

## A

Access Key Id 195–196, 198  
 acronym libraries 295  
 acronyms, mining for 295  
 ad hoc queries 247  
 ad networks 124  
 aggregated 13  
 Aggregate package 90–94, 99  
 aggregating operation 219  
 algebraic property 86  
 algorithms  
   computational  
     complexity 158  
     data mining 272  
     ES2 296  
     global analysis 293  
 allDocs 293  
 ALTER 257  
 ALTER TABLE,  
   limitations 276  
 Amazon AWS 101  
 Amazon Machine Image 194  
 Amazon Web Services. *See* AWS  
 AMI. *See* Amazon  
   Machine Image  
 anchor text analysis 291, 293  
 anomaly detection 131  
 Apache  
   log file collection 281  
   log files 42  
   log processing with  
     Cascading 282  
   Lucene 19  
   Top Level Project 20  
 Apriori 269  
   vs. BC-PDM 270  
 arrays 256  
 association analysis 269  
 Aster Data Systems 264  
 authentication 195  
 authority 39  
 average, computing 86–87,  
   93–98

AVG 225  
 AWS 194  
   account ID 195  
   command line tools 198  
   Import/Export 210  
   service recommendations  
     211  
   New York Times 267  
   regional support 199  
   setting up 194–201  
 AWS Account  
   Number 197–198

## B

Bag 222  
 bags  
   grouping operators 226  
   Pig and 213  
 basic arithmetic 224  
 BASS. *See* Business Analysis  
   Support System  
 BC-PDM 269  
   cost breakdown 270  
   four-dimensional  
     evaluation 269  
   Hadoop cluster 270  
   vs. Apriori 270  
 BerkeleyDB 281  
 Big Cloud-based Parallel Data  
   Mining. *See* BC-PDM  
 BIGINT 256  
 bigram 11  
 Bigtable 262  
   basic details 274  
 BitSet, Bloom filter 125  
 blocks 52  
   replica placement 184  
 Bloom filter 122–131, 158, 279  
   applications of 123–124  
   end-of-split output 128  
   implementing 124  
   overview 122  
   using Hadoop 128

BloomFilter 126, 129  
   Hadoop v 0.20 131  
 Boolean expressions 224  
 bottleneck 10  
 buckets 202  
   Hive 255  
   specified number 256  
 Business Analysis Support  
   System 269  
 bytearray 220, 222

## C

caching 278–279  
 call data record 268  
 candidate navigational  
   page 288  
 Capacity Scheduler 189  
 Carnegie Mellon  
   University 101  
 Cascading 263  
   Apache log processing 282  
   MapReduce 281  
 case studies  
   China Mobile 267, 269,  
     271–272  
   ES2 282–296  
   New York Times 267  
   StumbleUpon 272–273, 275,  
     277, 279, 281–282  
 cast 224  
 casting 236  
 cat command 39  
 cat, command 298  
 CDR. *See* call data record  
 certificate file 197  
 chaining 102  
   driver 105  
   jobs in sequence 103  
   MapReduce jobs 131  
   pre- and postprocessing  
     steps 104  
 chaining MapReduce  
   jobs 103–107

ChainMapper 104–107, 131  
 ChainReducer 104, 131  
 chararray 222  
 checksums, in HDFS files 139  
 chgrp, command 298  
 China Mobile 4, 267–272  
 chmod, command 299  
 chown, command 299  
 chunks 51  
 citation graph 65  
   cite\_count 254  
 cite\_grpd 242  
 cite table 251  
 classification 269  
 CLI 247  
   Hive 250  
 close 57  
 cloud 5  
 CloudBase 264  
 cloud computing 194  
   EC2 194  
   multiple reducers 49  
   S3 194  
   S3 data storage 205  
 Cloudera 263  
 ClueWeb09 101  
 cluster ID 142  
 clustering 269  
 clusters  
   accessing data from 204  
   busy 182  
   launching 200  
   metadata handling 181  
   moving code to 204  
   multicluster setup 186  
   of one 29  
   private key, public key 26  
   production, monitoring 140  
   SSH setup for Hadoop 25–27  
   topology 25  
   type configuration 202  
 CMU 4, 101  
 cocitation, symmetric 242  
 cocite 242  
 cocite\_bag 241  
 cocite\_flat 244  
 codecs 153  
 COGROUP 227  
 collaborative filtering 240  
 columns 254, 256  
 column store 275  
 combiner 76, 95–98, 152  
 command line interface.  
   *See* CLI

command line tools 198  
 comments 237  
 commutative property 99  
 comparison 224  
 compiler, Pig 233  
 complex statistical model 8  
 compression 153–155  
   Hive 248  
 CONCAT 225  
 conditional 224  
 -conf 70  
 Configuration 42, 161–162  
 configuration object, state  
   information 163  
 configuration properties  
   dfs.permissions.supergroup  
     178  
   Hadoop 28  
 configure 50, 161  
 Configured 70  
 constant 224  
 copyFromLocal, command 299  
 copyToLocal, command 299  
 core-site.xml 28  
 COUNT 220, 225, 253  
 COUNT(\*) 258  
 count, command 299  
 counters 98, 136, 146–148  
   missing values 147  
   names 146  
   summary 141  
   tracking records 146  
 counting 72–80, 86, 91–92  
 cp, command 299  
 CrawlDB 287  
 crawling 287  
 cross-product 110  
   flattening 242  
   record combinations 112  
 Cutting, Doug 4, 19  
 cygwin 14

## D

daemons  
   DataNode 22  
   Hadoop 22  
   JobTracker 24  
   logs 34  
   NameNode 22  
   Secondary NameNode 23  
   shut down 31  
   TaskTracker 24  
 data, semistructured 223

databases  
   accessing 170  
   bulk loading into 171  
   input and output 169  
 data block index 280  
 data flow language 213  
 DataInput 47  
 DataInputStream 52  
 data joining 103  
   cross-product 114  
   from different sources 107  
 DataJoinMapperBase 112, 163  
 datajoin package 131  
 DataJoinReducerBase 112  
 DataJoinReducerBase.  
   combine() 114  
 DataNode 140, 142  
   adding 180  
   daemon 180  
   dfsadmin 177  
   multiple drives 174  
   NameNode interaction 23  
   offline 180  
   overview 22  
   RAM 175  
   removing 180  
 DataOutput 47  
 data processing model 8  
 data sets 64–67, 101  
   Amazon AWS Public Data  
     Sets 101  
   analysis, China Mobile 272  
   collaborative filtering 101  
   development 140  
   human genome 101  
   maximum and minimum  
     computation 157  
   multiple 164  
   Netflix 101  
   patent citation 64  
   patent country codes 90  
   reducing 152  
   sparse 240  
   test 140  
   UCI 270  
   US Census 101  
 data skew 83, 94–95  
 data source, reduce-side  
   join 108  
 data types 213  
   nested 223  
 data warehousing 268  
 DayOfWeek 227  
 DBConfiguration 170  
 DBOutputFormat 170

DBWritable 170  
 Debian 263  
 debugger 136, 151  
 debugging 70, 145–152  
   arithmetic mistakes 137  
   local mode 135  
   modes 31  
   pre- and postprocessing 104  
 decommissioning 180  
 deflate 154  
 departureNode 49  
 Derby 249  
 DESCRIBE 219, 251, 257  
 development data set 64  
 DIFF 225  
 diff'ing data 138  
 directories, adding 39  
 distributed applications 4  
 distributed cache 118, 136  
   reduce-side join 121  
 DistributedCache 118  
   replicated joins 117  
 distributed system 6  
   HDFS 22  
 distributive functions 97  
 distributive property 84, 86,  
   95, 100  
 document similarity 99  
 Dogear 286  
 dot notation 223  
 dot operator 224  
 double 222  
 DOUBLE 256  
 DoubleValueSum 90  
 DRBD 184  
 driver 103, 161  
   MapReduce program 128  
 DROP TABLE 253, 257  
 du, command 299  
 DUMP 217  
 dus, command 299

## E

---

 EC2 4, 101  
   command line tools 195,  
   198  
   concurrent instances 204  
   GUI tools 199  
   images 203  
   instance handling 209  
   MapReduce programs  
   and 203, 205–209  
   New York Times 267  
   SSH key pair 195  
   supported OSs 194

Edge 47, 49  
 EditLog 182  
 Elastic Compute Cloud.  
   *See* EC2  
 Elastic MapReduce. *See* EMR  
 emit() function 14  
 Employee Referral Bonus  
   Program 289  
 EMR 209  
 equijoins 259  
 Ercegovac, Vuk 282  
 ES2 282–297  
   analytics 288  
   architecture 285  
   crawler 287  
   global analysis algorithms 292  
   Hadoop implementation  
   293  
   IBM 283  
   idp query 283  
   JSON 286  
   local analysis 288  
   mining tasks 295  
   offline analysis 285  
   redirection resolution 289  
   variant generation 285  
 ETL operations 269, 270  
   speed-up 271  
 evaluation mechanisms 213  
 Excite query log 217  
 exercises  
   advanced MapReduce  
   techniques 131–132  
   MapReduce paradigm  
   98–100  
 expressions, constant  
   value 223  
 expunge, command 299  
 EXTERNAL 257

## F

---

 Facebook 4  
   Hadoop and 247  
   Hive and 260  
   Scribe 281  
 failed tasks  
   recovering 183  
   rerunning 151  
 Fair Scheduler 187, 189  
 false negatives, Bloom  
   filter 122  
 false positives, Bloom  
   filter 122  
 Fast Fourier Transform 99  
 fetch list 287–288

FetchQueues 288  
 FFT 99  
 field, referencing 223  
 field delimiter 257  
 FIFO scheduler 186  
 file commands 38, 298–301  
 file management tasks 38  
 FileOutputFormat 57  
 -files 83, 120  
 files  
   adding 39  
   deleting 41  
   I/O 128, 129  
   merging 42  
   multiple 42  
   retrieving 41  
   storing multiple 267  
 file splits 154  
 FileStatus 43  
 filesystem  
   block replication 176  
   checking 176–178  
   corrupt block 176  
   fsck 176  
   metadata, handling 181  
   S3, default 208  
   unbalanced 181  
 FileSystem 42  
   operations 44  
 FILTER 237  
 filter and transform 12  
 FIR filter 99  
 FLATTEN 229  
 flattening 229  
   cross-product 242  
 float 222  
 FOREACH 219, 228  
   output/input schema 230  
 fsck 176  
 FSDataInputStream 43, 52  
 FSDataOutputStream 43  
 FsImage 182  
 fully distributed mode 31, 135,  
   145–152  
 FuncSpec 236  
 functions in Aggregate  
   package 90

## G

---

 garbage collection, Java 75  
 General Packet Radio  
   Service 268  
 GenericOptions 70  
 GenericOptionsParser 70, 82,  
   120, 149

- geo-classification 295
  - geographical data 64
  - get 39
  - get() 236
  - getAll() 236
  - getCollector 168
  - get, command 300
  - getmerge 42
  - getmerge, command 300
  - getPartition 50
  - getPos 57
  - getProgress 57
  - getRecordReader 55
  - getSplits 55
  - get(String) 162
  - getter methods 163
  - GFS. *See* Google File System
  - global analysis
    - algorithms 293
    - ES2 291
    - Jaql and Hadoop 294
  - Global Analysis 284
  - Google File System 19
  - Gottfrid, Derek 267
  - GPRS. *See* General Packet Radio Service
  - graph data 64
  - Gray, Jon 280
  - Greenplum 264
  - GROUP 242
  - GROUP BY 219, 253, 259
  - group key
    - join key 109
    - reduce-side join 108
  - Grunt
    - managing shell 216
    - Pig Latin and 217
    - shell 215
  - GUI tools 199
- H**
- 
- Hadoop 4–5
    - building blocks 21–25
    - cluster performance 270
    - clusters, key pairs 200
    - command line utilities 38
    - configuration 173
    - configuration directory 27
    - datajoin package 112
    - data processing modules 281
    - data types 46
    - default settings 28
    - distributed systems 6
    - EC2 setup 201–203
    - EMR 209
    - ES2 293
    - file API 42
    - file commands 38
    - file operations 40–41
    - fully distributed cluster 216
    - history 19, 20
    - kill command 216
    - modes 135, 140
    - linear scalability 152
    - production cluster
      - properties 174
    - related projects and vendors 262–265
    - ResolveSimple 290
    - running 27–33
    - SQL databases 7–8
    - subprojects list 246
  - HADOOP\_CLASSPATH 187
  - Hadoop cluster 5, 25
  - Hadoop Core 212
  - hadoop-default.xml 28
  - hadoop-env.sh 28
  - Hadoop-related projects
    - Aster Data 264
    - Cascading 263
    - CloudBase 264
    - Cloudera 263
    - Greenplum 264
    - Hama 264
    - HBase 262
    - Katta 263
    - Mahout 264
    - ZooKeeper 262
  - Hadoop site properties
    - dfs.balance.bandwidthPerSec 181
    - dfs.block.size 181
    - dfs.hosts.exclude 180
    - dfs.http.address 182
    - dfs.name.dir 183
    - fs.checkpoint.dir 183
    - fs.trash.interval 179
    - mapred.job.tracker 186
    - mapred.jobtracker.taskScheduler 187
    - topology.script.file.name 185
    - topology.script.number.args 185
  - hadoop-site.xml 28, 135
  - Hama 264
  - hash partitioner 172
  - HashPartitioner 49
  - hash table 10
  - HBase 262
    - added features 274
    - HFile 279
    - introduction 274
    - parallelism 280
    - region server 275, 278
    - SATA disks 280
    - StumbleUpon 274–275
    - table splitting 278
    - transcending single machine 277
    - ZooKeeper 263
  - HDFS 6, 129, 170, 248
    - balancer 181
    - blocks 52
    - data nodes 278
    - dataset storage 38
    - directory creation 39
    - file permission 178
    - filesystem 216
    - free space 175
    - HBase 262
    - metadata snapshots 23
    - moving data into 204
    - NameNode 22, 214
    - quota 179
    - reading and writing to 42
    - replication factor 32
    - security 178
    - shell commands 83
    - web interface 34
    - working directory 39
    - working with files 38–44
  - hdfs-site.xml 28
  - heap memory 175
  - help, looking up 41
  - help, command 300
  - heuristics 242
  - HFile 279
    - internal platform 280
    - keys and values 280
    - write path 279
  - histogram 73, 93
  - Hive 247–250
    - buckets 255
    - complex types 256
    - database 249
    - directory structure
      - example 254
    - installing and configuring 248–250
    - loading data 257
    - MapReduce 247, 252
    - metadata 249
    - metastore 247
    - queries 250–253
    - relational database 247

running queries 258  
 structured data 247  
 tables, managing 255  
 table storage 254  
 tarball 248  
 HiveQL 247  
   built-in aggregate functions 262  
   built-in functions 260–262  
   data model 254  
   joins 259  
   queries 258  
   SELECT COUNT 252  
   standard operators, list 259  
   statements 250  
   usage 254–260  
 hive-site.xml 249  
 holistic property 86  
 Holstad, Erik 280

---

## I

IBM 4  
 IBM intranet 282–283  
 idempotent 156  
 identity 243  
 IdentityReducer 83, 85, 88, 104  
 idp query 285, 295  
   ES2 283  
 ILLUSTRATE 220, 241  
 indexes 254  
 Individual Development Plan.  
   *See* IDP  
 information retrieval 104  
 INNER 227  
 inner join 108, 112, 115  
 inner product of vectors 99  
 InputFormat 52  
   classes 53  
   custom class 54  
   joining 117  
   KeyValueTextInputFormat 75  
   SequenceFileInputFormat 155  
 InputSplit 54  
 input splits 51  
 INSERT 258  
 INSERT OVERWRITE  
   TABLE 253  
 int 222  
 INT 250, 256  
 interactive mode 217  
 intermediate data 11  
 intermediate records 166

in total 172  
 intranet vs. web issues 283  
 int type 140  
 IntWritable 17  
 inverted index 67, 73,  
   103, 136  
 IsEmpty 225  
 IsolationRunner 151  
 isSplittable 55

---

## J

Jaql 285  
   ES2 283  
   global analysis 292  
 jar files 15  
 Java 14  
   generics 47  
   regular expression 257  
 Java Advanced Image  
   Extension 267  
 JAVA\_HOME 28  
 java.io.DataInputStream 52  
 java.lang.Comparable<T> 46  
 JBOD 181  
 jdb 151  
 JDBC, interface 247  
 JetS3t 267  
 joining data 107–122  
 Job 103  
 JobClient 70  
 JobClient.runJob() 103  
 JobConf 54, 70, 103, 161  
   chaining 106  
   object 129  
   set() 70  
   setCombinerClass() 97  
   setCompressMapOutput() 153  
   setInputFormat() 72, 155  
   setJobName() 70, 143  
   setMapOutputCompressorClass() 153  
   setMaxMapAttempts() 149  
   setMaxReduceAttempts() 149  
   setNumTasksToExecutePerJvm() 156  
   setOutputKeyClass() 72  
   setOutputValueClass() 72  
 JobConf object 106  
 JobConf properties  
   keep.failed.tasks.files 151  
   key.value.separator.in.input.line 75  
   mapred.job.name 143

mapred.job.reuse.jvm.num.  
   tasks 155  
 mapred.local.dir 151  
 mapred.map.max.attempts 149  
 mapred.output.dir 129  
 mapred.reduce.max.attempts 149  
 mapred.reduce.tasks 70, 138  
   table 150  
 JobConf.setOutputFormat 128  
 JobControl 103  
 job ID 142, 151  
   killing jobs 145  
 jobs  
   chaining in sequence 103  
   EMR 209  
   ID 142  
   kill 145, 216  
   maximum per user 188  
   name 142  
   pools 187  
   scheduling 186–189  
   tracking 142  
 JobTracker 70, 140, 142, 269  
   administration page 144  
   counter information 147  
   multiple 186  
   overview 24  
   TaskTracker interaction 24  
   Web UI 143  
 join  
   datajoin package 112  
   HiveQL 259  
   inner 112  
   outer 112  
   reduce-side 109, 121  
   repartitioned 110, 111  
   join, reduce-side 102, 108  
   join key, group key 109  
   join process 107  
   jpo.x.properties 249  
 JSON, ES2 286  
 JVM  
   mapred.job.reuse.jvm.num.  
     tasks 156  
   reusing 155  
   TaskTracker 24  
 JVM usage 69

---

## K

Katta 264  
   ZooKeeper 263  
 key pair, RSA 26  
 key pairs 200

KeyValueLineRecordReader  
     55, 57  
 key/value pairs 7, 12, 38  
     Configuration 42  
     data flow 44  
     list 46  
     splits and 55  
 KeyValueTextInputFormat  
     53, 55, 72–73  
 key/value types 72  
     in data flow 71  
 klass 106  
 K-means 269  
     Shanghai Branch 272  
 Krishnamurthy, Rajasekar 282

## L

LAMP 273  
 LIKE 259  
 LIMIT 218  
 linear scalability 152, 158  
 LineRecordReader 55  
 LinkedIn 4  
 Linux 14  
 lists 12  
 listStatus 43  
 LOAD DATA 257  
 Local Analysis 284  
 local mode 135–140  
 local reduce 50  
 log files 141–142  
 logging 141  
     messages 142  
 login, validating 27  
 log-normal distribution 76  
 long 222  
 LongSumReducer 51  
 long tail problem 287  
 long type 140  
 LongValueMax 90  
 LongValueMin 90  
 LongValueSum 90  
 LongWritable 17, 54  
 ls, command 300  
 lsr 40  
 lsr, command 300  
 Lucene 264  
     ES2 283  
 Lucene engine 19  
 Luo, Zhiguo 267

## M

Mac OS X 14  
 Mahout 264  
 map 54  
 Map 222  
 MapClass 71, 161  
 map\_input\_file 164  
 -mapper 81  
 Mapper 47, 69, 160  
     class list 48  
     close() 71  
     configure() 71  
     map() 71  
     model 106  
 Mapper.configure() 119  
 Mapper.map() 107  
 mappers 8, 47  
     configuration properties  
         description 153  
     mapred.compress.map.  
         output 153  
     mapred.map.output.  
         compression.  
             codec 153  
     with Bloom 102  
 mapping 12  
 mapred.join package 117  
 mapred.min.split.size 55  
 mapred-site.xml 28  
 MapReduce 8–14  
     algorithms, efficient 157  
     chaining jobs 103  
     configuration details 35  
     data flow 45  
     EC2 and 203, 205  
     framework 19  
     key/value pairs 38  
     partitioning and shuffling  
         50  
     program anatomy 44–51  
     reading and writing  
         51–58  
     tasks 156  
     tuning for performance 152  
     web interface 35  
     writing to database 170  
 MapReduceBase 47, 71, 119  
 maps 256  
 map-side join 117  
 master files 32  
 master/slave architecture 22  
     copying public key 27  
     Hadoop cluster 25  
 MAX 225  
 maximum 84, 86, 90,  
     93–95, 98  
 median 86, 93–94  
 memcached 133  
 MemStore 275  
 messaging queues 8  
 meta block index 280  
 metadata  
     block size 181  
     partitioning 167  
 metastore 247  
 MIN 225  
 minimum 86, 93–94  
 missing values 65, 88  
 mkdir 40  
 mkdir, command 300  
 Modulo 224  
 Moore's law 6  
 move 6  
 moveFromLocal, command  
     300  
 moveToLocal, command 300  
 moving average 99  
 multiple machines 11  
 MultipleOutputFormat  
     164, 166  
 MultipleOutputs 166  
     filename generation 168  
     naming structure 169  
 multiple parameter files 238  
 multiple relations 227  
 MultipleTextOutputFormat  
     164  
 multiquery 239  
 multiset 10  
 mutates 278  
 mv, command 300  
 MySQL database  
     at StumbleUpon 277  
     legacy data 275  
     metadata store  
         configuration 250

## N

NameNode 140, 142  
     backup node 184  
     DataNode interaction 23  
     dfsadmin 178  
     failure, recovering 183–184  
     multiple hard drives 184  
     name quotas 179  
     overview 22  
     RAID 181  
     name quotas 179  
 National Bureau of Economic  
     Research 64  
 natural language processing  
     11, 99, 118

navigational 283  
 NBER 64, 90  
 nested data types 223  
 nested schema 223  
 nesting 222  
 Netflix 101  
 network topology 184–185  
 New York Times 4, 267  
 next 55, 57  
 nodes 50  
 null 224  
 NullOutputFormat 58  
 nulls 251  
 NullWritable 58  
 numSplits 55  
 Nutch 19, 263  
   ES2 283  
   generate, fetch, update 287  
   StumbleUpon 282  
   web crawling 287

## O

---

offline analysis 285  
 offline processing 8  
 open source framework 4  
 operators, grouping 226  
 org.apache.hadoop.fs 42  
 org.apache.hadoop.  
   io.compress 153  
 outer join 112  
   datajoin package 114  
 output  
   files 57  
   in total sorting 172  
   sorting 171  
 OutputCollector 48, 71, 107,  
   128–129, 166  
 OutputCollector.collect  
   (K k, V v) 107  
 OutputFormat 57  
   classes 58  
   FileOutputFormat 154  
   NullOutputFormat 129  
   SequenceFileOutput  
   Format 154  
   TextOutputFormat 75, 87  
 OutputFormat class 164  
 outputs, multiple 164

## P

---

page\_view 255  
 pair-wise combinations 242  
 patents example 240

pair-wise computations 240  
 parallelization, granular 52  
 PARALLEL n 231  
 parameters  
   for practical use 174–176  
   multiple files 238  
   passing custom 160  
   security 201  
   substitution 238  
 partitioner 87, 166, 172  
   hash 172  
   Mapper output redirecting  
   49  
   TotalOrderPartitioner 172  
 partitioning 11  
   data 254  
   Hive 248  
   into multiple output  
   files 164  
   LOAD DATA 258  
   metadata 165  
 pass by reference, in output  
   collector 107  
 pass by value, in output  
   collector 107  
 patent citation data 64–65,  
   107, 131  
 patent description data  
   64–65, 107  
 patents  
   cocitation 240  
   cocitation counts 244  
 path 39  
 Path 43  
 pattern matching 224  
 performance improvement  
   107  
 Perl 282  
 permissions, setting 178  
 petabytes 3  
 Pig  
   and Cascading 263  
   Cygwin 214  
   data types 213  
   Hadoop cluster 214  
   installing 214–215  
   interactive shell 215  
   Java API 235  
   JAVA\_HOME 214  
   JobTracker 214  
   JVM 215  
   like SQL queries 215  
   multiquery execution 239  
   NameNode 214  
   parameters, setting 216  
   philosophy 213, 215  
   running 215–217  
   similar patents 240–244  
   UDFs and 214  
 PIG\_CLASSPATH 214  
 PiggyBank 233  
   UDF 236  
 piggybank.jar 234  
 PIG\_HADOOP\_VERSION 214  
 Pig Latin 213, 217–233  
   APIs 233  
   atomic data types 222  
   built-in expressions 225  
   comments 237  
   complex data types 222  
   data flow language 213  
   data read/write operators  
   218  
   data types and schemas  
   221–223  
   diagnostic operators 221  
   execution optimization 233  
   expressions 224  
   expressions and functions  
   223–225  
   Hadoop mode 215  
   Java equivalent classes 235  
   like SQL query 219  
   local mode 215  
   relational operators 225–233  
   schema 219  
   script 213  
   scripts 237–239  
   type 235  
   UDF statements 234  
   user-defined functions 233,  
   235–237  
 pig.properties file 214  
 PigServer 215  
 pipelines 8  
 pipes, Cascading 263  
 pirate-speak 100  
 pools 187  
 POSIX 178  
 pound operator 224  
 power law distribution 76  
 prefix 238  
 primitives 8  
 private key 197, 198  
   SSH 26, 200  
 production cluster 174  
 pseudo-code 9  
 pseudo-distributed  
   mode 29, 39, 41, 64,  
   135, 140

public key 197  
   cryptography 26  
   distribution 27  
   EC2 200  
   SSH 200  
 put 39  
 put, command 300  
 PutMerge 42, 43

## Q

QOS 268  
 quadratic complexity 240  
 quality of service. *See* QOS  
 queries, column 254  
 query processing  
   Hive 248  
   Hive examples 250  
 quotas, managing 179

## R

rack awareness 184–185  
 Raghavan, Sriram 282  
 RAID 181  
 RAM storage 10  
 Rawson, Ryan 272  
 RDBMS 276  
 readFields 47  
 read-many-times 8  
 read-slaves 278  
 RecordReader 54–55  
   joining 117  
 RecordReader<LongWritable,  
   Text> 55  
 records  
   processing guaranteed 52  
   skipping bad 148  
 RedHat, Cloudera 263  
 reduce 48, 139  
 Reduce, class 71  
 reducer 8, 48, 139  
   multiple 49  
   partitioner and 172  
   ResolveSimple 290  
 Reducer 48, 69, 97, 160  
   class list 49  
   reduce() 71, 97  
 reduce-side join 108–115  
 reducing 12  
 Reflection 236  
 REGEXP 259  
 region server 275  
   HBase 278  
 REGISTER 234

regression testing 138–140  
 regular expression patterns  
   295  
 Reiss, Frederick 282  
 relational data 64–65, 108  
   aggregate analysis 227  
   model 223  
   nonmatching 227  
 relational database 7, 152  
   interfacing with 169  
 remote storage 184  
 repartitioned join 108  
 replicated join 118–121  
 replication factor 40  
 Reporter 48, 142, 146  
   incrCounter 146  
   setStatus 142, 145  
 Resolve2Step 290  
   resolution table 291  
 ResolveSimple 290  
 results 293  
 rm 41  
 rm, command 300  
 rmr, command 300  
 ROW FORMAT 256  
 row store 275  
 RPM 263

## S

S3 194  
   accessing data directly 207  
   dataflow model 206  
   datapath 207  
   GUI tools 199  
   Hadoop options 207  
   moving data through 205  
   New York Times 267  
   storage service 202  
 S3 Block FileSystem 206  
 S3 bucket 202, 206  
 s3n 208  
 sampling data 64, 82, 84,  
   135, 152  
 sanity checking 137–138  
 scheduler 189  
 schema 7  
   data source records 108  
   fixed 219  
   nested 223  
 scheme 39  
 Scribe 281  
 search engine optimization  
   281  
 search-hadoop.com 264  
 Secondary NameNode  
   23, 181–183  
   confusing name 182  
   location specification 30  
   server 183  
   snapshots 23, 182  
 SecondaryNameNode 140,  
   142  
 Secret Access Key 198  
   slashes and 195  
   slash problem 208  
 Secure Shell. *See* SSH  
 security, parameters 201  
 segment 287  
 segment file 289  
 SELECT 251  
 Sematext 264  
 semijoin 108, 121–122, 131  
 semistructured data 223  
 SequenceFileInputFormat 54  
 SequenceFileOutputFormat  
   58  
 sequence file 54, 154–155  
   configuration properties  
   155  
   SequenceFileOutput-  
   Format 154  
 set command 216  
 SETI@home 6  
 setOutputFormat 57  
 setrep, command 301  
 setSpaceQuota command 179  
 Shanghai Branch 270  
   K-means 272  
 Shekita, Eugene 282  
 SHOW TABLES 251, 257  
 shuffle 45  
 shuffle, MapReduce 87  
 shuffling 11, 50  
 sign 224  
 signal processing 99  
 Simple Storage Service. *See* S3  
 site root analysis 291, 293  
 SIZE 225  
 SkipBadRecords 149, 150  
 skipping bad records 148–150  
 slave machines 25  
 slave nodes 186  
   location specification 30  
 slaves files 32  
 SMALLINT 255  
 SNN. *See* Secondary NameNode  
 Snoop Dogg 100  
 social networks 64  
 Solaris 14

- sorted order 171
  - spam filter 10
  - sparse data 240
  - sparse vector representation 99, 132
  - spatial join 132
  - speculative execution 156
    - configuration properties 157
  - SPLIT 225, 237
  - splits
    - Bloom filters and 128
    - input 52
    - InputFormat class 55
  - splittable file formats 154
  - SQL 7
    - and Hive 213
    - engine 8
    - vs. Pig Latin 213
  - Squid, Web proxy server 124
  - SSH
    - channels 30
    - installation 26
  - SSH login 200
  - SSN. *See* Secondary NameNode
  - standalone mode 28, 64, 135
  - standalone node 25
  - standard deviation 93
  - Stanford 4
  - stat, command 301
  - status messages 142, 145
  - STDERR 142, 148
  - STDOUT 142
  - stemming 104
  - stop words 104
  - STORE 217
  - Streaming 80–94
    - configuration properties 164
    - outputting logs 142
    - PHP scripting 84
    - Python scripting 82, 89
    - rewriting for Hadoop
      - Java 157
    - setting job name 143
    - skipping bad records 150
    - split processing 102
    - Unix commands 81–82
    - updating status 142
    - use of combiner 96
    - use of compressed files 155
    - use of counters 148, 150
  - Streamy.com 280
  - String 161
  - STRING 256
  - StringValueMax 90
  - StringValueMin 90
  - structs 256
  - structured query language. *See* SQL
  - Stumblers 272
  - Stumbles, per user 276
  - StumbleUpon 272–273, 275, 277, 279, 281–282
    - architecture 273
    - data handling 281
    - HBase 274–275
    - ratings 272
  - SUM 225
  - summarization 247
  - summing 86, 90
  - Sun, Shaoling 267
  - supergroup 178
  - superuser 178
  - symbolic links 33
  - sync markers in files 154
  - system malfunction 182
- T**
- 
- tab character 58
  - tables, splitting, HBase 278
  - tag, reduce-side join 108
  - TaggedMapOutput 112, 115
  - TaggedWritable datajoin 113
  - tail 41
  - tail, command 301
  - taps 263
  - tarball, Hive 248
  - task attempt ID 151
  - task ID 145
  - task initialization 155
  - tasks
    - CPU-intensive loads 175
    - failed, rerunning 156
    - parallel running 156
    - passing parameters 160
    - placement 185
    - reduce, side effects 157
    - retrieving information 163
  - TaskTracker 140, 269
    - bad record tracking 148
    - configuration object 161
    - default tasks 175
    - JobTracker interaction 24
    - overview 24
    - RAM 175
    - Web UI 145
  - Tata, Sandeep 282
  - template, for MapReduce program 67, 118
  - terabytes 3
  - test, command 301
  - Text 17, 54, 72
    - datajoin 112
  - text, command 301
  - TEXTFILE 257
  - TextInputFormat 52
  - TIFF, image processing 267
  - time series 64, 99
  - TimeUrlLineRecordReader 56
  - TINYINT 255
  - TokenCountMapper 51
  - tokenization 16
  - TOKENIZE 225
  - TokyoCabinet 281
  - Tool 70, 162
  - Tool.run() 70
  - ToolRunner 70, 162
  - toString 58
  - touchz, command 301
  - trash 179
  - trigram 11
  - Tuple 222
  - tuples
    - Cascading 263
    - Pig and 213
  - Twitter 4
  - type casting 224
- U**
- 
- Ubuntu 263
  - UCI data sets 270
  - UDFs 213
    - eval 234
    - filter functions 237
    - jar file 234
    - load/store 234
    - Pig Latin and 233
    - simple scalar 236
  - unique values, computing 92–93
  - UniqValueCount 90
  - Unix
    - commands, parameter values 238
    - cost breakdown 270
    - Hive CLI 258
    - pipes 8, 41, 103
    - server 268
    - server performance 270
    - wc-1 136
  - Unix command 137
  - UPPER UDF 235
  - URI, format 39

user-defined functions. *See* UDFs  
username 178

## V

---

Vaithyanathan, Shivakumar 282  
ValueHistogram 90, 93  
variance, computing 86, 158  
    variant generation 285  
visualization 65, 77, 92  
void close 47  
void configure(JobConf job) 47

## W

---

web  
    interfaces 34  
    search engine 19  
Web GUI 247

web server, log analysis 99, 131  
Web UI 98, 142, 147, 189  
    cluster 34–36  
    Hadoop 216  
whitespace characters 16  
Windows 14  
word count 9, 72  
    predefined classes 51  
    with Hadoop 14–19  
Writable 72, 170  
    Bloom filter 126  
    datajoin 113  
    IntWritable 73, 140  
    LongWritable 140  
    Text 73  
WritableComparable 72  
    example 47  
WritableComparable<T> 46  
write-once 8

## X

---

X.509 Certificate 195, 198  
Xu, Meng 267

## Y

---

Yahoo 4, 212

## Z

---

Zhu, Huaiyu 282  
ZooKeeper 262, 278