SYMBOLES

: (colon)
  object members, 522
  parameters, 322
, (comma)
  arrays, 522
  row metadata, 28
. (dot)
  row metadata, 28
  UNIX metacharacter, 507
* (asterisk), RSS Output step, 556
@ (at symbol), tokens, 536
^ (caret sign), UNIX metacharacter, 507
{ (curly brackets)
  object literals, 522
  object members, 522
- (dash)
  parameters, 322
  tokens, 536
$ (dollar sign), UNIX metacharacter, 507
$ [] (dollar sign/square brackets), variable
  hexadecimal values, 45
= (equals sign), parameters, 322
# (hash mark), chat channels, 633
%% (percent sign-double), variable names, 45
| (pipe sign), FIFO, 247–248
? (question mark), XPath, 543
' (quotes-double)
  object members, 522
  parameters, 322
  string literals, 522
' (quotes-single)
  parameters, 322
  row metadata, 28
/ (slash), parameters, 322
[ ] (square brackets), arrays, 522
_ (underscore), tokens, 536

A
Abort job utility, 185–186
Abort step, 186
acc, 354
Access
  Database Connection, 38, 93
  file extraction, 134
  Find Duplicate Query Wizard, 117
accumulating snapshot fact table, 260–261
  fact table loader, 121
  loading, 264–265
  partitioning, 264–265
action sequence
  schedule, 333
  transformations, 328–330
ActiveMQ, 459
Add a mini-dimension, SCD, 119
Add constant step, 178
  JavaScript, 396
add crc_src, 484
add crc_vault, 484
Add File to result, RSS Output step, 567
Add new column, SCD, 119
Add sequence step
  internal counters, 211–213
  rows, 405
Spoon, 211–217
surrogate keys, 211–217
Add to result filename, Get data from XML step, 535
Add validation msg in output, XSD Validator step, 529
Add XML step, 518, 540
   export_xml_from_db, 538
   streams, 538
   XML, 538–541
addExport, 416
Additional fields, Split field to rows step, 499
Additional steps, Enterprise Edition, 636
addJob, 416
Addresses tab, Mail, 337
addResultFile(), StepInterface, 618
addTrans, 415
ADempiere, 15
adjacency list, 242
ADLER 32, 484
Administration Console, 330–333
aggregate builder, ETL, 123
aggregate tables, 266–267
Aggregation Designer, Mondrian, 123, 267
Agile BI, 12–14, 302
   Enterprise Edition, 636
agile development
   BI, 12–14
   ETL, 301–306
   Spoon, 301–302
Agile Manifesto, 12–13
allocateSocket, 415
Amazon, EC2, 125, 427–447
Amazon Machine Image (AMI), 439–442
master, 442–443
slaves, 443–444
Ubuntu, 442
AMI. See Amazon Machine Image
analyseImpact(), StepMetaInterface, 599
Ant build tool, Apache, 571
Apache
   ActiveMQ, 459
   Ant build tool, 571
   DBCP, 400
   Log4j, 366
   Subversion, 343, 570
   VFS, 42, 349, 517, 619
   variables, 641–642
API. See application programming interface
application programming interface (API), 569–571
documentation, 596
Java, 570–574
   jobs, 573–574
   parameters, 579–580
   transformations, 572–573
   variables, 579–580
web services, 516
Arbor Essbase, 269
architecture
   DWH, 8
   logging, 364–367
   plugins, 593–599
   transformations, 452
archive files, 60–61
arrays, JSON, 522
ASCII, 15, 18
staging area, 8
at, 327
atlassian.com, 632
Attached Files tab, Mail, 339–340
attribute domain constraints, data quality, 168
attributes
   HTML, 520
   hubs, 468
   multi-valued, 498–500
   Palo, 285
   Root XML element, 541
   rows, 497
   SQL, 484
Attributes, satellites, 470
Attunity Stream, Oracle, 163
audit dimension, 117, 192
auditing. See also logging
data quality, 191–192
ETL, 22
authentication, Carte, 67
authkeypassphrase, 641
authorization, Palo, 285
auto_increment, 217
B
babbelirc.com, 633
backtracking, 32–33
Backup System, 124
Badard, Thierry, 591
balanced hierarchy, 120
balanced hierarchy, 120
BAPI. See Business Application Programming Interface
   BasePartitioner, 625
   BaseStep, 614
   .bat, 58
batch processing, 449
batch run, 449
batch_id (=audit key), 192
batch-level lineage extraction, 358–359
BCNF. See Boyce-Codd normal form
Beginning XML (Hunter), 518
behavioral testing, 306
BI. See business intelligence
The Big Debate, 465
BigNumber, 28
Binary, 28
BIREady, 6
black box testing, 21
Blocking Step, multi-threading, 410
Boolean, 28
Database Connection, 38
flags, 94
job entries, 366
JSON, 522
String, 30
valid, 160
boolean isUnconditional(),
JobEntryInterface, 622
bottlenecks, transformations, 379–382
Bouman, Roland, 228, 319, 327
Boyce-Codd normal form (BCNF), 218, 226
BPM. See Business Process Management
branches/, 342
bridge tables, 121–122
browsers, web, 413
buffers
logging, 364, 365–366
performance, 380
RowSet, 579
transformations, 406–407
Build Model, Spoon, 302
built-in variables and properties, 637–642
Bulk Loader, MySQL, 248, 249
bulk loading
database, 390
fact tables, 246–251
LucidDB, 249
PostgreSQL, 250
Table output step, 250
bundling, 442
messages, 601
Bus Architecture, SCD, 118–119
Business Application Programming
Interface (BAPI), 140
business intelligence (BI), 2
Agile BI, 12–14, 302, 636
ETL, 12
Pentaho BI, 322, 327–333
real-time, 450
Business key, 468
business keys, 209–210
dimension tables, 210, 527
DV, 467
DWH, 209
looking up, 210
natural keys, 210
Sakila, 527
storing, 210
surrogate keys, 210
XML, 527
Business Objects, SQL, 9
Business Process Management (BPM), 344
C
caching, data, 453
Calculate Dimension Attributes,
transformations, 85–86
calculations
casing, 170
Validator step, 180–181
Calculator step, 108–109, 170, 171
Formula step, 201
keys, 214
callback, 550
canvas
jobs, 56
Spoon, 318
transformations, 56
capture groups, regular expressions, 200, 205
Carlton, 6
Carte, 41, 55, 57
authentication, 67
clustering, 57
dynamic clustering, 434
master, 441
slave servers, 411–416, 435
TCP/IP, 57, 417
carte-config-master-8080.xml, 434–435
Cartesian join step, RSS Output step, 564
Cartesian product, 87
Caserta, Joe, 113
case-sensitivity, 507
casing, calculations, 170
Casters, Matt, 167, 315
cat, 248
Catalog location, Mondrian Input step, 274
catalogs, Database Connection, 39
CDC. See Change Data Capture
Central storage, 41
Change Data Capture (CDC), 16
  data extraction, 450–451
  database triggers, 157–158
  dimension table keys, 80–81
  ETL, 115, 154–163
  logging, 162–163
  MySQL, 162–163
  real-time data integration, 450–451
  relational databases, 450
  RFCs, 146
  snapshots, 146, 158–162
  source data, 155–157
  star schema, 227–228
  timestamps, 155–157, 163, 450
  triggers, 450
  Change number of copies to start, User Defined Java Class step, 404

CHANNEL, 641
  channel, RSS, 558–559
  Channel tab, RSS Output step, 565
  channel-log-table, 346
  channels
    log table, 372
    logging, 366
    chat channels, 633
  check()
    JobEntryInterface, 622
    StepMetaInterface, 599
  Check if XML file is well formed, job entries, 519
  checksums, 309
    JavaScript, 395
  partitioning, 426
  child key, 242
  chmod, UNIX, 322
  CI. See Continuous Integration
  classpath, 70–71
  cleansing. See data cleansing
  clear-box testing, 306
  clone(), StepInterface, 616
  Closure Generator step, 242–243
  closure table, 242
  cloud computing, 433–447
  cluster, kettle.pwd, 67
  clustering, 18–20. See also dynamic clustering
    Carte, 57
    data pipelining, 425
    database, 40
    schema, 417–418
    sorting, 394
    TCP/IP, 423
  transformations, 417–425
    partitioning, 430
  CMSs. See content management systems
  Codd, E.F. (Ted), 269
  Cognos, Powerplay, 269
  column profiling, 17, 146
  Combination lookup / update step, 99
    import_xml_into_db.ktr, 527
    Insert / Update step, 241
    junk dimensions, 241
    Spoon, 241
  ComboVar, 609
  Comma Separated Values (CSV), 47–50, 128
    dynamic transformations, 580–583
    sets, 498
  command-line
    jobs, 322–326
    parameters, 323–324, 325–326
    transformations, 322–326
  commands.jar, 596
  commentary, ETL, 298–299
  comments, item, 559
  commit size, Table output step, 390
  common.jar, 596
  Community Edition, 635–636
  Community Wiki, 631
  compatibility mode, JavaScript, 395
  complex join condition, XML Join step, 543
  Compliance Reporter, ETL, 125
  concatenation
    denormalization, 100–101
    Validator step, 179
  Concurrent Versions System (CVS), 343
  ConditionEditor, 609
  configuration, 63–72
    slave servers, 411–412
  conformation. See data conformation
  connect by prior, Oracle, 20
  Connection, Mondrian Input step, 274
  Connection Name, Database Connection, 38, 92–93
  connection pools, 400
  Connection Type, Database Connection, 38, 92–93
  constraints
    dependency, data validation, 183
    domain attribute constraints category, 179
    performance, 392
  content management systems (CMSs), 344, 626
  Content tab
    Add XML step, 539
<table>
<thead>
<tr>
<th>Index</th>
<th>C–D</th>
<th>647</th>
</tr>
</thead>
<tbody>
<tr>
<td>Get data from XML step</td>
<td>532–535</td>
<td></td>
</tr>
<tr>
<td>Regex Evaluation step</td>
<td>506–507</td>
<td></td>
</tr>
<tr>
<td>RSS Input step</td>
<td>561–562</td>
<td></td>
</tr>
<tr>
<td>Continuous Integration (CI)</td>
<td>13, 450, 630</td>
<td></td>
</tr>
<tr>
<td>testing</td>
<td>311</td>
<td></td>
</tr>
<tr>
<td>control files</td>
<td>247</td>
<td></td>
</tr>
<tr>
<td>COPY, PostgreSQL</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>Copy rows to result step</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>CDC</td>
<td>164</td>
<td></td>
</tr>
<tr>
<td>job entries</td>
<td>399</td>
<td></td>
</tr>
<tr>
<td>transformations</td>
<td>576–577</td>
<td></td>
</tr>
<tr>
<td>Copy Table</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Copy Tables</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Copy tables wizard, Spoon</td>
<td>584</td>
<td></td>
</tr>
<tr>
<td>Core, Java API</td>
<td>571</td>
<td></td>
</tr>
<tr>
<td>Counter name field, Step sequence step</td>
<td>212</td>
<td></td>
</tr>
<tr>
<td>CPU performance</td>
<td>394–398</td>
<td></td>
</tr>
<tr>
<td>CRC. See Cyclic Redundancy Check</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Create custom RSS, RSS Output step</td>
<td>565, 567</td>
<td></td>
</tr>
<tr>
<td>Create new rows, SCD</td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>Create Parent folder, RSS Output step</td>
<td>567</td>
<td></td>
</tr>
<tr>
<td>CRM. See Customer Relationship Management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crockford, Douglas</td>
<td>520</td>
<td></td>
</tr>
<tr>
<td>cron, UNIX</td>
<td>326–327</td>
<td></td>
</tr>
<tr>
<td>crontab, UNIX</td>
<td>326–327</td>
<td></td>
</tr>
<tr>
<td>CSV. See Comma Separated Values</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSV File Input step</td>
<td>384–385</td>
<td></td>
</tr>
<tr>
<td>dynamic templates</td>
<td>584</td>
<td></td>
</tr>
<tr>
<td>partitioning</td>
<td>428</td>
<td></td>
</tr>
<tr>
<td>rows</td>
<td>394</td>
<td></td>
</tr>
<tr>
<td>StepInterface, 617</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CsvFileReader</td>
<td>584</td>
<td></td>
</tr>
<tr>
<td>CsvFileReader.java</td>
<td>580–583</td>
<td></td>
</tr>
<tr>
<td>cubes</td>
<td>123</td>
<td></td>
</tr>
<tr>
<td>dimensions</td>
<td>270–271</td>
<td></td>
</tr>
<tr>
<td>OLAP</td>
<td>270–271</td>
<td></td>
</tr>
<tr>
<td>XML/A</td>
<td>278</td>
<td></td>
</tr>
<tr>
<td>current_flag</td>
<td>265</td>
<td></td>
</tr>
<tr>
<td>CURRENT_TIMESTAMP</td>
<td>480</td>
<td></td>
</tr>
<tr>
<td>Customer Relationship Management (CRM)</td>
<td>14–15, 138–146</td>
<td></td>
</tr>
<tr>
<td>deduplication</td>
<td>192–199</td>
<td></td>
</tr>
<tr>
<td>customizations.jar</td>
<td>591</td>
<td></td>
</tr>
<tr>
<td>CVS. See Concurrent Versions System</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyclic Redundancy Check (CRC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter rows step</td>
<td>485</td>
<td></td>
</tr>
<tr>
<td>Merge Join step</td>
<td>485</td>
<td></td>
</tr>
<tr>
<td>NULL</td>
<td>484</td>
<td></td>
</tr>
<tr>
<td>Update step</td>
<td>485</td>
<td></td>
</tr>
</tbody>
</table>

### D

Damerau-Levenshtein algorithm | 171 |
--- | --- |
data |
--- | --- |
 acquisition challenges | 14–16 |
 caching | 453 |
 delivery | 118 |
 federation | 10–11 |
 governance | 168 |
 late-arriving | 255–260 |
 dimensions | 256–260 |
 ETL | 122 |
 facts | 256 |
 migration | 9 |
 paths | 25 |
 SAP | 140–145 |
 semi-structured | 501–508 |
 sorting | 392–394 |
 static | 397–398 |
 streams | 577 |
 synchronization | 9 |
 traceability | 467 |
 transformations | 576–580 |
 unstructured | 501–508 |

Data Cleaning and Quality Screen Handler System, ETL | 116–117 |
--- | --- |
Data Cleanse step | 169 |
--- | --- |
data cleansing |
--- | --- |
data governance | 168 |
data quality | 168 |
data validation | 179–183 |
 ETL | 168–183 |
 reference tables | 172–179 |
 regular expressions | 203–205 |
 source data | 173 |
data conformation |
--- | --- |
 ETL | 118 |
 lookup tables | 172–175 |
 reference tables | 175–179 |
data conversion |
--- | --- |
 JavaScript | 395 |
 transformations | 29–30 |
data extraction | 127–165 |
--- | --- |
 Access files | 134 |
 CDC | 450–451 |
 database | 134–136 |
 ETL | 114–116 |
 Excel files | 134 |
 Freebase | 553–558 |
 HTML | 520 |
 HTTP client | 137–138 |
in input steps | 128 |
 lineage | 358–359 |
metadata, 359
parallelism, 538
real-time, 138
SOAP, 138
Spoon, 128
streams, 138
text files, 128–132
Web, 137
Web, 137–138
XBase files, 134
XML, 525–536
XML files, 133
data formats
HTTP, 517
non-relational, 498
non-tabular, 498
web services, 517–523
XML, 518–520
Data Grid step, 135, 173
SAP Input step, 142
Table input step, 132
testing, 311
data integration, 8, 569–592
challenges, 11–17
continuous, 450
ETL, 123
import_xml_into_db.ktr, 527
near real-time, 450
real-time, 449–461
Spoon, 302
streaming, 450
streams, 450
Data Integration Server, Enterprise Edition, 636
data lineage, 21, 357–363
data extraction, 358–359
impact analysis, 361–363
Data Manipulation Language (DML), 246
data mapping, 297–298
fields, 25
Sakila, 524–525
XML, 524–525
data mart
DV, 486–495
data pipelining
clustering, 425
multi-threading, 407–408
Data Profiler, Talend, 154
data profiling, 16–17, 127–128, 146–154
metadata, 17
Data Profiling System, 115
data quality
auditing, 191–192
categories, 168–169
challenges, 16–17
data cleansing, 168
Data Quality Assessment (Maydanchik), 169
Data Quality Lifecycle, 191
data types, Validator step, 179
data validation
data cleansing, 179–183
DataCleaner, 148, 153
dates, 182
deduplication, 183
dependency constraints, 183
error handling, 187–190
metadata, 182
NULL, 17, 180–181
rules, 180–183
Unknown, 17
XML Schema, 530
XSD Validator step, 530
Data Validator step, 179, 182, 187–188
Data Vault (DV), 9, 168, 465–495
business keys, 467
data mart, 486–495
database accounts, 477
dependency, 471
ETL, 477
extensibility, 471
hubs, 467–468
links, 468–469
NULL, 471
Sakila, 472–486
satellites, 469–471
tables, 485–486
3NF, 469
timestamps, 480
traceability, 471
data virtualization, EII, 10–11
data warehouse (DWH), 2
architecture, 8
business keys, 209
EDW, 465
jobs, 31
response time, 4
surrogate keys, 210
The Data Warehouse ETL Toolkit (Kimball and Caserta), 113, 191
Data Warehouse Lifecycle Toolkit (Kimball), 11, 113–114
Data Warehouse Toolkit (Kimball and Ross), 221, 228
Database, Java API, 571
database
accounts, 82
DV, 477
bulk loading, 390
CDC, 155
clustering, 40
connection pools, 400
connections, multi-threading, 408–409
extraction, 134–136
metadata, 588–590
OLTP, 75
partitioning, 40, 429–430
performance, 388–392
plugins, 627–628
repositories, 348–349
Sakila, 73–110
sequence, 211
surrogate keys, 217
sharding, 40
shared objects, 589
sorting, 393
time-outs, 453
triggers, CDC, 157–158
Database Connection, 37–41
DataCleaner, 150–151
Enable Connection Pooling, 400
Sakila, 90–95
transformations, 37, 90–95
Database Name, Database Connection, 38
Database repository, 41
DatabaseMeta, 589
DataCleaner
data validation, 153
Database Connection, 150–151
dependency, 153–154
dictionary, 153–154
eobjects.org, 147–154
JavaScript, 153
regular expressions, 151–152
Run profiling, 152
data-integration, 61
date
data validation, 182
dimensions, 239
Date, 28
Integer, conversion, 30
String, conversion, 29
Date mask matcher, DataCleaner, 149
Date of last insert (without stream field as source), Dimension lookup / update step, 238
Date of last insert or update (without stream field as source), Dimension lookup / update step, 238
Date of last update (without stream field as source), Dimension lookup / update step, 239
datetime-stamp, 484
DBCP, Apache, 484
dbhost, 354
deadlock, 383
Debian, 59
Debug option, 312–315
debugging
ETL, 21, 312–315
jobs, 56
logging, 364
real-time transformation streaming, 457–478
rows, 314
transformations, 56
decision support systems (DSS), 2
deduplication, 104
CRM, 192–199
data validation, 183
ETL, 117–118
exact duplicates, 193–194
non-exact duplicates, 194–195
transformations, 195–199
delays, parameters, 457
DELETE, 157
deleted records, CDC, 155
delimited text files, 128
DeMarco, Tom, 12
denormalization, 99
concatenation, 100–101
constraints, data validation, 183
data quality, 168
DataCleaner, 153–154
dimension tables, 219
DV, 471
profiling, 146
description, metadata, 37
Description, 332
description, 346
channel, 559
item, 559
descriptive fields, 318
design
building blocks, 25–42
flexibility, 19
principles, 23–25
Design ETL, Spoon, 302
develop, 354
dictionary, DataCleaner, 153–154
Dictionary matcher, DataCleaner, 149
dimension(s)
cubes, 270–271
hierarchies, 270
junk, 120, 241
late-arriving data, 256–260
mini-dimensions, 120, 239–240
Palo, 285
special dimension builder, 120
static, 84–87
user maintained, 120
Dimension lookup / update step, 99, 238
data caching, 453
history, 235–236
indexes, 390–391
Keys tab page, 234
late-arriving data, 259
SCD, 232–237
SOF, 266
surrogate keys, 234–235
dimension manager system, ETL, 122
dimension tables
business keys, 210, 527
dependency, 219
ETL, 207–244
fact tables, 251–260
keys, 109, 208–217
CDC, 80–81
loading, 218–228
natural keys, 99
OLTP, 226
rental star schema, 79–80
rows, 90
SCD, 118
snowflakes, 97, 218–225
star schema, 226–228
static, 84–87
surrogate keys, 209, 251–260
surrogate primary keys, 80
dimension(s) static dimensions, special
dimension builder, 120
dir, 324
directory, metadata, 36
dispose(), StepInterface, 614
distributed version control systems (DVCS),
344
DML. See Data Manipulation Language
Do Not Proceed, 90
Do not raise an error if no files, Get data
from XML step, 534
docs/api, 571
Document template XML step
export_xml_from_db, 537
XML Join step, 542
document type definition (DTD), 519
document-all, 319
documentation, API, 596
domain attribute constraints category, 179
dotall mode, 507
Double Metaphone algorithm, 171
DQGuru, 118
drill-down, 314
driver, jdbc.properties, 65
dSS. See decision support systems
DTD. See document type definition
DTD Validator step
job entries, 519
XML, 133
dV. See Data Vault
dVCS. See distributed version control
systems
DVD rental business, 73–110
DWH. See data warehouse
dynamic clustering
Carte, 434
cloud computing, 433–447
master, 434
schema, 434
dynamic ETL, 586–587
dynamic jobs, 584–586
Dynamic Slave, 445
dynamic templates, 583–584
dynamic testing, 307
dynamic transformations
CSV, 580–583
Spoon, 580–583
DynamicJob.java, 584
E
E4X. See ECMAScript for XML
EBS. See Elastic Block Service
Index

E-Business Suite, Oracle, 18
EC2. See Elastic Computing Cloud
e2-ami-tools, 440
e2-bundle-vol, 442
e2-describe-instances, 444
ECCD. See Extract, Cleanse, Conform, and Deliver
Eclipse, 302, 597–598, 607
ECMAScript, 394–396
ECMAScript for XML (E4X), 520
ecosystem, 629–634
Edit button, Database lookup step, 222
Edit mapping button, Insert / Update step, 231–232
EDW. See enterprise data warehouse
Elastic Block Service (EBS), 438
Elastic Computing Cloud (EC2), Amazon, 125, 427–447
elements
HTML, 520
user interface, 609
Elements name, Add XML step, 540
ELT. See extract, load, and transform
e-mail
HTML, 339
notifications, 336–340
Email Message tab, Mail, 339
embedding, 574–590
libraries, 574
Enable cache, Database lookup step, 253
Enable Connection Pooling, Database Connection, 400
Encoding list box, Get data from XML step, 533
Encr.bat, 58
encr.sh, 58
Engine, Java API, 571
EnterListDialog, 609
EnterMappingDialog, 609
EnterNumberDialog, 609
EnterPasswordDialog, 609
enterprise data warehouse (EDW), 465
Enterprise Information Integration (EII) data virtualization, 10–11
LucidDB, 10
Enterprise Repository, Enterprise Edition, 636
Enterprise Resource Planning (ERP), 127, 138–146
metadata, 14, 139
plugins, 140
EnterSelectionDialog, 609
entries, 346
environment, 354
environmentSubstitute(), StepInterface, 618–619
eobjects.org, DataCleaner, 147–154
ERP. See Enterprise Resource Planning
Error code, Data Validator step, 187
Error description, Data Validator step, 187, 188
Error fields, Data Validator step, 188
error handling
data acquisition, 15
data validation, 187–190
dynamic templates, 584
ETL, 117, 183–190
process errors, 184–187
StepInterface, 617
transformations errors, 186–187
XSD Validator step, 530
error messages, logging, 364
ErrorDialog, 609
/etc/init.d/carte, 441
ETL, 6
ETL. See extract, transform, and load
ETLT. See extract, transform, load, transform
eval(), JavaScript, 556
Excel
file extraction, 134
OLAP, 273
Excel Output step, 297
excludeFromCopyDistribute
Verification(), StepMetaInterface, 600
excludeFromRowLayoutVerification(), StepMetaInterface, 600
ExecTrans.java, 576
execute(), JobEntryInterface, 622
Execute a transformation, Spoon, 413
Execute SQL script job, Spoon Copy tables wizard, 584
Execute SQL script step
aggregate tables, 267
foreign keys, 252
import_xml_into_db.ktr, 526
Execute SQL step, multi-threading, 409–410
ExecuteJob.java, 573–574
ExecuteTrans.java, 572
export
metadata, StepMetaInterface, 600
repositories, 350–351
resource exporter, 444
exportResources()
JobEntryInterface, 622
StepMetaInterface, 600
export_xml_from_db, 537–538
expressions, Java, 70–71
extended description, metadata, 37
extensibility, 593–628. See also plugins
DV, 471
ETL, 19–20
eXtensible Markup Language (XML)
Add XML step, 538–541
business keys, 527
data extraction, 525–536
data format, 518–520
data mapping, 524–525
document construction, 538
document structure, 523–524
documents, generating, 537–544
ETL metadata, 24
examples, 523–544
file extraction, 133
job entries, 519–520
jobs, metadata, 346–347
JSON, 520
metadata, 345–347
repositories, 344
Sakila, 523–544
slave servers, 413
surrogate keys, 527
transformations, metadata, 345–346
VCS, 352
Version Migration System, 352
web services, 518–520
eXtensible Stylesheet Language (XSL), 133
Extension, RSS Output step, 567
external sort, 393
Extract, Cleanse, Conform, and Deliver
(ECCD), 167
extract, load, and transform (ELT), 9
extract, transform, and load (ETL)
aggregate builder, 123
agile development, 301–306
audit dimension assembler, 117
auditing, 22
Backup System, 124
best practices, 296–300
BI, 12
building blocks, 7–8
CDC, 115, 154–163
commentary, 298–299
Compliance Reporter, 125
connectivity, 17–18
Data Cleaning and Quality Screen Handler
System, 116–117
data cleansing, 168–183
data conformer, 118
data delivery, 118
data integration manager, 123
data migration, 9
data paths, 25
Data Profiling System, 115
data synchronization, 9
debugging, 21, 312–315
deduplication, 117–118
definition, 5
design flexibility, 19
development lifecycle, 295–320
dimension manager system, 122
dimension tables, 207–244
DV, 477
dynamic, 586–587
error handling, 117, 183–190
evolution, 5–6
extensibility, 19–20
extraction, 114–116
Extraction System, 115–116
fact table, loader, 121
fact table, provider system, 122
flow design, 300
hierarchy dimension builder, 119–120
impact analysis, 125
Job Scheduler, 124
jobs, 12, 30–36
late-arriving data handler, 122
Lineage and Dependency
Analyzer, 125
logging, 22
maintainability, 300–301
metadata, 21, 344–350
metadata, graphical user interface, 24
metadata, XML, 24
Metadata Repository Manager, 125
MOLAP, 123
monitoring, 333–340
multi-valued dimension bridge table
builder, 121–122
names, 24, 298–299
Parallelizing/Pipelining System, 125
platform independence, 18
Problem Escalation System, 125
RDBMS, 497
Recovery and Restart System, 124
reuse, 19, 300–301
Sakila, 73–110, 81–84
scalability, 18–19
SCD, 118–119
scheduling, 321–333
scripts, 5, 200–205
Security System, 125
solution documentation, 315–320
Sort System, 124–125
special dimension builder, 120
Spoon, 81–84
subsystems, 113–126
surrogate key creation system, 119
testing, 21, 306–312
tools, 6
tools, requirements, 17–22
transformations, 12, 25–30
transformations, challenges, 20
transparency, 24
Version Control System, 124
Version Migration System, 124
Workflow Monitor, 124
extration. See data extraction
Extraction System, 115–116
Extreme Programming (XP), 301

F
fact tables
accumulating snapshot fact table, 260–261
fact table loader, 121
loading, 264–265
partitioning, 264–265
bulk loading, 246–251
dimension tables, 251–260
Insert / Update step, 109
loader, 121
loading, 245–267
periodic snapshot fact tables, 260–261
fact table loader, 121
loading, 263–264
provider system, 122
rental star schema, 79
snapshots, 260–261
SOF, 261–263
loading, 265–266
transaction grain fact tables, 121
facts, late-arriving data, 256
failure hops, 90
Fetch Customer Address, 96–98
Field description, transformation log
tables, 369
Field Splitter step, 145
Field to split, Split field to rows step, 499
fields
JavaScript, 395
maps, 25
rows, 27
Select values step, 397
text files, 384
Fields grid, Regex Evaluation step, 506
Fields tab, 237
Add XML step, 539–541
Get data from XML step, 535–536
The fields that make up the grouping,
Denormalize Special Features, 104
FIFO. See first in, first out
file, 324
File locking, 42
File Output step, export_xml_from_db, 538
File repository, 41
File tab
Get date from XML step, 531–532
RSS Ouput step, 566–567
file-based version control systems, 342–344
filename, metadata, 36
filename, PostgresSQL, 250
Filename defined in a field, RSS Output step, 567
Filename field, RSS Output step, 567
FileObject, 619
Filter rows step, 198, 199
CRC, 485
UI, 609
Filter step, 108
Find Duplicate Query Wizard, Access, 117
first in, first out (FIFO), 247–248
Fixed File Input step, 385
StepInterface, 617
fixed width text files, 129
flags, Boolean, 94
floating point numbers, 28
JSON, 522
“follow when result is false” job hop, 32
“follow when result is true” job hop, 31–32
Force all to lower case, Database Connection, 38
Force all to upper case, Database Connection, 39
Foreign key, satellites, 470
foreign keys
Execute SQL script step, 252
parent key, 242
referential integrity, 251–252
Sakila, 105
SOF, 265
tables, 77, 208
forks, 591–592
Form/Layout, 607–608
Formula step, Calculator step, 201
forums, 631–632
Free Software Foundation (FSF), 570
Freebase, 549–558
data extraction, 553–558
MQL, 551–553
performance, 552
read service, 550–551
scalability, 552
web services, 550
Wikipedia, 549–550
FSF. See Free Software Foundation
full table scan, 391
Function Browser, SAP, 141
functional testing, 21, 306
fuzzy logic, 195
Fuzzy match step, 170, 171, 195–198

G
GA. See General Availability
gap penalty, 171
gender, coding for, 175–176
GENERAL, job entry results, 34
General Availability (GA), 629–630
Generate Row with URLs, Generate Rows step, 526
Generate Rows step, 314
Add sequence step, 213
export_xml_from_db, 537
Freebase, 554
import_xml_into_db.ktr, 526
RSS Output step, 564
SAP Input step, 142
SOAP, 547
generateJobMeta(), 586
GeoKettle, 591
GeoNames, 137, 172
GET, SOAP, 548
get(), User Defined Java Class step, 620
Get closer value, 196–197
Get data from XML step, 361, 518, 530–536
import_xml_into_db.ktr, 527
SOAP, 548
Get Fields button, 129
Insert / Update step, 231
Get File names step, 203
Get Lookup Fields button, Database lookup step, 223
Get rows from result step, 576–577
Get System Info step, 191
command-line parameters, 325–326
late-arriving data, 259
transformation log tables, 367, 370
Get update fields, Insert / Update step, 231–232
Get variables step, JavaScript, 396
Get XPath nodes, Get data from XML step, 522
getCopy(), 617
diaLogClassName(), 622
getEntryNr, 587
getExitStatus, 587
getFields()
StepInterface, 619
StepMetaInterface, 599, 604
getInputRowMeta(), 615
getLogChannelId, 588
getNrErrors, 587
getNrFilesRetrieved, 588
getNrLinesDeleted, 588
getNrLinesInput, 587
getNrLinesOutput, 587
getNrLinesRead, 587
getNrLinesRejected, 588
getNrLinesUpdated, 588
getNrLinesWritten, 588
getOptionalStreams(), 600
getPartition(), 625
getRequiredFields(), 600
getResourceDependencies()
JobEntryInterface, 622
StepMetaInterface, 600
getResult, 587
getResultFilesList, 588
getRow(), 614–615, 618
gRowFrom(), 616
gRows, 588
gRunThread(stepname, copy), 576
getSlaves, 416
getSQLStatements()
JobEntryInterface, 622
StepMetaInterface, 599
getStepIOMeta(), 600
getUniqueStepCountAcrossSlaves, StepInterface, 617
getUniqueStepNRAcrossSlaves(), 617
getUsedDatabaseConnections(), 622
getusedLibraries(), 600
gXML(), 583
JobEntryInterface, 622
StepMetaInterface, 599
GetXMLData - Different Options.ktr, 299
gXulOverlayFile(), 628
GIS. See Graphical Information Systems
Globally Unique identifiers (GUID),
    JavaScript, 395
globalreplace.sh, 347
GNOME, launchers, 62–63
GNU Public License (GPL), 343
GoldenGate, Oracle, 163
good-enough solutions, regular expressions, 501–502
Goodman, Nicholas, 19
Google Wave, 626
GPL. See GNU Public License
Graphical Information Systems (GIS), 591
graphical user interface (GUI), 24
    Java API, 571
grid-based services, 437
Guess button, Insert / Update step, 232
GUI. See graphical user interface
GUID. See Globally Unique identifiers
guid, item, 559
    .gzip, 517

H
handleStreamSelection(),
    StepMetaInterface, 600
hard disks, 386–387
heaps, maximum size, 71
HellowworldStepDialog.java, 609–613
hierarchy
    dimension builder, 119–120
    dimensions, 270
flattener, 20
ragged, 120
reursion, 120, 242–243
variables, 120
Hillyer, Mike, 74
history
    data quality, 168
    Dimension lookup / update step, 235–236
    transformation log tables, 367–368
HOLAP. See Hybrid OLAP
/home/ubuntu/runCarte.sh, 441
hop, 346
hops, 7
    failure, 90
    jobs, 31–32
    loops, 27
    rows, 26, 27
    success, 90
    transformations, 25, 26–27
    unconditional, 88
hops, 346
Host Name, Database Connection, 38
HTML
    attributes, 520
data extraction, 520
    elements, 520
e-mail, 339
    JavaScript, 520
    web pages, 520
    web services, 520
HTTP. See Hypertext Transfer Protocol
Http Authentication, Web services lookup step, 545
HTTP client step, 516
    extraction, 137–138
Freebase, 555
    import_xml_into_db.ktr, 526–527
SOAP, 547–548
HTTP GET, 548
    Freebase, 550
HTTP Post step, 516, 548
Hub Surrogate Keys(), 469
hubs
    attributes, 468
    DV, 467–468
    Sakila, 472–473
    surrogate keys, 469
tables, 467
Hudson, 60
Hunter, David, 518
Hybrid, SCD, 119
Hybrid OLAP (HOLAP), 272
Hyde, Julian, 458
Hypertext Transfer Protocol (HTTP), 515–517
data formats, 517
XML/A, 277, 278

I
IaaS. See Infrastructure as a Service
    .//id, XPath, 535
    .//@id, XPath, 535
IDE. See integrated development environment
identifiers, 317
IDENTITY, 217
identity, 641
id_existing, 480
Ignore comments?, Get data from XML step, 533
Ignore empty file, Get data from XML step, 534
il8n, 590
image, 601
Imhoff, Claudia, 296
impact analysis, 22
date lineage, 361–363
ETL, 125
StepMetaInterface, 599
import, repositories, 350–351
Import partitions button, Partitioning schema, 429
import_xml_into_db, Get data from XML step, 532
import_xml_into_db.ktr, 525–527
Execute SQL script step, 526
“in” tab, Web services lookup step, 545–546
Include date in filename, RSS Output step, 567
Include filename in result and Filename fieldname, Get data from XML step, 534
Include rownum in output, RSS Input step, 561–562
Include rownum in output?, Split field to rows step, 500
Include stepnr in filename, RSS Output step, 567
Include time in filename, RSS Output step, 567
Include url in output, RSS Input step, 561–562
incoming hops, 26
Increment by field, Step sequence step, 212
incrementLinesInput(), 618
incrementLinesOutput(), 618
incrementLinesRead(), 618
incrementLinesRejected(), 618
incrementLinesSkipped(), 618
incrementLinesUpdated(), 618
incrementLinesWritten(), 618
indexes
   performance, 390–392
tables, 392
IndexOfValue(), 606–607
InfiniDB, 123
info, 345
InfoBright, 123
Informatica, 9
Infrastructure as a Service (IaaS), 437
init()
   StepInterface, 614
   User Defined Java Class step, 459
InjectDataIntoTransformation.java, 578–579
Injector step, 578
Inmon, Bill, 465
Input source step, 479
   SQL, 483
Input step extraction, 128
Mondrian, 274–275
OLAP, 274, 278–279, 281
process, 278
Input Table step, 297
Input vault step, 479
input_id, 101
input/output, 380
$, 67
inputRowMeta, StepMetaInterface, 604
Insert, Dimension lookup / update step, 238
INSERT, 157
Insert, bulk loading, 251
Insert / Update step, 101–102
   accumulating snapshot fact tables, 264
   CDC, 155, 163
   Combination lookup / update step, 241
   fact table, 109
   keys, 230–231
   SCD, 229–230
   Update fields, 231–232
installation, 58–63
   rental star schema, 81
Sakila database, 77
Installer, Enterprise Edition, 636
Integer, 27
   Date, conversion, 30
integrated development environment (IDE) plugins, 596–597
Spoon, 55–57
Integrated Scheduling, Enterprise Edition, 636
integration. See also data integration testing, 307
internal counters, Add sequence step, 211–213
internal variables, 428–429
Internal.Cluster.Master, 639
Internal.Cluster.Size, 639
Internal.Job.Name, 638
Internal.Kettle.Build.Date, 638
Internal.Kettle.Version, 637
Internal.Slave.Server.Name, 639
Internal.Slave.Transformation.Number, 639
Internal.Step.CopyNr, 639
Internal.Step.Name, 639
$(Internal.Step.Partition.ID), 429
Index

Internal.Step.Partition.ID, 638
$(Internal.Step.Partition.Number), 429
Internal.Step.Partition.Number, 639
Internal.Step.Unique.Count, 639
Internal.Step.Unique.Number, 639
$(Internal.Transformation.Filename.Directory), 530
Internal.Transformation.Name, 638
internationalization, 590
Internet Relay Chat (IRC), 633
inter-table dependencies, 183
interval logging, 453
intra-table dependencies, 183
intrusive CDC, 16, 155
IRC. See Internet Relay Chat
Is a file
filename is defined in a field, XSD
Validator step, 529–530
let me specify filename, XSD Validator step, 529–530
Is defined inside XML, XSD Validator step, 529–530
ISNULL, 484
ISO8601, 170
isStopped, Result, 588
item, RSS, 559–560
Item tab, RSS Ouput step, 565–566

J

J2EE. See Java 2 Enterprise Edition
JAAS. See Java Authentication and Authorization Service
Jackrabbit, 350
jar, 587
.jar, 70, 517
libext/, 587
Jaro and Jaro-Winkler algorithm, 171
Jaro-Winkler algorithm, 198
Java
AMI, 440
API, 570–574
jobs, 573–574
parameters, 579–580
transformations, 572–573
variables, 579–580
expressions, 70–71
installation, 58–59
user-defined expressions and classes, 520
Java 2 Enterprise Edition (J2EE), 459
Java Authentication and Authorization Service (JAAS), 414
Java Content Repository (JCR), 350, 626
Java Development Kit (JDK), 58
jar, 587
Java Message Service (JMS), 449, 459–461
Java Naming and Directory Interface (JNDI), 64–65, 93
Java Runtime Environment (JRE), 58
variables, 642
Java Virtual Machine (JVM), 19, 58
logging, 453
rows, 397
variables, 43
directory, 642
Java.io.tmpdir, 642
JavaScript, 20, 202
DataCleaner, 153
eval(), 556
HTML, 520
job entries, 35–36
logging, 366
Mondrian, 276–277
performance, 394–396
variables, 43
XML/A, 281
JavaScript Object Notation (JSON)
examples, 549–558
Modified Java Script Value step, 522
plugins, 522
syntax, 521–522
transformations, 523
web services, 520–523
XML, 520
java.security.auth.login.config, 414
kettle.properties
java.version, 642
JCR. See Java Content Repository
JDBC
Database Connection, 93
drivers, 72
MySQL, 127
jdbc.properties, 64–65
kettle.properties, 67
JD/Edwards, 18, 139
JDK. See Java Development Kit
jedox.com, 282
jface.jar, 596
Jira, 632
JMS. See Java Message Service
JNDI. See Java Naming and Directory Interface.

JOB, 640
job, 324
job(s), 7
canvas, 56
command line, 322–326
Database Connection, 37
debugging, 56
DWH, 31
dynamic, 584–586
ETL, 12, 30–36
hops, 31–32
Java API, 573–574
Kitchen, 57
log tables, 373–374
loops, 399–400
metadata, 36–37, 574
Pan, 57
parallelism, 33–34, 411
performance, 399–400
Run button, 83–84
shared objects, 69
slaves, 445
Spoon, 82
variables, 89
XML, metadata, 346–347
job entries, 31
backtracking, 32–33
Boolean, 366
Copy rows to result step, 399
flow of execution, 90
JavaScript, 35–36
logging, 366
log table, 373–374
Mail, 90, 301, 337–340
plugins, 570, 621–624
results, 34–36
serial execution, 90
START, 88
transformations, 88
XML, 519–520
Job Scheduler, ETL, 124
JOBENTRY, 641
@JobEntry, 622
JobEntryDialogInterface, 622–624
JobEntryInterface, 622–624
jobentry-log-table, 346
job-log-table, 346
JobMeta, 574, 586
jobStatus, 416
Join comparison field, XML Join step, 543
Join condition properties, XML Join step, 542
join profile, 146
Join Rows step, 214
Main step to read from, 398
jpalo.com, 283
JRE. See Java Runtime Environment
JSON. See JavaScript Object Notation
JSONP, 550
jtwitter, 454
junk dimensions, 241
special dimension builder, 120
JVM. See Java Virtual Machine

K
Kalido, 6
.kettle, repositories.xml, 68
Kettle Logging Level, 330
KettleDatabaseRepository, 627
kettle-database-types.xml, 595
KETTLE_EMPTY_STRING_DIFFERS_FROM_NULL, 640
NULL, 28
Kettle.exe, 55
/kettle/getSlaves, 444
KETTLE_HOME, 64
$KETTLE_HOME/.kettle/
languageChoice, 601
$KETTLE_HOME/.kettle/shared.xml, 589
kettle-job-entries.xml, 595
KETTLE_..._LOG_DB, 641
KETTLE_..._LOG_SCHEMA, 641
KETTLE_MAX_LOG_SIZE_IN_LINES, 365, 454, 640
KETTLE_MAX_LOG_TIMEOUT_IN_MINUTES, 454, 640
kettle-partition-plugins.xml, 595
KETTLE_PASSWORD, 67
KETTLE_PLUGIN_CLASSES, 619, 640
kettle.properties, 58, 66–67, 414
logging, 365
variables, 43
kettle.pwd, 58, 67
kettle-repositories.xml, 595
KETTLE_REPO, 67
KETTLE_SHARED_OBJECTS, 589, 640
KETTLE_STEP_PERFORMANCE_SNAPSHOT_LIMIT, 640
kettle-steps.xml, 595
KETTLE_USER, 67
KETTLEVFS, 619
K–L  659

The key field, Denormalize Special Features, 104
keys. See also specific key types
  Calculator step, 214
dimension table, CDC, 80–81
dimension tables, 109, 208–217
  Insert / Update step, 230–231
JSON, 522
SCD, 217
  source systems, 209
Keys tab page, Dimension lookup / update step, 234
key/value pairs, 508–513
  object members, 522
  Regex Evaluation step, 510–511
  text files, 509–510
Kimball, Ralph, 11, 113, 167, 191, 221, 228, 295, 465
Kitchen, 41, 44, 54, 322–326
  jobs, 57
  level, 336
  logfile, 334
  logging, 364
  transformations, 57
  Kitchen.bat, 57
  kitchen.sh, 57
  .ktr, 345
libraries
  embedding, 574
  plugins, 596
  sapjco3.jar, 141
  StepMetaInterface, 600
  libswt/, 598
lightweight principle, 446–447
Limit, Get data from XML step, 534
Lindstedt, Dan, 9
lineage. See data lineage
Lineage and Dependency Analyzer, 125
link
  channel, 559
  item, 559
links
  DV, 468–469
  Sakila, 473–474
  link-to-link, 472, 474
Linstedt, Dan, 466
listdir, 324
listjobs, 324
listrep, 324
listtrans, 325
  .lnk, 62
Load all date from table, 253
Load DTS, 468, 469, 470
Load End DTS, 470
load_data, 298
loading. See also bulk loading
  lazy, 605
loadXML(), 599, 603
location outriggers, 97
Locking, Enterprise Edition, 636
logging, 333–336, 363–374
  architecture, 364–367
  avoiding, 398
  buffers, 364, 365–366
  CDC, 162–163
  channels, 366
  debugging, 364
  error messages, 364
  ETL, 22
  interval, 453
JavaScript job entries, 366
JVM, 453
kettle.properties, 365
Kitchen, 364
levels, 335–336
memory, 365
Pan, 364
parameters, 364
rows, 363
Spoon, 57, 333–334, 364, 365
transformations, 453–454
variables, 367
log data change processing, 451
log tables, 367–374
channels, 372
job, 373–374
job entries, 373–374
performance, 371
step log tables, 370–371
transformation, 367–370
Log4j, Apache, 366
logfile, 324, 334
loginmodulename, 414
Look and Feel (LAF), 590
lookup cascade, 100
Lookup Language, 103
lookup mode, 232
Lookup Original Language, 103
Lookup schema field, Database lookup step, 222–223
lookup tables, 172–175
lookup values, Validator step, 180
Loop Xpath, Get data from XML step, 532–533, 535
loops
Freebase, 557
hops, 27
jobs, 399–400
ls, 248
LucidDB, 10, 123
bulk loading, 249
EII, 10
SQL, 10
wrappers, 10

M
Mail, 301, 336–340
Addresses tab, 337
Attached Files tab, 339–340
Email Message tab, 339
job entries, 90, 301, 337–340
Server tab, 337–338
Mail Failure step, 90, 185–186, 336–337
Mail Success step, 90, 336–337
Main step to read from, Join Rows step, 398
maintainability, ETL, 300–301
Make transformation database transactional, Database Connection, 40
man crontab, 327
Manage thread priorities?, Transformation Settings, 397
Management Console, Enterprise Edition, 636
Manufacturing Requirements Planning (MRP), 138
mapping. See data mapping
Mark Attribute rows with id of header row, Modified Java Script Value step, 511
master, 417
AML, 442–443
Carte, 441
dynamic clustering, 434
transformations, 421–422
<master>, 435
Mastering Data Warehouse (Imhoff), 296
Max % errors allowed, Data Validator step, 188
Max nr errors allowed, Data Validator step, 188
Max number of articles, RSS Input step, 561–562
MAX(id) FROM test_sequence, 213–214
maximum heap size, 71
Maximum nr of lines in logging windows, Spoon, 365
max_log_lines, 412
max_log_timeout_minutes, 412
Maydanchik, Arkady, 168–169
MD5, 484
MDA. See Model Driven Architecture
MDX. See Multi Dimensional eXpressions
measures
Palo, 289
performance, 380–382
SOE, 265
Mechanical Turk, 437
memory
logging, 365
lookups, 253
performance, 393
Sort rows step, 453
Stream lookup step, 453
streams, 577
transformations, 452, 453
Merge join step, 479
CRC, 485
Merge Rows step, 160–161
MERGE/UPDATE, 249
message bundles, 601
metadata
data extraction, 359
data profiling, 17
data validation, 182
database, 588–590
Index n M–N 661

description, 37
directory, 36
ERP, 14, 139
ETL, 21, 344–350
  graphical user interface, 24
XML, 24
export, StepMetaInterface, 600
extended description, 37
filename, 36
jobs, 36–37, 574
names, 36
replacing, 588–590
repositories, 348–350
rows, 557, 606–607
  steps, 28
spreadsheets, 297
StepMetaInterface, 599
transformations, 36–37, 421–425, 572–573
User Defined Java Class step, 620
values, 605–606
XML, 345–347
  jobs, 346–347
  transformations, 345–346
Metadata Repository Manager, 125
Metaphone algorithm, 171
Metaweb Query Language (MQL), 551–553
methods
  partitioning, 425
  plugins, partitioning, 624–626
micro-batches, 450
Microsoft SQL Server 2008 Analysis Services (MSAS), 271, 277–280
Milestone, 630
Min nr of rows to read before doing %
evaluation, Data Validator step, 188
mini-dimensions, 239–240
  special dimension builder, 120
mirc.com, 633
Model, Spoon, 302–303
Model Driven Architecture (MDA), 6
Modified JavaScript Value step, 314
  DynamicJob, 586
Freebase, 556–557
JSON, 522, 549
  Mark Attribute rows with id of header row, 511
MOLAP. See Multi-dimensional OLAP
Mondrian, 242
  Aggregation Designer, 123, 267
Input step, 274–275
JavaScript, 276–277
OLAP, 271, 273–277
Split Time step, 276
Split-field step, 276
Strings cut step, 276
MonetDB, 123
monitoring, ETL, 333–340
Monitoring tab, Transformations settings, 381
MQL. See Metaweb Query Language
MRP. See Manufacturing Requirements Planning
MSAS. See Microsoft SQL Server 2008 Analysis Services
Multi Dimensional eXpressions (MDX), 269–270
Query, Mondrian Input step, 274
Multi-dimensional OLAP (MOLAP), 123, 269, 272
multiline mode, 507
multi-paths, backtracking, 32–33
multiple updates, CDC, 155
multi-threading, 403–411
  Blocking Step, 410
data pipelining, 407–408
database connections, 408–409
  Execute SQL step, 409–410
order of execution, 409–410
  row distribution, 404–407
row merging, 405–406
multi-valued attributes, 498–500
multi-valued dimension bridge table builder, ETL, 121–122
-Mxx, 60
MySQL, 73–74
  Bulk Loader, 248, 249
  CDC, 162–163
  JDBC, 127
NOW (), 484
RDBMS, 77, 134
SET, 103, 498
SUPER, 82
mysqlbinlog, 163
mysql_native.xul, 628

N
name, 346
names
  ETL, 24, 298–299
  job entry results, 34
metadata, 36
parameters, 44
pipes, 248
Namespace aware?, Get data from XML step, 533
natural keys
  business keys, 210
  dimension tables, 99
  junk dimensions, 241
near real-time data integration, 450
Needleman-Wunsch algorithm, 171
network latency, 369–370
network speed, 390
New validation button, 179
NIO buffer size, 386
non-intrusive CDC, 16, 155
non-relational data formats, 498
non-relational tabular formats, 498–501
non-tabular data formats, 498
norep, 323
normalization, 218
Normalize Special Features, 104
notepads, 346
notes, 318
  transformations, 25
NOW (), 484
Nr of errors fieldname, Data Validator step, 188
NULL
  Add XML step, 539, 541
  CRC, 484
data profiling, 146
data validation, 17, 180–181
  Database lookup step, 225
  DV, 471
KETTLE_EMPTY_STRING_DIFFERS_FROM_NULL, 28
  source data, 179
String, 28
Number, 27
Palo, 285
  String, 29–30
Number analysis, DataCleaner, 149
numbers, JSON, 522

O
OASI. See One Attribute Set interface
OBF, 414
obfuscated passwords, 67–68, 414
object literals, JSON, 522
object members, JSON, 522
object_timeout_minutes, 412
OCI. See Oracle Call Interface
ODBC, 93
ODS. See operational data store
OEM version, PDI, 590–591
OLAP. See online analytical processing
OLTP. See OnLine Transaction Processing
  Omit null values from XML result, Add XML step, 539
  Omit XML, Add XML step, 539
  One Attribute Set interface (OASI), 303
online analytical processing (OLAP), 123
  aggregate tables, 266
cubes, 270–271
  Input step, 274, 278–279, 281
  process, 278
  Mondrian, 271, 273–277
  multidimensional, 269
Palo, 282–291
  positioning, 272–273
  storage types, 272
XML/A, 277–282
OnLine Transaction Processing (OLTP), 2, 269–291
database, 75
dimension tables, 226
Open Office Calc, 297
OLAP, 273
Open Symphony, 327
OpenERP, 15
operating systems, scheduling, 322
operational data store (ODS), 4, 10
Oracle
  Attunity Stream, 163
  connect by prior, 20
  E-Business Suite, 18
  GoldenGate, 163
  RDBMS, 134
  Spatial, 591
  SQL*Loader, 247, 249
  Warehouse Builder, 6
ETLT, 9
Oracle Call Interface (OCI), 247
order, 346
ORDER BY, 225, 389, 479
org.pentaho.di.core.database
  .DatabaseInterface, 627
org.pentaho.di.trans.step
  .StepInterface, 614
original transformations, 421
os.arch, 642
os.name, 642
os.version, 642
OUTPUT_DIR, 319
Out of Memory, 71
outgoing hops, 26
output directory, 319
Output Fields, XSD Validator step, 529
Output one row, concatenate errors with separator, Data Validator step, 187
Output String Field, XSD Validator step, 529
Output Value, Add XML step, 539
Overwrite, SCD, 119

P
PAD. See Pentaho Aggregation Designer pagila, 74
Pair letters similarity algorithm, 171
Palo, 123, 273, 274, 282–291
Palo Cell Output step, 289–291
Palo Cells Input step, 285–289
Palo Dimension Input step, 285–289
Palo Dimension Output step, 289–291
Pan, 41, 44, 54, 322–326
jobs, 57
level, 336
logfile, 334
logging, 364
transformations, 57
Pan.bat, 57
pan.sh, 57
parallelism, 18–19
data extraction, 538
jobs, 33–34, 411
performance, 385–386
sorting, 393–394
text files, 385–386, 387
transformations, 27, 404
Parallelizing/Pipelining System, ETL, 125
parameters, 318
command-line, 323–324, 325–326
delays, 457
Java API, 579–580
logging, 364
named, 44
queries, 135–136
SQL, 99
transformation log tables, 369
transformations, 579–580
Validator step, 179
Version Migration System, 353–355
Parameters tab, 44
parent key, foreign keys, 242
Partitioner, 625–626
partitioning, 18–19, 425–430
accumulating snapshot fact tables, 264–265
checksums, 426
clustered transformations, 430
CSV File Input step, 428
database, 40, 429–430
methods, 425
plugins, 624–626
plugins, 426
methods, 624–626
round robin, 425
schema, 425–427
tables, performance, 392
Partitioning schema, Import partitions button, 429
pass, 323
password, jdbc.properties, 65
passwords, 58
obfuscated, 67–68, 414
UI, 609
Pattern finder, DataCleaner, 149
patterns, 501
pauseTrans, 415
PDL. See Pentaho Data Integration
Pearson, William, 270
peer/expert reviews, 297
##pentaho, 633–634
Pentaho Aggregation Designer (PAD), 123
Pentaho BI, Quartz scheduler, 322, 327–333
Pentaho Data Integration (PDI), 60, 328–330
AMI, 440
DataCleaner, 148
enterprise repository, 350
Java API, 571
OEM version, 590–591
slave servers, 413
Pentaho Report Designer (PRD), 574–575
Pentaho repository, 41
Pentaho Solutions (Bouman and van Dongen), 228, 327
PentahoSystemVersionCheck, 330–331
Peoplesoft, 15
perf-log-table, 345
performance
buffers, 380
constraints, 392
CPU, 394–398
data sorting, 392–394
database, 388–392
Freebase, 552
hard disks, 386–387
indexes, 390–392
JavaScript, 394–396
jobs, 399–400
log table, 371
measures, 380–382
memory, 393
parallelism, 385–386
relational databases, 390
rows, 382–383
SQL, 388
table partitioning, 392
text files, 384–387
transformations, 377–398
triggers, 392
tuning, 377–401
periodic snapshot fact tables, 260–261
fact table loader, 121
loading, 263–264
perspective, Spoon, 302
pipes, named, 248
PIT. See Point-In-Time
pivot fields, Palo, 288
platform independence, ETL, 18
Plug-in Registry, 595
@Step, 601
plugins, 20
architecture, 593–599
database, 627–628
ERP, 140
IDE, 596–597
JavaScript, 395
job entries, 570, 621–624
@JobEntry, 622
JSON, 522
LGPL, 570
libraries, 596
methods, partitioning, 624–626
partitioning, 426
repositories, 626–627
steps, 570, 619
transformation step, 599–619
types of, 594–595
plugins, 440
plugins/, 595
plugins/steps, 619
Point-In-Time (PIT), 472
Port Number, Database Connection, 38
POST
  Freebase, 550
  SOAP, 548
PostGIS, 591
PostgresSQL, 74
  bulk loading, 250
Power*Architect, 79
Powerplay, Cognos, 269
PRD. See Pentaho Report Designer
prd, 354
preparation of statements, 388
prepareExecution, 415
Preview option, 312–315
PreviewRowsDialog, 609
Primary key, 468, 469, 470
primary keys
  satellites, 470
  source system, 210
  surrogate primary keys
dimension tables, 80
tables, 77
  UPDATE, 470
Prism, 6
privacy, 308
private schedule, 331
Problem Escalation System, 125
process, 295
error handling, 184–187
process, OLAP Input step, 278
processRow(), 614, 615
processRows(), 456, 460
profiling
column profiling, 17, 146
data profiling, 16–17, 127–128, 146–154
  metadata, 17
dependency profiling, 146
  join profile, 146
properties
  built-in, 637–642
  JSON, 522
Proxy Host, Web services lookup step, 545
Proxy Port, Web services lookup step, 545
Prune Path to handle large files, Get data
  from XML step, 534
pubDate, item, 559
public schedules, 331
Punch through, Dimension lookup / update
  step, 238
putError(), 617
putRow(), 362, 557, 579, 616, 618
putRowTo(), 616
pwd/, 434

Q
Quartz scheduler, Pentaho BI, 322, 327–333
queries
  aggregate tables, 266
  parameters, 135–136
SQL, SELECT, 553
Query, MDX, Mondrian Input step, 274
Quote all in database, Database Connection, 38

R
ragged hierarchy, 120
RC. See Release Candidate
  -RCxx, 60
RDBMS. See Relational Database Management System
RDS. See Relational Database Service
Read articles from, RSS Input step, 561–562
read service, Freebase, 550–551
Read source as Url, Get data from XML step, 532
readRep(), StepMetaInterface, 599, 603
Really Simple Syndication (RSS), 18, 558–567
channel, 558–559
item, 559–560
transformations, 563
web services, 517
real-time business intelligence, 450
real-time data integration, 449–461
CDC, 450–451
source system, 451
transformation streaming, 452–461
real-time extraction, 138
CDC, 155, 163
real-time transformation streaming,
  debugging, 457–478
Record source, 468, 469
  satellites, 470
Recovery and Restart System, ETL, 124
Recurrence, 332
recursion, hierarchies, 120, 242–243
reference tables
  data cleansing, 172–179
  data conformation, 175–179
Referencing, 42
referential integrity, 42
  data quality, 168
  foreign keys, 251–252
RefinedSoundEx algorithm, 171
Regex Evaluation step, 204, 504–508
  key/value pairs, 510–511
Regex matcher, DataCleaner, 149
registerSlave, 416
regression tests, 307
regular expressions, 503–508
  capture groups, 200, 205
  data cleansing, 203–205
DataCleaner, 151–152
  good-enough solutions, 501–502
  Validator step, 180
Relational Database Management System (RDBMS), 134
ETL, 497
MySQL, 77
Relational Database Service (RDS), 438
relational databases, 39, 127
CDC, 450
  performance, 390
  transformations, 497
Relational OLAP (ROLAP), 242, 272, 274
Release Candidate (RC), 60, 630
remote execution, slave servers, 413
Remote Function Calls (RFCs), 140, 146
Remote Steps, 422
Rename fields step, XML/A, 280
rental star schema
  dimension tables, 79–80
  fact table, 79
  installation, 81
  Sakila, 78–81
rep, 323
repeating groups, 500–501
Replace in string step, 170, 203
Report all errors, not only the first, Data
  Validator step, 187
<report_to_masters>, 436
repositories, 41–42
  database, 348–349
  export, 350–351
  files, 349
  import, 350–351
  managing, 350–352
  metadata, 348–350
  plugins, 626–627
  upgrade, 351–352
  Version Migration System, 352–353
XML, 344
RepositoriesMeta.readData(), 573
repositories.xml, 68, 573
Repository, 626–627
Repository.loadTransformation(), 572
RepositoryMeta, 573
resetStepIOMeta(), StepMetaInterface, 600
resource exporter, 444
response time, DWH, 4
Result, 576–577, 587–588
Result Fieldname, XSD Validator step, 529
Result stream properties, XML Join step, 543
results tab, Web services lookup step, 546
Return/remove digits, data cleansing, 170
reuse
  ETL, 19, 300–301
  shared objects, 589
Revision management, 42
RFC_READ_TABLE, 143–144
RFCs. See Remote Function Calls
ROLAP. See Relational OLAP
root, 82
Root XML element, Add XML step, 539, 540–541
Ross, Margy, 221, 228
round robin, 386
partitioning, 425
sorting, 394
roundtrips, 388
row(s)
Add sequence step, 405
attributes, 497
CSV File Input step, 394
debugging, 314
dimension tables, 90
fields, 27
hops, 26, 27
JavaScript, 395
job entry results, 34
JVM, 397
logging, 363
metadata, 557, 606–607
steps, 28
multi-threading, 404–407
performance, 382–383
Sort rows step, 419
static data, 397–398
Table input step, 424
Text File Output step, 405
UI, 609
User Defined Java Class step, 404
Row denormaliser step, 511–512
Palo, 288
Row normaliser step, 500–501
RowDataUtil, 616
RowListener, 576
RowMetaInterface, 604, 606–607
Rownum fieldname, Split field to rows step, 500
Rownum in output and Rownum fieldname,
Get data from XML step, 535
RowProducer, 577, 579
RowSet, 579, 617
RSS. See Really Simple Syndication
rss, 558–559
RSS Input step, 561–562
RSS Output step, 562–567
R_STEP, 348
R_TRANSFORMATION, 348
Run button, 83–84
Run profiling, 152
running, 439
runtime.jar, 596
S
SaaS. See Software as a Service
Sakila
business keys, 527
CDC, 108
data mapping, 524–525
database, 73–110
installation, 77
subject areas, 75–76
Database Connection, 90–95
DV, 472–486
ETL, 73–110, 81–84
foreign keys, 105
hubs, 472–473
links, 473–474
rental star schema, 78–81
satellites, 474
snowflakes, 219–221
Spoon, 81–84
surrogate keys, 527
XML, 523–544
SalesForce.com input step, 140
SalesForce.com output steps, 140
SAP
data, 140–145
Function Browser, 141
SAP Input step, 140
Data Grid step, 142
Generate Rows step, 142
sapjco3.jar, 141
SAP Java Connector library (sapjco3.jar),
SAP Input step, 141
sapjco3.jar. See SAP Java Connector
library
SAP/R3, 14, 18, 141
Sarbanes-Oxley Act, 308
satellites
DV, 469–471
primary keys, 470
Sakila, 474
WHERE, 484
saveRep()
JobEntryInterface, 622
StepMetaInterface, 599, 603
scalability
ETL, 18–19
Freebase, 552
SCD. See Slowly Changing Dimension
Schedule Creator, 331–332
Scheduling, Spoon, 302
scheduling
action sequence, 333
ETL, 321–333
operating systems, 322
schema. See also XML Schema
clustering, 417–418
Database Connection, 39
DataCleaner, 148
dynamic clustering, 434
partitioning, 425–427
Schema name field, Add sequence step, 217
SCM. See software configuration
management
screens, 191
Script Values step, 394–395
scripts, 20. See also JavaScript
ETL, 5, 200–205
startup, 70
Scrum, 13, 301
searchInfoAndTargetSteps(),
StepMetaInterface, 600
Secure Sockets Layer (SSL), 337
Security repository, 42
Security System, 125
sed, 347
SELECT, 553
Select values step, 94, 100, 397
semi-additive, 260
SOE, 265
semi-structured data, 501–508
Separate history table, SCD, 119
sequence_value, 213
serial execution, job entries, 90
Serialize to file step, CDC, 164
Server tab, Mail, 337–338
services. See also web services
grid-based, 437
slave servers, 414–416
SET, 499
MySQL, 103, 498
Set Environment Variables step, 354–355
Set Variables step, 216
setDefault(), 599, 604
SETI@Home, 433
setOuputDone(), 615
sets, CSV, 498
Settings tab page, Regex Evaluation step,
504–506
.sh, 58
SHA-1, 484
shadow copies, 31
sharding, database, 40
shared objects, 68–69
database, 589
jobs, 69
Spoon, 69
transformations, 69
shared.xml, 68–69
shortcuts, Spoon, 62
shrunken or rolled dimensions, special
dimension builder, 120
Simple Object Access Protocol (SOAP)
accessing services directly, 546–549
examples, 544–549
addition, 138
OLAP, 274
WDSL, 517
web services, 517
Web services lookup step, 544–546
XML/A, 277
slave(s)
AMI, 443–444
jobs, 445
transformations, 421–422
Slave Browser tab, Spoon, 457
slave servers
Carte, 411–416, 435
configuration, 411–412
PDI, 413
remote execution, 413
services, 414–416
Sort rows step, 419
Spoon, 413
Table input step, 424
XML, 413
slices, 271
Slowly Changing Dimension (SCD), 20,
228–239
Bus Architecture, 118–119
Dimension lookup / update step, 232–237
dimension tables, 118
Dimensional Data Warehouse, 118–119
ETL, 118–119
hybrid, 238–239
Insert / Update step, 229–230
keys, 217
type 1, 229–232
type 2, 232–237
type 3, 237–238
Small and Medium Business (SMB), 139
small periodic batches, 450
smart keys, 80, 108
SMB. See Small and Medium Business
SMTP, 337
snapshots
CDC, 146, 158–162
fact tables, 121, 260–261, 263–264
Sniff test during execution, Spoon, 457–478
sniffing, 314–315
sniffStep, 416
snippets, User Defined Java Class step, 620
snowflakes
dimension tables, 97, 218–225
Sakila, 219–221
SOAP. See Simple Object Access Protocol
soapUI.org, 547
SOF. See state-oriented fact tables
Software as a Service (SaaS), 437
software configuration management (SCM), 626
sorting
clustering, 394
data, performance, 392–394
database, 393
parallelism, 393–394
round robin, 394
Sort rows step, 479
memory, 453
rows, 419
slave servers, 419
Sort size (rows in memory), 393
Sort size (rows in memory), 393
Sort System, 124–125
Sorted Merge step, 419
Soundex algorithm, 171
source code
Java API, 570
plugins, 594
source data
CDC, 155–157
data cleansing, 173
NULL, 179
PRD, 574
RSS Output step, 564
tabular format, 497
source system
database lookup step, 222
keys, 209
primary keys, 210
real-time data integration, 451
Source XML field, XML Join step, 542
sourceforge.net, 59–60, 570
source_system, 178
Spatial, Oracle, 591
special dimension builder
dimensions, 120
ETL, 120
special_features, 103–104
Split field to rows step, 104, 499–500
Split Time step, Mondrian, 276
Split-field step, Mondrian, 276
Spoon, 41, 54
Add sequence step, 211–217
agile development, 301–302
canvas, 318
Combination lookup / update step, 241
Copy tables wizard, 584
dynamic transformations, 580–583
ETL, 81–84
Execute a transformation, 413
extraction, 128
IDE, 55–57
jobs, 82
logging, 57, 333–334, 364, 365
perspective, 302
Sakila, 81–84
shared objects, 69
shortcuts, 62
Slave Browser tab, 457
slave servers, 413
Sniff test during execution, 457–478
transformations, 57, 82
variables, 44
Spoon.bat, 55, 62
.spoonrc, 64
spoon.sh, 55
spreadsheets
data acquisition, 15
metadata, 297
testing, 311
SQL
attributes, 484
Business Objects, 9
dynamic jobs, 584
ELT, 9
Informatica, 9
Input source step, 483
LucidDB, 10
ORDER BY, 225, 479
parameters, 99
performance, 388
query, SELECT, 553
StepMetaInterface, 599
streams, 99
WHERE, 553
SQL Server
RDBMS, 134
XML/A, 278
SQL statements to execute after connecting,
Database Connection, 39
SQLEditor, 609
SQL*Loader, Oracle, 247, 249
SQLPower, 118, 154
SQLStream, 458
src/, 597
SSL. See Secure Sockets Layer
standard input (STDIN), 247–248, 250
standard measures, DataCleaner, 149
star schema, 78–81. See also rental star schema
CDM, 227–228
denormalization, 226
dimension tables, 226–228
tables, 495
START, job entries, 88
Start at value field, Step sequence step, 212, 216
STARTDATE, 369
STARTDATE-ENDDATE, 369–370
startExec, 415
startJob, 416
startTrans, 415
startup scripts, 70
state-dependent objects, data quality, 168
state-oriented fact tables (SOF), 261–263
loading, 265–266
static data, rows, 397–398
static dimensions
special dimension builder, 120
tables, 84–87
static testing, 307
static values, JavaScript, 396
status, 415
STDIN. See standard input
STEP, 640
step, 346
@Step, Plug-in Registry, 601
_step_, 557
Step name, transformation log tables, 369
Step name field, Step sequence step, 212
StepInputDialogInterface getStepData(), 600
StepDialogInterface, 607–613
step_error_handling, 346
StepInterface, 614–619
StepInputDialogInterface getStep(), 600
step_log-table, 346
stepMetaInterface, 599–607
StepMetaInterface check, 599
steps, 7. See also specific steps
outgoing hops, 26
plugins, 570, 619
row metadata, 28
shared objects, 589
transformations, 26
VPLs, 47–49
stopJob, 416
stopTrans, 415
stream(s), 83
Add XML step, 538
data, 577
data integration, 450
editor, 347
extraction, 138
memory, 577
SQL, 99
StepMetaInterface, 600
Table output step, 538
transformations, 452–461, 577
Web services lookup step, 517
XML Join step, 541
Stream Datefield, Dimension lookup / update step, 235–236
Stream lookup step, 173, 178, 253–255, 383
import_xml_into_db.ktr, 527
memory, 453
StrictHostKeyChecking, 641
String, 27
Boolean, 30
Date, 29
NULL, 28
Number, 29–30
Palo, 285
string(s), 384
JSON, 522
UI, 609
String analysis, DataCleaner, 149
String getDialogClassName(),
StepMetaInterface, 600
string literals, JSON, 522
Strings cut step, Mondrian, 276
structural testing, 21
Stylus Studio, 523
subscription, 635
subsystems, ETL, 113–126
subtransformation interface, 101
Subversion, Apache, 343, 570
success hops, 90
SugarCRM, 15
SUPER, 82
supportsErrorHandling(), 600
surrogate key(s), 118
Add sequence step, 211–217
business keys, 210
creation system, 119
database sequence, 217
Dimension lookup / update step, 234–235
dimension tables, 209, 251–260
DWH, 210
generating, 210–217
hubs, 469
import_xml_into_db.ktr, 527
pipeline, 121, 252–255
Sakila, 527
SOF, 266
XML, 527
surrogate primary keys
dimension tables, 80
tables, 77
UPDATE, 470
Switch/Case step, 189–190
SWT, Eclipse, 607
swt.jar, 596
synchronization, data, 9
Synchronize after merge step, 160–161
sysdate, 354

T
tab-delimited files, 128
table(s). See also specific table types
  DataCleaner, 148
  DV, 485–486
  foreign keys, 77, 208
  hubs, 467
  indexes, 392
  link-to-link, 472, 474
logging, 367–374
  channels log tables, 372
  job entries log table, 373–374
  job log table, 373–374
  performance log tables, 371
  step log tables, 370–371
  transformation log tables, 367–370
partitioning, performance, 392
star schema, 495
static dimensions, 84–87
surrogate primary keys, 77
Table daterange end, Dimension lookup / update step, 236
Table input step, 103, 596
aggregate tables, 266
CDC, 160
Data Grid step, 132
rows, 424
slave servers, 424
Stream lookup step, 254
Table output step, 216, 397
bulk loading, 250
CDC, 164
commit size, 390
data lineage, 358
dynamic templates, 584
export_xml_from_db, 538
import_xml_into_db.ktr, 527
streams, 538
Use batch updates for inserts, 389
TableInput, 595
table_params, 355
TableView, 609
tabular format
  non-relational, 498–501
  source data, 497
tags/, 342, 352
Talend, 6
  Data Profiler, 154
.tar, 517
Target fields
  Denormalize Special Features, 105
  Insert / Update step, 230
Target XML field, XML Join step, 542
.tar.gz, 60
Task Scheduler, 327
TCP/IP
  Carte, 57, 417
  clustering, 423
templates, dynamic, 583–584
testing
  automation, 311
  CI, 311
  Data Grid step, 311
dynamic, 307
  ETL, 21, 306–312
  integration, 307
  spreadsheets, 311
  static, 307
  transformations, 311
  upgrade, 312
test_sequence.ktr, 212–213, 215
text file(s)
  extraction, 128–132
  Web, 137
  fields, 384
  key/value pairs, 509–510
  parallelism, 385–386, 387
  performance, 384–387
  reading, 384–387
  writing, 387
Text file input step, 203, 384
Text file output step
  CDC, 164
  rows, 405

S–T
TextVar, 609
third normal form (3NF), 218
DV, 469
threads, 397. See also multi-threading
RowProducer, 579
3NF. See third normal form
time analysis, DataCleaner, 149
time dimensions, 239
time-outs, databases, 453
TIMESTAMP, 80
timestamps
CDC, 155–157, 163, 450
DV, 480
title, 559
TLS. See Transport Layer Security
/tmp/cartelog, 441
tokens, Get data from XML step, 536
tools, 41
ETL, 6
requirements, 17–22
top-down level-wise loading, 219
Tortoise SVN, 570
TPC-H, 253
traceability
of data, 467
DV, 471
TRANS, 640
Trans, 577
trans, 325
transaction grain fact tables, 121
<transformation>, 345
transformation(s), 7
action sequence, 328–330
architecture, 452
bottlenecks, 379–382
buffers, 406–407
Calculate Dimension Attributes, 85–86
canvas, 56
clustering, 417–425
partitioning, 430
command line, 322–326
data, 576–580
data conversion, 29–30
Database Connection, 37, 90–95
deployment, 46
deduplication, 195–199
dynamic
CSV, 580–583
Spoon, 580–583
error handling, 186–187
ETL, 12, 25–30
challenges, 20
Get data from XML step, 532
hops, 25, 26–27
Java
API, 572–573
expressions, 70–71
job entries, 88
JSON, 523
Kitchen, 57
logging, 453–454
master, 421–422
memory, 452, 453
metadata, 36–37, 421–425, 572–573
notes, 25
Pan, 57
parallelism, 27, 404
parameters, 579–580
performance, 377–398
phases, 452
relational databases, 497
RSS, 563
Run button, 83–84
shared objects, 69
slave, 421–422
Spoon, 57, 82
steps, 26
streams, 452–461, 577
testing, 311
variables, 89, 579–580
VPLs, 46
XML, metadata, 345–346
Transformation File, 330
Transformation Inputs, 330
transformation log tables, 367–370
Get System Info step, 367
history, 367–368
parameters, 369
Transformation Settings, Manage thread
priorities?, 397
Transformation Step, 330
transformation step plugins, 599–619
Transformations Settings, 368
Monitoring tab, 381
transitive closure table, 242
trans-log-table, 345
TransMeta, 572, 577
TransMeta.getSQLStatements(), 566
transparency, ETL, 24
TRANS_PERFORMANCE, 640
Transport Layer Security (TLS), 337
transStatus, 415
triggers
CDC, 163, 450
database, 157–158
performance, 392
Truncate, bulk loading, 251
Trunk/, 342
Trunk version, 630
trunks, 283
tst, 354
Tungsten Replicator, 163
Twitter, 454–457
type, 65
TYPE_BIGNUMBER, 606
TYPE_BINARY, 606
TYPE_BOOLEAN, 606
TYPE_DATE, 606
TYPE_INTEGER, 606
TYPE_NUMBER, 606
TYPE_STRING, 606

U
UA. See User Acceptance test
Ubuntu, 439
AMI, 442
UI. See user interface
ui/laf.properties, 591
uname, 354
unbalanced hierarchy, 120
unconditional hops, 88
unconditional job hop, 31
UniCode, 15, 507
Uniform Resource Locators (URLs), 516
Web services lookup step, 545
Unique rows step, 193–194
unit tests, 307
UNIX, 12, 507
chmod, 322
cron, 326–327
crontab, 326–327
Kitchen, 57
Pan, 57
running programs, 62
Unknown, 17
unstructured data, 501–508
Unzip, AMI, 440
UPDATE, 157, 230
surrogate primary keys, 470
Update fields, Insert / Update step, 231–232
update mode, 232
Update step
CRC, 485
Dimension lookup / update step, 238
upgrade
repositories, 351–352
testing, 312
url, jdbc.properties, 65

URLs. See Uniform Resource Locators
Use batch updates for inserts, Table output step, 389
Use Kettle Repository, 330
Use tokens, Get data from XML step, 533, 535
user, 323
jdbc.properties, 65
User Acceptance test (UA), 307
User Console, 333
User Defined Java Class step, 620–624
Change number of copies to start, 404
DynamicJob, 586
get(), 620
init(), 459
JavaScript, 395
metadata, 620
rows, 404
snippets, 620
variables, 43
User Defined Java Expressions step, 70–71, 202–205
data cleansing, 202–203
user interface, 24. See also graphical user interface
elements, 609
StepMetaInterface, 600
user maintained dimensions, 120
User Name and Password, Database Connection, 38
user-defined expressions and classes, Java, 520
user.dir, 642
user.home, 642
user.name, 642
UTF-8, 129
Add XML step, 539
RSS Output step, 565

V
Vaillencourt, Luc, 591
valid, 160
Validate msg field, XSD Validator step, 529
Validate XML?, Get data from XML step, 533
validation. See data validation
Validator step, 179–180
valid_from, 265
valid_to, 265
value(s)
JSON, 522
metadata, 605–606
static, 396
Value distribution, DataCleaner, 149
Value mapper step, 94, 99, 170
Value when XML is invalid, XSD Validator step, 529
Value when XML is valid, XSD Validator step, 529
ValueMetaInterface, 605–606
van der Lek, Harm, 303
van Dongen, Jos, 228, 327
VARCHAR, 29
variables, 43
Apache VFS, 641–642
built-in, 637–642
hierarchy, 120
internal, 428–429
Java API, 579–580
JavaScript, 396
jobs, 89
JRE, 642
kettle.properties, 66
logging, 367
Spoon, 44
StepInterface, 618–619
transformations, 89, 579–580
using, 44–45
VariableSpace, 618–619
VCS. See Version Control System
version, 324
Version Control System (VCS), 341–344
ETL, 124
XML, 352
Version field, Dimension lookup / update step, 235
Version Migration System, 352–355
ETL, 124
parameters, 353–355
repositories, 352–353
XML, 352
Versioning, Enterprise Edition, 636
VFS. See Virtual File System
Virtual File System (VFS), 41, 42
Apache, 42, 349, 517, 619
variables, 641–642
virtual machines (VM), 438
visual programming languages (VPLs), 45–51
steps, 47–49
transformations, 46
Visualize, Spoon, 302–303
VM. See virtual machines
VPLs. See visual programming languages

W
Warehouse Builder, 6, 9
warnings, 405
waterfall model, 12
Wavemaker, 120
WDSL, SOAP, 517
web
browsers, slave servers, 413
extaction, 137–138
pages
HTML, 520
web services, 515–517
text files extraction, 137
web services, 515–568
Apache VFS, 517
API, 516
data formats, 517–523
Freebase, 550
HTML, 520
JSON, 520–523
RSS, 517
SOAP, 517
web pages, 515–517
XML, 518–520
Web Services Description Language (WSDL), 544
Web services lookup step, 517
SOAP, 544–546
streams, 517
Web services tab, Web services lookup step, 545
wget, 60
WHERE, 230
satellites, 484
SQL, 553
white box testing, 306
whitespace, 507
widgets, 607–608
wiki, 631
Wikipedia, 549–550
Windows, 61–62
Wintner, Robert, 140
WMS. See Workflow Management Systems
Workflow Management Systems (WMS), 344
Workflow Monitor, 124
wrappers, LucidDB, 10
write back, 271
WSDL. See Web Services Description Language
X
XBase, 134
XChat, 633
xchataqua.sourceforge.net, 633
XML. See eXtensible Markup Language
XML Join step, 519, 541–544
  streams, 541
XML output step, 518
  CDC, 164
XML Schema, 518, 528
  data validation, 530
  XSD Validator step, 519
XML Schema Definition, XSD Validator step, 529–530
XML source, XSD Validator step, 529
XML source from field, Get data from XML step, 532
XML source is a filename?, Get data from XML step, 532
XML source is defined in field, Get data from XML step, 532
XML source is defined in field?, Get data from XML step, 532
XML/A
  JavaScript, 281
  MSAS, 279–280
  OLAP, 277–282
  Rename fields step, 280
xml=Y, 414
-Xmx, memory, 253
XP. See Extreme Programming
XPath, 518, 532
  Get data from XML step, 535
XSD Filename, XSD Validator step, 529–530
XSD Source, XSD Validator step, 529–530
XSD Validator step, 519, 528–530
  data validation, 530
  error handling, 530
  job entries, 519
  XML, 133
xsi:schemaLocation, 529
XSL. See eXtensible Stylesheet Language
XSL Transformation job entry, 519
XSL Transformation step, 518–519
XSL Transformations (XSLT), 133
XSLT. See XSL Transformations
xstream.codehaus.org, 603
XUL, 628
Y
Yourdon, Ed, 12
YouTube, 315
Z
.zip, 60, 517