation  537
Generating XML with the Add XML Step  538
Using the XML Join Step  541
Chapter 22  Kettle Integration

The Kettle API

The LGPL License

The Kettle Java API

Source Code

Building Kettle

Building javadoc

Libraries and the Class Path

Executing Existing Transformations and Jobs

Executing a Transformation

Executing a Job

Embedding Kettle

Pentaho Reporting

Putting Data into a Transformation

Dynamic Transformations

Dynamic Template

Dynamic Jobs

Executing Dynamic ETL in Kettle

Result

Replacing Metadata

Direct Changes with the API

Using a Shared Objects File
Chapter 23 Extending Kettle

Plugin Architecture Overview

Plugin Types

Architecture

Prerequisites

Kettle API Documentation

Libraries

Integrated Development Environment

Eclipse Project Setup

Examples

Transformation Step Plugins

StepMetaInterface

Value Metadata

Row Metadata

StepDataInterface

StepDialogInterface

Eclipse SWT

Form Layout

Kettle UI Elements

Hello World Example Dialog

StepInterface

Reading Rows from Specific Steps

Writing Rows to Specific Steps

Writing Rows to Error Handling

Identifying a Step Copy

Result Feedback

Variable Substitution

Apache VFS

Step Plugin Deployment

The User-Defined Java Class Step

Passing Metadata

Accessing Input and Fields

Snippets

Example

Job Entry Plugins

JobEntryInterface

JobEntryDialogInterface

Partitioning Method Plugins

Partitioner

Repository Type Plugins

Database Type Plugins

Summary