

Index

34 subsystems of ETL, 430–434
64 bit architectures for data warehouse, 554, 558, 582
2NF (second normal form), 177–178
3NF (third normal form), 133
 business rules, 145–147
 Chris Date criticisms, 147
 complex schemas, 146
 versus dimensional modeling, 140–141
 incompleteness, 146
 primary criticism, 137–139
 query complexity, 138
 performance for BI queries, 138
 real data, 146
 redundancy, 138
 uniqueness, 146
 usability, 138
 CIF use of 3NF, 173–178

A

abstract dimensions, reasons to avoid, 311
abused users, 119
accumulating snapshot fact table, 194. *See* grain (fact tables)
 combining with periodic and transaction grains, 249
 comparison to other grains, 244–246
 date dimension roles, 300
 fact table loader, ETL subsystem #13, 432
 nulls to be expected, 276
 pipelines and short processes, 246
 procurement example, 241–242
 real time partition, 509
 university and admissions example, 247–248

accurate counting, combining CASE and SUM, 314–315
actions, tracking, 590, 593
activity based costing
 difficult environments for, 265
 modeling income statements, 401
ad hoc attack, 63
Adaptive Software Development, 109
additive facts, 12, 139–142, 182
 examples, 31, 213, 264
 declaring in metadata, 227
 non-additive example, 281
 semi-additive example, 182, 227
address cleaning and standardizing, 374–388, 439
 international addresses, 274, 378–383, 475
administration criteria for dimensional DWs, 228–229
administrative costs, 59
admissions, university, accumulating snapshot example, 247
affinity grouping
 data mining, 617
 market basket example, 421
aggregate builder, ETL subsystem #19, 432
aggregate data quality measures, 470
aggregate fact table definition, 188
aggregate navigation
 criterion for dimensional DW, 228
 dimensional modeling advantages, 142
 of dissimilar fact table grains, 545
 example, 32
 main architecture articles, 536–546
 main algorithm, 542

- metadata and, 539
 - minimum metadata requirement, 542
 - OLAP considerations, 548, 554
 - query tool discipline, 542
 - query tools, 639
 - recommended data warehouse architecture, 537
- aggregate navigator, 185, 188
- aggregate processing during ETL, 440, 486
- aggregated data
 - anticipates the business question, 236
 - characteristics, 239
 - data mining and, 44
 - data quality reporting, 470
 - drilling down from, 53
 - prematurely, 92, 143
- aggregated dimensional models, 239
- aggregates
 - administration with Type 1 SCD, 25
 - design requirements, 540
 - fact provider responsibilities, 164
 - goals for data warehouse, 539
 - metadata requirements, 536, 569
 - objection removers, 78–79
 - positive and negative impacts, 536
 - removing from real time partition, 495, 508
 - server configurations, 583
 - shrunk dimension tables, 540
- aggregation. *See* aggregated data
 - when premature defeats drill down, 143
- agile development approach, 107–111
- Agile Manifesto, 109
- AI (artificial intelligence), 616
- airline customer satisfaction dimension, 371
- airline flight segment database design, 393–395
- airline yield KPI use case, 22–24
- airport role playing dimensions, 396
- Alda, Alan, interviewing skills, 113
- allocating costs
 - conflicting requirements, 263–265
 - danger of implementing, 4–5, 63, 71–72
- allocation, environments, 265
- allocation rules for calculating profit, 72
 - compliance requirements, 426
- allocations
 - computing on the fly, 523
 - implementing in OLAP, 549
- income statement fact tables, 401–403
- profitability fact tables, 402
 - substituting rules of thumb, 44, 402
 - version number in audit dimension, 466, 469
- alternate reality, type 3 SCD, 27
- An Introduction to Database Systems* (Chris Date), 36, 137
- analytic application lifecycle, five stages, 22, 590–596
 - analytics matrix tracking, 158
- analytic application reports, 602
 - build versus buy, 603
- analytic requirements, identifying, 126
- analytic tools, 62, 63
- analytics matrix, 158
- Analytics Workshop, 127
- AND queries, 349–350
- architecture
 - address matching and standardizing, 385–386
 - aggregate navigation articles, 536–546
 - archiving, long term preservation, 579–582
 - BI architecture articles, 560–565, 607–610
 - BI comparison queries, 631–634
 - BI portal, dashboards, 610–612
 - BI upgrading unsuccessful, 674–676
 - bus architecture, 38–45, 51–52, 150–151
 - catastrophe protection, 576–578
 - change data capture, 452–453
 - criteria, dimensional DWs, 226–228
 - data architecture chapter, 133–178
 - data mining articles, 615–629
 - data quality, 460–467
 - distributed EDW, 56
 - drilling across, 189–191, 629–631
 - drilling down, 22–24, 186–189
 - EDW diagram, 51
 - ETL, 105
 - 34 subsystems of, 430–434
 - FTP-based integration, 450
 - integrated EDW, 13–21
 - late arriving data handling, 491–495
 - Lifecycle place for, 97
 - master data management (MDM), 516–520
 - metadata 567–571

- Microsoft SQL Server 2005 data architecture, 554–559
 - real time, 503–510
 - ROLAP versus OLAP, 549–553
 - SCDs and time variance of dimensions, 24–27
 - security, 83, 575
 - separating IT systems, 50–51
 - service oriented architecture (SOA), 513–515
 - storage area network (SAN), 585–587
 - surrogate key processing pipeline, 481–485
 - time handling, 192–194
 - architecture phase, data marts, 39–40
 - archiving, 2, 8
 - encapsulating and emulating strategy, 581
 - examples of very long term requirements, 579
 - historical letters case study, 347–351
 - limitations of media, formats, software, hardware, 580
 - metadata examples, 568–569
 - migrate and refresh strategy, 581
 - requirements affecting ETL design, 428
 - very long term digital preservation, 579–582
 - Atkinson, Toby, multinational name and address resource, 381
 - atomic data, 61
 - advantages, 43
 - aggregations, 235–236
 - as basis of dimensional models, 196
 - drilling down, 47, 188
 - normalized form, 200
 - storage architectures, CIF versus Kimball, 174
 - atomic fact tables, 43
 - as core foundation, 239
 - atomic grain, dimensionality, 235
 - atomic-level behavior data, 55
 - audit columns for change data capture, 452–453
 - audit dimension, 465–467
 - assembler, ETL subsystem #6, 431
 - in data mining, 619
 - data quality measures, 468
 - detailed design, 469–471
 - environmental descriptors, 468
 - fact tables, 187
 - automobile collisions, factless fact table, 257
 - automobile policy coverages, insurance case study, 278, 392
 - availability of data warehouse, 48, 53, 558
 - minimizing offline time, 502
 - taking aggregates offline, 440
 - averaging over time, 182
 - awkward formats, 92
- B**
- B-tree indexes, 37, 269, 508
 - back pointers to operational systems, 487–488
 - back room, 653. *See* ETL.
 - backup and recovery use cases, 79
 - backup system, ETL subsystem #23, 433, 578
 - backups
 - data staging, 8
 - objection removers, 79
 - balance transactions, 279
 - BEEP, 237
 - begin- and end-effective time stamps. *See* time stamps
 - behavior analysis, 598–600, 621–625
 - from clickstream, 410–413, 415–417
 - market basket analysis, 420–424
 - purchase behavior security risks, 573
 - behavior dimension, 231, 324, 643
 - behavior tags, 368–371
 - recency, frequency, intensity, 337–338, 368
 - behavioral queries, 640–644
 - non-behavior, 26–262
 - Berry, Michael, 600, 625
 - best practices
 - building DW/BI systems, 103
 - establishing operating procedures, 655
 - BI (business intelligence)
 - applications, chapter 13, 589–650
 - architecture
 - unsuccessful, 675
 - upgrading, 674–676
 - compliance, 596–597
 - CRM, 599–600
 - custom tools, 520–522
 - dimension browsing, 28
 - drilling across, accreting measures, 29
 - drilling down, 28
 - ease of use, 29

- environment
 - launching, 652
 - monitoring operations, 653–654
 - pervasive, 532–533
 - portal, 610–612
 - queries, 28
 - improving performance, 78
 - reports, 28. *See* reporting
 - sequential behavior analysis, 597–598
 - tools, 20–21
 - licenses, 682
 - sequential computation difficulties, 635
 - user interface, 28
 - value with, 589–600
- BI tool interfaces affecting ETL design, 429
- bitmap indexes, 81, 269, 325, 559, 562
- blended development approach, top-down and bottom up, 103
- book references
 - building interpersonal skills, 94
 - building public speaking skills, 95
 - building written communication skills, 95
 - understanding the business world, 94
- bottlenecks
 - authentication and access, 578
 - memory, 582
 - scalability, 507
- bottom-up approach, Kimball Lifecycle, 100, 128
- bottom-up market basket algorithm, 423
- boundaries with finance, IT, legal, and end users, 4–6
- bridge tables
 - account to customer in banking, 343, 344
 - begin- and end-effective time stamps, 345
 - correctly weighted report, 342
 - definition, 335
 - diagnosis tracking in health care, 342
 - ETL subsystem #15, bridge table builder, 432
 - for multiple alternate hierarchies, 366. *See* hierarchies
 - for variable depth hierarchies, 336, 357–359. *See* hierarchies
 - for satisfaction tracking, 373
 - impact report, 342
 - keyword tracking, 348
 - Microsoft Analysis Services alternative, 554
 - natural keys, 362–363
 - need for surrogate keys, 344
 - reports-to dimension, separate, 361
 - SIC codes, 343
 - surrogate keys, 344, 360–361
 - updating, 346
 - weighting factor, 342, 345
- Brin, David (*The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?*), 574
- browse a dimension, BI tool user interface design, 28, 135, 638
- budgeting case study, 403–407
- budgeting data aligned with planning data, 545
- bug tracking system, 601
- bus architecture, 46, 51, 150–151, 172. *See* architecture.
 - distributed systems, 151
 - independent from centralization, 151
- bus matrix
 - analytics, 158
 - consolidated processes in, 240
 - detailed implementation matrix, 159–160
 - drill down into, 159–161
 - executive communication, 15, 154
 - extensions, 158
 - feasibility grid, benefit versus feasibility, 131
 - grain, altering, 159
 - for integrated EDW, 15, 129
 - for manufacturing, 16
 - mishaps, 157
 - opportunity, 158
 - processes versus departments, 130
 - preliminary bus matrix and bubble chart, 218
 - primary introductions, 151–159
 - strategic initiatives versus business processes, 127, 158
- business acceptance, 88, 99, 113, 664–670
- Business Dimensional Lifecycle (Kimball Lifecycle), 96–99
- business intelligence. *See* BI (business intelligence)
- business needs affecting ETL design, 47, 53, 204, 426
- business phase of data mining, 626
- business processes

- as basis of dimensional models, 197
 - versus departments, 123
 - fact table grain, 126
 - identifying, 124, 125–127
 - consequences of incorrect, 126
 - subject areas, 61
 - tying to strategic initiatives with matrix, 127
 - business realignment, 667
 - business reengineering, 2, 461
 - driven from poor data quality, 459
 - organizational steps, 459
 - business requirements gathering, 2, 3, 5, 83, 113
 - conversationality, 114–115
 - curiosity, 114
 - data audits, 116
 - difficult users, 119
 - listening skills, 116
 - preparing beforehand, 115
 - wants/needs determination, 118–119
 - business rule screens, 122, 463
 - business rules, 145
 - screens, in ETL architecture, 463
 - supported by data models, 145
 - business sponsor, 86–89, 149–150, 655, 662–670
 - business user's responsibilities, 216
- C**
- calendar date dimension design, 291
 - calendar dimension, 293–294. *See* date dimension; time dimension
 - design, 435
 - multi-enterprise, 339
 - primary key, date format, 288
 - calendars
 - international dates, 476
 - multinational designs, 376
 - case studies
 - budgeting, 403–407
 - clickstream, 409–413, 413–417
 - growth scenario, 658–661
 - human resources, 396–400
 - insurance, 389–393
 - profitability, 400–403
 - text document searching, 417–420
 - travel, 393–396
 - catastrophic failures, 576
 - catastrophic SCD type 1 invalidation using OLAP, 552
 - causal dimensions
 - describing promotions or behavior, 235, 308–311, 674
 - design recommendations, 310
 - sourcing the data, 309
 - causal factors
 - analytic application, 22
 - determining, 590, 592
 - CDI (customer data integration), 105, 155
 - central data warehouse team, 80–83
 - centralization, 168
 - decentralized reality, 47, 52
 - inappropriate, 60
 - logical design and integration, 169
 - objection remover false promise, 76, 78, 79
 - risks, 103–104
 - risks of physical but not logical, 169
 - steps to migrate from disparate data, 170
 - centralized architecture comparison to planned economy, 178
 - centralized customer management system, 78
 - centralized DW/BI systems, risks, 59
 - change
 - anticipating, 47, 53
 - continuous, 60
 - source data changes, 7, 195–196
 - change data capture, 6–7, 452–453
 - with CRC (cyclic redundancy checksum), 486–487
 - with diff compare, 453
 - ETL subsystem #2, 431
 - change impact on dimensional models, 9. *See* graceful extensibility
 - checkups, 661–667
 - choice presentation in web-oriented data warehouse, 563
 - CIF (Corporate Information Factory), 99
 - compared to Kimball bus architecture, 171
 - hybrid with Kimball approach, 175
 - and Kimball approaches, fundamental differences, 174
 - claims periodic snapshot fact table, insurance case study, 391

- claims transactions fact table, insurance case study, 390
- classification queries, 642
- classifying, data mining activity, 616–617
- cleaning data, 431, 439, 443, 454–459
 - as prelude to data mining, 618–619
 - using regular expressions, 477–481
- clickstream case study, 409–413, 413–417
- clickstream data source challenges, 411
- clickstream dimension, 410–412
 - events, 413
 - session type, 415–417
 - visitor dimension, 414–415
 - web page object, 415
 - web sessions, 413
- clickstream facts, 412
- clueless users, 121
- clustering, data mining activity, 600, 616
- codes
 - expand as verbose text in dimensions, 199
 - interpreting into text, 618
- cognitive models, 91
- column screens in data quality architecture, 122, 463
- comatose users, 120
- combinatorial explosion in market basket analysis, 422
- comment fields, removing from fact tables, 275
- commitments fact table as part of budgeting value chain, 404
- common labels in conformed dimensions, 149
- communication, bus matrix as a vehicle for, 154
- comparisons
 - difficulty using SQL, 631–634
 - implementing with drill across, 23
- compliance, 2, 3
 - affecting ETL design, 426
 - international data, 476
 - as part of data steward role, 166
 - requirements for data warehouse, 407–409, 596–597
- compliance-enabled fact and dimension tables, 408
- compliance reporting, ETL subsystem #33, 434
- compression of fact tables with careful design, 274
 - dictionary compression in SQL Server 2008, 557
 - improving query performance with, 556–558
- conceptual models as part of user interface design, 91
- configuration choices for data warehouse servers, 583
- conformed dimension assembler, ETL subsystem #8, 431
- conformed dimensions, 15, 51–54, 141, 244, 516
 - affecting ETL design, 428
 - anonymous data warehouse key, 41
 - bus matrix, 154
 - commitment to use, 42
 - cost if not conformed, 91
 - criterion for dimensional DW, 227
 - data warehouses and, 41
 - definition, 40, 149, 190
 - designing, 41
 - establishing, 40
 - executive support required, 149
 - fixing stovepipe data marts, 202
 - grain, 41
 - importance of, 40–41
 - integrated EDW, 198
 - integration and, 105, 108
 - no need for, 45
 - replication, 485–486
 - unconformed and, 91
 - variations, 42
- conformed facts, 15, 42
 - cost if not conformed, 91
 - cross-process calculations, 201
 - definition, 150
 - unconformed, 91
 - using analytically, 162
- conforming data, ETL subsystem #8, 431
- conforming dimensions at query time, unrealistic goal, 523
- conforming nonconformed dimensions, 676–677
- consolidated fact tables, 240
 - combining processes, 240
 - example, 240–241
- constraint targets, always in dimension tables, 198
- constraints on data warehouse design, 58–60
- cookies in clickstream data, 411

corporate data model, 93
 Corporate Information Factory. *See* CIF
 correctly weighted report using bridge table, 342, 345–346. *See* impact report
 correlated dimensions, splitting or combining, 307
 correlated subqueries, 641
 cost allocations. *See* allocations
 costs, 90
 administrative, 59
 hardware, 59
 implementation costs, 59
 software, 59
 sources, 90
 surprises, 59
 coverage dimension in insurance examples, 243–246, 390–393
 coverage fact table, promotion tracking example, 257–258
 coverage tables, finding what didn't happen, 260–262
 CRC (cyclic redundancy checksum), 8
 in change data capture, 486–487
 criteria for dimensional DWs, 226
 critical thinking, 58
 critique data warehouse, 673–674
 CRM (customer relationship management), 599–600
 cross-browsing, 638–639
 cultural correctness, multinational name and addresses, 379
 currencies
 conversion version in audit dimension, 466, 469
 design, 435
 international data, 476
 in multinational designs, 377
 custody of data, compliance responsibility, 407
 custom tool development for ETL and BI, 520
 customer dimension extensions, 367
 customer modeling issues, 366–374
 customer profiling, factless fact table, 259
Customers.Com (Seybold), 525
 cyber warfare, 576–577
 cyclic redundancy checksum. *See* CRC

D

Dangermond, Jack on GIS systems, 384
 dashboards, 612–613
 data
 aggregated prematurely, 92
 awkward formats, 92
 as both fact and dimension, 277
 delivery, slow, 92
 integration, external, 449–450
 integration manager, ETL subsystem #21, 432
 locked, 92
 profiling, 2, 3, 121
 affecting ETL design, 427
 business rule screens, 122
 column screens, 122
 role in organization, 462
 structure screens, 122
 quality
 1996 perspective, 454
 affecting ETL design, 427
 aggregate data quality reporting, 470–471
 comprehensive architecture, 460
 critically dependent applications, 454
 culture steps, 461
 error event handler, ETL subsystem #5, 431
 error responses, 465
 estimating from historical data, 471
 international design issues, 474
 measures in audit dimension, 468
 no history, 473–474
 predictable changes, 474
 six sigma, 467
 standard deviation, 472
 X-11 ARIMA, 474
 staging, 8, 50. *See* ETL
 area, 437
 stewardship, 156, 165
 communications, 167
 goal of program, 165
 master data management, 517
 need for, 165
 qualifications necessary, 167
 responsibilities, 166–167
 transformations, 618–620
 tool-dependent, 620–621
 wrangling, 7

- data audits during requirements gathering, 116.
 - See data profiling
- data cleaning, 439
 - applications, 458
 - ETL subsystem #4, 431
 - regular expressions, 477–481
 - steps, 456
- data conformer, ETL subsystem #8, 431
- data flow, ETL system, 446–447
- data governance as foundation for MDM, 520.
 - See governance
- data marts, 39
 - architecture phase, 39–40
 - avoid departmental definition, 123
 - bus architecture, 46
 - business process subject areas, 61
 - data warehouse bus architecture, 51–54
 - dimensional modeling and, 50–51
 - higher level, 44
 - presentation area, 51
 - quick and dirty data warehouse myth, 202
 - stovepipe data marts, 39
- data mining, 44, 63, 615–617
 - affinity grouping, 617
 - aggregated data and, 44
 - business phase, 626
 - categories, 616–617
 - classifying, 616
 - clustering, 616
 - data mining phase, 626–628
 - data transformations, 618–620
 - tool-dependent, 620–621
 - data warehouse responsibilities, 623
 - database architecture and, 624
 - estimating, 617
 - explain variance of KPI, 24
 - metadata, 629
 - observations, 622–624
 - operations phase, 628–629
 - origins, 615–616
 - predicting, 617
 - process flow chart, 625
 - references, 24
- data profiling, 2–3, 121–123, 462
 - data quality driver, 427
 - ETL subsystem #1, 430
- data quality architecture articles, 460–481
 - Data Quality: The Accuracy Dimension* (Jack Olson), 427
- data warehouse bus architecture. See bus architecture
- data warehouse manager responsibilities, 70–73
- data warehouse not needed, objection remover, 77
- data warehouses
 - building in 15 minutes, 80
 - bus architecture, data marts, 51–54
 - central data warehouse team, 80–83
 - costs, sources, 90
 - mission, 73–74
 - planning, 38
 - publishing results, 48–49, 54
 - securing results, 49, 54
 - as Web-enabled system, 55
- data webhouses, 55–56, 410
- data wrangling, 6–8. See change data capture
- database market split, 35
- Date, Chris
 - An Introduction to Database Systems*, 36, 137
 - criticisms of E/R models and business rules, 147
 - on dimensional models, 137, 147
- date dimension
 - activity date versus booking date, 492
 - advantages, 225
 - attached to every dimensional model, 41, 197, 225
 - conformed EDW, 82, 154
 - design, 291
 - hierarchies in, 352
 - incompatible rollups, 292
 - keys, 198, 288, 297
 - latest thinking, 295
 - multiple dates in accumulating snapshot, 246
 - as outrigger dimension, 292
 - as outriggers, 299
 - recommended design, 250, 294
 - role playing, 298
 - used in SCD2 processing, 26, 193
- DBAs (database administrators), 7
- decentralized development, 47, 52, 60. See distributed architecture

- decodes needed in dimension tables, 198, 224
 - deduplicating source data in ETL system, 18, 439, 487
 - deduplication system, ETL subsystem #7, 431
 - degenerate dimensions, 182, 271
 - airline flight segment example, 394
 - bill of lading example, 396
 - grouping fact table rows, 271
 - health care billing example, 235
 - invoice header example, 264, 267
 - market basket analysis, 271
 - multiple keys in reference dimension, 272
 - order line item example, 300
 - parent-child fact tables, 264
 - shipment invoice example, 465–466
 - storing control numbers, 225
 - tie back to operational system, 271
 - web page event example, 414
 - demographic tracking, factless fact table, 259
 - demographics mini-dimension table, 320–322, 367
 - ETL processing steps, 321
 - permissible snowflake table, 337–338
 - denormalized dimension tables, 136, 181, 197, 334, 352
 - denormalized models, 9
 - departmental data marts to be avoided, 10, 47, 86–88, 123, 128
 - dependency analysis during ETL, 442. *See* impact analysis
 - deployment of data warehouse, 97, 651–661
 - back room, 653
 - BI applications, 609
 - dimensional relational (ROLAP) versus OLAP, 549–552
 - front room, 651–653
 - monitoring operations, 653–654
 - rapid, 47, 53, 60
 - descriptive attributes, verbose, 199
 - descriptive model, 60–61. *See* normative model
 - design drivers, 74
 - design review, 221, 223–231
 - design steps, 210, 405
 - 1 Choose the process, 211
 - 2 Choose the grain, 211, 223
 - 3 Choose the dimensions, 211
 - 3b Confirm the dimensions, 212
 - 4 Choose the facts, 213
 - 5 Store precalculations, 213
 - 6 Round out dimensions, 214
 - 7 Choose database duration, 215
 - 8 Specify SCDs, 215
 - 9 Decide physical design, 215
 - review and validation, 221–223
- design team roles, 3, 40–41, 80, 93, 216
- destruction of facility, 576
- diagnosis bridge table dimension in health care, 341
- diff compare, 453. *See* change data capture
- digital preservation, 579–582
- dimension design response for new attributes, 195
- dimension independence, 180
- dimension keys, durable, natural, surrogate, 18
- dimension limitations in OLAP, 325
- dimension manager, 17–19
 - joint responsibilities with fact provider, 21, 514
 - LDAP, 21
 - MDM resource, 514
 - responsibilities, 17, 163–164
- dimension manager system, ETL subsystem #17, 432
- dimension notification criterion for dimensional DW, 229
- dimension processing in ETL, 438. *See* SCDs
- dimension size limitations, 325
- dimension tables
 - accurate counting, 314–315
 - conformed, 141
 - decodes, 224
 - design process, 214–215
 - many-to-one relationships, 197
 - primary keys, surrogate keys, 198
 - replicating to fact providers, 18
 - row labels, 198
 - shrunken dimensions, 19
 - snowflaked, 181, 333
 - source of constraints and row headers, 139
 - text facts, 199
 - version numbers, 19, 20
- dimension update strategies, 325

- dimensional attributes, overwrites, 317. *See* SCDs
- dimensional criterion for dimensional DW, 228
- dimensional designs, graceful modifications, 194–195
- dimensional DWs
 - administration criteria, 228–229
 - architecture criteria, 227–228
 - criteria, 226
 - expression criteria, 229–231
 - rating scheme, 226
- dimensional models. *See* DM
 - aggregated, 239
 - atomic data, 196
 - based on reports, 200
 - business processes, 197
 - date dimensions, 197
 - departmental data marts, 144
 - extensible designs, 203
 - graceful extensibility, 9, 43, 142, 194–195, 228
 - motivation and advantages, 139
 - normalized model comparison, 134–137
 - versus normalized models, information content, 140
 - null usage, 276–277. *See* nulls
 - populating, 238
 - query evaluation strategy, 135
 - relational models and, 9, 181
 - source data changes, 195–196
 - summarized information, 238
 - symmetrical approach, 141
 - themes, 278
- dimensional queries, processing, 136
- dimensional relational versus OLAP, 550–551
- dimensional replication criterion for dimensional DW, 228
- dimensional scalability criterion for dimensional DW, 228
- dimensional star schema, fact tables, 181. *See* star join, star schema
- dimensional symmetry criterion for dimensional DW, 228
- dimensions, 10, 87, 180
 - abstract, 311–312
 - behavior, 324
 - causal, 308–311
 - conformed, 15, 244
 - correlated, splitting/combining, 307
 - degenerate, market basket analysis, 271
 - degenerate dimensions. *See* degenerate dimensions
 - design process, 211–212
 - facts, data as both, 277–278
 - generic, 311–312
 - hierarchies, multiple, 184
 - hot-swappable, 312–314
 - independence, 192
 - joins, avoiding, 307
 - junk dimensions. *See* junk dimensions
 - keyword dimension, 347–351
 - mini, 326–327, 367
 - missing, 181
 - reference dimensions, 272–273
 - retaining headers as, 266
 - smart keys, 199
 - user interface, 11
 - verbose description attributes, 199
- directory server, 427
- dirty data, cleaning up hierarchies, 354
- disorders of data warehouse. *See* DW/BI checkups
- distraction avoidance in web-oriented data warehouse, 564
- distributed architecture, 56, 60, 103, 151. *See* integration
 - catastrophic failure and, 577
- distributed systems, 151
- DM (dimensional modeling), 133–134. *See* dimensional models
 - 3NF comparison, 140–141
 - data marts and, 50–51
 - defending, 144
 - Microsoft Analysis Services, 553–554
 - myths. *See* myths
 - overview, 139–140
 - retail databases, 143
 - rules, 196
 - snowflaking, 143–144. *See* snowflaked dimension tables
 - stovepiping, 143
 - strengths, 141–142
 - symmetrical approach, 141–142

- top-down design, 135
 - understanding of, 143
 - drill-across reports, 14, 20, 42. *See* integration
 - grouping columns, 185
 - drilling across, 14, 150, 162, 240. *See* integration
 - article, 189–191
 - conformed dimensions, 82
 - danger using two fact tables in same query, 189
 - definition, 185
 - detailed implementation, 190
 - different grains, 33
 - fact tables, 185
 - fact tables of dissimilar grains, 191
 - implementing, 190–191
 - outer join, 190
 - queries, 190
 - sort-merge, 190
 - SQL, 629–631
 - drilling down, 186–189
 - atomic data, 47, 53, 188
 - BI tool user interface design, 28
 - computation, 187
 - data quality attributes, 187
 - definition, 183
 - grouping columns, 184
 - not in a hierarchy, 184, 187
 - predetermined hierarchy, 23
 - precise technical comments, 187
 - row headers, 186
 - user interface, 188
 - drilling up, 184
 - durable key, 18–19, 27, 328–331, 337. *See*
 - natural key
 - duration of data storage, 215
 - DW/BI (data warehouse/business intelligence), 1
 - business acceptance, 667–670
 - business realignment, 667
 - business representatives, 668–669
 - centralized, risks, 59
 - checkups
 - business acceptance disorder, 664–665
 - business sponsor disorder, 662–663
 - cultural/political disorder, 666
 - data disorder, 663–664
 - infrastructure disorder, 665–666
 - custom tools, 520–522
 - failure points, 93
 - feedback, 669–670
 - interview team, 668
 - isolationist approach, 84
 - management education, 670–673
 - marketing system, 656–658
 - performance, 682
 - projects, listing, 106
 - system operations planning, 654
- E**
- E/R modeling, Chris Date on, 147
 - EAI (enterprise application integration), 523
 - early arriving facts in real time applications, 494–495
 - ease of use, 49, 54
 - BI tool acceptability test, 30
 - EDM (enterprise data model), 9–10
 - education of senior staff, 672
 - EDW (enterprise data warehouse), 106. *See* CIF
 - architectural requirements, 47
 - architectures, normalized versus dimensional, 176–178
 - bus matrix, 15–16
 - conformed dimensions and facts, 17
 - integrated, 13–21, 108, 161–164
 - MDM and, 15
 - practical approach, 530
 - reports, 14–15
 - employee dimension
 - best practice design, 359–365
 - bridge table using natural keys, 362
 - dimension outrigger example, 335
 - fixed depth hierarchy compromise, 363
 - human resources example, 396–400
 - insurance matrix example, 160
 - normalized design example, 495–497
 - pathstring attribute for ragged hierarchy, 364
 - reports-to bridge table, 360
 - SCD processing example, 25–27
 - separate reports-to dimension, 361
 - telecomm matrix example, 152
 - time stamps, 399–400
 - enterprise application integration (EAI), 523

- ER (entity-relationship) models, 50, 133, 147
 - normalized, 57, 62
- ERP (enterprise resource planning)
 - data warehouse limitations, 526
 - role of, 526–528
 - systems
 - as primary data warehouses, 56
 - relationship to data warehouse, 525
 - vendors, 56
- error event handler, ETL subsystem #5, 431
- error event schema, recording data quality
 - events, 463–464
- error responses, 465
- ESRI, GIS vendor evaluation, 384–386
 - address standardizing, 385
 - extending SQL for geographic queries, 386
- estimating (data mining), 617
- ETL (extract, transform, and load) staging
 - systems, 62
 - aggregate processing, 440
 - architecture, 105
 - audit dimension, 465–466
 - bottlenecks, 683
 - business rule screens, 463
 - cleaning and conforming, 98
 - column screens, 463
 - custom tools, 520–522
 - data quality error event handler, 431. *See* data quality
 - data staging area, 437
 - deduplicating source data, 487
 - delivering, 98
 - dependency analysis, 442–443
 - design
 - archiving requirements, 428
 - BI tool interfaces, 429
 - business needs, 426
 - compliance, 426
 - conformed dimensions, 428
 - data profiling, 427
 - data quality, 427. *See* data quality
 - example, 445
 - foundations, 443
 - integration requirements, 428
 - latency, 428
 - licenses, 429–430
 - lineage requirements, 428
 - planning inputs, 446–447
 - security, 427–428
 - skills of staff, 429
 - staging, 428
 - tradeoffs, 434
 - designer's responsibilities, 216
 - documentation, 445
 - extract processing, 441
 - extracting, 98
 - hierarchy validation, 354, 440
 - householding, 439–440
 - impact analysis, 442–443
 - junk dimensions and, 307
 - lineage analysis, 442–443
 - managing, 98
 - operational resilience, 442
 - planning steps, 445
 - quality screens, 463
 - referential integrity, 438
 - requirements, 425
 - self-documentation, 442
 - snowflake design, 336
 - structure screens, 463
 - subsystems
 - accumulating snapshot grain fact table loader, 432
 - aggregate builder, 432
 - audit dimension assembler, 431
 - backup system, 433
 - change data capture system, 431
 - compliance reporting, 434
 - conformed dimension assembler, 431
 - data cleaning system, 431
 - data conformer, 431
 - data integration manager, 432
 - data profiling system, 430
 - deduplication system, 431
 - dimension manager system, 432
 - error event handler, 431
 - extract system, 431
 - fact table loader, 432
 - fact table provider system, 432
 - hierarchy dimension builder, 432
 - impact analyzer, 433
 - job scheduler, 433

- junk dimension builder, 432
- late-arriving data handler, 432
- lineage and dependency analyzer, 433
- metadata repository manager, 434
- multi-dimensional cube builder, 432
- multi-valued dimension bridge table loader, 432
- OLAP cube builder, 432
- parallelizing system, 433
- periodic snapshot grain fact table loader, 432
- pipelining system, 433
- problem escalation system, 433
- quality screen handler, 431
- recovery and restart system, 433
- SCD processor, 432
- security system, 433
- sort system, 433
- special dimension builder, 432
- surrogate key creation system, 432
- surrogate key pipeline, 432
- transaction grain fact table loader, 432
- version control system, 433
- version migration system, 433
- workflow monitor, 433
- time zones and, 434–435
- tool pros and cons, 442
- visual flow, 442
- euro, special business rules, 377
- event fact tables, 182
- exception reports, analytic applications, 22
- exceptions
 - analytic application, 22
 - handling, 451–452
 - identification, 590, 592
- explicit declaration criteria for dimensional DWs, 227
- expression criteria for dimensional DWs, 229–231
- extensibility
 - of dimensional models, 142, 203, 206
 - graceful, 9
 - new data source, 205
- extensible markup language. *See* XML
- external data integration, 449–450
- extract, transform, and load. *See* ETL
- extract processing during ETL, 441
- extract system, ETL subsystem #3, 431

F

- fact dimension, modeling sparse facts, 281
- fact provider, 17, 19–20, 164
 - joint responsibilities with dimension manager, 21
 - LDAP, 21
- fact table grains. *See* grain (fact tables)
 - transaction, periodic snapshot, accumulating snapshot, 193, 243–244
- fact table loader, ETL subsystem #13, 432
- fact table provider system, ETL subsystem #18, 432
- fact tables, 37
 - atomic, as core foundation, 239
 - audit dimensions, 187
 - combining types in periodic and accumulating snapshots, 249
 - consolidated
 - combining processes, 240
 - example, 240–241
 - cost cutting and, 682–683
 - departmental views, 9
 - design patterns, 273–283
 - design process, 213–214
 - dimensional star schema, 181
 - drilling across, 185, 240. *See* drilling across; integration
 - factless, 255–258
 - flexible width, 281
 - grain. *See* grain (fact tables)
 - declaring, 30
 - declaring before dimensions added, 199
 - uniformity, 197
 - granularity, 43–44
 - instantaneous transactions, 278
 - many-to-many relationships, 197
 - parent/child, 262–268
 - partitioning with smart date keys, 296–297
 - pivoting, 282–283
 - populating, 104
 - primary keys, 181
 - purpose of, 30
 - response to measurement events, 9, 11
 - rows, grouping with degenerate dimensions, 271

- scalable width, 281
 - second-level, 240
 - size reduction with careful design, 274
 - sparse facts, 280–282
 - surrogate keys, 33
 - reader suggestions, 269
 - where to use, 268
 - time stamps, 192
 - types. *See* grain (fact tables)
 - used as dimensions, 278
 - factless fact tables, 182, 255
 - attendance tracking, 256
 - automobile collisions, 257
 - customer profiling, 259
 - demographic tracking, 259
 - SCDs, 258
 - facts, 11, 134
 - additive facts, 12, 182, 227
 - conformed, 15, 42–43
 - design process, 213
 - dimensions, data as both, 277
 - measurement events, 11
 - non-additive, 31, 227, 281
 - nulls as, 277
 - numeric measurements, 179
 - semi-additive, 182, 227, 509, 548, 554, 639
 - unconformed, 91
 - user interface, 11
 - feedback from end users, 669–670
 - finance, boundaries with, 5
 - financial product dimensions, 82, 313, 338–339, 436
 - financial services date dimension roles, 298
 - first-level data marts, 44. *See* second-level; integration
 - first-level subject area, 153. *See* second-level; integration
 - fixed-depth hierarchy, 363–364
 - fixed-width databases, 338
 - FK (foreign key), 31
 - flat file, 62, 437, 440–441, 624
 - flexible width fact tables, 281
 - foreign keys, nulls as, 276–277
 - foreign keys in dimensional schemas, 31, 180–181
 - Friedmann, Thomas (*The World is Flat*), 474
 - front room, 32–33, 50–51, 651–653
 - architecture, 560–565
 - BI applications, 589–649
 - metadata, 569–570
 - FTP based integration, 450–452
 - fundamental fact table grains, 243–246. *See* grain (fact tables)
- ## G
- general ledger account dimension in budgeting value chain, 406
 - general ledger (GL), 4–5
 - fact table example, 336
 - tying to operational results, 5
 - generic dimensions, avoiding, 311
 - geocoder for ESRI GIS parsing of addresses, 386
 - geographic information system (GIS), link to data warehouse, 383
 - address standardizing, 385
 - evaluation of ESRI GIS vendor 384–386
 - extending SQL for geographic queries, 386
 - geography dimension in a conformed EDW, 82
 - GIS. *See* geographic information system
 - GL. *See* general ledger
 - governance, 108
 - driving MDM initiatives, 520
 - driving SOA initiatives, 513–514
 - graceful extensibility, 9, 43, 53, 59
 - graceful modification criterion for dimensional DW, 228
 - graceful modifications to dimensional designs, 140, 142, 194–195, 309
 - grain (fact tables), 30
 - accumulating snapshot grain, 32, 194
 - capture lowest possible, 10
 - clickstream data, 412
 - conformed dimensions, 41
 - declaration, 30, 182
 - before design begins, 200, 233
 - precedes key definition, 235
 - definition, 11
 - foundation of design, 223
 - as definition of business event, 237
 - design process, 211, 223
 - drilling across, 33

- fundamental grains comparison, 243–246
- mismatches, mixed grain in fact table, 200
- mixed grain problems, 223–224
- periodic snapshot grain, 32, 194
- transaction grain, 32, 193
- uniform throughout each fact table, 197
- grouping columns, 183
 - drill-across reports, 185
 - drilling down, 184
 - drilling up, 184
 - row headers, 186
 - in a SELECT list, 186
- growth management, 658–661

H

- hardware cost, 59. *See* costs
- header/line item designs, 262–268
- heterogeneous product design, 82, 313, 338–339, 436
- hierarchies, 351
 - alternate, 365–366
 - design for maintainability, 351
 - dimensions, multiple, 184
 - dirty sources, 354–355
 - drilling down, 23, 28, 187
 - fixed depth, 197, 363–364
 - mistakes in design, 199, 224, 334
 - multiple in a dimension, 229, 352
 - ragged, 229–230, 355–356
 - pathstring attribute, 364–365
 - shared ownership, 358
 - referential integrity, 352
 - single dimension, 224
 - splitting into multiple dimensions, 199
 - validation
 - during ETL, 440
 - in ETL system, 354
- hierarchy bridge table
 - design, 357
 - manufacturing parts explosion, 358–359
 - shared ownership, 358
 - time varying, 358
- hierarchy dimension builder, ETL subsystem #11, 432
- hierarchy management with custom tool, 521

- historical dimension rows, 488–490
- historical letter data warehouse, 347
- historically accurate attributes, lack of, 531
- history
 - preservation, data warehouse requirements, 191, 579–582
 - seamlessness, 61
- Holtzman, David, 115
- hot partition in real time systems. *See* real-time partition
- hot response cache in real time systems, 504
- hot-swappable dimensions, 312–314
 - criterion for dimensional DW, 231
 - multi-client security, 313
- householding during ETL, 379, 439, 455, 458
- hub and spoke architecture, 171. *See* CIF
- human resources case study, 396–400. *See* employee dimension
- hybrid approach, combining CIF and Kimball, 175
- hybrid SCDs (slowly changing dimensions)
 - combination type 1, 2, 3, 326
 - type 1, 2, 3, 326–328
 - type 1 + 2 tracking with natural keys in fact table, 328
 - type 1 fact and type 2 mini-dimension, 327
 - type 6 combination of all three types, 327

I

- impact analysis during ETL, 442
- impact analyzer, ETL subsystem #29, 433
- impact report using bridge table, 343, 346. *See* correctly weighted report
- implementation cost, 59
- income statement fact table
 - allocations, 402
 - design, 401
- incompatible data, dealing with, 21, 26, 43, 45, 83, 91, 111, 142, 149, 184, 191, 207, 282, 292, 313, 339, 373, 391, 514, 523
- incompatible technologies, 14, 48, 53, 56, 60, 75, 78, 290, 380, 455
- increased granularity of dimension, design response, 196
- indexes for DW/BI databases
 - B-tree indexes, 37, 269, 508

- bitmap indexes, 81, 269, 325, 559, 562
 - substring index, pattern index for high speed searching, 350
 - Inf*Act, Nielson syndicated reporting, 8
 - inheritance, line items inheriting dimensionality, 267
 - insurance coverage limits example, 12
 - insurance data warehouse examples, 129–130, 243–245, 257, 320–322, 389–393
 - integration. *See* dimension manager; fact provider, drill-across
 - conformed dimensions and, 108, 198
 - definition, 161–162
 - drill-across as litmus test, 14
 - EDW, 13, 38, 161–164
 - MDM, 13
 - measures, 162–163
 - normalization and, 207–208
 - integration requirements affecting ETL design, 428
 - international. *See* multinational
 - internet. *See* web
 - interviews, 5
 - gathering requirements, tactic and objective, 210
 - interviewing techniques, 113–121, 668–670
 - investment banking
 - custom hot-swappable dimensions, 312–313
 - junk dimensions, 303
 - IT (information technology)
 - boundaries, 6
 - functions, centralizing, 79
 - licenses, 2
 - partnership with, 91
 - review, 222
- J**
- job scheduler, ETL subsystem #22, 433
 - joins between dimensions, avoiding, 307
 - junk dimension builder, ETL subsystem #12, 432
 - junk dimensions
 - advantages, 306
 - combining or separating, 305
 - creating, using, maintaining, 497–499
 - decided granularity, 305
 - ETL choices for creating, 307
 - investment banking example, 303
 - when to use, 275
- K**
- Key, Alan (father of personal computer), 560
 - key performance indicator. *See* KPI
 - keys in dimensional schemas, 180
 - keyword dimension, 347–351
 - Kimball Approach, 99
 - bus architecture. *See* bus architecture
 - CIF, fundamental differences, 174
 - enterprise versus departmental, 128
 - hybrid with CIF, 175
 - measurement processes versus departmental reports, 204
 - myths, 206
 - Kimball Lifecycle, 96–99
 - agile approach and, 110
 - bottom up approach, 100
 - business intelligence track, 99
 - business requirements, 98
 - data track, 98
 - deployment, maintenance, and growth, 99
 - diagram, 97
 - Metaphor Computer Systems, 96–97
 - program/project planning and management, 98
 - technology track, 98
 - kitchen metaphor for DW/BI system, 65–68
 - know-it-all users, 120–121
 - KPI (key performance indicator), 2–5
 - airline example, 22
 - compliance impact, 597
 - conformed facts, 428
- L**
- labels, integrating, 162. *See* integration
 - languages. *See* multinational
 - late-arriving
 - data, 19
 - data handler, ETL subsystem #16, 432
 - dimension records processing steps, 492
 - fact records, processing steps, 491
 - latency affecting ETL design, 428
 - latent semantic analysis for unstructured text search, 418

launching BI environment, 652
 LDAP (lightweight directory access protocol)
 managing role enabled security, 578
 server, dimension manager and fact provider
 responsibilities, 21
 lean times DW fitness program, 680–684
 legacy data formats, resolving inconsistent, 618
 legal department, boundaries, 6
 licenses
 cost cutting and, 682
 for software and systems, affecting ETL design,
 429
 lightweight methodologies, 109. *See* agile
 line items, inheriting dimensionality, 267
 lineage, 468
 analysis during ETL, 442
 and dependency analyzer, ETL subsystem #29,
 433
 requirements affecting ETL design, 428
 Linoff, Gordon, 600, 625
 locked data, 92
 log scraping for change data capture, 453
 LSA (latent semantic analysis), 418

M

management education, 670–673
 many-to-many bridge tables, 335
 many-to-many dimension relationships, resolve
 in fact table, 197
 many-to-many relationships, prevalence of, 146
 many-to-one relationships
 outriggers, 334
 resolve in dimension table, 197
 snowflaking and, 334
 MapObjects Visual Basic tool for GIS, 384
 market basket analysis, 420–424
 degenerate dimensions, 271
 proposed dimensional design, 420
 marketing of the DW/BI system, 656–658
 marquee applications, 598
 master data management. *See* MDM
*Mastering Data Mining, The Art and Science of
 Customer Relationship Management* (Berry
 and Linoff), 600, 625
 matrix. *See* bus matrix

MDM (master data management), 13, 353
 business value, 515
 centralized enterprise source, 519–520
 deployment steps, 520
 dimension manager role, 514
 EDW and, 15
 importance of data governance, 520
 integration hub, 517–518
 need for, 516
 and SOA with agile development, 111
 solving data disparity, 516
 source system disparities, 515
 supported from data warehouse, 516–517
 three approaches, 516
 MDX (multidimensional expressions), 648
 measured facts, new, design response, 195
 measurements, fact tables, 11, 104, 179
 reports and, 204
 snapshots. *See* grain (fact tables)
 measures, integration. *See* conformed facts
 media, formats in data warehouse 55
 archiving, preservation and, 580
 medical information privacy, 573
 MERGE command (SQL) for SCD processing,
 499
 merge-sort, drilling across, 190
 meta meta data data (data about metadata), 566
 metadata, 613–614
 complete list for data warehouse, 567
 data mining, 629
 data warehouse scope, 566
 management tasks, 567
 management tools, 570–572
 repository manager, ETL subsystem #34, 434
 strategy recommendations, 571
 Metaphor Computer Systems, 96–97
 Microsoft Analysis Services 2005, 553–555
 Microsoft SQL Server 2005, data warehouse
 architecture guidelines, 554
 Microsoft SQL Server 2008
 database compression, 556–558
 new features, 556
 star schema optimization, 559
 table partitioning, 558
 migrating from disparate data to centralized, 170
 mini-dimension tables, 367

- customer attributes, 367
- demographics example, 320
- linking to primary dimension through fact table, 322
- monster dimensions, 320
- overwrite and, 326–327
- mistakes in building a DW/BI system, 100
- mixed grain problems, double counting, 223, 227
- model alternatives, analytic applications step, 590, 592–593
- monolithic approach, 38
- monster dimensions, rapidly changing, 320
- multi-dimensional cube builder, ETL subsystem #20, 432
- multi-pass SQL, 150. *See* drill-across
- multi-valued dimension
 - health care diagnoses, 341
- multi-valued dimension bridge table loader, ETL subsystem #15, 432
- multinational data, 374–388
 - addresses, 475
 - calendars, 376–377, 476
 - character sets, 475
 - compliance, 476
 - consistency criterion for dimensional DW, 229
 - cultures, 475
 - currencies, 377–378, 476
 - customer information in real time applications, 379
 - data warehouse design considerations, 387
 - dimension translation, 387
 - euro, 377–378
 - geographies, 475
 - languages, 475
 - names, 475
 - names and addresses, 378–383
 - Atkinson, Toby, 381
 - cultural correctness, 379
 - numbers, 476
 - postal address formats, 380
 - quality architecture, 477
 - quality issues, 474
 - reporting issues, 566
 - salutations, 475
 - time zones, 476, 477

- multiple dimension hierarchies criterion for dimensional DW, 229
- multiple dimension roles criterion for dimensional DW, 230
- multiple hierarchies in a dimension, 184, 351
- multiple valued dimensions criterion for dimensional DW, 230
- myths, 8–10, 38, 143–144, 201–203
 - atomic data should be normalized, 239
 - dimensional models pre-suppose the business question, 238
 - facts and fables, 204–208

N

- N-tiling, 636
- name and address processing, 439
 - cultural correctness, 379
 - international design issues, 378–383
- naming conventions, 220–221, 450
- natural keys, 18. *See* durable keys
 - bridge table, 362–363
 - in fact table for type 2 and type 2 tracking, 328–331
 - problems with, 287
 - in surrogate key pipeline, 482
 - surrogate keys, 285
- navigation, aggregate navigation, 32–33
- network database design, 393
- Nielsen syndicated reporting. *See* Inf*Act
- non-additive numeric fact, 281
 - bad design example, 213
 - computing from additive facts, 264
 - handling in BI tool, 635–639
 - handling in OLAP, 548
 - summarizing across time, 293
- non-behavior, explicit records for, 260
- non-existence of events, techniques for querying, 259
- nonconformed dimensions, conforming, 676–677
- nonexistent users, 121
- normalization, integration and, 207–208
- normalized data models, 9, 12, 133, 137. *See* ER BI queries, 144
 - complexity and BI, 146

- compared to dimensional model, 134
- creating dimensional views, 77
- uniqueness or completeness, 57, 146
- normalized data warehouse. *See* CIF
- normalized data warehouse, lack of procedure
 - for slowly changing dimensions, 177
- normalized EDW not for business intelligence, 176
- normalized hierarchy disadvantages, 224
- normative model, 60–61. *See* descriptive model
- NOT EXISTS
 - missing attributes, 262
 - what didn't happen, 261
- nulls
 - as dimension attributes, 277
 - as fact table foreign keys, 276–277
 - as facts, 277
- numbers, international data, 476

O

- objection removers, 76, 77
 - aggregates, 78
 - applications integrators, 78
 - backups, 79
 - centralized customer management system, 78
 - centralizing IT functions, 79
 - larger problem and, 77
 - recognizing, 77
 - security, 79
 - solutions for, 77
- ODS, operational data store hot cache, 504
- offline delays during ETL processing, 502–503
- OLAP (online analytical processing), 17, 46
 - advantages versus dimensional relational, 551
 - analytic syntax, 551
 - catastrophic invalidation with SCD Type 1, 552
 - cube builder, ETL subsystem #20, 432
 - data cube, 63
 - desktop versus server, 547
 - dimension limitations, 325
 - versus dimensional relational advantages, 550
 - versus dimensional relational disadvantages, 551
 - dimensions comparison with ROLAP
 - dimensions, 547

- disadvantages versus dimensional relational, 552
- implementing aggregations via strong hierarchies, 548
- major advantages, 548–549
- as major data warehouse component, 546
- versus ROLAP, final deployment choice, 549–553
- SCDs contrasted with ROLAP SCDs, 548
- security scenarios, 551
- sensitivity to type 1 SCD, 25
- similarity to star schemas, 63
- SQL-99 extensions, 645–649
- time constraints contrasted with ROLAP, 548
- Olson, Jack (*Data Quality: The Accuracy Dimension*), 427
- OLTP (online transaction processing), 36
 - data warehouse systems, 37
 - models, 137
- on-the-fly behavior dimensions criterion for dimensional DW, 231
- on-the-fly fact range dimensions criterion for dimensional DW, 231
- online analytical processing. *See* OLAP
- online transaction processing. *See* OLTP
- operating procedures, 655–656
- operational systems back pointers, 487–488
- operations phase of data mining, 628–629
- operators, RegExp, 479
- opportunity matrix, 158
 - processes versus departments, 130
- OR queries, 349–350
- outrigger dimension, 135–136, 334–335
 - cautions, 224
 - date dimension as, 292, 299
 - time dimension as, 292
 - variation of snowflaking, 224–225, 336–339
- overbooked users, 120
- overwriting, type 1 SCD, 25–26, 317
- overzealous users, 120

P

- packaged applications
 - avoiding stovepipes, 522–523
 - data warehouses and, 522–524, 529

- page events in clickstream dimensional design, 412–417
 - parallel communication paths, catastrophic failure and, 577
 - parallelizing system, ETL subsystem #31, 433
 - paralysis of project, 84–85
 - parent-child fact tables, 262–268
 - degenerate dimensions, 264
 - design alternatives, 263
 - partitioning
 - fact tables with smart date keys, 296–297
 - real time design, 507–510
 - surrogate keys and, 297
 - table partitioning, 558
 - tricks to minimize offline time, 502
 - type 2 SCD, 316
 - partnership between IT and business, 91
 - parts adding up to whole, 48, 53. *See* distributed architecture
 - pathstring attribute for ragged hierarchy, 364
 - pattern index for high speed searching, 350
 - payments fact table as part of budgeting value chain, 404
 - performance guidelines of web-oriented data warehouse, 562
 - periodic snapshot grain. *See* grain (fact tables)
 - periodic snapshot grain fact table loader, ETL subsystem #13, 432
 - periodic snapshot grain real time partition, 509
 - personal data
 - ownership, 574
 - uses and abuses, 573
 - personnel, staffing team 70, 217–218
 - pipeline processes, accumulating snapshots, 246
 - See* grain (fact tables)
 - pipelining system, ETL subsystem #31, 433
 - pivoting fact table with fact dimension, 282–283
 - P&L (profit and loss) fact table, 401–402, 436
 - playbooks for all operations, 656
 - populating dimensional models, 238
 - predicting (data mining), 617
 - presentation area, 51, 62–63, 67
 - preservation. *See* digital preservation
 - primary keys in dimensional models, 181
 - prioritization grid, benefit versus feasibility, 131
 - privacy
 - concerns from RFID tags, 534–535
 - data warehouse architecture and, 575
 - information transfer and, 476
 - tradeoffs in data warehouses, 572
 - private attributes in conformed dimensions, 16, 19
 - problem escalation system, ETL subsystem #30, 433
 - problem resolution in web-oriented data warehouse, 565
 - process-centric rows in bus matrix, 156–157
 - process steps, data warehouse design, 210
 - process streamlining in web-oriented data warehouse, 564
 - processes versus departments, 123
 - procurement pipeline, accumulating snapshot example, 241–242
 - product dimension, conformed in an EDW, 82
 - production keys, problems with, 287
 - production (source) transaction processing systems, 62
 - profitability case study, 400–403
 - profitability fact tables, allocations, 402, 436
 - progressive subsetting queries, 642
 - promotion dimension
 - design example, 308
 - design recommendations, 310
 - promotion profitability, 311
 - promotion tracking, factless fact table, 257
 - provenance, lineage 468
 - pruning algorithm in market basket analysis, 423
 - publishing metaphor for data warehouse manager, 58, 70, 73
 - publishing reports, 590–591
 - purchase behavior privacy, 573
- Q**
- quality culture, 461–462. *See* data quality architecture articles
 - quality screen handler, ETL subsystem #4, 431
 - quality screens in ETL architecture, 463
 - queries, BI
 - AND, 349–350
 - behavioral, 642
 - browse queries, 135

- decomposition, 639. *See* drill-across reports; drilling across
- drill-across operations, 190. *See* drill-across reports; drilling across
- features for query tools needed, 638–649
- hot-swappable dimensions, 313
- OR, 349–350
- performance
 - cost when too slow, 92
 - priorities for improving, 201
- SQL, categories, 641–642
- query time dimension conforming, goals, 523

R

- ragged dimension hierarchies criterion for dimensional DW, 229
- ragged hierarchies. *See* hierarchies
 - bridge table solution, 355–358
 - pathstring attribute solution, 364–365
 - recursive pointer problems, 357
- rapid deployment, 47, 53
- rating scheme for dimensional DWs, 226
- real-time architectures, 503–509
 - customer information in multinational applications, 379
 - late arriving dimensions, 494
 - real-time partitions, 507
- real-time partition design, 507
 - accumulating snapshot grain, 509–510
 - periodic snapshot grain, 509
 - transaction grain, 508
- real-time triage, judging user requirements, 510–511
- realignment, business, 667
- reason code, SCD2, 330–332. *See* SCDs
- reassuring users, in web-oriented data warehouse, 565
- recency, frequency, intensity. *See* RFI
- recovery and restart system, ETL subsystem #24, 433
- recursive pointer
 - problems, modeling ragged hierarchies, 357
 - replaced by hierarchy bridge table, 357
- redundancy, data
 - 3NF and, 138
 - reducing, 679–680
- reference dimensions, 272–273
- referential integrity
 - in dimensional schemas, 181, 228
 - enforcing during ETL, 438
 - handling nulls, 276, 295
 - in hierarchies, 352
- regular expressions (RegExp)
 - for data cleaning, 477–481
 - operators, 479
 - uses, 480–481
- relational databases, business rules, Chris Date, 137, 147
- relational models
 - dimensional models and, 9, 181
 - EDM and, 10
- relational online analytical processing. *See* ROLAP
- replicating conformed dimensions, 18, 485–486
- reporting
 - accuracy testing, 608
 - analytic application, 22
 - custom tool, 521
 - dashboard development, 612–613
 - deployment, 609
 - development, 607
 - documentation, 604–605
 - EDW, 14–15
 - maintenance, 609
 - management, 609
 - measurements and, 204
 - navigation framework, 606
 - performance testing, 608
 - portal development, 610–612
 - presentation area, 51, 62–63, 67
 - publishing, 590, 591
 - replication, 606
 - report creation, 602–608
 - reporting portal, 652
 - specifications, 604–605
 - standard, 602
 - system design, 603–606
 - target report list, 603
 - template, 604
 - user review, 606
 - users' involvement, 610
- response time to data warehouse queries, 49, 54, 56. *See* queries, BI

- responsibilities of DW/BI team
 - data warehouse manager, 70
 - team members, 217
 - results, preventing irrelevant, 59
 - return on investment. *See* ROI
 - review and validate design, 221
 - RFI (recency, frequency, intensity)
 - behavior tags, 337, 368–369
 - definitions, 369
 - RFID (radio frequency identification) tags
 - application examples, 533
 - impacting personal privacy, 534
 - sequential behavior analysis, 534
 - smart dust, 535
 - tracked in data warehouse, 533
 - ROI (return on investment), data warehouse, 93
 - ROLAP (relational online analytical processing), 48
 - ROLAP versus OLAP, final deployment choice, 549–553
 - role playing dimensions, 10, 300, 312
 - telecomm example, 301
 - transportation example, 301
 - in voyage and network designs, 395
 - rolling date reporting, 252
 - rolling operational results, tying to GL, 5
 - row change reason code, ETL. *See* SCDs, type 2
 - row headers. *See* grouping columns
 - row labels in dimension tables, 198
 - rules for dimensional modeling, 196
- S**
- sabotage, 576
 - SANs (storage area networks)
 - as counter to security catastrophes, 578
 - data warehouse and, 585
 - typical configuration, 586
 - Sarbanes-Oxley Act, 596
 - satisfaction metrics
 - chaotic lists, 373
 - design alternatives, 371
 - simultaneous dimension and fact, 372
 - standard fixed list, 371
 - scalable width fact tables, 281
 - scaling out, scaling up a data warehouse, 584
 - SCD processor, ETL subsystem #9, 432
 - SCDs (slowly changing dimensions), 315–332
 - comprehensive overview, 24
 - criterion for dimensional DW, 230
 - delaying dealing with, 199
 - dimension manager responsibilities, 18
 - factless fact tables, 258
 - handling, 193
 - hybrid combinations
 - type 1 +2 tracking with natural keys in fact table, 328–329
 - type 1 fact and type 2 mini-dimension, 327
 - type 6 combination of all three types, 327–328
 - MERGE command (SQL), 499–501
 - place in dimensional modeling, 322
 - processing in ROLAP, 438
 - with OLAP, 548
 - rapidly changing, mini-dimensions, 323
 - slowly changing entities, normalized time
 - variance tracking, 495–497
 - strategies for, 225–226
 - too fast, 324
 - type 1 (overwrite), 25–26
 - type 2 (new dimension record), 26–27
 - begin- and end-effective time stamp, 193, 323
 - change description, 193
 - most recent flag, 193
 - reason codes, 330–332
 - type 3 (new field), 27
 - scorecards and dashboards, 612–613
 - screens, data quality. *See* data quality
 - architecture articles
 - SCRUM, 109
 - SDE (spatial database engine)
 - ESRI GIS semantics extender for SQL, 386
 - searches
 - pattern index, 350–351
 - substrings, 350
 - seasonal fluctuations, removing when testing
 - data quality, 474
 - second-level subject area, 44, 155
 - fact tables, 240
 - profitability design, 400
 - risks, profitability and satisfaction, 102
 - second normal form, dimension tables, 181

- security, 2, 3
 - architecture, 83
 - catastrophes
 - categories, 576–577
 - techniques for countering, 577–578
 - EDWs, 83
 - ETL design, 427–428
 - ETL subsystem #32, 433
 - management with custom tool, 521
 - objection removers, 79
 - scenarios with OLAP, 551
 - technique for multiple clients, hot-swappable dimensions, 313
- self-documenting code, 601
- semi-additive numeric facts, 182, 293
 - BI application handling techniques, 639
 - declaring in metadata, 227
 - OLAP handling advantages, 548, 554
 - real time partition handling, 509
- sequential behavior analysis using RFID tags, 24, 597–598
- sequential computations in BI tool, 635–638
- server configuration choices for data warehouse, 583
- service accounts versus personal DBA accounts, 655
- service oriented architecture. *See* SOA
- session type dimension in clickstream
 - dimensional design, 415
- Seybold, Patricia (*Customers.Com*), 525
- shadow functions, office anthropology, 115
- shapefiles, GIS data object for boundaries and areas, 385
- shared ownership, hierarchy bridge table, 358
- shrunk dimension tables in aggregate
 - architecture, 540–542
- similarity metrics for unstructured text, 417–420
- six sigma data quality, 467
- skills, for DW/BI team, 93
- SLA (service level agreement), 655
- slowly changing dimensions. *See* SCDs
- smart dust. *See* RFID tags
- smart keys
 - date keys for partitioning fact tables, 296–297
 - dimensions, not for fact table joins, 200
 - disadvantages, 286
 - problems in data warehouse, 288
- snapshots, periodic, accumulating. *See* grain (fact tables)
- snowflaked dimension tables, 135, 181
 - classic design, 336
 - complex calendar dimension, 339
 - context-dependent, 338
 - definition, 333
 - financial product dimension, 338
 - impact on usability, 104
 - large custom dimension, 337
- snowflaking
 - as alternative to dimensional model, 143
 - disk space and, 224
 - as DM alternative, 143–144
 - outriggers, 224
- SOA (service oriented architecture)
 - agile development, 111
 - data warehouse and, 513–515
 - services defined for dimension manager, 514
- software development manager, lessons learned, 601
- sort-merge, drilling across, 190
- sort system, ETL subsystem #28, 433
- sparse facts
 - fact dimension, 281
 - wide fact tables, 280–282
- sparsity tolerance criterion for dimensional DW, 228
- spatial database engine. *See* SDE
- special dimension builder, ETL subsystem #12, 432
- sponsor from business, 86–89
- SQL-92, flexibility of, 645
- SQL-99, OLAP extensions, 645–649
- SQL (Structured Query Language)
 - CASE expression, 633
 - comparisons, 631
 - drill across, 629–631
 - as interim language, 631
 - MERGE for SCD processing, 499
 - multi-pass SQL, 150
 - queries, categories, 641–642
- staffing dimensional modeling team, 70, 216–217, 429
- skills development, 93

- staging area, 62, 66. *See* archiving
 - affecting ETL design, 428
 - standard deviation used for data quality
 - estimating, 472
 - standard reports, 602. *See* reporting
 - star join model, relationship to dimensional model, 139
 - star schema optimization in Microsoft SQL Server 2008, 559
 - star schemas
 - fact tables, 181
 - OLAP data cubes and, 63–64
 - Star Workstation, Xerox, 57
 - statistical analysis as part of data mining, 616
 - steering committees, 672–673
 - stovepipes, 38–39
 - avoiding, 522–523
 - converting to architected dimensional data marts, 45
 - strategic business initiatives, 127
 - matrix, 158–159
 - street segment data, TIGER Census Department, 386
 - structure screens, 122
 - in ETL data quality architecture, 463
 - sub-types and super-types. *See* heterogeneous product design
 - subject area groups in conformed dimension design, 154
 - subject areas
 - first level, 153
 - second-level, 155
 - substring searching in keyword list, 350
 - subtransactions describing behavior, 368
 - sunsetting older environments, 681
 - super-types and sub-types. *See* heterogeneous product design
 - surrogate key administration criterion for dimensional DW, 229
 - surrogate key creation system, ETL subsystem #10, 432
 - surrogate key pipeline, 20, 26, 481–485
 - ETL subsystem #14, 432
 - inserting surrogate keys, 482
 - surrogate keys, 109, 285–289
 - advantages, 225, 285–286
 - bridge tables, 344–345, 360–361
 - creating, 677–678
 - dimension manager responsibilities, 18
 - dimension table primary keys, 198
 - example used incorrectly, 289
 - fact tables, 33
 - required by type 2 SCD, 26
 - fact tables
 - reader suggestions, 269
 - where to use, 268
 - natural keys, 285
 - partitioning and, 297
 - uncertainty, 287
 - surveillance privacy, 573
- T**
- table partitioning. *See* partitioning
 - tape recorders during requirements gathering, 115
 - TCO (total cost of ownership) of data
 - warehouse, 89
 - telecomm bus matrix, 152
 - telecomm dimensional roles example, 301
 - telephone system comparison, 60
 - text document searching, 417–420
 - text field problems in fact table, 224
 - text in fact tables, removal techniques, 275
 - text facts
 - recency, frequency, intensity behavior tags, 369
 - recommended design, 370
 - The Data Warehouse Lifecycle Toolkit* (Kimball, et al), 97
 - The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?* (Brin), 574
 - The World is Flat* (Friedman), 474
 - third normal form, fact tables, 181. *See* 3NF
 - TIGER census department data, USA street segments, 386
 - time constraints, ultra precise, 251
 - time dimension, 192
 - bad design, 293
 - incompatible rollups, 292
 - keys, 293
 - as outrigger dimension, 292

- recommended design, 294
 - role playing, 298
 - time spans created by transactions, 250
 - time stamps
 - begin- and end-effective, 251, 289
 - type 2 SCD, 323
 - bridge tables, 345–346
 - employee dimension table, 399
 - fact tables, 192
 - to nearest second, 490
 - time zones and, 375
 - time variance in dimensions. *See* slowly changing dimensions
 - time zone discovery (www.timezoneconverter.com), 477
 - time zones
 - ETL system tradeoff, 434
 - international data, 476, 477
 - synchronizing, 374–376
 - top-down design, dimensional modeling, 135.
 - See* bottom-up approach
 - total cost of ownership. *See* TCO
 - training
 - data subsets used in data mining, 620–629
 - DW/BI business users, 101, 652
 - transaction grain fact table, 32, 193, 243–244.
 - See* grain (fact tables)
 - transaction grain fact table loader, ETL subsystem #13, 432
 - transaction grain real time partition. *See* real-time partition design
 - transaction processing models, 137
 - Transaction Processing Performance Council, 37
 - transaction workloads in data warehouse, 532
 - translations in multinational data warehouse, 387, 477
 - transportation database design, 301, 393
 - travel case study, 393–396
 - trust building in web-oriented data warehouse, 61
 - type 1, type 2, type 3, type 6 SCDs. *See* SCDs
- U**
- uncertainty, encoding with surrogate keys, 287
 - unconformed dimensions and facts, 91
 - UNICODE character set, 475
 - multinational information, 380
 - units of measure, conflicts, 435
 - university admissions, accumulating snapshot example, 247
 - unstructured text applications, 420
 - LSA (latent semantic analysis), 418
 - similarity metrics, 417–418
 - unstructured text fact table, 417
 - user-focused cognitive and conceptual models, 91
 - user interface, 57
 - advances driven by the Web, 561
 - design, 56
 - BI tools, 28, 57, 91
 - dimensions, 11
 - drilling down, 188
 - facts, 11
 - guidelines for web-oriented data warehouse, 562
 - poorly performing, 92
 - urgency, 561
 - WYSIWYG (what you See is what you get), 560
 - user types
 - abused, 119
 - boundaries, 5
 - clueless, 121
 - comatose, 120
 - control, 107
 - know-it-all, 120–121
 - nonexistent, 121
 - overbooked, 120
 - overzealous, 120
- V**
- version control, 71, 215, 450
 - version control system, ETL subsystem #25, 433
 - version management
 - audit dimension, 466–470
 - fact and dimension tables, 19–21, 25, 163–164, 313, 344, 408, 450
 - version migration system, ETL subsystem #26, 433
 - voyage database design, 393

W

- waterfall development approach compared to agile approach, 107
- waterfall development risks, 102
- web-oriented data warehouse, 48–51, 55
 - choice presentation, 563–564
 - dimensional design, 410
 - distracted avoidance, 564
 - page object dimension, 415
 - performance guidelines, 562–563
 - problem resolution, 565–566
 - process streamlining, 564–565
 - reassuring users, 565
 - session modeling, protocol analysis 413, 416
 - user interface guidelines, 562–566
 - visitor dimension, 414
 - web page characteristics, 409–413

- weighting factor in bridge tables, 342
- what didn't happen, techniques for finding, 259
- what if analysis, analytic applications, 22
- workflow monitor, ETL subsystem #27, 433
- worksheets during design phase, 219
- WYSIWYG (what you See is what you get) user interfaces, 560

X

- X-11 ARIMA statistic for data quality testing, 474
- Xerox PARC, birthplace of personal computer, 560
- XML (extensible markup language), 8
 - data warehouse integration, 523
- XP (Extreme Programming), 109