



Contents

Introduction	xxxiii
Part I Getting Started with Pentaho	1
Chapter 1 Quick Start: Pentaho Examples	3
Getting Started with Pentaho	3
Downloading and Installing the Software	4
Running the Software	5
Starting the Pentaho BI Server	5
Logging in	6
Mantle, the Pentaho User Console	7
Working with the Examples	8
Using the Repository Browser	9
Understanding the Examples	9
Running the Examples	11
Reporting Examples	11
BI Developer Examples: Regional Sales - HTML	11
Steel Wheels: Income Statement	12
Steel Wheels: Top 10 Customers	13
BI Developer Examples:	
button-single-parameter.prpt	13
Charting Examples	14
Steel Wheels: Chart Pick List	15

Steel Wheels: Flash Chart List	15
BI Developer Examples: Regional Sales - Line/Bar Chart	16
Analysis Examples	16
BI Developer Examples: Slice and Dice	17
Steel Wheels Analysis Examples	18
Dashboarding Examples	19
Other Examples	20
Summary	20
Chapter 2 Prerequisites	21
Basic System Setup	22
Installing Ubuntu	22
Using Ubuntu in Native Mode	23
Using a Virtual Machine	23
Working with the Terminal	24
Directory Navigation	24
Command History	25
Using Symbolic Links	25
Creating Symbolic Links in Ubuntu	26
Creating Symlinks in Windows Vista	26
Java Installation and Configuration	27
Installing Java on Ubuntu Linux	27
Installing Java on Windows	28
MySQL Installation	29
Installing MySQL Server and Client on Ubuntu	29
Installing MySQL Server and Client on Windows	30
MySQL GUI Tools	31
Ubuntu Install	31
Windows Install	31
Database Tools	31
Power*Architect and Other Design Tools	31
Squirrel SQL Client	32
Ubuntu Install	32
Windows Install	33
SQLeonado	33
Summary	34

Chapter 3	Server Installation and Configuration	37
	Server Configuration	37
	Installation	38
	Installation Directory	38
	User Account	38
	Configuring Tomcat	39
	Automatic Startup	40
	Managing Database Drivers	44
	Driver Location for the Server	44
	Driver Location for the Administration Console	44
	Managing JDBC Drivers on UNIX-Based Systems	44
	System Databases	45
	Setting Up the MySQL Schemas	46
	Configuring Quartz and Hibernate	46
	Configuring JDBC Security	50
	Sample Data	51
	Modify the Pentaho Startup Scripts	51
	E-mail	52
	Basic SMTP Configuration	52
	Secure SMTP Configuration	54
	Testing E-mail Configuration	54
	Publisher Password	54
	Administrative Tasks	55
	The Pentaho Administration Console	55
	Basic PAC Configuration	55
	Starting and Stopping PAC	56
	The PAC Front End	56
	Configuring PAC Security and Credentials	57
	User Management	58
	Data Sources	60
	Other Administrative Tasks	61
	Summary	61
Chapter 4	The Pentaho BI Stack	63
	Pentaho BI Stack Perspectives	65
	Functionality	65
	Server, Web Client, and Desktop Programs	65

xviii Contents

Front-Ends and Back-Ends	66	
Underlying Technology	66	
The Pentaho Business Intelligence Server	67	
The Platform	67	
The Solution Repository and the Solution Engine	68	
Database Connection Pool Management	69	
User Authentication and Authorization	69	
Task Scheduling	69	
E-mail Services	70	
BI Components	70	
The Metadata Layer	70	
Ad hoc Reporting Service	72	
The ETL Engine	72	
Reporting Engines	72	
The OLAP Engine	72	
The Data Mining Engine	72	
The Presentation Layer	73	
Underlying Java Servlet Technology	74	
Desktop Programs	74	
Pentaho Enterprise Edition and Community Edition	76	
Creating Action Sequences with Pentaho Design Studio	77	
Pentaho Design Studio (Eclipse) Primer	78	
The Action Sequence Editor	80	
Anatomy of an Action Sequence	83	
Inputs	83	
Outputs	85	
Actions	85	
Summary	89	
Part II		
Dimensional Modeling and Data Warehouse Design	91	
Chapter 5	Example Business Case: World Class Movies	93
World Class Movies: The Basics	94	
The WCM Data	95	
Obtaining and Generating Data	97	
WCM Database: The Big Picture	97	

DVD Catalog	99
Customers	101
Employees	101
Purchase Orders	101
Customer Orders and Promotions	102
Inventory Management	104
Managing the Business: The Purpose of Business	
Intelligence	105
Typical Business Intelligence Questions for WCM	108
Data Is Key	109
Summary	110
Chapter 6 Data Warehouse Primer	111
Why Do You Need a Data Warehouse?	112
The Big Debate: Inmon Versus Kimball	114
Data Warehouse Architecture	116
The Staging Area	118
The Central Data Warehouse	119
Data Marts	121
OLAP Cubes	121
Storage Formats and MDX	122
Data Warehouse Challenges	123
Data Quality	124
Data Vault and Data Quality	125
Using Reference and Master Data	127
Data Volume and Performance	128
Open Source Database Window Support	132
Changed Data Capture	133
Source Data-Based CDC	133
Trigger-Based CDC	134
Snapshot-Based CDC	135
Log-Based CDC	136
Which CDC Alternative Should You Choose?	137
Changing User Requirements	137
Data Warehouse Trends	139
Virtual Data Warehousing	139
Real-Time Data Warehousing	140
Analytical Databases	142

Data Warehouse Appliances	143
On Demand Data Warehousing	144
Summary	144
Chapter 7 Modeling the Business Using Star Schemas	147
What Is a Star Schema?	147
Dimension Tables and Fact Tables	148
Fact Table Types	149
Querying Star Schemas	150
Join Types	153
Applying Restrictions in a Query	156
Combining Multiple Restrictions	157
Restricting Aggregate Results	157
Ordering Data	158
The Bus Architecture	158
Design Principles	160
Using Surrogate Keys	160
Naming and Type Conventions	162
Granularity and Aggregation	163
Audit Columns	164
Modeling Date and Time	165
Time Dimension Granularity	165
Local Versus UTC Time	165
Smart Date Keys	166
Handling Relative Time	166
Unknown Dimension Keys	169
Handling Dimension Changes	169
SCD Type 1: Overwrite	171
SCD Type 2: Add Row	171
SCD Type 3: Add Column	174
SCD Type 4: Mini-Dimensions	174
SCD Type 5: Separate History Table	176
SCD Type 6: Hybrid Strategies	178
Advanced Dimensional Model Concepts	179
Monster Dimensions	179
Junk, Heterogeneous, and Degenerate Dimensions	180
Role-Playing Dimensions	181



	Multi-Valued Dimensions and Bridge Tables	182
	Building Hierarchies	184
	Snowflakes and Clustering Dimensions	186
	Outriggers	188
	Consolidating Multi-Grain Tables	188
	Summary	189
Chapter 8	The Data Mart Design Process	191
	Requirements Analysis	191
	Getting the Right Users Involved	192
	Collecting Requirements	193
	Data Analysis	195
	Data Profiling	197
	Using eobjects.org DataCleaner	198
	Adding Profile Tasks	200
	Adding Database Connections	201
	Doing an Initial Profile	202
	Working with Regular Expressions	202
	Profiling and Exploring Results	204
	Validating and Comparing Data	205
	Using a Dictionary for Column Dependency	
	Checks	205
	Alternative Solutions	205
	Developing the Model	206
	Data Modeling with Power*Architect	208
	Building the WCM Data Marts	210
	Generating the Database	212
	Generating Static Dimensions	213
	Special Date Fields and Calculations	216
	Source to Target Mapping	218
	Summary	220
Part III	ETL and Data Integration	221
Chapter 9	Pentaho Data Integration Primer	223
	Data Integration Overview	223
	Data Integration Activities	224
	Extraction	226



xxii Contents

Change Data Capture	226
Data Staging	226
Data Validation	227
Data Cleansing	228
Decoding and Renaming	228
Key Management	229
Aggregation	229
Dimension and Bridge Table Maintenance	229
Loading Fact Tables	230
Pentaho Data Integration Concepts and Components	230
Tools and Utilities	230
The Data Integration Engine	232
Repository	232
Jobs and Transformations	232
Plug-in Architecture	235
Getting Started with Spoon	236
Launching the Spoon Application	236
A Simple “Hello, World!” Example	237
Building the Transformation	237
Running the Transformation	244
The Execution Results Pane	245
The Output	246
Checking Consistency and Dependencies	247
Logical Consistency	247
Resource Dependencies	247
Verifying the Transformation	247
Working with Database Connections	248
JDBC and ODBC Connectivity	248
Creating a Database Connection	249
Testing Database Connections	252
How Database Connections Are Used	252
A Database-Enabled “Hello, World!” Example	253
Database Connection Configuration Management	256
Generic Database Connections	257
Summary	258

Chapter 10 Designing Pentaho Data Integration Solutions	261
Generating Dimension Table Data	262
Using Stored Procedures	262
Loading a Simple Date Dimension	263
CREATE TABLE dim_date: Using the Execute SQL Script Step	265
Missing Date and Generate Rows with Initial Date: The Generate Rows Step	267
Days Sequence: The Add Sequence Step	268
Calculate and Format Dates: The Calculator Step	269
The Value Mapper Step	273
Load dim_date: The Table Output Step	275
More Advanced Date Dimension Features	276
ISO Week and Year	276
Current and Last Year Indicators	276
Internationalization and Locale Support	277
Loading a Simple Time Dimension	277
Combine: The Join Rows (Cartesian product) Step	279
Calculate Time: Again, the Calculator Step	281
Loading the Demography Dimension	281
Understanding the stage_demography and dim_demography Tables	283
Generating Age and Income Groups	284
Multiple Incoming and Outgoing Streams	285
Loading Data from Source Systems	286
Staging Lookup Values	286
The stage_lookup_data Job	287
The START Job Entry	288
Transformation Job Entries	288
Mail Success and Mail Failure	289
The extract_lookup_type and extract_lookup_value Transformations	292
The stage_lookup_data Transformation	293
Check If Staging Table Exists: The Table Exists Step	294
The Filter rows Step	294
Create Staging Table: Executing Dynamic SQL	295
The Dummy Step	296

The Stream Lookup Step	297
Sort on Lookup Type: The Sort Rows Step	299
Store to Staging Table: Using a Table Output Step to Load Multiple Tables	300
The Promotion Dimension	300
Promotion Mappings	301
Promotion Data Changes	301
Synchronization Frequency	302
The load_dim_promotion Job	302
The extract_promotion Transformation	303
Determining Promotion Data Changes	304
Saving the Extract and Passing on the File Name	306
Picking Up the File and Loading the Extract	306
Summary	308
Chapter 11 Deploying Pentaho Data Integration Solutions	309
Configuration Management	310
Using Variables	310
Variables in Configuration Properties	311
User-Defined Variables	312
Built-in Variables	314
Variables Example: Dynamic Database Connections	314
More About the Set Variables Step	318
Set Variables Step Gotchas	319
Using JNDI Connections	319
What Is JNDI?	319
Creating a JNDI Connection	320
JNDI Connections and Deployment	321
Working with the PDI Repository	322
Creating a PDI Repository	322
Connecting to the Repository	323
Automatically Connecting to a Default Repository	324
The Repository Explorer	325
Managing Repository User Accounts	327
How PDI Keeps Track of Repositories	328
Upgrading an Existing Repository	329
Running in the Deployment Environment	330



Running from the Command Line	330
Command-Line Parameters	330
Running Jobs with Kitchen	332
Running Transformations with Pan	332
Using Custom Command-line Parameters	333
Using Obfuscated Database Passwords	334
Running Inside the Pentaho BI Server	334
Transformations in Action Sequences	334
Jobs in Action Sequences	335
The Pentaho BI Server and the PDI Repository	336
Remote Execution with Carte	337
Why Remote Execution?	338
Running Carte	339
Creating Slave Servers	340
Remotely Executing a Transformation or Job	341
Clustering	341
Summary	343

Part IV Business Intelligence Applications 345

Chapter 12 The Metadata Layer 347

Metadata Overview	347
What Is Metadata?	347
The Advantages of the Metadata Layer	348
Using Metadata to Make a More User-Friendly Interface	348
Adding Flexibility and Schema Independence	348
Refining Access Privileges	349
Handling Localization	349
Enforcing Consistent Formatting and Behavior	350
Scope and Usage of the Metadata Layer	350
Pentaho Metadata Features	352
Database and Query Abstraction	352
Report Definition: A Business User's Point of View	352
Report Implementation: A SQL Developer's Point of View	353
Mechanics of Abstraction: The Metadata Layer	355



Properties, Concepts, and Inheritance in the Metadata Layer	355
Properties	355
Concepts	356
Inheritance	356
Localization of Properties	357
Creation and Maintenance of Metadata	357
The Pentaho Metadata Editor	357
The Metadata Repository	358
Metadata Domains	359
The Sublayers of the Metadata Layer	359
The Physical Layer	359
The Logical Layer	362
The Delivery Layer	365
Deploying and Using Metadata	366
Exporting and Importing XMI files	366
Publishing the Metadata to the Server	367
Refreshing the Metadata	367
Summary	368
Chapter 13 Using The Pentaho Reporting Tools	371
Reporting Architecture	371
Web-Based Reporting	373
Practical Uses of WAQR	375
Pentaho Report Designer	376
The PRD Screen	377
Report Structure	378
Report Elements	380
Creating Data Sets	381
Creating SQL Queries Using JDBC	382
Creating Metadata Queries	385
Example Data Set	386
Adding and Using Parameters	386
Layout and Formatting	389
Alternate Row Colors: Row Banding	390
Grouping and Summarizing Data	391
Adding and Modifying Groups	391
Using Functions	393
Using Formulas	395

Adding Charts and Graphs	397
Adding a Bar Chart	400
Pie Charts	400
Working with Images	401
Working with Subreports	404
Passing Parameter Values to Subreports	405
Publishing and Exporting Reports	406
Refreshing the Metadata	407
Exporting Reports	408
Summary	408

Chapter 14 Scheduling, Subscription, and Bursting 411

Scheduling	411
Scheduler Concepts	412
Public and Private Schedules	412
Content Repository	412
Creating and Maintaining Schedules with the Pentaho Administration Console	413
Creating a New Schedule	414
Running Schedules	416
Suspending and Resuming Schedules	416
Deleting Schedules	417
Programming the Scheduler with Action Sequences	417
Add Job	418
Suspend Job, Resume Job, and Delete Job	420
Other Scheduler Process Actions	420
Scheduler Alternatives	420
UNIX-Based Systems: Cron	421
Windows: The at Utility and the Task Scheduler	421
Background Execution and Subscription	422
How Background Execution Works	422
How Subscription Works	423
Allowing Users to Subscribe	423
Granting Execute and Schedule Privileges	424
The Actual Subscription	425
The User's Workspace	426
Viewing the Contents of the Workspace	426

xxviii Contents

The Waiting, Complete, and My Schedules	
Panes	427
The Public Schedules Pane	427
The Server Administrator's Workspace	428
Cleaning Out the Workspace	429
Bursting	430
Implementation of Bursting in Pentaho	430
Bursting Example: Rental Reminder E-mails	430
Step 1: Finding Customers with DVDs That Are Due This Week	431
Step 2: Looping Through the Customers	432
Step 3: Getting DVDs That Are Due to Be Returned	434
Step 4: Running the Reminder Report	434
Step 5: Sending the Report via E-mail	436
Other Bursting Implementations	438
Summary	439
Chapter 15 OLAP Solutions Using Pentaho Analysis Services	441
Overview of Pentaho Analysis Services	442
Architecture	442
Schema	444
Schema Design Tools	444
Aggregate Tables	445
MDX Primer	445
Cubes, Dimensions, and Measures	446
The Cube Concept	446
Star Schema Analogy	447
Cube Visualization	447
Hierarchies, Levels, and Members	448
Hierarchies	448
Levels and Members	449
The All Level, All Member, and Default Member	450
Member Sets	451
Multiple Hierarchies	451
Cube Family Relationships	451
Relative Time Relationships	452
MDX Query Syntax	453
Basic MDX Query	453

Axes: ON ROWS and ON COLUMNS	453
Looking at a Part of the Data	454
Dimension on Only One Axis	455
More MDX Examples: a Simple Cube	455
The FILTER Function	455
The ORDER Function	456
Using TOPCOUNT and BOTTOMCOUNT	457
Combining Dimensions: The CROSSJOIN Function	457
Using NON EMPTY	457
Working with Sets and the WITH Clause	458
Using Calculated Members	459
Creating Mondrian Schemas	460
Getting Started with Pentaho Schema Workbench	460
Downloading Mondrian	460
Installing Pentaho Schema Workbench	461
Starting Pentaho Schema Workbench	461
Establishing a Connection	462
JDBC Explorer	463
Using the Schema Editor	463
Creating a New Schema	463
Saving the Schema on Disk	464
Editing Object Attributes	465
Changing Edit Mode	465
Creating and Editing a Basic Schema	466
Basic Schema Editing Tasks	466
Creating a Cube	466
Choosing a Fact Table	468
Adding Measures	469
Adding Dimensions	470
Adding and Editing Hierarchies and Choosing Dimension Tables	471
Adding Hierarchy Levels	474
Associating Cubes with Shared Dimensions	476
Adding the DVD and Customer Dimensions	478
XML Listing	480
Testing and Deployment	481
Using the MDX Query Tool	481
Publishing the Cube	482

Schema Design Topics We Didn't Cover	483
Visualizing Mondrian Cubes with JPivot	484
Getting Started with the Analysis View	484
Using the JPivot Toolbar	485
Drilling	486
Drilling Flavors	486
Drill Member and Drill Position	487
Drill Replace	488
Drill Through	488
The OLAP Navigator	488
Controlling Placement of Dimensions on Axes	489
Slicing with the OLAP Navigator	490
Specifying Member Sets with the OLAP Navigator	492
Displaying Multiple Measures	493
Miscellaneous Features	493
MDX Query Pane	493
PDF and Excel Export	494
Chart	494
Enhancing Performance Using the Pentaho	
Aggregate Designer	496
Aggregation Benefits	496
Extending Mondrian with Aggregate Tables	497
Pentaho Aggregate Designer	500
Alternative Solutions	502
Summary	502
Chapter 16 Data Mining with Weka	503
Data Mining Primer	504
Data Mining Process	504
Data Mining Toolset	506
Classification	506
Clustering	507
Association	507
Numeric Prediction (Regression)	508
Data Mining Algorithms	508
Training and Testing	509
Stratified Cross-Validation	509
The Weka Workbench	510

Weka Input Formats	511
Setting up Weka Database Connections	512
Starting Weka	514
The Weka Explorer	516
The Weka Experimenter	517
Weka KnowledgeFlow	518
Using Weka with Pentaho	519
Adding PDI Weka Plugins	520
Getting Started with Weka and PDI	520
Data Acquisition and Preparation	521
Creating and Saving the Model	523
Using the Weka Scoring Plugin	525
Further Reading	527
Summary	527
Chapter 17 Building Dashboards	529
The Community Dashboard Framework	529
CDF, the Community, and the Pentaho Corporation	529
CDF Project History and Who's Who	530
Issue Management, Documentation, and Support	531
Skills and Technologies for CDF Dashboards	531
CDF Concepts and Architecture	532
The CDF Plugin	534
The CDF Home Directory	534
The plugin.xml File	535
CDF JavaScript and CSS Resources	536
The .xcdf File	537
Templates	538
Document Template (a.k.a. Outer Template)	538
Content Template	541
Example: Customers and Websites Dashboard	542
Setup	544
Creating the .xcdf File	544
Creating the Dashboard HTML File	545
Boilerplate Code: Getting the Solution and Path	545
Boilerplate Code: Dashboard Parameters	546
Boilerplate Code: Dashboard Components	546

xxxii Contents

Testing	547
Customers per Website Pie Chart	548
Customers per Website: Pie Chart Action Sequence	548
Customers per Website: XactionComponent	551
Dynamically Changing the Dashboard Title	553
Adding the website_name Dashboard Parameter	553
Reacting to Mouse Clicks on the Pie Chart	554
Adding a TextComponent	555
Showing Customer Locations	557
CDF MapComponent Data Format	557
Adding a Geography Dimension	558
Location Data Action Sequence	559
Putting It on the Map	561
Using Different Markers Depending on Data	562
Styling and Customization	565
Styling the Dashboard	566
Creating a Custom Document Template	568
Summary	569
Index	571