

Index

A

- absolute positioning, PRD, 398
- access control
 - metadata layer refining privileges, 349
 - schema design and, 483
- accumulating periodic snapshot fact table, 149–150
- action definitions, 68
- action sequence editor, 80–82
- Action Sequence Wizard, 80, 86
- action sequences
 - adding as templates for Design Studio, 88
 - creating with PDS. *See* PDS (Pentaho Design Studio)
 - Customers per Website pie chart, 548–551
 - executed by solution engine, 68
 - executing in background, 422–423
 - functionality of, 78
 - inputs for, 83–85
 - for location data, 559–561
 - outputs for, 85
 - process actions and, 85–89
 - programming Scheduler with, 412, 417–420
 - running jobs within, 335
 - solution repository containing, 68
 - subscribing to, 423–426
 - using transformations in, 334–336
- actions, process, 85–89
- Active Directory (AD), and EE single sign-on, 77
- Active@ ISO Burner, 22
- Ad Hoc Report Wizard, 373–375
- Add Cube button, 466–467
- Add Job process action, Scheduler, 418–419
- Add Levels, 474–476
- Add Parameter function, PRD, 386–389
- Add Sequence step, PDI
 - loading date dimension, 268–269
 - loading demography dimension, 283, 284
- additive measures, 150
- Addresses tab page, Mail job, 290
- Ad-Hoc Report component, 192
- admin user
 - creating slave servers, 340
 - managing PDI repository accounts, 326–327
 - PDI repository, 324
- administration-console, 38, 44
- administrative tasks
 - data sources, 60–61
 - managing schedules and subscriptions, 61
 - Pentaho Administrative Console. *See* PAC (Pentaho Administrative Console)
 - user management, 58–60
- Administrator profile, PDI repository, 327
- Advanced category, Database Connection dialog, 250
- Advisor button, PAD, 501
- Age and Age Sequence step, demography dimensions, 282

- Age Group step, demography dimensions, 282, 284–285
- Aggregate Designer, 130, 442
- aggregate tables
 - creating manually, 500
 - drawbacks of, 502
 - extending Mondrian with, 497–500
 - generating and populating, 445
 - Pentaho Analysis Services, 445
- aggregation
 - alternatives to, 502
 - benefits of, 496–497
 - data integration process, 229
 - data warehouse design, 163–164
 - data warehouse performance, 130
 - Mondrian, 496
 - PRD reports, 393–395
 - restricting results of, 157
 - Slice and Dice pivot table example of, 17–18
 - using subreports for different, 404–406
 - WAQR reports, 374
- AJAX technology, CDF dashboards, 529–530
- algorithms, as data mining tool, 508–509
- aliases, 152–153, 384–385
- all level, MDX hierarchies, 450–451
- all member, MDX hierarchies, 450
- All Schedules panel, server admin's workspace, 428
- Alves, Pedro, 530
- analysis
 - examples of, 16–19
 - views in Pentaho BI Server, 484–485
 - views in user console, 73–74
- analytical databases, 142–143
- analytics, business, 503
- AND operator, multiple restrictions, 157
- appliances, data warehouse, 143–144
- Application Service Providers (ASPs), 144
- architecture
 - Community Dashboard Framework, 532–534
 - data warehouse. *See* data warehousing architecture
 - Pentaho Analysis Services, 442–444
 - Pentaho BI, 64
 - reporting, 371–373
- archiving
 - data warehouse performance and, 132
 - transaction data, 128
- ARFF (Attribute Relation File Format), 511, 519
- AS. *See* action sequences
- ASPs (Application Service Providers), 144
- assignments, user management, 59
- association, as data mining tool, 507–508
- at utility, job scheduling, 421
- Attribute Relation File Format (ARFF), 511, 519
- attributes
 - dimensions, 476–477
 - global data mart data model, 206–207
 - hierarchies, 472–473
 - level, 474–475
 - measures, 469–470
 - Mondrian cubes, 467
 - not fitting into dimension tables, 180–181
- audit columns, 163–164
- authentication
 - hibernate database storing data on, 45, 47
 - JDBC security configuration, 50
 - Mail job entry configuration, 290–291
 - Pentaho Administrative Console configuration, 57–58
 - slave server configuration, 340
 - SMTP configuration, 53–54
 - Spring Security handling, 60, 69
- authorization
 - hibernate database storing data on user, 47, 60
 - JDBC security configuration, 50
 - managing user accounts, 327–328
 - Spring Security handling, 60, 69
 - user configuration, 58–60
- automated knowledge discovery, 503. *See also* data mining
- automatic startup, UNIX/Linux systems, 40–41
- availability, remote execution, 338–339
- averages, calculating, 217–218
- axes
 - controlling dimension placement on, 489–490
 - dimensions on only one axis, 455
 - MDX representing information on multiple axes, 453
- Azzurri Clay, 32

B

- back office
 - data warehouse architecture, 117
 - database support for, 95–96
- back-end programs, Pentaho BI stack, 66
- background directory, content repository, 413
- background execution, 422–423, 426–429
- backup, PDI repository, 329
- banded report tools, 376
- bar charts, 400–402
- Base concept, metadata layer, 357
- batch-wise ETL, 118
- BC (business case), 192. *See also* World Class Movies
- BDW (Business Data Warehouse), 111, 115
- BETWEEN . . . AND operators, 151–152
- BI (business intelligence). *See also* Pentaho BI Server
 - analytics and, 503
 - components, 70–73
 - dashboards and, 529
 - data mart design and. *See* data marts, design
 - data mining and, 505–506
 - definition of, 107
 - example business case. *See* WCM (World Class Movies), example business case
 - importance of data, 108–109
 - platform, Pentaho BI stack, 64
 - purpose of, 105–109
 - real-time data warehousing and, 140–142
 - reports and. *See* reports
 - using master data management, 127–128
- BI Developer examples
 - button-single-parameter.prpt, 13–14
 - CDF section, 530
 - overview of, 8–9
 - Regional Sales - HTML reporting, 11–12
 - Regional Sales - Line/Bar Chart, 16–17
 - Slice and Dice Analysis, 17–18
- BIRT reports, 72, 372
- biserver-ce, 38. *See also* Pentaho home directory
- bitmap indexes, and data warehousing, 129–130
- blob field, reports with images, 401–403
- BOTTOMCOUNT function, MDX queries, 457
- bridge tables
 - maintenance of, 229
 - multi-valued dimensions and, 182–183
 - navigating hierarchies using, 184–185
- browsers, logging in, 6–7
- Building the Data Warehouse* (Inmon), 113–114
- built-in variables, 314
- bum, managing Linux init scripts, 42
- Burst Sales Report, 54
- bursting
 - defined, 430
 - implementing in Pentaho, 430
 - other implementations, 438
 - overview of, 430
 - rental reminder e-mails example, 430–438
- bus architecture, 179–189
- business analysts, 193–195
- business analytics, 503
- business case (BC), 192. *See also* WCM (World Class Movies), example business case
- Business Columns, at logical layer, 362–363
- Business Data Warehouse (BDW), 111, 115
- business intelligence. *See* BI (business intelligence)
- Business Intelligence server. *See* Pentaho BI Server
- business layer, metadata model, 71
- business modeling, using star schemas. *See* star schemas
- Business Models, logical layer, 362
- business rules engines, 141
- Business Tables, 362–365
- Business Views, 71, 362
- button-single-parameter.prpt sample, 13–14

C

- c3p0 connection pool, 49–50
- C4.5 decision tree algorithm, 508, 512
- C5.0 decision tree algorithm, 508–509
- caching, Join Rows (Cartesian product) step, 280
- Calculate and Format Dates step, 269–273
- Calculate Time step, 277, 281
- calculated members, 459–460, 483
- calculations, PRD functions for, 395

- Calculator step, PDI
 - age and income groups for demography dimension, 285
 - Calculate Time step, 281
 - current and last year indicators, 276
 - loading date dimension, 269–273
 - loading demography dimension, 282
- calendar days, calculating date dimensions, 270
- Carte server
 - clustering, 341–342
 - creating slave servers, 340–341
 - as PDI tool, 231
 - remote execution and, 337–339, 341
 - running, 339–340
- carte.bat script, 339–340
- carte.sh script, 339–340
- Cartesian product, 153
- Cascading Style Sheets. *See* CSS (Cascading Style Sheets)
- catalog, connecting DataCleaner with, 202
- Catalog of OMG Modeling and Metadata Specifications, 352
- CategorySet collector function, 399–400
- causality, vs. correlation in data mining, 508
- CDC (Change Data Capture), 133–137
 - choosing option, 137
 - intrusive and non-intrusive processes, 133
 - log-based, 136–137
 - methods of implementing, 226
 - overview of, 133
 - snapshot-based, 135–136
 - source data-based, 133–134
 - trigger-based, 134–135
- CDF (Community Dashboard Framework), 542–569
 - concepts and architecture, 532–534
 - content template, 541–542
 - customer and websites dashboard. *See* customer and websites dashboard
 - dashboarding examples, 19–20
 - document template, 538–541
 - history of, 530–531
 - home directory, 534–535
 - JavaScript and CSS resources, 536–537
 - overview of, 529
 - plug-in, 534
 - plugin.xml file, 535–536
 - skills and technologies for, 531–532
 - summary, 569–570
 - synergy with community and Pentaho corporation, 529–530
 - templates, 538
 - .xcdf file, 537–538
- central data warehouse, 117, 119–121
- CEP (complex event processing), 141
- Change Data Capture. *See* CDC (Change Data Capture)
- CHAPTERS, axes, 453
- Chart editor, 398–399
- charts. *See also* Customers per Website pie chart
 - adding bar charts to PRD reports, 400
 - adding images to PRD reports, 401–404
 - adding pie charts to PRD reports, 400–402
 - adding to PRD reports, 397–400
 - bursting. *See* bursting
 - examples, 14–16
 - including on dashboards, 548
 - JPivot, 494–496
 - not available in WAQR, 375
 - reacting to mouse clicks on pie, 554–555
- Check if Staging Table Exists step, 293–294
- child objects, and inheritance, 356
- Citrus, 13–14
- class path, Weka data mining, 512–515
- classification
 - algorithms, 508–509
 - as data mining tool, 506–507
 - with Weka Explorer, 524
- client, defined, 65
- cloud computing, 144
- clustering
 - as data mining tool, 507
 - database connection options for, 250
 - remote execution with Carte using, 337, 341–342
 - snowflaking and, 186–187
- collapsed dimensions, 498
- collector functions, 397, 399–400
- colors, for PRD reports, 390–391
- column profiling, 197–198, 199
- columnar databases, 142–143
- column-oriented databases, 502
- columns
 - date dimension, 213–216
 - meaningful names for, 163
 - quickly opening properties for, 210
 - SCD type 3, 174

- using dictionary for dependency checks on, 205
- COLUMNS, axes, 453–454
- Columns section, OLAP Navigator, 19
- columns stores, analytical databases, 142–143
- Combine step, PDI
 - Join Rows (Cartesian product) step, 278–281
 - loading demography dimension, 283
 - loading time dimension, PDI, 277
- command line
 - creating symbolic links from, 26
 - installing Java SDK from, 27–28
 - Linux systems using, 24–25
 - running jobs and transformations from, 330–334
 - setting up MySQL schemas from, 46
 - starting desktop programs from, 76
- comment lines, startup scripts, 52
- Common Warehouse Metamodel (CWM), 70, 352
- Community Dashboard Framework. *See* CDF (Community Dashboard Framework)
- Community Edition of Pentaho
 - default document templates with, 539
 - Java servlet technology, 74
 - overview of, 76–77
- comparing data, DataCleaner, 199, 205
- Competing on Analytics*, 503
- Complete pane, Workspace, 427
- complex event processing (CEP), 141
- compression, analytical databases and, 142
- concepts, Pentaho metadata, 356–357
- conditional formatting, schema design and, 483
- conflicting data, data warehouses, 124
- conformed dimensions, 115, 158–160
- conformed rollup, 498
- connect by prior statement, SQL, 184
- connection pools
 - adding to Hibernate configuration, 49–50
 - database connection options, 250
 - managing, 69
- connections. *See* database connections
- consistency, and transformations, 247
- consolidated fact tables, 189
- consultants, hiring external analysts, 193–194
- content repository, 412–413, 429
- content template, CDF, 533, 541–542
- contents, of document templates, 540
- conversion functions, PRD, 393
- Copy Tables Wizard, 210
- correlation, vs. causality in data mining, 508
- count_rows transformation, 315
- cousins, cube family relationships, 452
 - Create dim_date step, date dimension, 265–267
 - Create dim_demography step, demography dimension, 282
 - Create dim_time step, time dimension, 277
 - Create Kettle Job option, 210
- Create stage_demography step, demography dimension, 282
- Create Staging Table step, 293, 295–296
- CREATE TABLE statement
 - Create dim_date, loading date dimension, 265–267
 - creating staging table, 296
 - dim_demography dimension table, 283–284
 - simplified date dimension table, 263–265
 - stage_promotion table, 305
- create_quartz_mysql.sql script, 46
- create_repository_mysql.sql script, 46
- create_sample_datasource.sql script, 46
- credentials, Pentaho Administrative Console, 57–58
- CRISP-DM (CRoss Industry Standard Process for Data Mining), 505
- cron implementations, 415, 421
- CROSS JOIN, 153, 155–156
- CROSSJOIN function, MDX queries, 457
- cross-tables, cube visualization, 447–448
- crosstabs, cube visualization, 447–448
- cross-validation
 - compensating for biases, 509
 - stratified, 509–510
 - with Weka Explorer, 524
- CSS (Cascading Style Sheets)
 - building CDF dashboards, 532
 - CDF and, 536–537
 - styling dashboards, 566–567
- Ctrl+Alt+N, 287
- CTRL+C, 25
- CTRL+R, 25
- Ctrl+Spacebar, 312, 317

- cubes
 - adding dimensions to Mondrian schemas, 470
 - adding measures to cube fact tables, 469–470
 - analyzing. *See* JPivot
 - associating with shared dimensions, 476–477
 - creating, 466–467
 - fact tables for, 468
 - family relationships, 451–452
 - FILTER function, 455–456
 - MDX queries operating on, 445–446
 - overview of, 446–447
 - publishing, 482–483
 - testing, 481–482
 - visualization of, 447–448
 - curly braces {}, set syntax, 458–459
 - current job scope, 313
 - current year indicator, loading date dimension, 276–277
 - current_record column, 172–173
 - current_week column, 167–168, 216–218
 - customer and websites dashboard, 542–569
 - adding TextComponent, 555–557
 - boilerplate code for dashboard components, 546–547
 - boilerplate code for dashboard parameters, 546
 - boilerplate code for dashboard solution and path, 545–546
 - custom document template for, 568–569
 - Customers per Website pie chart, 548–553
 - dynamically changing dashboard title, 553–554
 - .html file for, 545
 - MapComponent data format, 557–562
 - marker options for data, 562–565
 - reacting to mouse clicks on pie chart, 554–555
 - setting up, 544
 - showing customer locations, 557
 - styling the dashboard, 565–568
 - testing, 547
 - .xcdf file for, 544
 - customer dimensions, Mondrian schemas, 478–480
 - customer locations. *See also* MapComponent
 - marker options for showing, 563–565
 - showing on customer and websites dashboard, 557–562
 - customers, WCM example
 - customer order fulfillment, 94
 - developing data model, 101–102
 - main process flows, 96
 - orders and promotions, 102–105
 - websites targeting, 94–95
 - Customers per Website pie chart
 - action sequences, 548–551
 - overview of, 548
 - XactionComponent, 551–553
 - Customize Selection screen, WAQR, 374
 - CWM (Common Warehouse Metamodel), 70, 352
- D**
- DaaS (DataWarehouse as a Service), 144
 - Dashboard Builder, Enterprise Edition, 77
 - dashboard content template, CDF, 533, 541–542
 - dashboard document template. *See* document templates (outer templates), CDF
 - dashboards
 - CDF. *See* CDF (Community Dashboard Framework)
 - components, 543
 - customers and websites. *See* customer and websites dashboard
 - examples of, 19–20
 - overview of, 529
 - summary, 569–570
 - Dashboards object, 534
 - Dashboards.js, 536
 - data, marker options, 562–565
 - data acquisition and preparation, Weka
 - data mining, 521–522
 - data analysis
 - data profiling for, 197–198
 - data warehousing and, 195–197
 - using DataCleaner. *See* DataCleaner
 - data changes, promotion dimension, 301–302, 304–306
 - data cleansing, source data, 228
 - data governance, 125
 - data integration
 - activities. *See* ETL (Extraction, Transformation, and Loading)
 - defined, 223

- engine, 230, 232
- overview, 223–224
- using PDI. *See* PDI (Pentaho Data Integration)
- data lineage, 114
- data marts
 - bus architecture, 119–121
 - data warehouse architecture, 117
 - independent, 119
 - Inman vs. Kimball approach, 115–116
 - OLAP cubes, 121–122
 - overview of, 121
 - storage formats and MDX, 122–123
- data marts, design, 191–220
 - data analysis, 195–198. *See also* DataCleaner
 - data modeling with Power*Architect, 208–209
 - developing model, 206–208
 - requirements analysis, 191–195
 - source to target mapping, 218–219
 - WCM example. *See* WCM (World Class Movies), building data marts
- data mining. *See also* Weka data mining
 - algorithms, 508–509
 - association in, 507–508
 - classification in, 506–507
 - clustering in, 507
 - defined, 504
 - engine, 72–73
 - further reading, 527
 - numeric prediction (regression), 508
 - process, 504–506
 - stratified cross-validation, 509–510
 - summary, 527
 - toolset, 506
 - training and testing, 509
- data models
 - developing global data mart, 206–208
 - as forms of metadata, 114
 - normalized vs. dimensional, 115–116
 - with Power*Architect, 208–209
 - reference for understanding, 208
- Data Profiler, Talend, 206
- data profiling
 - alternative solutions, 205–206
 - overview of, 197–198
 - using DataCleaner. *See* DataCleaner
 - using Power*Architect tool, 208
- data quality. *See* DQ (data quality)
- data sets, PRD reports, 381–386
- data sources
 - managing, 60–61
 - working with subreports for different, 404–406
- data staging, 226–227
- data timeliness, 114
- data validation, ETL and, 227–228
- data vault (DV), 125–127
- The Data Warehouse Lifecycle Toolkit* (Kimball), 158, 170, 194
- data warehousing, 111–145
 - analytical databases, 142–143
 - appliances, 143–144
 - Changed Data Capture and, 133–137
 - changing user requirements, 137–139
 - data quality problems, 124–128
 - data volume and performance problems, 128–133
 - debate over Inman vs. Kimball, 114–116
 - on demand, 144
 - example of. *See* WCM (World Class Movies), example business case
 - need for, 112–114
 - overview of, 113–114
 - real-time, 140–142
 - using star schemas. *See* star schemas
 - virtual, 139–140
- data warehousing architecture, 116–123
 - central data warehouse, 119–121
 - data marts, 121–123
 - overview of, 116–118
 - staging area, 118–119
- The Data Warehousing Institute (TDWI), 119
- database connection descriptors, 252–253, 359–360
- Database Connection dialog, 249–252, 257–258, 360
- database connections
 - adding to data mart, 201–202
 - configuring, 256–257
 - creating, 249–252
 - establishing to PSW, 462
 - generic, 257–258
 - “Hello World” example, 253–256
 - JDBC and ODBC, 248
 - JNDI, 319–322
 - managing in Repository Explorer, 326
 - managing with variables, 314–318
 - to PDI repository, 322–324
 - to PDI repository, automatically, 324–325

- database connections (*continued*)
 - at physical layer of metadata domain, 359–360
 - testing, 252
 - using, 252–253
 - for Weka data mining, 512–514
- database segmentation (clustering), 507
- database-based repository, 358–359
- databases
 - column-oriented databases, 502
 - connection pool management, 69
 - managing drivers, 44–45
 - policies prohibiting extraction of data, 226
 - system, 45–52
 - tools for, 31–34
 - used by WCM example, 95–97
- DataCleaner, 198–206
 - adding database connections, 201–202
 - adding profile tasks, 200–201
 - alternative solution, 205–206
 - doing column dependency checks, 205
 - doing initial profile, 202
 - overview of, 198–200
 - profiling and exploring results, 204–205
 - selecting source tables, 218–219
 - validating and comparing data, 205
 - working with regular expressions, 202–204
- datacleaner-config.xml file, 201
- DataWarehouse as a Service (DWaaS), 144
- date and time, modeling, 165–168
- date dimension
 - generating, 213–216
 - role-playing, 182
 - special date fields and calculations, 216–218
- date dimension, loading, 263–277
 - Add sequence step, 268–269
 - Calculator step, 269–273
 - current and last year indicators, 276–277
 - Execute SQL script step, 265–267
 - Generate Rows step, 267–268
 - internationalization and locale support, 277
 - ISO week and year attributes, 276
 - overview of, 263–265
 - Table Output step, 275–276
 - using stored procedures, 262–263
 - Value Mapper step, 273–275
- Date Mask Matcher profile, DataCleaner, 200
- date_julian, relative time, 167–168
- day of week number, date dimensions, 271–272
- Days Sequence step, date dimensions, 268–269
- Debian-based Linux, automatic startup in, 42
- decision tree algorithms, 508–509
- decoding, data integration and, 228–229
- default member, MDX hierarchies, 450–451
- Define process tab, action sequence editor, 81–83
- degenerate dimensions, 181
- Delete Job process action, Scheduler, 420
- Delete link, Public Schedules pane of Workspace, 428
- deleting schedules, 417
- delivery layer, metadata model, 355, 365–366
- demography dimension, loading, 281–286
 - generating age and income groups, 284–285
 - multiple ingoing and outgoing streams, 285–286
 - overview of, 281–283
 - stage_demography and dim_demography tables, 283–284
- demography dimension, Orders data mart, 212
- Demography Key sequence step, 283, 285–286
- de-normalization, 115–116
- dependencies, “Hello World” transformation, 247
- dependency profiling, 197–198
- deployment
 - of PDI. *See* PDI (Pentaho Data Integration), deployment
 - of Pentaho metadata, 366–368
- descendants, cube family relationships, 452
- descriptive text, for schedules, 415
- design tools, schema, 444
- desktop programs, 65, 74–76, 196
- Details Body, PRD reports, 378
- development skills, CDF dashboards, 531
- Devlin, Barry, 113–114
- dicing, defined, 441
- dictionary, DataCleaner, 205

- Dictionary Matcher profile, DataCleaner, 200
- `dim_demography` table, 282–284
- `dim_promotion` dimension table. *See* promotion dimension
- dimension keys, 162, 169
- dimension tables
 - aggregating data, 229
 - attributes in small vs. large, 206–207
 - choosing for Mondrian schemas, 471–474
 - fact tables vs., 148–150
 - loading demography dimension, 281–286
 - loading simple date dimension. *See* date dimension, loading
 - loading simple time dimension, 277–281
 - maintenance of, 229
 - overview of, 262
 - star schema and, 148
 - using stored procedures, 262–263
- dimension usage, 476–477
- dimensional model, advanced concepts, 179–189
 - building hierarchies, 184–186
 - consolidating multi-grain tables, 188–189
 - junk, heterogeneous and degenerate dimensions, 180–181
 - monster dimensions, 179–180
 - multi-valued dimensions and bridge tables, 182–183
 - outriggers, 188
 - overview of, 179
 - role-playing dimensions, 181–182
 - snowflakes and clustering dimensions, 186–187
- dimensional model, capturing history, 170–179
 - overview of, 169–170
 - SCD type 1: overwrite, 171
 - SCD type 2: add row, 171–173
 - SCD type 3: add column, 174
 - SCD type 4: mini-dimensions, 174–176
 - SCD type 5: separate history table, 176–178
 - SCD type 6: hybrid strategies, 178–179
- dimensional model, defined, 17
- dimensions
 - adding to Mondrian schemas, 470–471
 - associating cubes with shared, 476–477
 - calculated members, 459–460
 - controlling placement on axes, 489–490
 - cubes and, 446
 - data marts with conformed, 115
 - defined, 17
 - DVD and customer, 478–480
 - on only one axis, 455
 - slicing with OLAP Navigator, 490–491
 - static, 213–216
- directories
 - navigation commands for, 24–25
 - Repository Explorer managing, 326
 - server installation, 38
 - for UNIX-based systems, 39
- Disconnect repository option, 324
- disks, saving schemas on, 464–465
- distribution of class values, prediction models, 509
- Document structure, styling dashboard, 566
- document templates (outer templates), CDF
 - contents of, 540
 - customizing, 568–569
 - default templates shipping with Community Edition, 539
 - examples of reusable content, 538–539
 - naming conventions, 540–541
 - overview of, 533
 - placeholders, 540
- documentation
 - CDF, 531
 - Pentaho Data Integration, 261–262
- DOM specification, HTML standards, 532
- domains, metadata, 359
- DQ (data quality)
 - categories of, 124–125
 - data vault and, 125–127
 - using reference and master data, 127–128
- Dresner, Howard, 107
- drill down, OLAP and, 441
- drill member action, 487
- drill position, 488
- drill replace method, 488
- drill through action, 488
- drill up, OLAP and, 441
- drilling, 486–488
- drivers
 - for data profiling in DataCleaner, 201–202
 - for JNDI connections, 320–321
 - managing database, 44–45
- Drools, 141

DROP TABLE statement, date dimensions, 265–267
 dual-boot systems, 23
 Dummy step, 296–297
 duplicate data, and data warehouses, 124
 DV (data vault), 125–127
 DVD dimensions, Mondrian schemas, 478–480
 DVDs
 inventory management, 104–105
 rental reminder e-mails example, 430–438
 WCM data model, 99–102
 DWaaS (DataWarehouse as a Service), 144

E

ECHO command, 29
 Eclipse IDE, 77–80
 ECMAScript, 532
 Edit link, Public Schedules pane of Workspace, 427–428
 edit modes, schema editor, 465–466
 EE (Enterprise Edition) of Pentaho, 76–77, 530
 elements, PRD report, 380–381
 ELT (extract, load, transform), 224
 e-mail
 configuring, 52–54
 example of bursting. *See* rental reminder e-mails example
 Pentaho BI Server services, 70
 testing configuration, 54
 Email Message tab, Mail job entry, 290–291
 EMAIL process action, 436–437
 email_config.xml, 52, 70
 employees
 collecting requirements from, 194
 developing data model, 101, 103
 inventory management, 105
 Encr.bat script, 334
 encr.sh script, 334
 end user layer (EUL), 117
 end users, defined, 192
 Enterprise Edition (EE) of Pentaho, 76–77, 530
 Enterprise Resource Planning (ERP), in data analysis, 195
 environment variables, installing Java, 28

eobjects.org. *See* DataCleaner
 ERD (Entity Relationship Diagram), 102, 104
 ERMaster, 32
 error execution path, 288
 Esper, 141
 E-stats site, U.S. Census, 109
 ETL (Extraction, Transformation, and Loading). *See also* PDI (Pentaho Data Integration)
 for back office, 118
 building “Hello World!”. *See* Spoon, “Hello World!”
 data warehouse architecture, 117–118
 engine, 72
 overview of, 224
 scheduling independent of Pentaho BI Server, 420
 staging area optimizing, 118–119
 ETL (Extraction, Transformation, and Loading), activities
 aggregation, 229
 Change Data Capture, 226
 data cleansing, 228
 data staging, 226–227
 data validation, 227–228
 decoding and renaming, 228–229
 dimension and bridge table maintenance, 229
 extraction, 226
 key management, 229
 loading fact tables, 229
 overview of, 225
 ETLT (extract, transform, load, and transform), 224
 EUL (end user layer), 117
 examples, shipped with Pentaho
 analysis, 16–19
 charting, 14–16
 dashboarding, 19–20
 other types of, 20
 overview of, 8–9
 reporting, 11–14
 understanding, 9–11
 using repository browser, 9
 Excel exports, JPivot, 494
 Execute a Job dialog, Carte, 341
 Execute a Transformation dialog, Carte, 341
 Execute privileges, 424

Execute SQL Script step, PDI
 creating date dimension table, 265–267, 277
 creating demography dimension table, 282
 creating staging table, 295–296
 Execution Results pane, Spoon, 245–246, 255
 Experimenter, Weka, 510, 517–518
 Explorer, Weka
 creating and saving data mining model, 523–524
 overview of, 510
 working with, 516–517
 exporting
 data to spreadsheet or CSV file with WAQR, 375
 jobs and transformations, 326
 metadata to server, 367
 PRD reports, 408
 XMI files, 366
 expression string function, `TextComponent`, 556
 expressions
 MQL Query Builder limitations, 385
 in Pentaho reporting formulas, 396
 eXtensible attribute-Relation File Format (XRFF), 511–512
 Extensible Markup Language. *See* XML (Extensible Markup Language) files
 extract, 306–307
 extract, load, transform (ELT), 224
 extract, transform, load, and transform (ETLT), 224
`extract_lookup_type` transformation, 287, 292–293
`extract_lookup_value` transformation, 287, 292–293
`extract_promotion` transformation, 302–304
 Extraction, Transformation, and Loading. *See* ETL (Extraction, Transformation, and Loading)
 extraction process, ETL
 Change Data Capture activities, 226
 data staging activities, 226–227
 defined, 224
 overview of, 226
 supportive activities of, 225

F

F3 keyboard shortcut, 249
 fact tables
 conformed dimensions linking to, 115
 consolidated, 189
 creating Orders data mart, 212
 developing global data mart data model, 208
 dimension tables vs., 148–150
 loading, 230
 Mondrian schemas, 468
 types of, 149–150
 using smart date keys to partition, 166
 Fake Name Generator, 97, 101–102
 family relationships, cubes, 451–452
 Feature List, JNDI connections, 320–321
 federated architecture, data warehousing, 119–120
 fields
 creating SQL queries using JDBC, 383
 developing global data mart data model, 207–208
 formatting when using images in reports, 404
 for Missing Date step, dimension tables, 267–268
 special date, 216–218
 WAQR report, 374
 file formats, Weka, 511–512
 file-based repository, 358–359
`FILTER` function, MDX queries, 455–456
 Filter rows step, 293–295
 filters
 OLAP Navigator, 19
 schedules belonging to same group, 414–415
 WAQR reports, 374
 Flash Charts, Pentaho, 15–16
 flow-oriented report tools, 376
 folder contents pane, user console, 8
 folders, 68
 folds, cross-validation and, 509–510
 foreign key constraints, 150–151, 383
 Format Row-Banding option, PRD reports, 391
 formatting
 dates, 271
 email body in HTML, 292
 metadata layer applying consistency of, 360

formatting (*continued*)
 PRD report with images, 404
 PRD reports, 389–390
 report previews, 376
 WAQR reports, 374, 375
 Formula step, loading date dimensions, 276–277
 formulas, Pentaho reporting, 395–396
 Forums link, PRD Welcome screen, 377
 forward engineering, 208
Freakonomics, 503
 Free Java Download button, 28
 Free memory threshold, staging lookup values, 299
 FROM statement, SQL, 151–152
 front office, data warehouse architecture, 117–118
 front-end
 Pentaho BI stack, 66
 user console as, 73
 FULL OUTER JOIN, 155
 full table scans, 129
 functionality, Pentaho BI stack, 65
 functions, report, 393–395

G

GA (generally available) release, 4
 Gender and Gender Sequence step,
 demography dimension, 282
 Gender label step, demography
 dimension, 282
 General category, Database Connection
 dialog, 249–251
 General tab, action sequence editor, 81,
 86–89
 Generate 24 hours step, time dimension,
 277
 Generate 60 Minutes step, time dimension,
 277
 Generate Rows step
 date dimension, 267–268
 demography dimension, 282–283
 Generate Rows with Initial Date step, date
 dimension, 267–268
 generating age and income groups,
 demography dimension, 284–285
 generic database connections, 257–258, 316
 geography dimension table,
 MapComponent, 558–559

Get System Info step, 333
 getting started, 3–20
 downloading and installing software,
 4–5
 logging in, 6–7
 Mantle (Pentaho user console), 7–8
 overview of, 3
 starting Pentaho VI server, 5–6
 working with examples. *See* examples,
 shipped with Pentaho
 global data mart data model, 206–208
 global functions, PRD reports, 393–395
 global input source, 84
 global variables, user-defined, 312
 Gmail, configuring, 70
 GNOME terminal, 24
 grand-parent job scope, 313
 granularity
 consolidating multi-grain tables, 188–189
 data warehouse design, 163–164
 dimension tables and fact tables, 149
 global data mart data model, 208
 star schema modeling, 163–164
 time dimensions, 165
 graphs
 adding to PRD reports. *See* charts
 not available in WAQR, 375
 Greenplum, 143
 grids, field, 244
 GROUP BY statement, SQL, 151–153
 Group Header/Footer, PRD reports, 378
 groups
 creating schedule using, 414–415
 MQL Query Builder limitations, 385
 for PRD reports, 378, 391–393
 for UNIX-based systems, 39
 for WAQR reports, 373–374, 375
 guest user, PDI repository, 324, 326–327
 GUI tools, MySQL, 31

H

HAVING statement, SQL, 151, 157
 headers, PRD reports, 378
 “Hello World!”. *See* Spoon, “Hello
 World!”
 helper tables, 184
 heterogeneous dimensions, 181
 Hibernate database
 configuring JDBC security, 50–51

- defined, 45
- overview of, 47–50
- hierarchies
 - adding, 471
 - adding levels, 474–476
 - attributes, 472–473
 - cube aggregation and, 448–449
 - cube family relationships, 451–452
 - levels and members, 449–451
 - multiple, 451
 - navigating through data using, 184–186
- history
 - capturing in data warehouse. *See* dimensional model, capturing history
 - creating separate table for, 176–178
 - data warehouse advantages, 113
 - data warehouse retaining, 128
 - storing commands, 25
- HOLAP (Hybrid OLAP), 122
- home directory, CDF, 534–535
- Home Theater Info site, 100–101
- home-grown systems, for data analysis, 195
- hops
 - connecting transformations and jobs, 233–234, 287
 - creating in “Hello World”, 242–243
 - Filter row steps in staging lookup values, 295
 - splitting existing, 254
- horizontal partitioning, 180
- Hour Sequence step, time dimension, 277
- HSQLDB database server
 - managing system databases by default, 45
 - migrating system databases from, 46–52
 - start-pentaho script holding, 6
- HTML (HyperText Markup Language)
 - building CDF dashboards, 531
 - DOM specification, 532
 - formatting mail body in, 292
 - for layout, 566–567
- .html file, 545
- hub and spoke architecture, 119–121
- hubs, for data vault models, 125
- Hybrid OLAP (HOLAP), 122
- hybrid strategies, 178–179

I

- icon, variable, 311
- identification, for user accounts, 327–328
- IDs, defining custom, 387
- images
 - adding to PRD reports, 401–404
 - creating bootable CDs from downloaded files, 22
 - virtualization software and, 23–24
- importing
 - parameter values to subreports, 405–406
 - XMI files, 366
- IN operator, 151
- Income and Income Sequence step,
 - demography dimension, 282
- Income Group step, demography
 - dimension, 282, 284–285
- Income Statement sample report, 12
- incomplete data, data warehouses, 124
- incorrect data, data warehouses, 124
- independent data marts, 119–120
- indexes
 - analytical databases not needing, 142
 - creating in SQL, 212–213
 - improving data warehouse performance, 129
- InfoBright, 23, 143, 502
- inheritance
 - Pentaho metadata layer and, 356–357
 - PRD reports using style, 389–390
- init script (`rc` file), 40–42
- initialization phase, transformations, 266–267
- inline styles, HTML e-mail, 436
- Inmon, Bill, 113–116
- INNER JOIN, 153–154
- input formats, Weka data mining, 511–512
- input streams, PDI transformation sets, 285–286
- inputs, action sequence, 83–85
- installation, Pentaho BI Server
 - directory, 38
 - overview of, 4
 - setting up user account, 38–39
 - subdirectories, 38
- integer division, date dimensions, 272
- Interface address parameter, Carte, 340
- International Organization for
 - Standardization. *See* ISO (International Organization for Standardization)

internationalization, 277, 484
 Internet Movie Database, 99
 interview process, collecting requirements, 194
 intrusive CDC, 133–135
 invalid data, cleansing, 228
 inventory, WCM example, 94–95, 104–105
 ISO (International Organization for Standardization)
 images, 22
 standards, 97
 week and year attributes, 276
 _iso columns, date dimension, 182–183
 issue management, CDF, 531
 Item Band, running reminder report, 435

J

J48 tree classifier, 508–509
 .jar files, 44–45
 .jar files, 201
 JasperReports, 72, 372
 Java
 installing and configuring, 27–29
 Pentaho programmed in, 4, 66
 servlet technology, 67, 74
 java - java weka.jar command, 514
 Java Database Connectivity. *See* JDBC (Java Database Connectivity)
 Java Naming and Directory Interface. *See* JNDI (Java Naming and Directory Interface)
 Java Virtual Machine (JVM), 43, 314, 333–334
 JAVA_HOME variable, 28
 JavaMail API, 52, 54
 JavaScript
 building CDF dashboards, 532
 CDF and, 536–537
 evaluation, 205
 JDBC (Java Database Connectivity)
 configuring security, 50–51
 connection parameters, 462
 connection pooling, 69
 creating and editing data sources, 60–61
 creating database connections, 250–251
 creating generic database connections, 257–258
 creating SQL queries, 382–385
 managing database drivers, 44–45

overview of, 248–249
 Weka reading data from databases using, 512–513
 JDBC Explorer, 463
 JDBC-ODBC bridge, 249
 jdbc.properties file, 320–321
 Jetty, 55, 57–58
 JFreeChart, 15, 397–404
 JFreeReports, 72. *See also* PRD (Pentaho Report Designer)
 JNDI (Java Naming and Directory Interface)
 configuring connections for metadata editor, 360
 creating and editing data sources, 60–61
 creating database connections, 319–322
 job entries
 defined, 287
 hops connecting, 287–288
 Mail Success and Mail Failure, 289–292
 START, 288
 Transformation, 288–289
 using variables. *See* variables
 jobs
 adding notes to canvas, 264–265
 creating, 287–288
 creating database connections, 249
 data integration engine handling, 232
 exporting in Repository Explorer, 326
 overview of, 235
 remotely executing with Carte, 341–342
 running from command line, 330–334
 running inside Pentaho BI Server, 334–337
 running with Kitchen, 332
 running within action sequences, 335
 storing in repository, 232
 transformations vs., 232–233, 287
 user-defined variables for, 312
 using database connections, 252–253
 with variable database connection, 317–318
 JOIN clauses, 151–154
 join paths, 364–365
 join profiling, 197–198
 Join Rows (Cartesian product) step, 277–281, 283
 JPivot
 analysis view, 484–485
 charts, 494–496
 drilling with pivot tables, 486–488

- MDX query pane, 493
- NON EMPTY function and, 458
- overview of, 442, 484
- PDF and Excel exports, 494
- toolbar, 485
- jQuery, 532
- jquery.<name>.js, 536
- js directory, CDF, 535
- jsquery.js, 536
- junk dimension table, 181
- JVM (Java Virtual Machine), 314, 333–334

K

- K.E.T.T.L.E. *See* Pentaho Data Integration
- Kettle jobs, 73, 210
- Kettle.exe script, 236
- kettle.properties file, 312–313, 324–325
- key management, data integration and, 229
- Kickfire appliance, 144
- Kimball, Ralph
 - back and back office definitions, 117–118
 - data warehouse bus architecture, 158
 - Inmon data warehouse model vs., 114–116
 - SCD strategies, 170
 - snowflaking and, 186–187
- Kitchen tool
 - generic command-line parameters, 330–332
 - as PDI tool, 230–231
 - running jobs/transformations from command line, 330–334
 - using PDI repository with. *See* repository, PDI
- Klose, Ingo, 530
- KnowledgeFlow, Weka, 510, 518–519
- known values, data mining outcomes, 506

L

- last day in month, date dimensions, 273
- last year indicator, date dimensions, 276–277
- last_week column, 167–168, 216–218
- latency, remote execution reducing, 339
- layout
 - dashboard, 566
 - PRD reports, 389–390, 393
- LEFT OUTER JOIN, 154

- levels
 - adding hierarchy levels, 474–476
 - attributes, 474–475
 - MultiDimensional eXpressions, 449–451
- lib directory, CDF, 535
- Library for Support Vector Machine (LibSVM), 512
- LibSVM (Library for Support Vector Machine), 512
- line charts, 543
- links, for data vault models, 125
- Linux, 24–25, 40–42
- list, picking variables from, 312, 317
- List Scheduled Jobs process action, Scheduler, 420
- listeners, TextComponent, 556–557
- Live mode, running Ubuntu in, 23
- Load dim_demography step, 283
- Load dim_time step, 277
- LOAD FILE command, MySQL, 402
- Load stage_demography step, 283
- load_dim_promotion job, 302–303, 306–307
- loading process, ETL. *See also* ETL (Extraction, Transformation, and Loading)
 - defined, 224
 - dimension table maintenance, 229
 - loading fact tables, 230
 - supportive activities of, 225
- local time, UTC (Zulu) time vs., 165–166
- locale support, date dimensions, 277
- localization, in metadata layer, 349, 357
- locations, customer. *See* customer locations; MapComponent
- log-based CDC, 136–137
- logging in, getting started, 6–7
- logical consistency, 247
- logical layer, metadata model
 - Business Models, 362
 - Business Tables and Business Columns, 362–363
 - defined, 355
 - purpose of, 362
 - relationships, 364–365
- Login button, getting started, 6–7
- login modules, Jetty, 57–58
- logos, reports with, 401–404
- lookup values, staging, 286–300
 - Check if Staging Table Exists step, 294
 - Create Staging Table step, 295–296

lookup values, staging, (*continued*)
 Dummy step, 296–297
 extract_lookup_type/extract_lookup_value transformations, 292–293
 Filter rows step, 294–295
 Mail Success and Mail Failure job entries, 289–292
 overview of, 286–287
 Sort on Lookup Type step, 299–300
 stage_lookup_data job, 287–288
 stage_lookup_data transformation, 293–294
 START job entry, 288
 Stream Lookup step, 297–299
 Table Output step, 300
 transformation job entries, 288–289
 lookup_value table, promotion mappings, 301
 looping, bursting example, 432–434
 lost dimensions, 498
 LucidDB analytical database, 140, 142–143, 502

M

machine learning, 503. *See also* data mining
 Mail Failure job entries, 288–292
 Mail Success job entries, 288–292
 mainframe systems, data analysis using, 195–196
 managers, collecting requirements from, 194
 Mantle. *See* Pentaho user console (Mantle)
 MapComponent, 557–562
 adding geography dimension table, 558–559
 code for including on dashboard, 561–562
 data format, 556–557
 location data action sequence, 559–561
 marker options for showing distribution of customer locations, 563–565
 overview of, 557
 mapping
 adding schema independence, 348–349
 dim_promotion table, 301
 planning loading of dimension table, 300
 relational model to multi-dimensional model, 444
 source to target, 218–219

maps, dashboard, 543
 marker options, for customer locations, 562–565
 Market Analysis By Year example, 18–19
 market basket analysis, and association, 507
 massive parallel processing (MPP) cluster, 142
 master data management (MDM), 126–128
Mastering Data Warehouse Design (Imhoff et al.), 208
 MAT (moving annual total), 217–218
 materialized views, data warehouse
 performance, 130–131
 MDM (master data management), 126–128
 MDX (MultiDimensional eXpressions), 445–460
 calculated members, 459–460
 WITH clause for working with sets, 458–459
 CROSSJOIN function, 457
 cube family relationships, 451–452
 cubes, 446–448
 FILTER function, 455–456
 hierarchies, 448–449, 451
 levels and members, 449–451
 NONEMPTY function, 457–458
 ORDER function, 456–457
 overview of, 445–446
 query syntax, 453–455
 storage formats and, 122–123
 TOPCOUNT and BOTTOMCOUNT functions, 457
 MDX query pane, JPivot, 493
 MDX query tool, PSW, 481–482
 measure dimensions, 447
 measures
 adding to cube fact tables, 469–470
 cubes and, 447
 OLAP Navigator displaying multiple, 493
 ORDER function, 456
 Slice and Dice pivot table example, 17
 transaction and snapshot fact tables, 150
 member sets, OLAP Navigator specifying, 492
 members, MDX, 449–451
 menu bar, user console, 7
 message fields, images in reports, 404
 message queues, real-time data
 warehousing, 141

- Message Template process action, 438
- metadata. *See also* Pentaho metadata layer
 - in data warehouse environment, 113–114
 - displaying in DataCleaner, 200
 - refreshing after publishing report to server, 407–408
 - transformation, 233
 - unclear, 124
 - using ERP systems for data analysis, 195
- Metadata Data Source Editor, 385
- metadata domains, 359
- Metadata Query Language. *See* MQL (Metadata Query Language)
- metadata repository, 358–359
- metadata.xml file, 367
- MetaEditor.bat file, 358
- metaeditor.sh file, 358
- MetaMatrix solution, 140
- metrics. *See* measures
- Microsoft Windows
 - automatic startup in, 43
 - creating symbolic links in Vista, 26
 - how PDI keeps track of repositories, 328–329
 - installing Java on, 28–29
 - installing MySQL GUI tools on, 31
 - installing MySQL server and client on, 30
 - installing Pentaho Server, 38
 - installing Squirrel on, 33
 - job scheduling for, 421
 - obfuscated database passwords in, 334
 - Pentaho Metadata Editor in, 357–358
 - running Carte, 339–340
 - starting and stopping PAC in, 56
 - starting Pentaho BI Server, 5–6
 - starting Pentaho Design Studio, 10
 - starting Spoon application, 236
- mini-dimensions
 - improving monster dimensions with, 174–176
 - junk dimension vs., 181
 - vertical partitioning vs., 179–180
- Minute Sequence step, time dimensions, 277
- missing data, data warehouses, 124
- Missing Date step, date dimensions, 267–268
- mission, managing business from view of, 105–106
- models, creating/saving Weka, 523
- Modified JavaScript Value step, PDI, 277
- Mogwai ERDesigner, 32
- MOLAP (Multidimensional OLAP), 122
- Mondrian
 - Aggregate Designer, 130
 - aggregation and, 496
 - alternatives to aggregation, 502
 - benefits of aggregation, 496–497
 - data warehousing with, 123
 - downloading, 460
 - extending with aggregate tables, 497–500
 - Pentaho Aggregate Designer and, 500–502
 - as Pentaho's OLAP engine, 72
 - types of users working with, 192
 - using aggregate tables, 229
- Mondrian schemas
 - adding measures to cube fact tables, 469–470
 - creating and editing basic, 466
 - creating with PSW, 444
 - cube fact tables, 468
 - cubes, 466–467
 - cubes, associating with dimensions, 476–477
 - dimension tables, 471–474
 - dimensions, adding, 470–471
 - DVD and customer dimensions, 478–480
 - editing tasks, 466
 - hierarchies, 471–476
 - other design topics, 483–484
 - overview of, 444, 460
 - Pentaho Schema Workbench and, 460–463
 - publishing cubes, 482–483
 - testing, 481–482
 - using schema editor, 463–466
 - XML source for, 480–481
- MonetDB, 142, 502
- Moneyball, 503
- monitoring, defined, 309
- monster dimensions
 - mini-dimensions handling, 174–176
 - partitioning, 179–180
- month number, date dimensions, 271–272
- mouse clicks, reacting to on pie chart, 554–555
- moving annual total (MAT), 217–218
- MPP (massive parallel processing)
 - clusters, 142

MQL (Metadata Query Language)
 generating SQL queries with, 351, 355–356
 Pentaho Metadata Layer generating SQL from, 70–71
 storing query specifications as, 352
 MQL Query Builder, 385–386
 MSI installer, 30
 multi-click users (builders), 192–193
 multi-database support, 208
 MultiDimensional eXpressions. *See* MDX (MultiDimensional eXpressions)
 multi-dimensional model, mapping to relational model, 444
 Multidimensional OLAP (MOLAP), 122
 multi-disciplinary data warehouse team, 192
 multiple hierarchies, MDX, 451
 multiple ingoing and outgoing streams, demography dimensions, 285–286
 multi-valued dimensions, and bridge tables, 182–183
 Murphy, Paul, 113–114
 My Schedules pane, Workspace, 427
 MySQL
 GUI tools, 31
 installation, 29–31
 Kickfire for, 144
 migrating system databases to. *See* system databases
 nonsupport for window functions, 132
 Rollup function, 132
 setting up database connections for Weka data mining, 512–515
 MySQL Administrator, 31
 mysql command-line tool, 46
 MySQL Query Browser, 31
 MySQL Workbench, 32
 mysqlbinlog, 137

N

NAICS (North American Industry Classification System), 128
 <name>Components.js, 537
 naming conventions
 data integration process, 228–229
 data warehouses, 162–163
 document templates (outer templates), 540–541
 steps, 240

native mode, Ubuntu in, 23
 natural keys, 161, 229
 Nautilus, 26
 navigation, of data mart data, 184–186
 network traffic, reduced by remote execution, 339
 New Action Sequence wizard, 80
 New option, PRD, 378
 New Project Wizard, PDS, 80–82
 New Task panel, DataCleaner, 199
 No Data, PRD reports, 379
 NON EMPTY function, MDX queries, 457–458
 non-additive facts, 150
 nonintrusive CDC, 133, 136–137
 normalization, 115–116, 186–187
 North American Industry Classification System (NAICS), 128
 notes, transformation or job canvas, 264–265
 NULL values, 124, 169
 Number Analysis profile, DataCleaner, 200, 202
 numeric prediction (regression), data mining, 508

O

OASI (One Attribute Set Interface), 184–186
 obfuscated passwords, 314, 334
 object attributes, editing with schema editor, 465
 Object Management Group (OMG), 352
 ODBC (Open Database Connectivity)
 creating database connections, 251
 creating generic database connections, 258
 overview of, 249
 offsets, and relative time, 167
 OLAP (Online Analytical Processing)
 cubes, 121–122
 data mining compared with, 503
 Navigator, 18–19
 as Pentaho BI Server component, 72
 storage formats and MDX, 122–123
 OLAP Navigator
 controlling placement of dimensions on axes, 489–490
 displaying multiple measures, 493

- overview of, 488–489
- slicing with, 490–491
- specifying member sets, 492
- OMG (Object Management Group), 352
- on demand data warehousing, 144
- One Attribute Set Interface (OASI), 184–186
- one-click users (consumers), 192–193
- Online Analytical Processing. *See* OLAP (Online Analytical Processing)
- online data, and data analysis, 196
- Open Database Connectivity. *See* ODBC (Open Database Connectivity)
- Open Symphony project, 69–70, 411–412
- OpenRules, 141
- Operational BI, 140–141
- Options category, Database Connection dialog, 250
- OR operator, 157
- ORDER BY statement, SQL, 151, 158
- ORDER function, MDX queries, 456–457
- order numbers, as degenerate dimensions, 181
- ordering data, 158
- Orders data mart, creating, 210–212
- organizational performance, analytics and, 503
- OUTER JOIN, 153–155
- outer templates, CDF. *See* document templates (outer templates), CDF
- output
 - action sequence, 85
 - formats, 11–14
 - “Hello World” example, 246
 - streams, PDI, 285–286
- outriggers, 188
- OVER clause, window functions, 131–132
- override, 356

P

- P*A (Power*Architect) tool
 - building data marts using, 210–212
 - creating database connections to build data marts, 210
 - data modeling with, 208–209
 - generating databases, 212–213
 - overview of, 30–31
- PAC (Pentaho Administrative Console)
 - basic configuration, 55–56
 - creating data sources with, 60–61
 - home directory, 38
 - home page, 56–57
 - overview of, 55
 - pluggable authentication, 58
 - security and credentials, 57–58
 - starting and stopping, 56
 - testing dashboard from, 547
 - user management, 58–60
- PAC (Pentaho Administrative Console), schedules
 - creating new schedule, 414–416
 - deleting schedules, 417
 - overview of, 413
 - running schedules, 416
 - suspending and resuming schedules, 416–417
- Package Manager, 29–30
- PAD (Pentaho Aggregate Designer)
 - benefits of aggregation, 496–497
 - defined, 75
 - enhancing performance with, 496
 - extending Mondrian with aggregate tables, 497–500
 - generating and populating aggregate tables, 445
 - overview of, 500–502
 - using aggregate tables, 229
- page breaks, WAQR reports, 375
- Page Header/Footer, PRD reports, 378
- PAGES, axes, 453
- Pan tool
 - as PDI tool, 230–231
 - running jobs/transformations from command line, 330–334
 - using PDI repository. *See* repository, PDI
- parameters
 - custom command-line, 333–334
 - dashboard, 546
 - job scheduler, 418–419
 - report, 13–14, 386–389
 - running Carte, 340
 - running jobs with Kitchen, 332
 - running transformations with Pan, 332
 - specifying value of, 330–331
 - subreport, 405–407
 - when to use instead of variables, 316–317
- parent job scope, variables with, 313
- parent object, and inheritance, 356
- parent-child relationship, cubes, 452

- Partition By clause, window functions, 131–132
- partitioning
 - for data warehouse performance, 129
 - database connection options, 250
 - horizontal, 180
 - using smart date keys, 166
 - vertical, 179
- PAS (Pentaho Analysis Services). *See also*
 - OLAP (Online Analytical Processing)
 - aggregate tables, 445
 - architecture, 442–444
 - components, 442–443
 - overview of, 441
 - schema, 444
 - schema design tools, 444
- passwords
 - connecting to repository, 324
 - creating slave servers, 340
 - installing MySQL on Linux, 29–30
 - installing MySQL on Windows, 30
 - not storing in plain-text files, 202
 - obfuscated database, 314, 334
 - PAC home page, 56
 - publisher, 54–55
 - publishing report to Pentaho BI Server, 407
 - user account, 327–328
- path global variable, dashboard path, 545–546
- Pattern Finder profile, DataCleaner, 200
- Pause button, PAC, 416
- PDF files
 - generated by Steel Wheels, 12
 - implementing bursting in reports, 438
 - JPivot exports, 494
- PDI (Pentaho Data Integration), 223–259
 - adding Weka plugins, 520
 - checking consistency and dependencies, 247–248
 - concepts, 224–230
- PDI (Pentaho Data Integration)
 - data integration engine, 232
 - data integration overview, 223–224
 - defined, 76
 - designing solutions, 261–262
 - Enterprise Edition and, 77
 - generating dimension table data. *See*
 - dimension tables
 - getting started with Spoon. *See* Spoon
 - jobs and transformations, 232–235
 - loading data from source systems. *See*
 - lookup values, staging; promotion dimension
 - plug-in architecture, 235
 - repository, 232
 - Reservoir Sampling, 520
 - tools and utilities, 230–231
 - Weka, getting started with, 520–521
 - Weka and, 519–520
 - Weka data acquisition and preparation, 521–522
 - Weka model, creating and saving, 523
 - Weka scoring plugin, 523–524
 - working with database connections, 248–258
- PDI (Pentaho Data Integration),
 - deployment, 309–343
 - configuration, using JNDI connections, 319–322
 - configuration, using PDI repository, 322–330
 - configuration, using variables, 310–319
 - configuration management, 310
 - overview of, 309
 - remote execution with Carte, 337–342
 - running from command line, 330–334
 - running inside Pentaho BI Server, 334–337
 - using variables. *See* variables
- PDI (Pentaho Data Integration), designing, 261–308
 - generating dimension table data. *See*
 - dimension tables
 - loading data from source systems. *See*
 - lookup values, staging; promotion dimension
 - overview of, 261–262
- PDI repository. *See* repository, PDI
- PDM (Pentaho Data Mining). *See* Weka
 - data mining
- PDS (Pentaho Design Studio), 77–89
 - Action sequence editor, 80–82
 - anatomy of action sequence, 83–89
 - defined, 75
 - Eclipse, 78–80
 - overview of, 77–78
 - studying examples using, 10
- Peazip, 4–5
- Pentaho Administrative Console. *See* PAC
 - (Pentaho Administrative Console)

- Pentaho Aggregate Designer. *See* PAD (Pentaho Aggregate Designer)
- Pentaho Analysis Server. *See* Mondrian
- Pentaho Analysis Services. *See* OLAP (Online Analytical Processing); PAS (Pentaho Analysis Services)
- Pentaho BI Server, 66–74
 - analysis view, 484–485
 - building dashboards for, 529
 - charting examples, 14–16
 - configuring e-mail, 70
 - example solutions included in, 8–9
 - incorporating jobs in action sequences, 336
 - incorporating transformations in action sequences, 334–336
 - installing, 4, 38–43
 - logging in, 6–7
 - overview of, 67
 - and PDI repository, 336–337
 - Pentaho user console, 7–8
 - platform, 67–70
 - presentation layer, 73–74
 - publishing cubes to, 482–483
 - publishing metadata to, 367
 - publishing reports to, 406–407
 - reporting examples, 11–14
 - response to dashboard requests, 533
 - starting, 5–6
 - underlying Java servlet technology, 74
- Pentaho BI Server, components
 - data mining engine, 72–73
 - ETL engine, 72
 - OLAP engine, 72
 - PML (Pentaho Metadata Layer), 70–72
 - reporting engines, 72
 - Web Ad Hoc Query and Reporting Service (WAQR), 72
- Pentaho BI Server, configuring
 - administrative tasks. *See* administrative tasks
 - e-mail, 52–54
 - installation, 38–43
 - managing database drivers, 44–45
 - overview of, 37–38
 - publisher password, 54–55
 - system databases. *See* system databases
- Pentaho BI stack, 63–90
 - creating action sequences with PDS. *See* PDS (Pentaho Design Studio)
 - desktop programs, 74–76
 - front-end/back-end aspect, 66
 - functionality, 65
 - overview of, 63–65
 - Pentaho BI Server. *See* Pentaho BI Server
 - Pentaho EE and Community Edition, 76–77
 - server, client and desktop programs, 65
 - underlying technology, 66–67
 - Weka data mining. *See* Weka data mining
- Pentaho community, maintaining CDF dashboards, 529–530
- Pentaho Corporation, CDF and, 530
- Pentaho Data Integration. *See* PDI (Pentaho Data Integration)
- Pentaho Data Mining (PDM). *See* Weka data mining
- Pentaho Design Studio. *See* PDS (Pentaho Design Studio)
- Pentaho home directory, 38
- Pentaho Metadata Editor. *See* PME (Pentaho Metadata Editor)
- Pentaho metadata layer, 347–369
 - advantages of, 348–350
 - concepts, 356
 - creating metadata queries, 385–386
 - creating PRD data sets from, 381–382
 - database and query abstraction, 352–355
 - defined, 347–348
 - delivery layer, 365–366
 - deploying and using metadata, 366–368
 - inheritance, 356–357
 - localization of properties, 357
 - logical layer, 362–365
 - metadata domains, 359
 - metadata repository, 358–359
 - overview of, 70–72
 - Pentaho Metadata Editor, 357–358
 - physical layer, 359–362
 - PRD using as data source, 373
 - properties, 355–356
 - scope and usage of, 350–352
- Pentaho Metadata Layer (PML), 70–72
- Pentaho Report Designer. *See* PRD (Pentaho Report Designer)
- Pentaho Schema Workbench. *See* PSW (Pentaho Schema Workbench)
- Pentaho user console (Mantle)
 - getting started, 7–8
 - presentation layer, 73
 - refreshing metadata with, 367–368

- Pentaho user console (Mantle) (*continued*)
 - testing dashboard from, 547
 - welcome screen and login dialog, 6–7
- Pentaho User Portal, 376
- `pentaho-init.sh`, 41–42
- `PentahoSystemVersionCheck` schedule, 413
- performance
 - analytics and, 503
 - data volume and, 128–133
 - file-based vs. database repositories, 358–359
- periodic snapshot fact table, 150
- periodical data loading, 118
- permissions, managing, 327
- perspectives, Eclipse IDE, 80
- Petabyte, 128
- Physical Columns, 360–362
- physical data warehousing, 139
- physical layer, metadata model
 - connections, 359–360
 - defined, 71, 355
 - overview of, 359
- Physical Tables and Column Tables, 360–362
- Physical Tables, 360–362
- pie charts
 - adding to PRD reports, 400–402
 - Customers per Website, 548–553
 - as dashboard component, 543
 - reacting to mouse clicks on, 554–555
- `pieclicked()`, 554–555
- `PieSet` collector function, 399–402
- pivot tables
 - for cube visualization, 447–448
 - drilling and, 486–488
 - OLAP Navigator and. *See* OLAP Navigator
 - Slice and Dice Analysis example, 17–18
 - Steel Wheels Analysis examples, 18
- `PivotCategorySet` collector function, 399
- placeholders, document templates, 540
- platform, Pentaho BI Server, 67–70
- plug-ins
 - adding Weka, 520
 - dashboard requests handled by, 534–536
 - installing Squirrel on Ubuntu with, 33
 - PDI, 235–236
 - Weka scoring, 523–524
- `plugin.xml` file, CDF, 535–536
- PME (Pentaho Metadata Editor)
 - defined, 75
 - defining metadata with, 350
 - editing contents of metadata repository, 358–359
 - overview of, 357–358
- PML (Pentaho Metadata Layer), 70–72
- policies, scheduling tool, 420
- Pooling category, Database Connection dialog, 250
- port 8080, Tomcat, 39–40
- port 8099, PAC, 56
- ports, running Carte, 340
- PostgreSQL, 132, 143
- power users (analysts), 192–193
- Power*Architect. *See* P*A (Power*Architect) tool
- PRD (Pentaho Report Designer)
 - adding and modifying groups, 391–393
 - adding and using parameters, 386–389
 - adding charts and graphs, 397–404
 - as banded report editor, 376
 - creating data sets, 381–386
 - creating metadata-based report, 366
 - defined, 76
 - exporting reports, 408
 - layout and formatting, 389–390
 - modifying WAQR reports in, 375
 - overview of, 376–377
 - publishing reports, 406–408
 - report elements, 380–381
 - report structure, 378–380
 - row banding, 390–391
 - using formulas, 395–396
 - using functions, 393–395
 - using subreports, 404–406
 - Welcome screen, 377–378
- PRD Query Designer, 397–398
- predictions
 - data mining for, 506
 - non-numeric, 506–508
 - numeric, 508
- prerequisites, 21–36
 - basic system setup, 22–25
 - database tools, 31–34
 - Java installation and configuration, 27–29
 - MySQL installation, 29–31
 - overview of, 21–22
 - using symbolic links, 25–26
- Preview menu option, PRD reports, 408

- primary key indexes, 129
- primary keys, 160–161, 229
- private schedules, 412, 418–419
- privileges
 - creating repository, 323
 - refining access, 349
- Product Line Analysis example, 18–19
- product management, databases for, 95–96
- `product_type`, 217
- profiling
 - alternative solutions, 205–206
 - authorization of user accounts, 327–328
 - overview of, 197–198
 - using DataCleaner. *See* DataCleaner
 - using Power*Architect tool, 208
- Program Files directory, 5
- projects, Eclipse IDE, 79–82
- promotion dimension
 - data changes, 301–302
 - determining promotion data changes, 304–306
 - `extract_promotion job`, 303–304
 - `load_dim_promotion job`, 302–303
 - mappings, 301
 - overview of, 300–301
 - picking up file and loading extract, 306–307
 - saving extract and passing on file name, 306
 - synchronization frequency, 302
- `promotion table`, 301–302
- promotions, WCM example, 94–95, 102–105
- prompt, running in background, 422–423
- Prompt/Secure Filter action, process actions, 85
- properties
 - accessing report, 435
 - Column Tables, 361–362
 - database connections, 249–251
 - generic database connections, 257–258
 - Pentaho metadata layer, 355–356
 - Physical Tables, 361
 - PRD report, 378
 - quickly opening for tables or columns, 210
 - slave servers, 340
 - SMTP e-mail configuration, 53
 - subscribing to public schedules, 423–424
 - variables, 311–312
 - `.prpt` file format, PRD reports, 376

- pruning, analytical databases, 142
- PSW (Pentaho Schema Workbench), 460–463
 - creating Mondrian schemas, 444
 - defined, 75
 - downloading, 460
 - establishing connection to, 462
 - installing, 461
 - JDBC Explorer and, 463
 - MDX query tool in, 481–482
 - overview of, 442
 - specifying aggregate tables, 499
 - starting, 461
- public schedules
 - allowing users to subscribe to, 423–424
 - creating, 414
 - defined, 412
- Public Schedules pane, Workspace, 427
- Publish to Server dialog, 367
- publisher password, 54–55
- `publisher_config.xml` file, 55, 407
- publishing
 - cubes, Mondrian schemas, 482–483
 - defined, 68
 - directly to Pentaho BI Server, 372, 406–408
- purchase orders, WCM example, 101–105

Q

- quarter number, date dimensions, 272, 273
- Quartz
 - configuring, 47
 - defined, 45
 - Enterprise Job Scheduler, 411–412
 - task scheduling with, 69–70
- Query Designer, SQL, 382–384
- query governor, 164
- query performance, data warehousing, 128–133
 - aggregation, 130
 - archiving, 132
 - bitmap indexes, 129–130
 - indexes, 129
 - materialized views, 130–131
 - partitioning, 129
 - window functions, 131–132
- query redirection, with materialized views, 130
- query syntax, MDX, 453–455

querying star schemas. *See* star schemas, querying
 Quick Start Guide, PRD, 377

R

Raju, Prashant, 51
 randomness of data, prediction models, 509
 rc file (init script), 40–42
 Read-Only profile, PDI repository, 327
 real-time data warehousing, 140–142
 real-time ETL, 118
 record streams, transformation, 233–234
 record types, transformation, 233–234
 records
 distributing through output streams, 286
 transformation, 233–234
 Recurrence options, schedules, 415–416
 recursion, 184
 Red Hat-based Linux, automatic startup, 42
 reference data, master data vs., 128
 refreshing metadata
 after publishing report, 407–408
 with user console, 367–368
 Regex Matcher profile, DataCleaner, 200, 202–204
 regexes (regular expressions), in Data Cleaner, 202–204
 RegexSwap, 203
 region, Orders data mart, 212
 Regional Sales - HTML reporting example, 11–12
 Regional Sales - Line/Bar Chart example, 16–17
 regression (numeric prediction), data mining, 508
 relational model, mapping, 444
 Relational OLAP (ROLAP), 122, 442
 relationships
 Pentaho metadata, 362, 364–365
 relative time, 452
 relative time
 handling, 166–168
 relationships, 452
 reminder report, running, 434–436
 remote execution, 337–339
 rental reminder e-mails example, 430–438
 finding customers with DVDs due this week, 431–432
 getting DVDs due to be returned, 434
 looping through customers, 432–433
 overview of, 430–431
 running reminder report, 434–436
 sending report via e-mail, 436–438
 replication, data warehousing and, 134
 report bursting, 85–89
 Report Header/Footer, PRD, 378
 Report Wizard, PRD, 378
 reporting engines
 defined, 72
 Pentaho Report Designer, 76
 reporting architecture, 372
 reports. *See also* PRD (Pentaho Report Designer)
 alternatives for creating, 64
 attaching output to e-mail message, 436–438
 bursting. *See* bursting
 collecting requirements with existing, 194
 examples of, 11–14
 multi-click users working with, 192
 Pentaho handling JasperReports or BIRT, 372
 Pentaho metadata layer for. *See* Pentaho metadata layer
 power user tools for, 192
 practical uses of WAQR, 375–376
 reminder, 434–435
 reporting architecture, 371–373
 WAQR, 72
 Web-based reporting, 373–375
 repositories
 content, 412–413
 content, managing, 429
 launching Spoon with database, 236
 metadata, 358
 solution, 68
 storing jobs and transformations in database, 232
 repositories.xml file, 328–329
 repository, PDI, 322–330
 automatically connecting to default, 324–325
 configuring for Pentaho BI Server, 336–337
 connecting to, 323–324
 creating, 322–323
 keeping track of, 328–329
 managing user accounts, 326–328

- opening Repository Explorer, 325–326
- overview of, 322
- upgrading existing, 329–330
- repository browser, 8–9
- Repository dialog, 322–324
- Repository Explorer, PDI, 325–328
- request input source, action sequences, 83–84
- requirements analysis
 - collecting requirements, 193–195
 - for data warehouse solution, 191–192
 - getting right users involved, 192–193
- reserved words, avoiding for databases, 163
- Reservoir Sampling, Weka, 520
- Resource option, PRD Welcome screen, 377
- resources directory, CDF, 535
- Resources option, PRD Welcome screen, 377
- resources.txt, CDF, 537
- restrictions, 156–157
- Resume Job process action, Scheduler, 420
- Resume Scheduler process action, Scheduler, 420
- resuming schedules, 416–417, 420
- reverse engineering, 208
- RIGHT OUTER JOIN, 155
- ROLAP (Relational OLAP), 122, 442
- role-playing dimensions, 181–182
- roles
 - managing server, 59–60
 - Mondrian supporting security, 72
 - schema design and, 483
- Rollup function, MySQL, 132
- root job scope, variables, 313–314
- row banding, reports, 390–391
- rows
 - color banding in PRD reports, 390–391
 - OLAP Navigator, 19
 - SCD type 2, 171–173
 - working with field grids, 244
- ROWS, axes, 453–454
- RTF files, exporting reports, 408
- rule algorithm, for classification, 508
- Run Now link, Public Schedules pane of Workspace, 427
- running functions, PRD reports, 393–395
- runtime input source, action sequences, 84

S

- SaaS (Software as a Service), 144
- Sakila sample database, 97
- sampladata database, 45
- samples directory, PDI, 261
- satellites, data vault models based on, 125
- SBI (Serialized Binary Instances), 512
- scalability
 - of remote execution, 338
 - scale-out, 338
 - scale-up, 338
 - working with chart, 401–402
- SCDs (Slowly Changing Dimensions)
 - creating Orders data mart, 212
 - developing global data mart data model, 207
 - overview of, 170
 - type 1: overwrite, 171
 - type 2: add row, 171–173
 - type 3: add column, 174
 - type 4: mini-dimensions, 174–176
 - type 5: separate history table, 176–178
 - type 6: hybrid strategies, 178–179
- Schedule privileges, 424
- Scheduler
 - concepts, 412–413
 - creating and maintaining schedules with PAC, 413–417
 - programming action sequences, 417–420
- Scheduler Status process action, Scheduler, 420
- schedules, defined, 412
- scheduling, 411–422
 - alternatives to using Scheduler, 420–422
 - background execution and, 422–423
 - content repository, 412–413
 - creating/maintaining with PAC, 413–417
 - how subscriptions work, 423–426
 - managing, 61
 - overview of, 411–412
 - programming Scheduler with action sequences, 417–420
 - public and private schedules, 412
 - Scheduler concepts, 412
 - user's workspace, 426–429
- schema. *See also* Mondrian schemas
 - clustering and, 342
 - design tools, 444
 - metadata layer limiting impact of changes to, 348–349

- schema. (*continued*)
 - Pentaho Analysis Services and, 444
 - setting up MySQL, 46
 - using schema editor, 463–466
- schema editor, 463–466
 - changing edit modes, 465–466
 - creating new schema with, 463–464
 - editing object attributes with, 465
 - overview of, 463
 - saving schema on disk, 464–465
- scope
 - choosing variable, 313–314
 - of metadata layer, 350–351
- scoring plugin, Weka, 519, 525–526
- SECTIONS, axes, 453
- Secure Sockets Layer (SSL), 290–291
- security
 - automatically connecting to PDI
 - repository and, 325
 - JDBC configuration, 50–51
 - Mondrian supporting roles, 72
 - PAC configuration, 57–58
 - slave server configuration, 340
 - SMTP configuration, 54
 - user configuration, 58–60
 - using obfuscated database passwords
 - for, 314
- security input source, action sequences, 84–85
- Select a repository dialog, Spoon, 322–323
- SELECT statement, SQL, 151, 156–157
- self-join, 184
- semi-additive measures, 150
- sequence number, relative time, 167
- sequence with offset 0, relative time, 167
- Serialized Binary Instances (SBI), 512
- server
 - administrator's workspace, 428
 - Business Intelligence. *See* Pentaho BI Server
 - Carte. *See* Carte server
 - client and desktop programs, Pentaho BI stack, 65
 - defined, 65
 - HSQLDB database, 6, 45–52
 - slave, 337, 340–341
 - Tomcat. *See* Tomcat server
- Server Administration Console, 368
- Service Manager, Windows, 43
- service.bat script, 43
- servlet container, 74
- servlets, Java, 67, 74
- session input source, action sequences, 84
- SET command, 29
- Set Environment Variables dialog, 312
- set up, customer and websites dashboard, 544
- Set Variables step
 - choosing variable scope, 313–314
 - dynamic database connection example, 314–318
 - limitations of, 319
 - user-defined variables for, 312
 - working with, 318–319
- sets
 - in MDX queries, 458–459
 - specifying member, 492
- settings.xml file, 336–337
- shared dimensions, 471, 476–477
- shared objects files, 256–257
- shared.xml file, 256–257
- shell scripts, starting PSW, 461
- Show repository dialog, Spoon, 322–323
- siblings, cube family relationships, 452
- Simple Mail Transfer Protocol. *See* SMTP (Simple Mail Transfer Protocol)
- single sign-on, Enterprise Edition, 77
- Single Version of Truth, 126
- sizing
 - data warehouse advantages, 112–113
 - report objects, 398
- slave servers, and Carte, 337, 340–341
- Slice and Dice analysis example, 17–18
- slices, pie chart, 400–401
- slicing/slicers
 - defined, 441
 - looking at part of data with, 454
 - with OLAP Navigator, 490–491
- Slowly Changing Dimensions. *See* SCDs (Slowly Changing Dimensions)
- smart date keys, 166
- SMTP (Simple Mail Transfer Protocol)
 - authenticating request, 53–54
 - basic configuration, 52–53
 - e-mailing using, 52, 70
 - Mail job entry configuration, 289–291
 - secure configuration, 54
- snapshot-based CDC, 135–136
- snowflake technique
 - creating Orders data mart, 212
 - outriggers, 188

- overview of, 186–187
- schema design and, 483
- software, downloading and installing, 4–5
- Software as a Service (SaaS), 144
- solution engine, Pentaho BI Server, 68
- `solution` global variable, dashboard, 545–546
- solution repository, Pentaho BI Server, 68
- Sort directory property, staging lookup values, 300
- Sort on Lookup Type step, 293, 299–300
- Sort Rows step, 299–300
- sorting
 - PRD reports, 385–386
 - WAQR reports, 374
- source code, jobs and transformation vs., 233
- source data-based CDC, 133–134
- Source Forge website, 4
- source systems, 286–307
 - data vault reproducing information stored in, 126
 - data warehouse architecture, 117
 - extraction of data, 226
 - loading data from. *See* lookup values, staging; promotion dimension
 - mappings to target data warehouse, 218–219
 - staging area for, 118–119
- SourceForge website, 460
- splitting existing hops, 254
- Spoon
 - Pentaho Data Integration utility, 230–231
 - user-defined variables, 312
 - using PDI repository with. *See* repository, PDI
- Spoon, “Hello World!”
 - building transformation, 237–244
 - Execution Results pane, 245–246
 - launching application, 236–237
 - output, 246
 - overview of, 237
 - running transformation, 244–245
 - working with database connections, 253–256
- `Spoon.bat` script, 236
- `spoon.sh` script, 236
- spreadsheets, and data analysis, 196
- Spring Security, 69–70
- SQL (Structured Query Language)
 - applying query restrictions, 156–158
 - building blocks for selecting data, 151–153
 - Create `dim_date`, loading date dimensions, 265–267
 - creating custom scripts for profiling, 206
 - creating date dimension table, 265–267, 277
 - creating demography dimension table, 282
 - creating queries using JDBC, 382–385
 - creating staging table, 295–296
 - creating tables and indexes, 212–213
 - examining metadata layer, 353–354
 - join types, 153–156
 - loading aggregate table, 498
 - MDX compared with, 453
 - Pentaho Metadata Layer generating, 70
 - querying star schemas and, 151–152
- SQL Editor, 255
- SQLLeonardo, 33–34
- SQLite database connections, 254, 323
- SQLPower, 205–206, 208
- Squirrel, 32, 46
- SSL (Secure Sockets Layer), 290–291
- stack. *See* Pentaho BI stack
- staff members. *See* employees
- `stage_demography` table, 282, 283–284
- `stage_lookup_data` transformation, 287–288, 293–294
- `stage_promotion` table, 304–305
- staging area
 - for extraction of data, 226–227
 - overview of, 118–119
- Staging table Exists step, as Dummy step, 296–297
- Standard Measures profile, DataCleaner, 200, 202
- standards
 - data warehousing, 113, 139
 - ISO, 97
- star schemas
 - building hierarchies, 184–186
 - consolidating multi-grain tables, 188–189
 - creating SQL queries using JDBC, 383
 - dimension tables and fact tables, 148–150
 - junk, heterogeneous and degenerate dimensions, 180–181
 - MDX compared with, 447–448
 - monster dimensions, 179–180
 - multi-valued dimensions and bridge tables, 182–183

- star schemas (*continued*)
 - outriggers, 188
 - overview of, 147–148, 179
 - role-playing dimensions, 181–182
 - snowflakes and clustering dimensions, 186–187
- star schemas, capturing history, 170–179
 - overview of, 169–170
 - SCD type 1: overwrite, 171
 - SCD type 2: add row, 171–173
 - SCD type 3: add column, 174
 - SCD type 4: mini-dimensions, 174–176
 - SCD type 5: separate history table, 176–178
 - SCD type 6: hybrid strategies, 178–179
- star schemas, design principles, 160–169
 - audit columns, 164
 - granularity and aggregation, 163–164
 - modeling date and time, 165–168
 - naming and type conventions, 162–163
 - unknown dimension keys, 169
 - using surrogate keys, 160–162
- star schemas, querying, 150–158
 - applying restrictions, 156–157
 - combining multiple restrictions, 157
 - join types, 153–156
 - ordering data, 158
 - overview of, 150–153
 - restricting aggregate results, 157
- starflake, 186
- START job entry, 287, 288
- start-pentaho.bat script, 5–6, 51
- start-pentaho.sh script, 5–6, 52
- start.sh, 56
- startup
 - automatic, 40–43
 - desktop program, 76
 - modifying scripts when discarding HSQLDB database, 51–52
 - overview of, 5–6
- startup.bat, 56
- Steel Wheels examples
 - analysis, 18–19
 - Chart Pick List, 15
 - Flash Chart List, 15–16
 - Income Statement report, 12–13
 - overview of, 8–9
 - Top 10 Customers report, 13
- Step Metrics grid, Execution Results pane, 245–246
- steps, transformation
 - building transformation in “Hello World!”, 238–244
 - creating, moving and removing, 239
 - hops connecting, 287
 - horizontally or vertically aligning, 243
 - job entries vs., 287
 - overview of, 233–235
 - types of, 239–240
 - using variables. *See* variables
- stop.bat, 56
- stop-pentaho.bat script, 51
- stop-pentaho.sh script, 52
- stop.sh, 56
- Store to Staging Table step, 294
- stored procedures, date dimensions, 262–263
- stovepipe solution, 119–120
- stratified cross-validation, data mining, 509–510
- Stream Lookup step, 293, 297–299
- String Analysis profile, DataCleaner, 200, 202
- structure, PRD report, 378–380
- structured external data, data analysis, 196
- Structured Query Language. *See* SQL (Structured Query Language)
- style inheritance, PRD reports, 389
- <style> tag, .xcdxf file, 538
- stylesheets, forcing for HTML e-mail, 436
- styling, dashboards, 565–568
- sublayers, Pentaho metadata layer, 359–366
- subreports, 404–406
- subscriptions, 423–430
 - creating, 425–426
 - granting Execute and Schedule privileges, 424
 - managing, 61
 - overview of, 423
 - for users, 423–424
 - viewing in Public Schedules pane of Workspace, 427–428
- supervised learning. *See* classification
- support, CDF, 531
- surrogate keys
 - data integration and, 229
 - star schema modeling with, 160–162
- Suspend Job process action, Scheduler, 420

- Suspend Scheduler process action, Scheduler, 420
- suspending schedules, 416, 420
- symbolic links (symlinks), 25–26
- Synaptic Package Manager
 - installing `bum`, 42
 - installing Java on Ubuntu Linux, 27–28
 - installing MySQL GUI tools on Ubuntu, 31
 - installing MySQL on Linux, 29–30
- synchronization frequency, 301–302
- system databases, 45–52
 - configuring Hibernate, 47–50
 - configuring JDBC security, 50–51
 - configuring Quartz, 47
 - configuring sample data, 51
 - modifying Pentaho startup scripts, 51–52
 - overview of, 45–46
 - setting up MySQL schemas, 46
- system of entry, 127
- system of record, MDM, 127
- system setup, prerequisites, 22–25

T

- Table Datasource editor, PRD, 387
- Table Exists step, staging lookup tables, 294
- Table Input step, 316
- Table Output step
 - dynamic database connection example, 317
 - loading date dimension in PDI, 275–276
 - staging lookup values, 300
 - Store to Staging Table step, 294
- tables
 - bridge. *See* bridge tables
 - coloring Power*Architect, 212
 - conventions for, 162
 - creating in SQL, 212–213
 - creating SQL queries using JDBC, 383–385
 - defining relationships manually for report, 383
 - dimension. *See* dimension tables
 - quickly opening properties for, 210
 - tagging, loading invalid data after, 228
- Target Database connection, 253–256
- targets, data mining outcomes, 506
- `.tar.gz` file, 357

- Task Scheduler, 421
- task scheduling, Pentaho BI Server
 - platform, 69–70
- TDWI (The Data Warehousing Institute), 119
- team, multi-disciplinary data warehouse, 192
- technologies
 - Community Dashboard Framework, 531–532
 - Pentaho BI stack, 66–67
- Teiid, 140
- Temp directory property, Join Rows (Cartesian product) step, 280
- `<template>` tag, `.xcd` file, 538
- templates
 - creating action sequences as PDS, 88
 - modifying PRD report, 375–376
 - WAQR reports, 373
- templates, CDF
 - content, 533, 541–542
 - custom document, 568–569
 - document, 533, 538–541
 - overview of, 538
- terminal, 24–25
- Test tab, action sequence editor, 81–82
- testing
 - customer and websites dashboard, 547
 - data mining models, 507
 - database connections, 252
 - Mondrian schemas, 481–482
 - PRD report parameters, 388
 - schedules, 415
 - validating model derived from training process, 509
- text, adding to schedule, 415
- text analytics, data mining, 503
- text editors, Mondrian schemas, 444
- Text file input step, Spoon, 240–243
- Text file output step, Spoon, 243–245
- `TextComponent`, dashboard, 555–557
- “The 38 Subsystems of ETL” (Kimball), 225
- themes, dashboard, 566
- time
 - modeling date and, 165–168
 - relative time relationships, 452
- Time Analysis profile, DataCleaner, 200, 202
- time dimension
 - generating, 213–216
 - granularity, 165

- time dimension (*continued*)
 - loading simple, 277–281
 - role-playing, 182
 - TimeSeries collector function, 399
 - timestamps, inventory management, 105
 - <title> tag, .xcdxf file, 538
 - titles, dashboard, 543, 553–554
 - TMP-file prefix, staging lookup values, 300
 - Tomcat server
 - configuring, 39–40
 - start-pentaho script starting, 6
 - Tomcat5.exe, 43
 - toolbars
 - JPivot, 485
 - Pentaho user console, 7
 - toolset, data mining
 - algorithms, 508–509
 - association, 507–508
 - classification, 506–507
 - clustering, 507
 - numeric prediction (regression), 508
 - overview of, 506
 - Weka data mining, 510
 - Top 10 Customers report, 13
 - TOPCOUNT function, MDX queries, 457
 - training, data mining models, 507, 509
 - transactional fact tables, 149
 - transformation job entries, staging lookup values, 288–289
 - transformation process, ETL. *See also* ETL (Extraction, Transformation, and Loading)
 - aggregation activities, 229
 - data cleansing activities, 228
 - data validation activities, 227–228
 - decoding and renaming activities, 228–229
 - defined, 224
 - key management activities, 229
 - supportive activities, 225
 - Transformation Properties dialog, 253
 - transformations
 - adding database support to “Hello, World”, 253–256
 - adding notes to canvas, 264–265
 - building in “Hello World!”, 238–244
 - checking consistency and dependencies, 247–248
 - creating database connections, 249
 - data integration engine and, 232
 - dynamic database connection example, 314–318
 - exporting in Repository Explorer, 326
 - extract_promotion transformation, 303–304
 - incorporating into action sequences, 334–336
 - jobs composed of, 235
 - jobs vs., 232–233, 287
 - loading demography dimension, 281–286
 - loading promotion dimension, 302–306
 - loading time dimension, PDI, 277–281
 - overview of, 233–235
 - Pentaho data integration tools and components, 231–232
 - remotely executing with Carte, 341–342
 - running from command line, 330–334
 - running in “Hello World!”, 244–245
 - running inside Pentaho BI Server, 334–337
 - running with Pan, 332
 - storing in repository, 232
 - user-defined variables for, 312
 - using database connections, 252–253
 - transformations, loading date dimension table
 - Calculate and Format Dates step, 269–273
 - Create dim_date, 264–265
 - Days Sequence step, 268–269
 - Generate Rows with Initial Date step, 267–268
 - Load dim_date, 275–276
 - Missing Date step, 267–268
 - overview of, 264–265
 - Value Mapper step, 273–275
 - tree view, repository browser. *See* repository browser
 - trigger-based CDC, 134–135
 - 2000 Census Zip Code data set, 109
 - two-click users (refreshers), 192–193
- ## U
- Ubuntu
 - automatic startup in, 42
 - creating symbolic links in, 26–27
 - downloading, 22
 - installing Java on, 27–28

- installing MySQL GUI tools on, 31
- installing MySQL server and client on, 29–30
- installing Squirrel on, 32–33
- running as virtual machine, 23
- using in native mode, 23
- Ubuntu Linux Toolbox* (Negus and Caen), 24–25
- unbalanced hierarchies, 185–186
- UNIX-based systems
 - automatic startup in, 40–41
 - job scheduling for, 421
 - keeping track of PDI repositories in, 328–329
 - location for Pentaho Server installation, 38
 - managing JDBC drivers in, 44–45
 - Pentaho BI Server software placement in, 5
 - running Carte in, 339–340
 - setting up user account, group and directory in, 39
 - starting and stopping PAC in, 56
 - starting Pentaho BI Server in, 5–6
 - starting Pentaho Design Studio in, 10
 - starting Spoon in, 236
 - using CRON job scheduler in, 415
 - using obfuscated database passwords in, 334
 - using Pentaho Metadata Editor in, 357–358
- Unknown Demography step, 283
- unknown dimension keys, 169
- Unknown value, unknown dimension keys, 169
- unsupervised learning. *See* clustering
- `update-rc.d` utility, Linux automatic startup, 42
- updates, data warehousing, 112
- upgrading, PDI repository, 323, 329–330
- US-EN (date dimension), 213–216
- user console. *See* Pentaho user console (Mantle)
- User Information dialog, Repository Explorer, 327–328
- User profile, PDI repository, 327
- user-defined functions, schema design, 484
- user-defined variables, 312–314
- usernames
 - log in, 6–7
 - PAC home page, 56

- slave servers, 340
- user account, 327–328
- users
 - account setup, 38–39
 - authentication and authorization, 69
 - collecting requirements from interviews with, 194
 - connecting to repository, 323–324
 - examining metadata layer, 353–354
 - managing, 58–60
 - managing repository accounts, 326–328
 - requirement changes, data warehousing, 137–139
 - subscribing to public schedules, 423–424
 - types involved in data mart design, 192–193
 - using metadata for friendly interface for, 348
- user's Workspace, schedules/background execution, 426–429
- UTC (Zulu) time, 165–166

V

- `valid_from` timestamp, history preservation, 171–172
- `valid_to` timestamp, history preservation, 171–173
- validation
 - DataCleaner, 199, 205
 - ETL and data, 227–228
 - “Hello World” example transformation, 247–248
- Value Distribution profile, DataCleaner, 201
- Value Mapper step, PDI
 - loading date dimension, 273–275
 - loading demography dimension, 282
- values
 - adding and using in PRD reports, 386–389
 - changing into relative values using % sign, 398
 - creating slave servers using, 340
 - overview of, 330–331
 - passing to subreports, 405–407
- variables, 310–319
 - built-in, 314
 - in configuration properties, 311–312
 - creating slave servers using, 340

variables, (*continued*)
 dynamic database connections example,
 314–318
 icon for, 311
 overview of, 310–311
 picking from list, 312
 Set Variables step, 318–319
 user-defined, 312–314
 vertical partitioning, 179
 views
 Eclipse IDE, 79
 user's Workspace, 426–427
 virtual data warehousing, 139–140
 virtual machines, 23, 67
 VirtualBox, 23–24
 visibility, toggling Execution Results
 pane, 246
 visualization, data mining and, 503

W

Waikato Environment for Knowledge
 Analysis. *See* Weka data mining
 Waiting pane, Workspace, 427
 WAQR (Web Ad Hoc Query and
 Reporting Service)
 creating report views in user console, 73
 modifying in PRD, 373, 375
 as Pentaho BI Server component, 72
 practical uses of, 375–376
 Web-based reporting using, 373–375
 warehouse, WCM example, 95, 96, 104–105
 Watermark, PRD reports, 379
 WCM (World Class Movies), building data
 marts, 210–218
 generating database, 212–213
 generating static dimensions, 213–216
 overview of, 210–212
 special date fields and calculations,
 216–218
 WCM (World Class Movies), example
 business case, 95–105
 basics, 94–95
 big picture, 97–99
 customer orders and promotions,
 102–104
 customers, 101
 DVD catalog, 99–101
 employees, 101
 inventory management, 104–105
 main process flows, 96

obtaining and generating data, 97
 overview of, 93–94, 95–97
 purchase orders, 101–102
 purpose of business intelligence, 105–109
 Web Ad Hoc Query and Reporting
 Service. *See* WAQR (Web Ad Hoc
 Query and Reporting Service)
 Web development skills, CDF dashboards,
 531
 web pages, CDF dashboards as, 532
 Web Publish URL, 406–407
 Web references, 49
 analytical database products, 142–143
 available regular expressions, 203
 Azzurri Clay, 32
 connection pools, 49–50
 creating data sources with PAC, 60–61
 cron expressions, 415
 crontab and cron implementations, 421
 CWM information, 352
 data vault, 126
 DataCleaner, 198
 date format masks, 271
 downloading nightly builds of Pentaho, 4
 e-commerce revenue data, 109
 ERMaster, 32
 ETL activities, 225
 image burner to create bootable CD from
 downloaded file, 22
 InfoBright, 23
 JavaMail API, 54
 Jetty, 55
 JFreeChart, 397
 JNDI, 320
 Kickfire appliance, 144
 Modified JavaScript Value step, PDI, 277
 modifying report templates, 376
 Mogwai ERDesigner, 32
 moving sample data to MySQL, 51
 MySQL GUI tools downloads, 31
 MySQL SELECT syntax, 158
 MySQL Workbench, 32
 obtaining cinematic information, 99
 open source business rules engines, 141
 PAC pluggable authentication, 58
 Pentaho Data Integration documentation,
 261–262
 Pentaho Data Integration tools
 download, 230
 Pentaho Data Integration transforma-
 tions and jobs download, 261

Pentaho Metadata Editor, 357
 Pentaho reporting formulas, 396
 Pentaho's released software downloads, 4
 Power*Architect, 32
 Power*Architect tool, 208
 Quartz and Open Symphony project, 412
 real-time data warehousing, 141–142
 service.bat script and Tomcat5.exe, 43
 snowflaking, 186–187
 SQLLeonardo, 34
 Squirrel SQL client, 32–33
 TDWI (The Data Warehousing Institute) report, 121
 Tomcat manual, 40
 Ubuntu, how to install, 23
 Ubuntu, running as virtual machine, 23
 Ubuntu download, 22
 at utility and Task Scheduler, 422
 virtual data warehousing, 140
 Windows installers for MySQL 5.1, 30
 Web-based reporting, 373–375
 Weblogs, data analysis using, 196
 website table, 301, 302
 website_name Dashboard parameter, 553–554
 websites, WCM example, 94–95
 Weka data mining, 503–527
 engine, 72–73, 76, 192
 Experimenter, 517–518
 Explorer, 516–517
 further reading, 527
 input formats, 511–512
 KnowledgeFlow, 518–519
 setting up database connections for, 512–514
 starting Weka, 514–516
 summary, 527
 toolset, 510
 Weka data mining, using with PDI
 adding PDI plugins, 520
 creating/saving model, 523
 data acquisition and preparation, 521–522
 getting started with Weka and PDI, 520–521
 overview of, 519–520
 scoring plugin, 519, 523–524
 Welcome page, Spoon, 236–237
 Welcome screen, Pentaho Report Designer, 377–378

What You See Is What You Get (WYSIWYG) editor, and PRD, 376–377
 WHERE clauses, SQL, 151–152
 where condition, 383–384
 WHERE statement, SQL, 151, 153
 window functions, data warehouse performance, 131–132
 Windows installer, 30–31
 WITH clause, MDX sets, 458–459
 workspace
 Eclipse IDE, 78–79
 Pentaho user console, 8
 user's, for schedules/background execution, 426–429
 World Class Movies. *See* WCM (World Class Movies), building data marts;
 WCM (World Class Movies), example business case
 WYSIWYG (What You See Is What You Get) editor, and PRD, 376–377

X

X terminal, 24
 xaction editor, 80–82
 .xaction extension, 10–11, 68. *See also* action sequences
 XactionComponent, 551–553
 .xcdf file, 533, 537–538, 544
 XMI (XML Metadata Interchange) format
 building reports on metadata, 366
 creating metadata queries, 385–386
 publishing metadata to server, 367
 .xmi files, storing metadata as, 350
 XML (Extensible Markup Language) files
 action sequences as, 68
 adding database connections in DataCleaner, 201
 data analysis using, 196–197
 desktop programs based on, 76
 dumping repository directory and contents into, 326
 exporting objects in PDI repository to, 329–330
 Tomcat configuration using, 39–40
 XML editor, Mondrian schemas, 444
 XML for Analysis (XML/A) specification, 123
 XML Metadata Interchange. *See* XMI (XML Metadata Interchange) format

XML source, Mondrian schemas, 480–481
XML/A (XML for Analysis) specification,
123
XRFF (eXtensible attribute-Relation File
Format), 511–512
XYZSeries collector function, 400

Y

`year()` function, PRD reports, 397
`ytd_cy`, 217
`ytd_cyvalue`, 217

`ytd_ly`, 217
`ytd_lyvalue`, 217
YYSeries collector function, 399

Z

.zip file
PDI tools as, 230
Pentaho Metadata Editor as, 357
Pentaho’s released software as, 4–5
Zipcensus data sets, 196
Zulu (UTC) time, 165–166