

# Index

## Symbols

- 3NF (third normal form) models, 7
  - ERDs (entity-relationship diagrams), 8
  - normalized 3NF structures, 8
- 4-step dimensional design process, 38, 70–72

## A

- abnormal scenario indicators, 255–256
- abstract generic dimensions, 66
  - geographic location dimension, 310
- accessibility goals, 3
- accidents (insurance case study), factless fact tables, 396
- accounting case study, 202
  - budgeting, 210–213
  - fact tables, consolidated, 224–225
  - G/L (general ledger), 203
    - chart of accounts, 203–204
    - currencies, 206
    - financial statements, 209–210
    - fiscal calendar, multiple, 208
    - hierarchies, 209
    - journal entries, 206–207
    - period close, 204–206
    - periodic snapshot, 203
    - year-to-date facts, 206
- hierarchies
  - fixed depth, 214
  - modifying, ragged, 221
  - ragged, alternative modeling approaches, 221–223
  - ragged, bridge table approach, 223
  - ragged, modifying, 220–221
  - ragged, shared ownership, 219
  - ragged, time varying, 220
  - ragged, variable depth, 215–217
  - variable depth, 214–215
- OLAP and, 226
- accumulating grain fact tables, 12
- accumulating snapshots, 44, 118–119, 194–196
  - claims (insurance case study), 393
  - complex workflows, 393–394
  - timespan accumulating snapshot, 394–395
- ETL systems, 475
- fact tables, 121, 326–329
  - complementary fact tables, 122
  - milestones, 121
  - OLAP cubes, 121–122
  - updates, 121–122
- healthcare case study, 343
- policy (insurance case study), 384–385
- type 2 dimensions and, 196
- activity-based costing measures, 184
- additive facts, 11, 42
- add mini dimension and type 1 outrigger (SCD type 5), 55
- add mini-dimension (SCD type 4), 55
  - multiple, 156–159
- add new attribute (SCD type 3), 55, 154–155
  - multiple, 156
- add new row (SCD type 2), 54, 150–152
  - effective date, 152–153
  - expiration date, 152–153
  - type 1 in same dimension, 153
- addresses
  - ASCII, 236
  - CRM and, customer dimension, 233–238
  - Unicode, 236–238
- add type 1 attributes to type 2 dimension (SCD type 6), 56
- admissions events (education case study), 330
- aggregate builder, ETL system, 481
- aggregated facts
  - as attributes, 64
  - CRM and, customer dimension, 239–240

- aggregate fact tables, 45
  - clickstream data, 366–367
- aggregate OLAP cubes, 8, 45
- aggregate tables, ETL system development, 519
- agile development, 34–35
  - conformed dimensions and, 137–138
- airline case study, 311
  - bus matrix, 311–315
  - calendars as outriggers, 321–323
  - class of service flown dimension, 319–320
  - destination airport dimension, 320–321
  - fact tables, granularity, 312–316
  - origin dimension, 320–321
  - passenger dimension, 314
  - sales channel dimension, 315
  - segments, linking to trips, 315–316
  - time zones, multiple, 323
- aliasing, 171
- allocated facts, 60
- allocating, 184–186
- allocations, profit and loss fact tables, 60
- ALTER TABLE command, 17
- analytics
  - big data management, 531
  - GA (Google Analytics), 367
  - in-database, big data and, 537
- analytic solutions, packaged, 270–271
- AND queries, skill keywords bridge, 275
- architecture
  - big data best practices
    - backflow, 535–536
    - boundary crashes, 536
    - compute resources, 537
    - data highway planning, 533–534
    - data quality planning, 535
    - data value, 535
    - ecosystems, 534
    - fact extractor, 534
    - in-database analytics, 537
    - performance improvements, 537
    - prototypes, 536
    - streaming data, 536
  - DW/BI alternatives, 26–29
  - enterprise data warehouse bus architecture, 22, 123–125
  - hub-and-spoke CIF architecture, 28–29
  - hybrid hub-and-spoke Kimball architecture, 29
  - independent data mart architecture, 26–27
  - MapReduce/Hadoop, 530
  - RDBMS, extension, 529–530
  - real-time processing, 522–524
- archiving, 447–448, 485–486
- artificial keys, 98

- ASCII (American Standard Code for Information Interchange), 236
- atomic grain data, 17, 74
- attributes
  - aggregated facts as, 64
  - bridge tables, CRM and, 247
  - changes, 514
  - detailed dimension model, 437
  - expiration, 266
  - flags, 48
  - indicators, 48
  - null, 48, 92
  - numeric values as, 59
  - pathstring, ragged/variable depth hierarchies, 57
  - product dimensions, 132
  - SCD type 3 (add new attribute), 154–155
    - multiple, 156
- audit columns, CDC (change data capture), 452
- audit dimensions, 66, 192–193, 284, 495
  - assembler, 460
  - insurance case study, 383
  - key assignment, 511–512
- automation, ETL system development
  - errors, 520
  - exceptions, 520
  - job scheduling, 520

## B

- backflow, big data and, 535–536
- backups, 495
- backup system, ETL systems, 485
  - archiving, 485–486
  - compliance manager, 493–495
  - dependency, 490–491
  - high performance, 485
  - lights-out operations, 485
  - lineage, 490–491
  - metadata repository, 495
  - parallelizing/pipelining system, 492
  - problem escalation system, 491–492
  - recovery and restart system, 486–488
  - retrieval, 485–486
  - security system, 492–493
  - simple administration, 485
  - sorting system, 490
  - version control system, 488
  - version migration system, 488
  - workflow monitor, 489–490
- banking case study, 282
  - bus matrix, 282–296
- dimensions
  - household, 286–287

- mini-dimensions, 289–291
  - multivalued, weighting, 287–289
  - too few, 283–286
- facts, value banding, 291–292
- heterogeneous products, 293–295
- hot swappable dimensions, 296
- user perspective, 293
- behavior
  - customers, CRM and, 249–251
  - sequential, step dimension and, 251–252
  - study groups, 64, 249
- behavior tags
  - facts, 241
  - time series, 63, 240–242
- BI application design/development (Lifecycle), 408, 423–424
- BI applications, 22
- BI (business intelligence) delivery interfaces, 448
- big data
  - architecture best practices
    - backflow, 535–536
    - boundary crashes, 536
    - compute resources, 537
    - data highway planning, 533–534
    - data quality planning, 535
    - data value, 535
    - ecosystems, 534
    - fact extractor, 534
    - in-database analytics, 537
    - performance improvements, 537
    - prototypes, 536
    - streaming data, 536
  - data governance best practices, 541
  - dimensionalizing and, 541
  - privacy, 541–542
  - data modeling best practices
    - data structure declaration, 540
    - data virtualization, 540
    - dimension anchoring, 539
    - integrating sources and confined dimensions, 538
    - name-value pairs, 540
    - SCDs (slowly changing dimensions), 539
    - structured/unstructured data integration, 539
    - thinking dimensionally, 538
  - management best practices
    - analytics and, 531
    - legacy environments and, 532
    - sandbox results and, 532–533
    - sunsetting and, 533
  - overview, 527–529
- blobs, 530
- boundary crashes, big data and, 536
- bridge tables
  - customer contacts, CRM and, 248
  - mini-dimensions, 290–291
  - multivalued
    - CRM and, 245–246
    - time varying, 63
  - multivalued dimensions, 63, 477–478
  - ragged hierarchies and, 223
  - ragged/variable depth hierarchies, 57
  - sparse attributes, CRM and, 247
- bubble chart, dimension modeling and, 435–436
- budget fact table, 210
- budgeting process, 210–213
- bus architecture, 124–125
  - enterprise data warehouse bus architecture, 52
- business analyst, 408
- Business Dimensional Lifecycle, 404
- business-driven governance, 136–137
- business driver, 408
- business initiatives, 70
- business lead, 408
- business motivation, Lifecycle planning, 407
- business processes
  - characteristics, 70–71
  - dimensional modeling, 39, 300
  - retail sales case study, 74
  - value chain, 111–112
- business representatives, dimensional modeling, 431–432
- business requirements
  - dimensional modeling, 432
  - Lifecycle, 405, 410
    - documentation, 414
    - forum selection, 410–411
    - interviews, 412–414
    - launch, 412
    - prioritization, 414–415
    - representatives, 411–412
    - team, 411
- business rule screens, 458
- business sponsor, 408
  - Lifecycle planning, 406
- business users, 408
  - perspectives, 293
- bus matrix
  - accounting, 202
  - airline, 311
  - banking, 282
  - detailed implementation bus matrix, 53
  - dimensional modeling and, 439
  - enterprise data warehouse bus matrix, 52
  - healthcare case study, 339–340
  - HR (human resources), 268–269
  - insurance, 378–389
    - detailed implementation, 390

- inventory, 113–119
- opportunity/stakeholder matrix, 127
- order management, 168
- procurement, 142–147
- telecommunications, 297–299
- university, 325–326
- web retailers, clickstream integration, 368–370

## C

- calculation lag, 196–197
- calendar date dimensions, 48
- calendars, country-specific as outriggers, 321–323
- cannibalization, 90
- cargo shipper schema, 317
- case studies
  - accounting, 202
  - budgeting, 210–213
  - consolidated fact tables, 224–225
  - G/L (general ledger), 203–210
  - hierarchies, 214–223
  - OLAP and, 226
- airline, 311
  - calendars as outriggers, 321–323
  - class of service flown dimension, 319–320
  - destination airport dimension, 320–321
  - fact table granularity, 312–316
  - origin dimension, 320–321
  - passenger dimension, 314
  - sales channel dimension, 315
  - time zones, multiple, 323
- CRM (customer relationship management)
  - analytic, 231–233
  - bridge tables, 245–248
  - complex customer behavior, 249–251
  - customer data integration, 256–260
  - customer dimension and, 233–245
  - fact tables, abnormal scenario indicators, 255–256
  - fact tables, satisfaction indicators, 254–255
  - fact tables, timespan, 252–254
  - low latency data, 260–261
  - operational, 231–233
  - step dimension, sequential behavior, 251–252
- education, 325–326
  - accumulating snapshot fact table, 326–329
  - additional uses, 336
  - admissions events, 330
  - applicant pipeline, 326–329
  - attendance, 335
  - change tracking, 330
  - course registrations, 330–333
  - facility use, 334
  - instructors, multiple, 333
  - metrics, artificial count, 331–332
  - research grant proposal, 329
  - student dimensions, 330
  - term dimensions, 330
- electronic commerce
  - clickstream data, 353–370
  - profitability, sales transactions and, 370–372
- financial services, 282, 287–295
  - dimensions, household, 286–287
  - dimensions, too few, 283–286
- healthcare, 339–340
  - billing, 342–344
  - claims, 342–344
  - date dimension, 345
  - diagnosis dimension, 345–347
  - EMRs (electronic medical records), 341–348
  - HCPSC (Healthcare Common Procedure Coding System), 342
  - HIPAA (Health Insurance Portability and Accountability Act), 341
  - ICD (International Classification of Diseases), 342
  - images, 350
  - inventory, 351
  - measure type dimension, 349–350
  - payments, 342–344
  - retroactive changes, 351–352
  - subtypes, 347–348
  - supertypes, 347–348
  - text comments, 350
- HR (Human Resources Management)
  - bus matrix, 268
  - employee hierarchies, 271–272
  - employee profiles, 263–267
  - hierarchies, 273–274
  - managers key, 272–273
  - packaged data models, 270–271
  - periodic snapshots, 267–268
  - skill keywords, 274–277
  - survey questionnaire, 277–278
- insurance, 375–377
  - accident events factless fact table, 396
  - accumulating snapshot, 384–385
  - bus matrix, 378, 389–390
  - claim transactions, 390–396
  - conformed dimensions, 386
  - conformed facts, 386
  - degenerate dimension, 383
  - dimensions, 380

- dimensions, audit, 383
- dimensions, low cardinality, 383
- dimensions, multivalued, 388
- junk dimensions, 392
- mini-dimensions, 381–382
- multivalued dimensions, 382
- NAICS (North American Industry Classification System), 382
- numeric attributes, 382
- pay-in-advance facts, 386–387
- periodic snapshot, 385
- policy transaction fact table, 383
- policy transactions, 379–380
- premiums, periodic snapshot, 386–388
- SCDs (slowly changing dimensions), 380–381
- SIC (Standard Industry Classification), 382
- supertype/subtype products, 384, 387
- timespan accumulating snapshot, 394
- value chain, 377–378
- inventory
  - accumulating snapshot, 118–119
  - fact tables, 115–116
  - periodic snapshot, 112–114
  - semi-additive facts, 114–115
  - transactions, 116–118
- order management, 167
  - accumulating snapshots, 194–196
  - audit dimension, 192–193
  - customer dimension, 174–175
  - deal dimension, 177–179
  - header/line pattern, 186
  - header/line patterns, 181–182
  - invoice transactions, 187
  - junk dimensions, 179–180
  - lag calculations, 196
  - multiple currencies, 182–184
  - product dimension, 172–173
  - profit and loss facts, 189–191
  - transaction granularity, 184–186
  - transactions, 168–171
  - units of measure, multiple, 197–198
- procurement, 141–142
  - bus matrix, 142–143
  - complementary procurement snapshot
    - fact table, 147
  - transactions, 142–145
- retail sales, 72–73
  - business process selection, 74
  - dimensions, selecting, 76
  - facts, 76–77
  - facts, derived, 77–78
  - facts, non-additive, 78
  - fact tables, 79
  - frequent shopper program, 96
  - grain declaration, 74–75
  - POS schema, 94
  - retail schema extensibility, 95–97
  - telecommunications, 297–299
- causal dimension, 89–90, 284
- CDC (change data capture)
  - ETL system, 451
  - audit columns, 452
  - diff compare, 452
  - log scraping, 453
  - message queue monitoring, 453
  - timed extracts, 452
- centipede fact tables, 58, 108–109
- change reasons, 266–267
- change tracking, 147–148
  - education case study, 330
  - HR (human resources) case study, embedded managers key, 272–273
  - SCDs, 148
- chart of accounts (G/L), 203–204
  - uniform chart of accounts, 204
- checkpoints, data quality, 516
- CIF (Corporate Information Factory), 28–29
- CIO (chief information officer), 377
- claim transactions (insurance case study), 390
  - claim accumulating snapshot, 393–394
  - junk dimensions and, 392
  - periodic snapshot, 395–396
  - timespan accumulating snapshot, 394–395
- class of service flown dimension (airline case study), 319–320
- cleaning and conforming, ETL systems, 450
  - audit dimension assembler, 460
  - conforming system, 461–463
  - data cleansing system, 456
    - quality event responses, 458
    - quality screens, 457–458
  - data quality improvement, 455–456
  - deduplication system, 460–461
  - error event schema, 458–460
- clickstream data, 353–354
- dimensional models, 357–358
  - aggregate fact tables, 366–367
  - customer, 361–362
  - date, 361–362
  - event dimension, 359
  - GA (Google Analytics), 367
  - page dimension, 358–359
  - page event fact table, 363–366
  - referral dimension, 360
  - session dimension, 359–360
  - session fact table, 361–363
  - step dimension, 366
  - time, 361–362
- session IDs, 355–356

- visitor identification, 356–357
- visitor origins, 354–355
- web retailer bus matrix integration, 368–370
- collaborative design workshops, 38
- column screens, 457
- comments, survey questionnaire (HR), 278
- common dimensions, 130
- compliance, ETL system, 445
- compliance manager, ETL system, 493–495
- composite keys, 12
- computer resources, big data and, 537
- conformed dimensions, 51, 130, 304
  - agile movement and, 137–138
  - drill across, 130–131
  - grain, 132
  - identical, 131–132
  - insurance case study, 386
  - limited conformity, 135
  - shrunk on bus matrix, 134
  - shrunk rollup dimensions, 132
  - shrunk with row subset, 132–134
- conformed facts, 42, 139
  - insurance case study, 386
  - inventory case study, 138–139
- conforming system, ETL system, 461–463
- consistency
  - adaptability, 4
  - goals, 3
- consolidated fact tables, 45
  - accounting case study, 224–225
- contacts, bridge tables, 248
- contribution amount (P&L statement), 191
- correctly weighted reports, 288
- cost, activity-based costing measures, 184
- COUNT DISTINCT, 243
- country-specific calendars as outriggers, 321–323
- course registrations (education case study), 330
- CRM (customer relationship management), 229
  - analytic, 231–233
  - bridge tables
    - customer contacts, 248
    - multivalued, 245–246
    - sparse attributes, 247
  - complex customer behavior, 249–251
  - customer data integration, 256
    - multiple customer dimension conformity, 258–259
    - single customer dimension, 256–258
  - customer dimension and, 233
    - addresses, 233–236
    - addresses, international, 236–238
    - counts with Type 2, 243
  - dates, 238
  - facts, aggregated, 239–240
  - hierarchies, 244–245
  - names, 233–236
  - names, international, 236–238
  - outriggers, low cardinality attribute set and, 243–244
  - scores, 240–243
  - segmentation, 240–243
- factless fact tables, 176
- hierarchies, 174–175
- multiple, partial conformity, 258–259
- single, 256–258
- single versus multiple dimension tables, 175–176
- customer matching, 257
- customer relationship management. *See* CRM, 230
- currency, multiple
  - fact tables, 60
  - G/L (general ledger), 206
  - order transactions, 182–184
- current date attributes, dimension tables, 82–83
- customer contacts, bridge tables, 248
- customer dimension, 158, 174–175
- clickstream data, 361–362
- CRM and, 233
  - addresses, 233–236
  - addresses, international, 236–238
  - counts with Type 2, 243
  - dates, 238
  - facts, aggregated, 239–240
  - hierarchies, 244–245
  - names, 233–236
  - names, international, 236–238
  - outriggers, low cardinality attribute set and, 243–244
  - scores, 240–243
  - segmentation, 240–243
- data architect/modeler, 409
- data bags, 530
- database administrator, 409

## D

- data cleansing system, ETL system, 456
  - quality event responses, 458
  - quality screens, 457–458
- data compression, ETL system, 454
- data governance, 135–136
  - big data best practices, 541
  - dimensionalizing, 541
  - privacy, 541–542
  - business-driven governance, 136–137
  - objectives, 137
- data handlers, late arriving, 478–479
- data highway planning, 533–534
- data integration
  - conformed dimensions, 130–138
  - CRM and, 256
    - multiple customer dimension conformity, 258–259
    - single customer dimension, 256–258
  - ETL system, 444–446
  - MDM (master data management), 256
  - structure/unstructured data, 539
  - value chain integration, 111–112
- data latency, ETL system, 447
- data mart, independent data mart
  - architecture, 26–27
- data mining
  - DW/BI system and, 242–243
  - null tracking, 92
- data modeling, big data best practices
- data structure declaration, 540
- data virtualization, 540
- dimension anchoring, 539
- integrating sources and conformed
  - dimensions, 538
- name-value pairs, 540
- SCDs (slowly changing dimensions), 539
- structured/unstructured data integration, 539
  - thinking dimensionally, 538
- data models, packaged, 270–271
- data profiling
  - ETL system, 450–451
  - tools, 433
- data propagation, ETL system, 482
- data quality
  - checkpoints, 516
  - ETL system, 445
  - improvement, 455–456
  - planning, big data and, 535
- data steward, 408
- data structure, analysis time, 540
- data value, big data and, 535
- data virtualization, big data and, 540
- data warehousing versus operational
  - processing, 2
- date dimension, 79–81, 284, 302
  - calendar date, 48
  - clickstream data, 361–362
  - current date attributes, 82–83
  - fixed time series buckets and, 302–303
  - healthcare case study, 345
  - populating, 508
  - relative date attributes, 82–83
  - role playing, 171
  - smart keys, 101–102
  - textual attributes, 82
  - time-of-day, 83
- dates
  - CRM and, customer dimension, 238
  - dimension tables, 89
  - timespan fact tables, 252–254
  - transaction fact table, 170–171
    - foreign key, 170
    - role playing, 171
- date/time
  - GMT (Greenwich Mean Time), 323
  - time zones, multiple, 323
  - UTC (Coordinated Universal Time), 323
- date/time dimensions, 470
- date/time stamp dimensions, 284
- deal dimensions, 177–178
- decision-making goals, 4
- decodes, dimensions, 303–304
- decoding production codes, 504
- deduplication system, 460–461
- degenerate dimension, 47, 284, 303
  - insurance case study, 383
  - order number, 178–179
  - retail sales case study, 93–94
  - surrogate keys, 101
  - telecommunications case study, 303
  - transaction numbers, 93–94
- demand planning, 142
- demographics dimension, 291
  - size, 159
- denormalized flattened dimensions, 47
- dependency analysis, 495
- dependency, ETL, 490–491
- deployment
  - Lifecycle, 424
  - OLAP, 9
- derived facts, 77–78
- descriptions, dimensions, 303–304
- descriptive context, dimensions for, 40
- destination airport dimension (airline case study), 320–321
- detailed implementation bus matrix, 53, 390
- detailed table design documentation, 437–439
- diagnosis dimension (healthcare case study), 345–347

- diff compare, CDC (change data capture), 452
- dimensional modeling, 7
  - 3NF (third normal form) models, 7–8
  - 4-step design process
    - business process, 70–71
    - dimensions, 72
    - facts, 72
    - grain, 71
  - atomic grain data, 17
  - benefits of thinking dimensionally, 32–33
  - business processes, 300
  - business representatives, 431–432
  - calendar coordination, 433–434
  - clickstream data, 357–367
  - data profiling tools, 433
  - design
    - bubble chart, 435–436
    - detailed model development, 436–439
    - documentation finalization, 441
    - validation, 440–441
  - dimension tables, 13
    - attributes, 13–14
    - hierarchical relationships, 15
    - snowflaking, 15
  - extensibility, 16
  - facts
    - additive facts, 11
    - composite keys, 12
    - FK (foreign keys), 12
    - grains, 10
    - numeric facts, 11
    - textual facts, 12
  - fact tables, 10–12
    - grain categories, 12
  - fundamentals
    - business processes, 39
    - business requirement gathering, 37–38
    - collaborative workshops, 38
    - data realities gathering, 37–38
    - descriptive context, 40
    - facts, 40
    - four-step dimensional design process, 38
    - grain, 39
    - model extensions, 41
    - star schemas, 40
  - Lifecycle data track, 420
  - mistakes to avoid, 397–401
  - myths, 30
    - departmental versus enterprise, 31
    - integration, 32
    - predictable use, 31–32
    - scalability, 31
    - summary data, 30
  - naming conventions, 433
  - OLAP (online analytical processing) cube, 8
    - deployment considerations, 9
    - overview, 429–431
    - participant identification, 431–432
    - reports, 17
    - simplicity in, 16
    - sources, 300
    - star schemas, 8
    - terminology, 15
    - tools, 432
  - dimensional thinking, big data and, 538
  - dimension manager system, 479–480
  - dimensions
    - anchoring, big data and, 539
    - attributes, 514
      - aggregated facts as, 64
      - bridge tables, CRM and, 247
      - changes, 514
      - detailed dimension model, 437
      - expiration, 266
      - flags, 48
      - indicators, 48
      - null, 48, 92
      - numeric values as, 59
      - pathstring, ragged/variable depth
        - hierarchies, 57
      - product dimensions, 132
      - SCD type 3 (add new attribute), 154–156
      - See also* attributes, 48
    - audit dimension, 66, 192–193, 284
      - assembler, 460
      - insurance case study, 383
    - average number in model, 284
    - calendar date, 48
    - causal, 89–90, 284
    - change reasons, 266–267
    - class of service flown (airline case study), 319–320
    - conformed, 51, 130, 304
      - agile movement and, 137–138
      - drill across, 130–131
      - grain, 132
      - identical, 131–132
      - insurance case study, 386
      - limited conformity, 135
      - shrunk, bus matrix and, 134
      - shrunk rollup dimensions, 132
      - shrunk with row subset, 132–134
    - customer dimension, 158, 174–175
      - conformity, 258–259
      - CRM and, 233–245
      - factless fact tables, 176
      - hierarchies, 174–175
      - single, 256–258
      - single versus multiple dimension tables, 175–176
    - data governance, big data and, 541
    - date dimension, 48, 284, 302

- fixed time series buckets and, 302–303
- healthcare case study, 345
- populating, 508
- role playing, 171
- date/time stamp, 284
- deal dimension, 177–178
- decodes, 303–304
- degenerate, 47, 284, 303
  - order number, 178–179
- demographic, 291
  - size, 159
- denormalized flattened, 47
- descriptions, 303–304
- destination airport (airline case study), 320–321
- detailed dimension model, 437
- diagnosis (healthcare case study), 345–347
- dimensional design models, 72
- drilling across, 51
- event dimension, clickstream data, 359
- generic, abstract, 66
- geographic location, 310
- granularity, hierarchies and, 301–302
- hierarchies
  - fixed depth position hierarchies, 56
  - ragged/variable depth with hierarchy
    - bridge tables, 57
  - ragged/variable depth with pathstring
    - attributes, 57
  - slightly ragged/variable depth, 57
- hot swappable, 66, 296
- household, 286–287
- insurance case study, 380
  - degenerate dimension, 383
  - mini-dimensions, 381–382
  - multivalued dimensions, 382
  - numeric attributes, 382
  - SCDs (slowly changing dimensions), 380–381
- junk dimensions, 49, 179–180, 284
- keys, natural, 162
- late arriving, 67
- low cardinality, insurance case study, 383
- measure type, 65
  - healthcare case study, 349–350
- mini-dimensions, 289–290
  - bridge tables, 290–291
  - insurance case study, 381–382
  - type 5 SCD and, 160
- multivalued
  - bridge table builder, 477–478
  - bridge tables and, 63
  - insurance case study, 382–388
  - weighting, 287–289
- origin (airline case study), 320–321
- outrigger, 50
- page dimension, clickstream data, 358–359
- passenger (airline case study), 314
- product dimension
  - characteristics, 172–173
  - operational product master, 173
  - order transactions, 172–173
- rapidly changing monster dimension, 55
- referral dimension, clickstream data, 360
- retail sales case study, 76
- role-playing, 284
- sales channel, airline case study, 315
- service level performance, 188–189
- session dimension, clickstream data, 359–360
- shrunk, 51
- shrunk rollup, 132
- special dimensions manager, ETL systems, 470
  - date/time dimensions, 470
  - junk dimensions, 470
  - mini-dimensions, 471
  - shrunk subset, 472
  - static, 472
  - user-maintained, 472–473
- static dimension, population, 508
- status, 284
- step dimension, 65
  - clickstream data, 366
  - sequential behavior, 251–252
- student (education case study), 330
- term (education case study), 330
- text comments, 65
- too few, 283–286
- transaction profile dimension, 49, 179
- transformations
  - combine from separate sources, 504
  - decode production codes, 504
  - relationship validation, 504–505
  - simple data, 504
  - surrogate key assignment, 506
  - value chain, 52
- dimension surrogate keys, 46
- dimension tables, 13
  - attributes, 13–14
  - calendar date dimensions, 48
  - changed rows, 513–514
  - date dimension, 79–81
    - current date attributes, 82–83
    - smart keys, 101–102
    - textual attributes, 82
    - time-of-day, 83
  - dates, 89
  - degenerate dimensions, 47
    - surrogate keys, 101
  - transaction numbers, 93–94

- denormalized flattened dimensions, 47
- drilling down, 47
- durable keys, 46
- extracts, 513
- fact tables, centipede, 108–109
- flags, 48, 82
- hierarchical relationships, 15
- hierarchies, multiple, 48, 88–89
- historic data population, 503–506
- holiday indicator, 82
- indicators, 48, 82
- junk dimensions, 49
- loading, 506–507
- loading history, 507–508
- natural keys, 46, 98–101
- new rows, 513–514
- null attributes, 48
- outrigger dimensions, 50
- outriggers, 106–107
- product dimension, 83–84
  - attributes with embedded meaning, 85
  - drilling down, 86–87
  - many-to-one hierarchies, 84–85
  - numeric values, 85–86
- promotion dimension, 89–91
  - null items, 92
- role-playing, 49
- snowflaking, 15, 50, 104–106
- store dimension, 87–89
- structure, 46
- supernatural keys, 46, 101
- surrogate keys, 46, 98–100
- transaction profile dimensions, 49
- weekday indicator, 82
- dimension terminology, 15
- dimension-to-dimension table joins, 62
- documentation
  - detailed table design, 437–439
  - dimensional modeling, 441
  - ETL development, 502–503
    - sandbox source system, 503
  - Lifecycle architecture requirements, 417
  - Lifecycle business requirements, 414
- draft design
  - exercise discussion, 306–308
  - remodeling existing structures, 309
- drill across, 51, 130–131
- drill down, 47, 86–87
  - ETL development, 500
    - hierarchies, 501
    - table schematics, 501
  - G/L (general ledger) hierarchy, 209
  - management hierarchies, 273–274
- dual date/time stamps, 254
- dual type 1 and type 2 dimensions (SCD type 7), 56
- duplication, deduplication system, 460–461
- durable keys, 46
  - supernatural keys, 101
- DW/BI, 1
  - alternative architecture, 26–29
  - data mining and, 242–243
  - goals, 3
  - international goals, 237–238
  - Kimball architecture, 18
    - BI applications, 22
    - ETL (extract, transformation, and load) system, 19–21
    - hybrid hub-and-spoke Kimball, 29
    - operational source systems, 18
    - presentation area, 21–22
    - restaurant metaphor, 23–26
  - publishing metaphor for DW/BI managers, 5–7
  - system users, 2
- dynamic value bands, 64, 291
- E**
  - ecosystems, big data and, 534
    - case study, 325–326
  - education
    - accumulating snapshot fact table, 326–329
    - additional uses, 336
    - admissions events, 330
    - applicant pipeline, 326–329
    - attendance, 335
    - bus matrix, 325–326
    - change tracking, 330
    - course registrations, 330–333
    - facility use, 334
    - instructors, multiple, 333
    - metrics, artificial count, 331–332
    - research grant proposal, 329
    - student dimension, 330–332
    - term dimension, 330
  - effective date, SCD type 2, 152–153
  - EHR (electronic health record), 341
  - electronic commerce case study, 353–372
  - embedded managers key (HR), 272–273
  - embedding attribute meaning, 85
  - employee hierarchies, recursive, 271–272
  - employee profiles, 263–265
    - dimension change reasons, 266–267
  - effective time, 265–266
  - expiration, 265–266
  - fact events, 267
    - type 2 attributes, 267
  - EMRs (electronic medical records),
    - healthcare case study, 341, 348

- enterprise data warehouse bus architecture, 22, 52, 123–125
- enterprise data warehouse bus matrix, 52, 125–126
  - columns, 126
    - hierarchy levels, 129
  - common mistakes, 128–129
  - opportunity/stakeholder matrix, 127
  - procurement, 142–143
  - retrofitting existing models, 129–130
  - rows
    - narrowly defined, 128
    - overly encompassing, 128
    - overly generalized, 129
  - shrunk conformed dimensions, 134
  - uses, 126–127
- ERDs (entity-relationship diagrams), 8
- error event schema, ETL system, 458–460
- error event schemas, 68
- ETL (extract, transformation, and load)
  - system, 19–21, 443
    - archiving, 447–448
    - BI, delivery, 448
    - business needs, 444
    - cleaning and conforming, 450
      - audit dimension assembler, 460
      - conforming system, 461–463
      - data cleansing system, 456–458
      - data quality, improvement, 455–456
      - deduplication system, 460–461
      - error event schema, 458–460
    - compliance, 445
    - data integration, 446
    - data latency, 447
    - data propagation manager, 482
    - data quality, 445
    - delivering, 450, 463
      - aggregate builder, 481
      - dimension manager system, 479–480
      - fact provider system, 480–481
      - fact table builders, 473–475
      - hierarchy manager, 470
      - late arriving data handler, 478–479
      - multivalued dimension bridge table builder, 477–478
      - SCD manager, 464–468
      - special dimensions manager, 470–473
      - surrogate key generator, 469–470
      - surrogate key pipeline, 475–477
  - design, 443
    - Lifecycle data track, 422
  - developer, 409
  - development, 498
    - activities, 500
    - aggregate tables, 519
    - default strategies, 500
    - drill down, 500–501
    - high-level plan, 498
    - incremental processing, 512–519
    - OLAP loads, 519
    - one-time historic load data, 503–512
    - specification document, 502–503
    - system operation and automation, 520
    - tools, 499
  - ETL architect/designer, 409
  - extracting, 450
    - CDC (change data capture), 451–453
      - data profiling, 450–451
      - extract system, 453–455
    - legacy licenses, 449
    - lineage, 447–448
    - managing, 450, 483
      - backup system, 485–495
      - job scheduler, 483–484
    - OLAP cube builder, 481–482
    - process overview, 497
    - security, 446
    - skills, 448
    - subsystems, 449
  - event dimension, clickstream data, 359
  - expiration date, type 2 SCD, 152–153
  - extended allowance amount (P&L statement), 190
  - extended discount amount (P&L statement), 190
  - extended distribution cost (P&L statement), 191
  - extended fixed manufacturing cost (P&L statement), 190
  - extended gross amount (P&L statement), 189
  - extended net amount (P&L statement), 190
  - extended storage cost (P&L statement), 191
  - extended variable manufacturing cost (P&L statement), 190
  - extensibility in dimensional modeling, 16
  - extracting, ETL systems, 450
    - CDC (change data capture), 451
      - audit columns, 452
      - diff compare, 452
      - log scraping, 453
      - message queue monitoring, 453
      - timed extracts, 452
    - data profiling, 450–451
    - extract system, 453–455
  - extraction, 19
  - extract system, ETL system, 453–455
- F**
  - fact extractors, 530
    - big data and, 534

- factless fact tables, 44, 97–98, 176
  - accidents (insurance case study), 396
  - admissions (education case study), 330
  - attendance (education case study), 335
  - course registration (education case study), 330–333
  - facility use (education case study), 334
  - order management case study, 176
- fact provider system
  - ETL system, 480–481
- facts, 10, 12, 72, 79
  - abnormal scenario indicators, 255–256
  - accumulating snapshots, 44, 121–122, 326–329
  - additive facts, 11, 42
  - aggregate, 45
    - as attributes, 64
    - clickstream data, 366–367
    - CRM and customer dimension, 239–240
  - allocated facts, 60
  - allocating, 184–186
  - behavior tags, 241
  - budget, 210
  - builders, ETL systems, 473–475
  - centipede, 58, 108–109
  - compliance-enabled, 494
  - composite keys, 12
  - conformed, 42, 138–139
  - consolidated, 45
  - currency, multiple, 60
  - derived, 77–78
  - detailed dimension model, 437
  - dimensional modeling process and, 40
  - drill across, 130–131
  - employee profiles, 267
  - enhanced, 115–116
  - FK (foreign keys), 12
  - grains, 10, 12
  - granularity, airline bus matrix, 312–315
  - header/line fact tables, 59
  - historic, 508
  - incremental processing, 515, 519
  - invoice, 187–188
  - joins, avoiding, 259–260
  - lag/duration facts, 59
  - late arriving, 62
  - loading, 512
  - mini-dimension demographics key, 158
  - multiple units of measure, 61
  - non-additive, 42, 78
  - normalization, order transactions, 169–170
  - null, 42, 92
  - numeric facts, 11
  - numeric values, 59, 85–86
  - page event, clickstream data, 363–366
  - partitioning, smart keys, 102
  - pay-in-advance, insurance case study, 386–387
  - periodic snapshots, 43, 120–122
  - policy transactions (insurance case study), 383
  - profitability, 370–372
  - profit and loss, 189–192
  - profit and loss, allocations and, 60
  - real-time, 68
  - referential integrity, 12
  - reports, 17
  - retail sales case study, identifying, 76–79
  - satisfaction indicators, 254–255
  - semi-additive, 42, 114–115
  - service level performance, 188–189
  - session, clickstream data, 361–363
  - set difference, 97
  - shrunk rollup dimensions, 132
  - single granularity and, 301
  - snapshot, complementary procurement, 147
  - structure, 41–42
  - subtype, 67, 293–295
  - supertype, 67, 293–295
  - surrogate keys, 58, 102–103
  - textual facts, 12
  - terminology, 15
  - time-of-day, 83
  - timespan, 252–254
  - timespan tracking, 62
  - transactions, 43, 120
    - dates, 170–171
    - single versus multiple, 143–145
  - transformations, 509–512
  - value banding, 291–292
  - year-to-date, 206
  - YTD (year-to-date), 61
- fact-to-fact joins, avoiding with multipass SQL, 61
- feasibility in Lifecycle planning, 407
- financial services case study, 281
  - bus matrix, 282
  - dimensions
    - hot-swappable, 296
    - household, 286–287
    - mini-dimensions, 289–291
    - multivalued, weighting, 287–289
    - too few, 283–286
  - facts, value banding, 291–292
  - heterogeneous products, 293–295
  - OLAP, 226
  - user perspective, 293
- financial statements (G/L), 209–210
- fiscal calendar, G/L (general ledger), 208
- fixed depth position hierarchies, 56, 214
- fixed time series buckets, date dimensions and, 302–303

FK (foreign keys). *See* foreign keys (FK), 12

flags

- as textual attributes, 48
- dimension tables, 82
- junk dimensions and, 179–180

flattened dimensions, denormalized, 47

flexible access to information, 407

foreign keys (FK)

- demographics dimensions, 291
- fact tables, 12
- managers employee key as, 271–272
- mini-dimension keys, 158
- null, 92
- order transactions, 170
- referential integrity, 12

forum, Lifecycle business requirements, 410–411

frequent shopper program, retail sales

- schema, 96

FROM clause, 18

## G

GA (Google Analytics), 367

general ledger. *See* G/L (general ledger), 203

generic dimensions, abstract, 66

geographic location dimension, 310

G/L (general ledger), 203

- chart of accounts, 203–204
- currencies, multiple, 206
- financial statements, 209–210
- fiscal calendar, multiple, 208
- hierarchies, drill down, 209
- journal entries, 206–207
- period close, 204–206
- periodic snapshot, 203
- year-to-date facts, 206

GMT (Greenwich Mean Time), 323

goals of DW/BI, 3–4

Google Analytics (GA), 367

governance

- business-driven, 136–137
- objectives, 137

grain, 39

- accumulating snapshots, 44
- atomic grain data, 74
- budget fact table, 210
- conformed dimensions, 132
- declaration, 71
  - retail sales case study, 74–75
- dimensions, hierarchies and, 301–302
- fact tables, 10
  - accumulating snapshot, 12
  - periodic snapshot, 12
  - transaction, 12

- periodic snapshots, 43
  - single, facts and, 301
  - transaction fact tables, 43
- granularity, 300
- GROUP BY clause, 18
- growth
  - Lifecycle, 425–426
  - market growth, 90

## H

Hadoop, MapReduce/Hadoop, 530

HCPSC (Healthcare Common Procedure Coding System), 342

HDFS (Hadoop distributed file system), 530

headcount periodic snapshot, 267–268

header/line fact tables, 59

header/line patterns, 181–182, 186

healthcare case study, 339–340

- billing, 342–344
- claims, 342–344
- date dimension, 345
- diagnosis dimension, 345–347
- EMRs (electronic medical records), 341, 348
- HCPSC (Healthcare Common Procedure Coding System), 342
- HIPAA (Health Insurance Portability and Accountability Act), 341
- ICD (International Classification of Diseases), 342
- images, 350
- inventory, 351
- measure type dimension, 349–350
- payments, 342–344
- retroactive changes, 351–352
- subtypes, 347–348
- supertypes, 347–348
- text comments, 350

heterogeneous products, 293–295

hierarchies

- accounting case study, 214–223
- customer dimension, 174–175, 244–245
- dimension granularity, 301–302
- dimension tables, multiple, 88–89
- drill down, ETL development, 501
- employees, 271–272
- ETL systems, 470
- fixed-depth positional hierarchies, 56
- G/L (general ledger), drill down, 209
- management, drilling up/down, 273–274
- many-to-one, 84–85
- matrix columns, 129
- multiple, 48
- nodes, 215

- ragged/variable depth, 57
- slightly ragged/variable depth, 57
- trees, 215–216
- high performance backup, 485
- HIPAA (Health Insurance Portability and Accountability Act), 341
- historic fact tables
  - extracts, 508
  - statistics audit, 508
- historic load data, ETL development, 503–512
  - dimension table population, 503–506
- holiday indicator, 82
- hot response cache, 238
- hot swappable dimensions, 66, 296
- household dimension, 286–287
- HR (human resources) case study, 263
  - bus matrix, 268–269
  - employee profiles, 263–265
    - dimension change reasons, 266–267
    - effective time, 265–266
    - expiration, 265–266
    - fact events, 267
    - type 2 attributes, 267
- hierarchies
  - management, 273–274
  - recursive, 271–272
- managers key
  - as foreign key, 271–272
  - embedded, 272–273
- packaged analytic solutions, 270–271
- packaged data models, 270–271
- periodic snapshots, headcount, 267–268
- skill keywords, 274
  - bridge, 275
  - text string, 276–277
- survey questionnaire, 277
- text comments, 278
- HTTP (Hyper Text Transfer Protocol), 355–356
- hub-and-spoke CIF architecture, 28–29
- hub-and-spoke Kimball hybrid architecture, 29
- human resources management case study. *See* HR (human resources), 263
- hybrid hub-and-spoke Kimball architecture, 29
- hybrid techniques, SCDs, 159, 164
  - SCD type 5 (add mini-dimension and type 1 outrigger), 55, 160
  - SCD type 6 (add type 1 attributes to type 2 dimension), 56, 160–162
  - SCD type 7 (dual type 1 and type 2 dimension), 56, 162–163
- hyperstructured data, 530

## I

- ICD (International Classification of Diseases), 342
- identical conformed dimensions, 131–132
- images, healthcare case study, 350
- impact reports, 288
- incremental processing, ETL system
  - development, 512
  - changed dimension rows, 513–514
  - dimension attribute changes, 514
  - dimension table extracts, 513
  - fact tables, 515–519
  - new dimension rows, 513–514
- in-database analytics, big data and, 537
- independent data mart architecture, 26–27
- indicators
  - abnormal, fact tables, 255–256
  - as textual attributes, 48
  - dimension tables, 82
  - junk dimensions and, 179–180
  - satisfaction, fact tables, 254–255
- Inmon, Bill, 28–29
- insurance case study, 375–377
  - accidents, factless fact tables, 396
  - accumulating snapshot, complementary policy, 384–385
- bus matrix, 378–389
  - detailed implementation, 390
- claim transactions, 390
  - claim accumulating snapshot, 393–394
  - junk dimensions and, 392
  - periodic snapshot, 395–396
  - timespan accumulating snapshot, 394–395
- conformed dimensions, 386
- conformed facts, 386
- dimensions, 380
  - audit, 383
  - degenerate, 383
  - low cardinality, 383
  - mini-dimensions, 381–382
  - multivalued, 382, 388
  - SCDs (slowly changing dimensions), 380–381
- NAICS (North American Industry Classification System), 382
- numeric attributes, 382
- pay-in-advance facts, 386–387
- periodic snapshot, 385
- policy transactions, 379–380, 383
- premiums, periodic snapshot, 386–388
- SIC (Standard Industry Classification), 382
- supertype/subtype products, 384, 387
- value chain, 377–378
- integer keys, 98
  - sequential surrogate keys, 101

- integration
  - conformed dimensions, 130–138
  - customer data, 256
    - customer dimension conformity, 258–259
    - single customer dimension, 256, 257, 258
  - dimensional modeling myths, 32
  - value chain, 122–123
- international names/addresses, customer dimension, 236–238
- interviews, Lifecycle business requirements, 412–413
  - data-centric, 413–414
- inventory case study, 112–114
  - accumulating snapshot, 118–119
  - fact tables, enhanced, 115–116
  - periodic snapshot, 112–114
  - semi-additive facts, 114–115
  - transactions, 116–118
- inventory, healthcare case study, 351
- invoice transaction fact table, 187–188

## J

- job scheduler, ETL systems, 483–484
- job scheduling, ETL operation and automation, 520
- joins
  - dimension-to-dimension table joins, 62
  - fact tables, avoiding, 259–260
  - many-to-one-to-many, 259–260
  - multipass SQL to avoid fact-to-fact joins, 61
- journal entries (G/L), 206–207
- junk dimensions, 49, 179–180, 284
  - airline case study, 320
  - ETL systems, 470
  - insurance case study, 392
  - order management case study, 179–180
- justification for program/project planning, 407

## K

- keys
  - dimension surrogate keys, 46
  - durable, 46
  - foreign, 92, 291
  - managers key (HR), 272–273
  - natural keys, 46, 98–101, 162
  - supernatural keys, 101
  - smart keys, 101–102
  - subtype tables, 294–295
  - supernatural, 46
  - supertype tables, 294–295
  - surrogate, 58, 98–100, 303
    - assigning, 506
    - degenerate dimensions, 101

- ETL system, 475–477
  - fact tables, 102–103
  - generator, 469–470
  - lookup pipelining, 510–511
- keywords, skill keywords, 274
  - bridge, 275
  - text string, 276–277
- Kimball Dimensional Modeling Techniques. *See* dimensional modeling
- Kimball DW/BI architecture, 18
  - BI applications, 22
  - ETL (extract, transformation, and load) system, 19–21
  - hub-and-spoke hybrid, 29
  - presentation area, 21–22
  - restaurant metaphor, 23–26
  - source systems, operational source systems, 18
- Kimball Lifecycle, 404
  - DW/BI initiative and, 404
- KPIs (key performance indicators), 139

## L

- lag calculations, 196–197
- lag/duration facts, 59
- late arriving data handler, ETL system, 478–479
- late arriving dimensions, 67
- late arriving facts, 62
- launch, Lifecycle business requirements, 412
- Law of Too, 407
- legacy environments, big data management, 532
- legacy licenses, ETL system, 449
- Lifecycle
  - BI applications, 406
    - development, 423–424
    - specification, 423
  - business requirements, 405, 410
    - documentation, 414
    - forum selection, 410–411
    - interviews, 412–413
    - interviews, data-centric, 413–414
    - launch, 412
    - prioritization, 414–415
    - representatives, 411–412
    - team, 411
  - data, 405
    - dimensional modeling, 420
    - ETL design/development, 422
    - physical design, 420–422
  - deployment, 424
  - growth, 425–426
  - maintenance, 425–426
  - pitfalls, 426

- products
    - evaluation matrix, 419
    - market research, 419
    - prototypes, 419
  - program/project planning, 405–406
    - business motivation, 407
    - business sponsor, 406
    - development, 409–410
    - feasibility, 407
    - justification, 407
    - planning, 409–410
    - readiness assessment, 406–407
    - scoping, 407
    - staffing, 408–409
  - technical architecture, 405, 416–417
    - implementation phases, 418
    - model creation, 417
    - plan creation, 418
    - requirements, 417
    - requirements collection, 417
    - subsystems, 418
    - task force, 417
  - lift, promotion, 89
  - lights-out operations, backup, 485
  - limited conformed dimensions, 135
  - lineage analysis, 495
  - lineage, ETL system, 447–448, 490–491
  - loading fact tables, incremental, 517
  - localization, 237, 324
  - location, geographic location dimension, 310
  - log scraping, CDC (change data capture), 453
  - low cardinality dimensions, insurance case study, 383
  - low latency data, CRM and, 260–261
- M**
- maintenance, Lifecycle, 425–426
  - management
    - ETL systems, 450, 483
      - backup system, 485–495
      - job scheduler, 483–484
  - management best practices, big data analytics, 531
    - legacy environments, 532
    - sandbox results, 532–533
    - sunsetting and, 533
  - management hierarchies, drilling up/down, 273–274
  - managers, publishing metaphor, 5–7
  - many-to-one hierarchies, 84–85
  - many-to-one relationships, 175–176
  - many-to-one-to-many joins, 259–260
  - MapReduce/Hadoop, 530
  - market growth, 90
  - master dimensions, 130
  - MDM (master data management), 137, 256, 446
  - meaningless keys, 98
  - measurement, multiple, 61
  - measure type dimension, 65
    - healthcare case study, 349–350
  - message queue monitoring, CDC (change data capture), 453
  - metadata coordinator, 409
  - metadata repository, ETL system, 495
  - migration, version migration system, ETL, 488
  - milestones, accumulating snapshots, 121
  - mini-dimension and type 1 outlier (SCD type 5), 160
  - mini-dimensions, 289–290
    - bridge tables, 290–291
    - ETL systems, 471
    - insurance case study, 381–382
    - type 4 SCD, 156–159
  - modeling
    - benefits of thinking dimensionally, 32–33
    - dimensional, 7–12
      - atomic grain data, 17
      - dimension tables, 13–15
      - extensibility, 16
      - myths, 30–32
      - reports, 17
      - simplicity in, 16
      - terminology, 15
  - multipass SQL, avoiding fact-to-fact table joins, 61
  - multiple customer dimension, partial conformity, 258–259
  - multiple units of measure, 61, 197–198
  - multivalued bridge tables
    - CRM and, 245–246
    - time varying, 63
  - multivalued dimensions
    - bridge table builder, 477–478
    - bridge tables and, 63
    - CRM and, 245–247
    - education case study, 325–333
    - financial services case study, 287–289
    - healthcare case study, 345–348
    - HR (human resources) case study, 274–275
    - insurance case study, 382–388
    - weighting factors, 287–289
  - myths about dimensional modeling, 30
    - departmental versus enterprise, 31
    - integration, 32
    - predictable use, 31–32
    - scalability, 31
    - summary data, 30

**N**

- names
  - ASCII, 236
  - CRM and, customer dimension, 233–238
  - Unicode, 236–238
- name-value pairs, 540
- naming conventions, 433
- natural keys, 46, 98–101, 162
  - supernatural keys, 101
- NCOA (national change of address), 257
- nodes (hierarchies), 215
- non-additive facts, 42, 78
- non-natural keys, 98
- normalization, 28, 301
  - facts
    - centipede, 108–109
    - order transactions, 169–170
    - outiggers, 106–107
    - snowflaking, 104–106
- normalized 3NF structures, 8
- null attributes, 48
- null fact values, 509
- null values
  - fact tables, 42
  - foreign keys, 92
- number attributes, insurance case study, 382
- numeric facts, 11
- numeric values
  - as attributes, 59, 85–86
  - as facts, 59, 85–86

**O**

- off-invoice allowance (P&L) statement, 190
- OLAP (online analytical processing) cube, 8, 40
  - accounting case study, 226
  - accumulating snapshots, 121–122
  - aggregate, 45
  - cube builder, ETL system, 481–482
  - deployment considerations, 9
  - employee data queries, 273
  - financial schemas, 226
  - Lifecycle data physical design, 421
  - loads, ETL system, 519
  - what didn't happen, 335
- one-to-one relationships, 175–176
- operational processing versus data
  - warehousing, 2
- operational product master, product
  - dimensions, 173
- operational source systems, 18
- operational system users, 2
- opportunity/stakeholder matrix, 53, 127
- order management case study, 167–168

- accumulating snapshot, 194–196
  - type 2 dimensions and, 196
- allocating, 184–186
- audit dimension, 192–193
- bus matrix, 168
- currency, multiple, 182–184
- customer dimension, 174–175
  - factless fact tables, 176
  - single versus multiple dimension tables, 175–176
- date, 170–171
  - foreign keys, 170
  - role playing, 171
- deal dimension, 177–178
- degenerate dimension, order number and, 178–179
- fact normalization, 169–170
- header/line patterns, 181–186
- junk dimensions, 179–180
- product dimension, 172–173
- order number, degenerate dimensions, 178–179
- order management case study, role playing, 171
- origin dimension (airline case study), 320–321
- OR, skill keywords bridge, 275
- outtrigger dimensions, 50, 89, 106–107
  - calendars as, 321–323
  - low cardinality attribute set and, 243–244
  - type 5 and type 1 SCD, 160
- overwrite (type 1 SCD), 54, 149–150
  - add to type 2 attribute, 160–162
  - type 2 in same dimension, 153

**P**

- packaged analytic solutions, 270–271
- packaged data models, 270–271
- page dimension, clickstream data, 358–359
- page event fact table, clickstream data, 363–366
- parallelizing/pipelining system, 492
- parallel processing, fact tables, 518
- parallel structures, fact tables, 519
- parent/child schemas, 59
- parent/child tree structure hierarchy, 216
- partitioning
  - fact tables, smart keys, 102
  - real-time processing, 524–525
- passenger dimension, airline case study, 314
- pathstring, ragged/variable depth hierarchies, 57
- pay-in-advance facts, insurance case study, 386–387
- payment method, retail sales, 93

- performance measurement, fact tables, 10, 12
  - additive facts, 11
  - grains, 10–12
  - numeric facts, 11
  - textual facts, 12
- period close (G/L), 204–206
- periodic snapshots, 43, 112–114
  - education case study, 329, 333
  - ETL systems, 474
  - fact tables, 120–121
    - complementary fact tables, 122
  - G/L (general ledger), 203
  - grain fact tables, 12
  - headcount, 267–268
  - healthcare case study, 342
  - insurance case study, 385
    - claims, 395–396
    - premiums, 386–387
  - inventory case study, 112–114
  - procurement case study, 147
- perspectives of business users, 293
- physical design, Lifecycle data track, 420
  - aggregations, 421
  - database model, 421
  - database standards, 420
  - index plan, 421
  - naming standards, 420–421
  - OLAP database, 421
  - storage, 422
- pipelining system, 492
- planning, demand planning, 142
- P&L (profit and loss) statement
  - contribution, 189–191
  - granularity, 191–192
- policy transactions (insurance case study), 379–380
  - fact table, 383
- PO (purchase orders), 142
- POS (point-of-sale) system, 73
  - POS schema, retail sales case study, 94
  - transaction numbers, 93–94
- presentation area, 21–22
- prioritization, Lifecycle business
  - requirements, 414–415
- privacy, data governance and, 541–542
- problem escalation system, 491–492
- procurement case study, 141–142
  - bus matrix, 142–143
  - snapshot fact table, 147
  - transactions, 142–145
- product dimension, 83–84
  - attributes with embedded meaning, 85
  - characteristics, 172–173
  - drilling down, 86–87

- many-to-one hierarchies, 84–85
- numeric values, 85–86
- operational product master, 173
- order transactions, 172–173
  - operational product master, 173
- production codes, decoding, 504
- products
  - heterogeneous, 293–295
  - Lifecycle
    - evaluation matrix, 419
    - market research, 419
    - prototypes, 419
- profit and loss facts, 189–191, 370–372
  - allocations and, 60
  - granularity, 191–192
- program/project planning (Lifecycle), 405–406
  - business motivation, 407
  - business sponsor, 406
  - development, 409–410
  - feasibility, 407
  - justification, 407
  - planning, 409–410
  - readiness assessment, 406–407
  - scoping, 407
  - staffing, 408–409
  - task list, 409
- project manager, 409
- promotion dimension, 89–91
  - null values, 92
- promotion lift, 89
- prototypes
  - big data and, 536
  - Lifecycle, 419
- publishing metaphor for DW/BI managers, 5–7

## Q

- quality events, responses, 458
- quality screens, ETL systems, 457–458
- questionnaire, HR (human resources), 277
  - text comments, 278

## R

- ragged hierarchies
  - alternative modeling approaches, 221–223
  - bridge table approach, 223
  - modifying, 220–221
  - pathstring attributes, 57
  - shared ownership, 219
  - time varying, 220
  - variable depth, 215–217
- rapidly changing monster dimension, 55

- RDBMS (relational database management system), 40
  - architecture extension, 529–530
  - blobs, 530
  - fact extractor, 530
  - hyperstructured data, 530
- real-time fact tables, 68
- real-time processing, 520–522
  - architecture, 522–524
  - partitions, 524–525
- rearview mirror metrics, 198
- recovery and restart system, ETL system, 486–488
- recursive hierarchies, employees, 271–272
- reference dimensions, 130
- referential integrity, 12
- referral dimension, clickstream data, 360
- relationships
  - dimension tables, 15
  - many-to-one, 175–176
  - many-to-one-to-many joins, 259–260
  - one-to-one, 175–176
  - validation, 504–505
- relative date attributes, 82–83
- remodeling existing data structures, 309
- reports
  - correctly weighted, 288
  - dimensional models, 17
  - dynamic value banding, 64
  - fact tables, 17
  - impact, 288
  - value band reporting, 291–292
- requirements for dimensional modeling, 432
- restaurant metaphor for Kimball architecture, 23–26
- retail sales case study, 72–73, 92
  - business process selection, 74
  - dimensions, selecting, 76
  - facts, 76–77
    - derived, 77–78
    - non-additive, 78
  - fact tables, 79
  - frequent shopper program, 96
  - grain declaration, 74–75
  - payment method, 93
  - POS (point-of-sale) system, 73
  - POS schema, 94
  - retail schema extensibility, 95–97
  - SKUs, 73
- retain original (SCD type 0), 54, 148–149
- retrieval, 485–486
- retroactive changes, healthcare case study, 351–352
- reviewing dimensional model, 440, 441
- RFI measures, 240
- RFP (request for proposal), 419
- role playing, dimensions, 49, 89, 171, 284
  - airline case study, 313
  - bus matrix and, 171
  - healthcare case study, 345
  - insurance case study, 380
  - order management case study, 170
- S**
  - sales channel dimension, airline case study, 315
  - sales reps, factless fact tables, 176
  - sales transactions, web profitability and, 370–372
  - sandbox results, big data management, 532–533
  - sandbox source system, ETL development, 503
  - satisfaction indicators in fact tables, 254–255
  - scalability, dimensional modeling myths, 31
  - SCDs (slowly changing dimensions), 53, 148, 464–465
    - big data and, 539
    - detailed dimension model, 437
    - hybrid techniques, 159–164
    - insurance case study, 380–381
    - type 0 (retain original), 54, 148–149
    - type 1 (overwrite), 54, 149–150
      - ETL systems, 465
      - type 2 in same dimension, 153
    - type 2 (add new row), 54, 150–152
      - accumulating snapshots, 196
      - customer counts, 243
      - effective date, 152–153
      - ETL systems, 465–466
      - expiration date, 152–153
      - type 1 in same dimension, 153
    - type 3 (add new attribute), 55, 154–155
      - ETL systems, 467
      - multiple, 156
    - type 4 (add mini-dimension), 55, 156–159
      - ETL systems, 467
    - type 5 (add mini-dimension and type 1 outrigger), 55, 160
      - ETL systems, 468
    - type 6 (add type 1 attributes to type 2 dimension), 56, 160–162
      - ETL systems, 468
    - type 7 (dual type 1 and type 2 dimension), 56, 162–164
      - ETL systems, 468
  - scheduling jobs, ETL operation and automation, 520
  - scoping for program/project planning, 407

- scoring, CRM and customer dimension, 240–243
- screening
  - ETL systems
    - business rule screens, 458
    - column screens, 457
    - structure screens, 457
    - quality screens, 457–458
- security, 495
  - ETL system, 446, 492–493
  - goals, 4
- segmentation, CRM and customer dimension, 240–243
- segments, airline bus matrix granularity, 313
  - linking to trips, 315–316
- SELECT statement, 18
- semi-additive facts, 42, 114–115
- sequential behavior, step dimension, 65, 251–252
- sequential integers, surrogate keys, 101
- service level performance, 188–189
- session dimension, clickstream data, 359–360
- session fact table, clickstream data, 361–363
- session IDs, clickstream data, 355–356
- set difference, 97
- shared dimensions, 130
- shipment invoice fact table, 188
- shrunk dimensions, 51
  - conformed
    - attribute subset, 132
    - on bus matrix, 134
    - row subsets and, 132–134
  - rollup, 132
  - subsets, ETL systems, 472
- simple administration backup, 485
- simple data transformation, dimensions, 504
- single customer dimension, data integration and, 256–258
- single granularity, facts and, 301
- single version of the truth, 407
- skill keywords, 274
  - bridge, 275
    - AND queries, 275
    - OR queries, 275
  - text string, 276–277
- skills, ETL system, 448
- SKUs (stock keeping units), 73
- slightly ragged/variable depth hierarchies, 57
- slowly changing dimensions. *See* SCDs, 148
- smart keys
  - date dimensions, 101–102
  - fact tables, partitioning, 102
- snapshots
  - accumulating, 44, 118–119, 194–196
  - claims (insurance case study), 393–395
  - education case study, 326
  - ETL systems, 475
  - fact tables, 121–122, 326–329
  - fact tables, complementary, 122
  - healthcare case study, 343
  - inventory case study, 118–119
  - order management case study, 194–196
  - procurement case study, 147
  - type 2 dimensions and, 196
- incremental processing, 517
- periodic, 43
  - education case study, 329
  - ETL systems, 474
  - fact tables, 120–121
  - fact tables, complementary, 122
  - G/L (general ledger), 203
  - headcounts, 267–268
  - insurance case study, 385, 395–396
  - inventory case study, 112–114
  - premiums (insurance case study), 386–388
- snowflaking, 15, 50, 104–106, 470
  - outriggers, 106–107
- social media, CRM (customer relationship management) and, 230
- sorting
  - ETL, 490
  - international information, 237
- source systems, operational, 18
- special dimensions manager, ETL systems, 470
  - date/time dimensions, 470
  - junk dimensions, 470
  - mini-dimensions, 471
  - shrunk subset, 472
  - static, 472
  - user-maintained, 472–473
- specification document, ETL development, 502–503
  - sandbox source system, 503
- SQL multipass to avoid fact-to-fact table joins, 61
- staffing for program/project planning, 408–409
- star joins, 16
- star schemas, 8, 40
- static dimensions
  - ETL systems, 472
  - population, 508
- statistics, historic fact table audit, 508
- status dimensions, 284
- step dimension, 65
  - clickstream data, 366
  - sequential behavior, 251–252
- stewardship, 135–136

- storage, Lifecycle data, 422
  - store dimension, 87–89
  - strategic business initiatives, 70
  - streaming data, big data and, 536
  - strings, skill keywords, 276–277
  - structure screens, 457
  - student dimension (education case study), 330
  - study groups, behavior, 64
  - subsets, shrunken subset dimensions, 472
  - subtypes, 293–294
    - fact tables
      - keys, 294–295
      - supertype common facts, 295
    - healthcare case study, 347–348
    - insurance case study, 384, 387
    - schemas, 67
  - summary data, dimensional modeling and, 30
  - sunsetting, big data management, 533
  - supernatural keys, 46, 101
  - supertypes
    - fact tables, 293–294
      - keys, 294–295
      - subtype common facts, 295
    - healthcare case study, 347–348
    - insurance case study, 384–387
    - schemas, 67
  - surrogate keys, 58, 98–100, 303
    - assignment, 506
    - degenerate dimensions, 101
    - dimension tables, 98–100
    - ETL system, 475–477
      - generator, 469–470
    - fact tables, 102–103
    - fact table transformations, 516
    - late arriving facts, 517
    - lookup pipelining, 510–511
  - survey questionnaire (HR), 277
    - text comments, 278
  - synthetic keys, 98
- T**
- tags, behavior, in time series, 63
  - team building, Lifecycle business requirements, 411
    - representatives, 411–412
  - technical application design/development (Lifecycle), 406
  - technical architect, 409
  - technical architecture (Lifecycle), 405, 416–417
    - architecture implementation phases, 418
    - model creation, 417
    - plan creation, 418
    - requirements
      - collection, 417
      - documentation, 417
      - requirements collection, 417
      - subsystems, 418
      - task force, 417
  - telecommunications case study, 297–299
  - term dimension (education case study), 330
  - text comments
    - dimensions, 65
    - healthcare case study, 350
  - text strings, skill keywords, 276–277
  - text, survey questionnaire (HR) comments, 278
  - textual attributes, dimension tables, 82
  - textual facts, 12
  - The Data Warehouse Toolkit (Kimball), 2, 80
  - third normal form (3NF) models, 7
    - entity-relationship diagrams (ERDs), 8
    - normalized 3NF structures, 8
  - time
    - GMT (Greenwich Mean Time), 323
    - UTC (Coordinated Universal Time), 323
  - timed extracts, CDC (change data capture), 452
  - time dimension, 80
    - clickstream data, 361–362
  - timeliness goals, 4
  - time-of-day
    - dimension, 83
    - fact, 83
  - time series
    - behavior tags, 63, 240–242
    - fixed time series buckets, date dimensions and, 302–303
  - time shifting, 90
  - timespan fact tables, 252–254
    - dual date/time stamps, 254
  - timespan tracking in fact tables, 62
  - time varying multivalued bridge tables, 63
  - time zones
    - airline case study, 323
    - GMT (Greenwich Mean Time), 323
    - multiple, 65
    - number of, 323
    - UTC (Coordinated Universal Time), 323
  - tools
    - dimensional modeling, 432
    - data profiling tools, 433
    - ETL development, 499
  - transactions, 43, 120, 179
    - claim transactions (insurance case study), 390
    - claim accumulating snapshot, 393–394
    - junk dimensions and, 392
    - periodic snapshot, 395–396
    - timespan accumulating snapshot, 394–395

- fact tables, 12, 143–145
- healthcare case study, 342
- inventory transactions, 116–118
- invoice transactions, 187–188
- journal entries (G/L), 206–207
- numbers, degenerate dimensions, 93–94
- order management case study
  - allocating, 184–186
  - date, 170–171
  - deal dimension, 177–178
  - degenerate dimension, 178–179
  - header/line patterns, 181–182, 186
  - junk dimensions, 179–180
  - product dimension, 172–173
- order transactions, 168
  - audit dimension, 192–193
  - customer dimension, 174–176
  - fact normalization, 169–170
  - multiple currency, 182–184
- policies (insurance case study), 379–380
- procurement, 142–143
- transaction profile dimension, 49, 179
- transportation, 311
  - airline case study, 311–323
  - cargo shipper schema, 317
  - localization and, 324
  - travel services flight schema, 317
- travel services flight schema, 317
- trees (hierarchies), 215
  - parent/child structure, 216
- type 0 (retain original) SCD, 54
  - retain original, 148–149
- type 1 (overwrite) SCD, 54
  - add to type 2 dimension, 160–162
  - ETL system, 465
  - overwrite, 149–150
  - type 2 in same dimension, 153
- type 2 (add new row) SCD, 54, 150–152
  - accumulating snapshots, 196
  - customer counts, 243
  - effective date, 152–153
  - employee profile changes, 267
  - ETL system, 465–466
  - expiration date, 152–153
  - type 1 in same dimension, 153
- type 3 (add new attribute) SCD, 55, 154–155
  - ETL system, 467
  - multiple, 156
- type 4 (add mini-dimension) SCD, 55, 156–159
  - ETL system, 467
- type 5 (add mini-dimension and type 1 outrigger) SCD, 55

- type 5 (add mini-dimension and type outrigger) SCD, 160
  - ETL system, 468
- type 6 (add type 1 attributes to type 2 dimension) SCD, 56, 160–162
  - ETL system, 468
- type 7 (dual type 1 and type 2 dimension) SCD, 56, 162–163
  - as of reporting, 164
  - ETL system, 468

## U

- Unicode, 236–238
- uniform chart of accounts, 204
- units of measure, multiple, 197–198
- updates, accumulating snapshots, 121–122
- user-maintained dimensions, ETL systems, 472–473
- UTC (Coordinated Universal Time), 323

## V

- validating dimension model, 440–441
- validation, relationships, 504–505
- value band reporting, 291–292
- value chain, 52
  - insurance case study, 377–378
  - integration, 122–123
  - inventory case study, 111–112
- variable depth hierarchies
  - pathstring attributes, 57
  - ragged, 215–217
  - slightly ragged, 214–215
  - variable depth/ragged hierarchies with bridge tables, 57
  - variable depth/slightly ragged hierarchies, 57
- version control, 495
  - ETL system, 488
- version migration system, ETL system, 488
- visitor identification, web sites, 356–357

## W

- weekday indicator, 82
- WHERE clause, 18
- workflow monitor, ETL system, 489–490
- workshops, dimensional modeling, 38

## X–Y–Z

- YTD (year-to-date) facts, 61
  - G/L (general ledger), 206

