
Index

A

- Abandoned basket statistics, 48
- Advanced analytics, 186–188, 241–245
 - analytic team responsibilities, 241–245
 - core analytics compared to, 186–188
- Advertising results, assessment of, 48–50
- Analysis, 87–176, 179–200
 - analytic data set (ADS), 133–145
 - business importance of, 194–195
 - “cherry picking” of findings, 188–189
 - cloud computing, 102–109
 - core versus advanced analytics, 186–188
 - determination of, 179–200
 - embedded scoring, 99–100, 145–151
 - enterprise analytic data set (EADS), 137–145
 - Enterprise Data Warehouse (EDW), 91–93
 - extract, transform, and load (ETL) process, 90–91
 - framing the problem, 189–191
 - G.R.E.A.T. criteria, 184–186
 - grid computing, 109–111
 - inferences versus computing statistics, 198–199
 - MapReduce, 110–117
 - massively parallel processing (MPP) database systems, 93–102
 - processes, 121–152
 - reporting compared to, 179–184
 - samples versus population, 195–198
 - sandbox environments, 108–109, 122–133
 - scalability, 87–119
 - statistical significance, 191–195
 - tools and methods for, 153–176
- Analytic data set (ADS), 133–145, 149. *See also* Enterprise analytic data set (EADS)
 - development, 134–135
 - embedded scoring, inputs for, 149
 - enterprise (EADS), 137–145
 - production, 134–135
 - traditional, 135–137
- Analytic innovation center, 259–269
 - commitment, 261
 - failures, dealing with, 267–269
 - guiding principles of, 263
 - innovation council, 262–263
 - scope of, 264–266
 - sponsorship, 261
 - team strength, 261–262
 - technology platform, 259–260
 - third-party products and services, 260–261
- Analytic methods, 153–162
 - collaborative filtering, 162
 - commodity models, 156–159
 - ensemble models, 154–156
 - page rank, 162
 - text analysis, 159–161
- Analytic professionals, 9–10, 201–225, 231–232, 239, 245–247, 283–289
 - analytic teams of, 231–232, 239, 245–247
 - business savvy and, 211–216

- business value of, 231–232
 - certification of, need for, 222–224
 - clean data and, 209–211
 - commitment of, 208
 - common misconceptions about, 203–204
 - communication skills, 216–220
 - creativity of, 208–211
 - cross training, 239
 - cultural awareness of, 214–216
 - data scientists as, 202–203
 - decisions, granularity of, 212–213
 - education of, 205
 - experience in industry, 205–206
 - focus on importance of data by, 213–214
 - information technology (IT)
 - compared to, 245–247
 - innovation and, 283–289
 - intuition of, 220–222
 - job description, avoiding the “list”, 207
 - presentation skills, 216–220
 - role of, 9–10
 - vision of, 283–289
- Analytic sandbox, 122–125
- Analytic tools, 163–175
- data visualization, 170–172
 - graphical user interfaces (GUI), 163–165
 - graphics and tables, 174–175
 - immersive intelligence, 173
 - open source software, 167–168
 - point solutions, 165–167
- R Project for Statistical Computing, 168–170
- user interfaces, 163–167
 - visualization, 170–175
- Analytics team, 227–248
- advanced analysis by, 241–245
 - centralized structures, 234–236
 - cross training analytic professionals, 239
 - decentralized/functional structures, 233–234
 - hybrid structures, 236–237
 - indecision and, 230
 - industry use of, 228–229
 - information technology (IT)
 - compared to, 245–247
 - management interaction with, 240–241
 - matrix approach, 238–239
 - skills, maintaining, 237–241
 - structures of, 232–237
 - talented analytic professionals, value of, 231–232
 - Asset tracking, RFID tags, 65
 - Attribution modeling, 44–45
 - Automated toll RFID tags, 65
 - Automotive insurance collection of telematics data, 54–57
- B**
- Big data, 3–27
- analysis of, need for, 12–14
 - changes from, 9–10
 - combined with traditional data, 21–22
 - defined, 4–5
 - differences from traditional data, 7–9
 - enterprise data warehouses (EDWs), 22
 - evolution of, 24–25
 - extract, transform, and load (ETL) process, 20
 - filtering, effectiveness of, 20–21
 - identification of, 16–17
 - impact of, 3–4
 - qualification of, 24–25
 - regulation of, 11–12
 - risks of, 10–12
 - standards for, 22–23
 - structure of, 14–16
 - traditional data and, 7–9, 21–22
 - use of, 5–7
 - value of, 8–9, 17–20
 - volume versus velocity and complexity of, 5–7
- Big data sources, 29–83
- casino chip tracking, 71–73
 - radio frequency identification data (RFID), 64–68, 71–73

- sensor data, 68, 73–76
- smart grid data, 68–70
- social network data, 78–82
- telematics data, 54–57
- telemetry data, 76–78
- text data, 57–60
- time and location data, 60–64
- web data, 29–51
- Black box, telematics data from, 54
- Business, 194–195, 211–216, 228–229, 231–232, 252–253
 - analytic professional
 - understanding of, 211–216
 - analytic teams used in, 228–229
 - data analysis, importance of, 194–195
 - innovation, need for, 252–253
 - value of analytic professionals, 231–232
- C**
- Capacity planning, sandbox used for, 131–133
- Casino chip tracking, 71–73
- Central processing units (CPU), MPP systems and, 94–96
- Centralized structures, 234–236
- “Cherry picking” of analysis findings, 188–189
- Clean data, analytic professional and, 209–211
- Clickstream data, 24–25
- Cloud computing, 102–109
 - criteria for environment, 102–103
 - National Institute of Standards and Technology (NIST) characteristics, 103–104
 - private clouds, 107–108
 - public clouds, 104–107
 - sandbox environment compared to, 108–109
 - scalability and analysis using, 102–109
- Collaborative filtering, 162
- Commodity models, 156–159
- Communication skills, 216–220
 - advertising and, 218–220
 - analytic professionals use of, 216–220
 - delivery, importance of, 218
 - presentation skills and, 216
 - results, success of analysis and, 217–218
- Core analytics, advanced analytics compared to, 186–188
- Customer behavior, 32–42
 - behavior types, 34–35
 - faceless customer data, 36
 - feedback behavior, 41–42
 - knowledge, use of, 32–33
 - privacy and, 35–36
 - purchase paths and preferences, 38–39
 - research behavior, 39–41
 - shopping behavior, 37–38
 - transaction types (location flags), 33–34
 - web data, 32–42
- Customer segmentation, 47–48
- D**
- Data preparation and scoring, 96–102
 - embedded processes, 99–100
 - massively parallel processing (MPP) and, 96–102
 - predictive modeling markup language (PMML) and, 100–101
 - structured query language (SQL) and, 96–100
 - user-defined functions, 99
- Data scientists, 202–203. *See also* Analytic professionals
- Data size, measurement of, 89
- Data storage, MPP systems and, 94–96
- Data visualization, 170–172
- Decentralized/functional structures, 233–234
- Development analytic data set (ADS), 134–135

Discovery, *see* Innovation

Diversification, analytic innovation
and, 258–259

E

Embedded scoring, 99–100, 145–151
access of, 146–147

analytic data set (ADS) inputs,
149

batch updates, 146

integration of, 147–148

massively parallel processing
(MPP) systems and, 99–100
model and score management,
148–151

model information, 149–150

model scoring output, 151

model validation and reporting,
150–151

predictive modeling markup
language (PMML) and, 148

real-time scoring, 146

routines, 145–146

structured query language (SQL)
and, 99–100, 147

Ensemble models, 154–156

Enterprise analytic data set (EADS),
137–145

characteristics of, 138–139

creation of, 139

data in, 140–141

logical versus physical structure,
141–142

process of, 137–138

table-based versus views,
143–144

updating, 142

use of, 144–145

Enterprise Data Warehouses
(EDWs), 22, 91–93

External sandbox, 126–128

Extract, transform, and load (ETL)
process, 20, 90–91

F

Faceless customer data, 36

Feedback behavior, 41–42

G

G.R.E.A.T. criteria, data analysis,
184–186

Graphical user interfaces (GUI),
163–165

Grid computing, 109–111

H

Hybrid sandbox, 128–130

Hybrid structures, 236–237

I

Industrial engines and equipment
use of sensor data, 73–76

Inferences versus computing
statistics, data analysis and,
198–199

Information technology (IT)
compared to analytic
professionals, 245–247

Innovation, 251–291. *See also*
Analytic innovation center
analytic, 251–269

applications of principles, 278–
279, 282–283, 289–290

“break out of the box”, 275–279

business need for, 252–253

center for combination of
concepts, 259–269

common vision for, 283–289

defined, 251

discovery and, creation of,
271–291

diversification and, 258–259

focus on the target, 283–289

iterative approach to, 256–257

key principles, 274–290

perspective changes and,
257–259

priorities and, 286–289

ripple effects from, 279–283

risk and, 254–255

setting the stage for, 272–274

traditional approaches hampering,
253–255

Internal sandbox, 125–126

Internet transactions, 7

Intuition of analytic professionals,
220–222

Iterative approach to analytic
innovation, 256–257

M

MapReduce, 110–117
parallel programming framework
of, 110–111
scalability and analysis using,
110–117
strengths and weaknesses of,
114–116
two-step process, 110,
112–114
unstructured text analysis,
111–112

Massively parallel processing (MPP),
93–102
central processing units (CPU)
and, 94–96
data preparation and scoring
using, 96–102
data storage and, 94–96
database systems, 93–102
embedded processes, 99–100
predictive modeling markup
language (PMML) and,
100–101
scalability for analysis using,
93–102
structured query language (SQL)
and, 96–100
user-defined functions, 99

Models, 148–151, 154–159
commodity, 156–159
embedded scoring, management
using, 148–151
ensemble, 154–156
scoring output, 151
validation and reporting, 150–151

N

National Institute of Standards and
Technology (NIST) cloud
characteristics, 103–104

Next best offer, 42–44

O

Open source software, 167–168

P

Page rank, 162

Parallel programming frameworks,
see MapReduce; Massively
parallel processing (MPP)

Passive RFID tags, 64

Point solutions, 165–167

Predictive modeling markup
language (PMML), 100–101,
148

embedded scoring and, 148

massively parallel processing

(MPP) systems, 100–101

Presentation skills, 216–220. *See also*
Communication skills

Privacy of data, 12, 35–36

big data and, 12

web sources, 35–36

Private clouds, 107–108

Problem statement, framing for data
analysis, 189–191

Production analytic data set (ADS),
134–135

Public clouds, 104–107

Purchase paths and preferences,
38–39

R

R Project for Statistical Computing,
168–170

Radio frequency identification data
(RFID), 18–19, 64–68, 71–73

asset tracking, 65

automated toll tags, 65

big data value and, 18–19

casino chip tracking, 71–73

data combined with, 66–67

fraud reduction from, 67

passive tags, 64

serial numbers, 64

tags, retail and manufacturing use

of, 18–19, 64–68

use of, 65–68

Real-time scoring, 146

Recency, frequency, and monetary (RFM) value, 31
 Relational database management systems (RDBMS), 91
 Reporting, analysis compared to, 179–184
 Research behavior, 39–41
 Response modeling, 45–47
 Retail and manufacturing use of RFID tags, 18–19, 64–68
 Risk, analytic innovation and, 254–255

S

Samples versus population, data analysis and, 195–198
 Sandbox environments, 108–109, 122–133
 analytic, 122–125
 benefits of, 123–125
 capacity planning using, 131–133
 cloud environment compared to, 108–109
 data analysis using, 108–109, 122–123
 external, 126–128
 hybrid, 128–130
 identification of new sources using, 130–131
 internal, 125–126
 workload management using, 131–133
 Scalability, 87–119
 centralization of data, 91–93
 cloud computing, 102–109
 combined analytical technologies, 117–118
 data size, measurement of, 89
 Enterprise Data Warehouse (EDW), 91–93
 extract, transform, and load (ETL) process, 90–91
 grid computing, 109–111

 history of, 88–89
 MapReduce, 110–117
 massively parallel processing (MPP) database systems, 93–102
 merging analytic and data environments, 90–93
 relational database management systems (RDBMS), 91
 structured query language (SQL) and, 96–100
 Semi-structured data, 14–16
 Sensor data, 7, 68, 73–76
 external effects of structure of, 75
 industrial engines and equipment monitoring, 73–76
 output, 7
 smart grid data and, 68
 use of, 74–76
 Serial numbers, RFID tags and, 64
 Shopping behavior, 37–38
 Smart grid data, 7–8, 21, 68–70
 big data used for, 7–8
 mixing data types, 21
 sensors and, 68
 smart meter readings, 7–8, 21
 use of, 69–70
 utilities (power) and, 68–70
 Social network data, 8, 78–82, 281–282
 complications with, 78–79
 ripple effects from innovation of, 281–282
 telecommunication industries and, 78–82
 total value of customer, 80
 use of, 79–82
 user interaction and, 8
 Statistics, 191–195, 198–199
 data analysis and, 191–195, 198–199
 inferences compared to, 198–199
 significance of, 191–195
 Structured query language (SQL), 96–100, 147
 embedded processes, 99–100, 147

- massively parallel processing (MPP) systems, 96–102
 - push down, 98
 - user-defined functions, 99
- T**
- Table-based EADS, views compared to, 143–144
 - Team structures, 232–237
 - centralized, 234–236
 - decentralized/functional, 233–234
 - hybrid, 236–237
 - Telecommunication industries and social network data, 78–82
 - Telematics data, 54–57
 - automotive insurance collection of, 54–57
 - black box, 54
 - use of, 56–57
 - Telemetry data, video games and, 76–78
 - Telephone to social media, ripple effects from innovation of, 280–281
 - Text data, 57–60, 111–112, 159–161
 - analysis of, 159–161
 - interpretation of, 58–59, 160–161
 - meaning, emphasis changes of, 160–161
 - text mining tools, 57–58
 - unstructured data, 111–112, 160–161
 - use of, 59–60
 - 360-degree view, 30–31
 - Time and location data, 60–64
 - global positioning systems (GPS), 60–61
 - interpretation of, 61
 - marketing and, 62–63
 - use of, 62–64
 - Traditional data, 7–9, 14, 21–22
 - combined with traditional data, 21–22
 - differences from big data, 7–9
 - structure of, 14
 - Transactional data, 24
- U**
- Unstructured data, 14, 16, 111–112, 160–161
 - MapReduce and, 111–112
 - sources, 14, 16
 - text analysis, 111–112, 160–161
 - User-defined functions, MPP systems, 99
 - User interfaces, analytic tools and, 163–167
 - Utilities (power) use of smart grid data, 68–70
- V**
- Video games, telemetry data and, 76–78
 - Vision, 283–289
 - common long-term perspective, 284–286
 - competition and, 288–289
 - innovation principle of, 284–289
 - priorities and, 286–289
 - Visualization, 170–175
 - analytic tools using, 170–175
 - data, 170–172
 - graphics and tables, 174–175
 - immersive intelligence, 173
- W**
- Web browsing history, 12, 21
 - mixing data types, 21
 - privacy and, 12
 - Web data, 29–51
 - abandoned basket statistics, 48
 - applications of, 42–50
 - assessing advertising results, 48–50
 - attribution modeling, 44–45
 - behavior types, 34–35
 - customer behavior, 32–42
 - customer segmentation, 47–48
 - faceless customer data, 36
 - feedback behavior, 41–42
 - knowledge, use of, 32–33
 - next best offer, 42–44
 - overview of, 30–31

- privacy and, 35–36
- purchase paths and preferences, 38–39
- recency, frequency, and monetary (RFM) value, 31
- research behavior, 39–41
- response modeling, 45–47
- shopping behavior, 37–38
- 360-degree view, 30–31
- transaction types (location flags), 33–34
- Web logs, 8–9, 15
 - semi-structure of, 15
 - value of data in, 8–9
- Workload management, sandbox used for, 131–133

<http://www.pbookshop.com>