Index

A

Advanced data visualization, 35
Apache Oozie, 216
Atomicity, consistency, isolation, and durability (ACID), 111
Average revenue per user (ARPU), 261

B

Banking industry
  applications and systems, 54
  insurance
    analytics domains and opportunities, 59
    applications and systems, 58
    customer-centric analytics, 60
    finance-centric analytics, 60
    risk-centric analytics, 59
  "Next best action," concepts of, 54
  predictive analytics, 55
  retail banking, 55
  risk management, 55
Big data
  Amazon, 14
  analysis
    heterogeneity and incompleteness, 197
    human collaboration, 199
    privacy, 198
    scale, 198
    system architecture, 199
    timeliness, 198
  analytics organization models, 17
  analytics process maturity, 16
  application paradigms, 2
  business models, 9
  corporate firewalls, 2
  cost-benefit analysis, 21
  customer information, 11
  customer intimacy, 14
  data democratization, 8
  data discovery/exploratory analytics, 5
  data management, 8
  data stores, 2
  decentralized model, 17
  definition, 2–3
  designing business models, 15
  e-commerce applications, 14
  election campaign, 14
  energy and utilities, 9
  enriching and contextualizing data, 5
  enterprises, 7
  evolution of, 12
  external data, 2
  financial services, 9
  future capabilities, 21
  health care and life sciences, 10
  independent model, 17
  industrial values, 10
  industry (*see* Industry)
  innovation, 7
  internal data, 2
  media and telecommunications, 9
  multimedia content, 4
  online businesses, 14
  online services and web analytics, 9
  operational analytics/embedded analytics, 6
  operations efficiency, 15
  polystructured nature of, 3
  retail and consumer products, 10
  retailers, 11
  scale measures, 3
  search objectives, 4
  sentiment analysis, 4

Big data (*cont.*)
shared services model, 17
social-media platforms, 14
technology investments, 18
telecom companies, 11
total cost of ownership, 3
transform raw data, 1
value drivers, 12
web 2.0 companies, 7
Big data analytics methodology
analytical method selection, 206
analytical models, 210
analytics approach
definition, 205
loan delinquency problem, 205
product mix optimization, 205
analytics outcomes, 206
business hypotheses
loan repayment delinquency problem, 204
product mix optimization problem, 204
business use case
loan repayment delinquency problem, 202
product mix optimization problem, 203
data sets
automatic right metadata generation, 208
data acquisition, 207
heterogeneity, 208
loan repayment delinquency problem, 209
production mix optimization problem, 209
designing big data scale, 211
gathering data
partition management with Apache Oozie, 216
querying complex data with Hive, 216
SerDe function, 217
tweeter data, 217
high-level view, 200
measuring and monitoring results, 218
production ready system, 212
setting up big data analytics system, 214
support team, 219
Big data management
advanced analytics, 35
advanced data visualization, 35
cost, 32
data discovery, 35
data integration, 32
data quality, 33
data services, 35
data types, 31
data virtualization, 35
enterprise data warehouse (EDW), 37
IT stack, 38
leading practices, 36
map-reduce technology, 40
MDM, 33
metadata management, 33
query, 42
rapid data insight, 35
skill, 34
SMAQ stack, 39
storage mechanism, 41
Big data scale, 86
Big data warehouse (BDW)
analytics community, 127
architecture
analytics models, 151
big data discovery, 149
big data ingestion, 148
big data quality, 150
big data sources, 148
cloud, 152
conceptual view, 146–147
database, 148
enterprise data platform ecosystem, 146
Hadoop distributions, 148
ILM, 151
information policy management, 150
master data management, 150
metadata, 150
reporting and advanced data visualization, 151
security and privacy, 152
streaming analytics, 149
text analytics, 149
Bonferroni principle, 127
data context, 127
data processing life cycle, 109
data profiling/quality analysis, 109
data quality (*see* Data quality management)

design principle
ACID, 112
BASE, 113
CAP, 113
scale out approach, 111
scale up approach, 111
vs. EDW, 109
enterprise data platform ecosystem, 109
enterprise data platform system
EDW analysis, 116
goal of, 116
hybrid architecture, 116
Hadoop
Avro, 119
components, 118
cost and time-effective manner, 124
filter/workload partition stage, 123
Flume, 119
framework, 120
HDFS, 118
Hive, 118
Mahout, 119
map-reduce function, 118, 121
map-reduce job, 121
map-reduce phase, 123
myriad components, 124
node, 121
Oozie, 119
Pig Latin, 119
Sqoop, 119
suitability test, 125
technical components, 123
unstructured/semi-structured, 121
Whirr, 119
low latency, 126
MDM
bulk data integration, 127
connectivity and interoperability layer, 137
data integration, 132
data model, 129
data repository, 128
enterprise data management principles, 128
external data, 137
external participants, 137
governance processes, 128
implementation, 128
interaction system, 134
logical architecture, 136
logical integration architecture, 140
MDM hub, 132
multi-domain interaction, 135
paradigm, 128
real-time integration, 128
requirements, 131
SEC filing documents, 131
service component, 139
tools, 129
traditional approaches, 129
sandboxes, 126
system requirement/hybrid architecture, 115
Bonferroni principle, 127
Business hypotheses
loan repayment delinquency problem, 204
product mix optimization problem, 204
Business intelligence (BI), 74, 83
Business use case
loan repayment delinquency problem, 202
product mix optimization problem, 203

C

Cassandra, 169
Cassandra data model, 187
cluster, 190
column, 188
column family, 188
counter logic, 186
data structure design, 191
concurrent writes, 193
de-normalization, 192
entities, 191
JSON, super column, 189
keyspace, 190
vs. relational data model, 190
super column, 188
Clinical Disease Repository (CDR), 67
Confirmatory Data Analysis (CDA), 5
Consistency, availability, and partition-tolerance (CAP), 112
CouchDB, 166
Customer relationship management (CRM) system, 5, 143

D

Database
  columnar database
    column-based data structure, 94
    complex queries, 95
    large table scans, 95
    time-based queries, 96
    unpredictable queries, 95
  column-store databases, 75
  complex analytics, 105
  CPU, 74
  distributed hash table, 81
  E-commerce retail application, 101
  flexibility, 104
  high availability, 104
  implementation process, 104
  implications, 73
  in-memory technology, 74
  key value store, 81
  loading capability, 105
  low-cost commodity hardware, 104
  migration, 104
  next generation data warehouses
    big data flow, 97
    definition, 97
    polyglot persistence approach, 98
  scalability, 105
  scale-out database architecture
    non-relational database, 78
    relational database
      (*see* Relational database)
    replication strategies, 76–77
    sharding approaches, 77
    structured data, 77
    unstructured data, 77
  Sybase, 74
  top-notch performance, 104
  workloads (*see* Workloads)
  XML, 82
Data discovery, 35
Data integration, 32
Data modeling, 155
  data marts
    ad-hoc queries, 195
    canned reports, 194
  integration patterns, 155
    data mash ups, 157
    forensic, data, 158
    high velocity integration, 157
    levels of, 155
    linkage analysis, 157
    rare event detection, 157
    streaming analytics, 157
    text analytics, 157
    time series analysis, 157
    workload design, 156
map-reduce
  algorithms, 158
  collate, 162
  combiner function, 161
  count and sum (pattern), 162
  cross correlation, 165
  distinct values, 164
  filtering (grepping), 163
  framework, 160
  function of, 159
  iterative message passing
    (graph processing), 164
  order illustration, pattern, 158
  parallel reduction, 160
  parsing, 163
  patterns, 158, 161
  reduce function, 159
  shuffling approach, 160
  sorting, 163
  task execution, distributed, 163
  use cases, 158
  validation, 163
NoSQL techniques, 165
  Cassandra model
    (*see* Cassandra data model)
  census data, column family, 174
  CFDB design, 175–176
  column family, 170–171, 173
  column family database, 173
  comparator and validator, 184
  composite columns *vs*. super
    columns, 187
  counter logic, 185
  database uses
    (application), 170
  data store types, 166
  de-normalize and duplicate, 178
  de-normalized entities, 179
  document databases, 166
  document store, 171
  event logic model, 182
  graph databases, 167, 171
  idempotent operations, 184
  JSON techniques, 172
  key value store, 171

model column families, query patterns, 177
normalized entities, 178
partially de-normalized entities, 179
peer store, 169
query patterns, 180
RDBMS logical data model, 169, 177
read and write (heavy data), 184
relational model, 178
shard key, 183
surrogate keys, 186
syntax, JSON, 172
timestamp, 180
unique key selection, 184
use cases, 168–169
user oriented, 165
value storage, 181
wide column store, 171
wide rows (order, group and filter), 182
XML databases, 167
XML, JSON, 172
NoSQL technologies
activities, data extraction, 194
activities, data preparation, 193
data preparation and extraction, 193
migration approach, 193
schema migration (ETL), 193
Data quality, 33
Data quality management
approach, 140
cleansing data, 140
data acquisition, 141
data element classification, 141
vs. high availability
analytical value, 146
analytic data platform, 144
core system, 143
CRM system, 143
data quality matters, 143
efforts, 144
fundamental aspects of, 142
profitability, 143
quality assess, 146
sparse/outlier records, 144
standard and shared method, 144
textual/unstructured data, 145
trade-off, 142
type of, 145
uniqueness/accuracy, 142
workload scaling, 142
ingestion and integration data, 141
metadata, 141
principles, 140
volatility and velocity, 140
Data scientist
actionability test, 285
activity, 259
algorithms, 252
analytics techniques, 252
big data, 251
business challenges, 260
business data visualization
bar graphs, 274
box plots, 277
detailed view, 272
graphical view, 271
hierarchical data, 271
line graphs, 273
multi-dimensional view, 272
scatter plots, 276
semi-structured and unstructured data, 279
summarized view, 271
visualization velocity, 279
characteristics, 251
conceptualizing data visualization, 270
conceptual modeling, 252
CSP
analytics techniques, 263
churn articulation, 266
customer usage, 263
data discovery activities, 262, 264
“gets hot” device, 265–266
sentiment analysis, 265
data analysis workflow, 256
data discovery platform, 283
data ingestion/foraging, 261
definition, 252
design principles
collaboration and reusability, 257
discover/seek patterns, 257
ingest and integrate data, 257
insight generation, 257
evaluation of
chi-square (χ^2) statistic, 268
coefficient, 267
histogram/frequency curve, 267

Data scientist (*cont.*)
independent *vs*. dependent variable, 267
one-tailed test, 269
R-Square (R^2)/Pseudo-R^2 statistic, 268
two-tailed test, 269
Hook visualization
barometric pressure, 282
Hurricane Sandy, 280–281
wind speeds, 282
hypothesis testing, 252
machine learning, 252
natural language processing, 252
needle movement test, 285
north pole test, 286
predictive modeling, 252
resonant story telling test, 284
result presentation, 270
skills, 255
small data, 251
sniff the domain out, 285
statistical analysis, 252
story, 279
string of pearls test, 284
telecom industry, 260
use case curation test, 286
variables, 251
Data services, 35
Data virtualization, 35
Distributed hash table (DHT), 81

E

E-commerce, 101
Enterprise. *See* Big data
Enterprise data modeling, 31
Enterprise data warehouse (EDW), 37
Enterprise information management (EIM)
big data (*see* Big data management)
business applications, 27
business model, 26
capabilities, 7
definition, 25
enterprise data model and data stores, 28
enterprise technology and architecture, 27
governance, 31
information lifecycle management, 28
information management and usage, 26
organization and culture, 27
regulations and compliance, 30
Exist, 167
Extract, transform, load (ETL) staging system, 120

F, G

Facebook, 14

H

Hadoop distributed file system (HDFS), 118
HBase, 169
Health care
applications and systems, 67
diagnosis and preventive actions, 70
drug tax, 65
location aware analytics application, 69
patient's care, 65
telemedicine analytics, 69
text mining and correlations, patient outcomes, 68
HFlame enhancement, 235
Hive, 216
Horizontal scaling, 111
Human-to-machine (H2M) interaction, 237

I, J, K, L

Industry
banking (*see* Banking)
benefits, 50
communication, media and technology, 49
data availability and utilization, 46
financial services, 48
Google and Amazon, 45
health and life sciences, 49
health-care companies (*see also* Health care)
hospitality and travel industry, 45
IT/operations
hardware and software vendors, 70
log analysis, 71
public sector, 49

resources, 50
telecommunication
applications and systems, 51
deliver real-time analytics, 52
network performance, 52–53
service quality, 53
video-based services, 53
uses, 13
Infinite Graph, 167
InfoGrid, 167
Information lifecycle
management (ILM), 28, 151
In-memory solutions
database, 225
in-memory analytics, 223
in-memory data grids, 222
in-memory technologies, 223
IT stack, 38

M

Machine-to-machine
(M2M) interaction, 237
Map-reduce technology, 40
MarkLogic, 167
Massively parallel processing (MPP), 92
Master data management (MDM), 29, 33
bulk data integration, 127
connectivity and interoperability
layer, 137
data integration, 132
data model, 129
data repository, 128
enterprise data management
principle, 128
external data, 137
external participants, 137
governance processes, 128
implementation, 128
interaction system, 134
logical architecture, 136
logical integration architecture, 140
MDM hub, 132
multi-domain interaction, 135
paradigm, 128
real-time integration, 128
requirements, 131
SEC filing documents, 131
service component, 139
tools, 129
traditional approaches, 129
Master-slave replication, 76
Maturity model, 36
Metadata management, 30, 33
MongoDB, 166
Multimedia content, 4

N

Natural language processing
(NLP) technologies, 4
Neo4J, 167

O

Online Analytical Processing
(OLAP) analysis, 5
Online transaction processing
(OLTP), 74, 83
Oracle, 167

P, Q

Polyglot persistence application
Cassandra, 101
Digg API, 100
Digg App, 100
DiggBar, 100
Digg Dialog, 100
HDFS, 101
MySQL, 101
Redis, 101

R

Real time analytics
CAP theorem, 225
collect real-time data, 227
explore, analyze,
and visualize data, 228
Hadoop and NoSQL Conundrum, 228
Hadoop's map-reduce model, 231
index and data mapping,
machine generated data, 240
in-memory data grid, 229
log processing
block compression, 241
CLOB field, 238
document indexing sample, 242
e-commerce site, 238
Google protocol buffer, 240
Hadoop solutions, 242

Real time analytics (*cont.*)
human-to-machine interaction, 237
index files, 241
index sharding, 244
Lucene index, 242
machine-to-machine interaction, 237
MapFiles, 240
message fields, 238
primary key field, 242
SequenceFile format, 240
textual data, 241
process streaming data, 228
Recommendation system
association rule based model, 247
classical model, 245
collaborative filtering approach, 246
content-based approach, 246
item-based collaboration filter, 247
singular value decomposition, 247
user-based collaboration filter approach, 247
Relational database
map-reduction and Hadoop ecosystem, 80
OldSQL, NewSQL, and NoSQL applicability, 78–79
un-modeled data, 82
Relational database management system (RDBMS), 74
Retails
applications and systems, 62
consumer behavior, 63
e-mail, 65
show rooming trend, 63
traditional and non-traditional channels, 63
Riak, 169

S

Segal's Law, 30
Semantic analysis, 4
Sensor data, 11
Sentiment analysis, 4
SerDe function, 217
SMAQ stack, 39
SoLoMoMe, 63
Structured query language (SQL), 74

T, U

Total cost of ownership (TCO), 3
Tracking hurricane sandy, 282

V

Vertical scaling, 111

W, X, Y, Z

Workloads
analytics, 83
big data scale, 86
business intelligence, 83
characteristics, 83
computation intensiveness, 89
consistency, 85
data latency, 84
data types, 85
hardware architectures, 92
online transaction processing, 83
predictability, 86
reads and writes, 84
response time, 85
updatability, 85
users and query concurrency, 88