

# Index

## A

Adaptive, Hash-partitioned Exact Window Join (AH-EWJ) 153  
 adaptive sampling 95, 105, 109  
 Apache Lucene 205, 213  
 Apache Software Foundation 197, 199, 203-205, 218, 220, 291  
 Application Programming Interface (API) 3, 276  
 articulated network 113  
 Atacama Large Millimeter Array (ALMA) 197-198  
 Australian Square Kilometre Array Precursor (ASKAP) 203  
 Australia Telescope National Facility (ATNF) 203  
 average length 390

## B

band-pass filter (BPF) 341  
 behavioral network 113  
 benchmarking 416, 418, 425  
 benchmarks 1, 12-14, 21, 155, 163, 178, 250, 255, 381, 384-385, 390-392, 394-395, 415-432  
 big data analytics 1-2, 12, 47-51, 54-55, 57, 63, 65-66, 107, 112, 223-224, 265, 268, 418, 429, 431  
 blog audience 114  
 blogosphere 110-112, 114  
 Burrows-Wheeler transform (BWT) 352, 366  
 bursty

## C

Cache Join (CACHEJOIN) 150  
 Capacity Utilization Factor 398  
 Center for Large-scale Data Systems Research (CLDS) 427  
 Central Limit Theorem 407, 412  
 centroid 24-25, 31-33, 40-42, 322, 324-327  
 cluster sampling 102-103

combiner 4-5, 32, 34, 39, 42-43, 322, 326  
 Common Warehouse Metamodel (CWM) 387  
 Commonwealth Scientific and Industrial Research Organization (CSIRO) 203  
 Complex Event Processing (CEP) 10  
 composite excess entropy (CEE) 407, 412  
 conceptual representation 149  
 constant memory 354, 368  
 Content Standard for Digital Geospatial Metadata (CSDGM) 188  
 CPU virtualization 384-385  
 Creative Commons Zero (CC0) 184  
 cyberinfrastructure 190, 193, 270-271, 273-276, 278-279, 282, 284, 286, 289-290, 292-293

## D

data access libraries (DA) 307  
 Data Condensation 298, 306  
 data dictionary 125-127, 137, 139  
 Data Documentation Initiative (DDI) 186, 189  
 datanode 26, 32, 347  
 Data Reduction 72-74, 77-82, 84, 88-89, 92-93, 215, 298, 300, 306-307  
 data warehouse 15, 19, 26, 115-119, 124-125, 128-129, 133, 135-136, 141-142, 144-151, 175, 204, 220, 383, 386-388, 390-393, 395  
 Digital Humanities 270-271, 273-274, 277-280, 282, 284, 290-293  
 Digital libraries 191, 270-272  
 Digital Signal Processing 318, 345  
 discrete convolution theorem 334  
 Discrete Fourier Transform (DFT) 79, 320  
 Discrete Wavelet Transform (DWT) 80  
 discretization 335  
 Distributed Memory System (DMS) 82, 92  
 Double Pipelined Hash Join (DPHJ) 153  
 Dryad 177, 183-184, 190-192, 194  
 DryadLINQ 25, 44

## **E**

EA and a Tabu Search heuristic (EA-TS) 228, 244  
Early Hash Join (EHJ) 153  
Enterprise Data Warehouse (EDW) 19  
eScience 179, 190, 192, 205  
evolutionary algorithm (EA) 228  
excess entropy (EE) 400, 412  
Expanded Very Large Array (EVLA) 197-198, 200, 218, 220  
Extract, Transform and Load (ETL) 19  
extract, transform, load (ETL) 298

## **F**

Fast Fourier Transform (FFT) 334  
Federal Geographic Data Committee (FGDC) 188  
federated cloud 54, 71  
federation 50, 57, 61-62, 66, 208-209, 214, 218  
Fine Grained Tournament Selection (FGTS) 240  
Flexible Extensible Digital Object Repository Architecture 190  
frequency domain 80, 320-321, 347

## **G**

genome-wide association studies (GWAS) 349-350  
GeoBase 295-297, 300-309, 311  
Geospatial Data Abstraction Library (GDAL) 296  
global conceptual schema 119, 126-127, 141  
Google's File System (GFS) 425  
Graphics and Workstation Performance Group (GWPG) 420  
Graphics Processing Units (GPUs) 73, 354  
GrHyMM 118, 146

## **H**

Hadoop file system (HDFS) 302  
Hash-Merge Join (HMJ) 153  
HathiTrust Research Center (HTRC) 270-271, 274, 277  
HBase 8-11, 20-21, 284, 296-298, 301-306, 309, 311, 313, 344-345  
HFile 8  
High Performance Computing (HPC) 107  
High Performance Group (HPG) 420  
Hilbert curves 92, 300, 312  
hybrid approach 115, 124, 135, 147, 149, 241

## **I**

Independent Publishers Guild (IPC) 100  
In-Memory Database 17, 22  
Intel Developers Forum (IDF) 424  
Interuniversity Consortium for Political and Social Research (ICPSR) 177  
iRODS 190, 212

## **J**

JobTracker 3-4, 11

## **K**

Karoo Array Telescope (KAT-7) 202  
k-means 23-25, 31-33, 37-38, 40, 42-46, 322-324, 326-328

## **L**

lexical analyzer (LA) 129  
Linear Time-Invariant 339  
load balancing 8, 49-56, 58, 64-65, 67-69, 71, 305-306, 308-310  
Local Search (LS) 242  
logical program 115, 118, 129-131, 134-135, 142, 149  
low-end local worker (LLW) 62  
low-frequency component 80-81, 87-89  
low-pass filter 331

## **M**

mapper 2-3, 11-12, 25, 27, 32, 41-43, 321-322, 325-326, 332-334  
Map phase 4, 46, 321  
Maximally Overlapped Bin-packing (MOB) 58, 64-65  
Mesh Join (MESHJOIN) 150, 153  
metaheuristic 222, 224, 228, 235, 242, 250-251, 255, 262, 269  
Metaheuristic 222, 269  
METAHEURISTIC 234  
metaheuristics 228, 234-235, 250, 257, 262, 267-268  
Metaheuristics 269  
micro-blog 95, 109, 112, 114  
mid-end remote worker (MRW) 62  
Minimum Description Length (MDL) 334  
minor allele frequency (MAF) 349-350, 352

## N

namenode 26, 32  
 National Aeronautics and Space Administration (NASA) 182  
 National Climate Data Center (NCDC) 182  
 National Institutes of Health (NIH) 178  
 National Oceanographic Data Center (NODC) 177, 183  
 National Radio Astronomy Observatory (NRAO) 196-198, 200, 203, 206-207, 217  
 National Science Foundation (NSF) 178  
 NetCDF-Java library 306  
 Network Attached Storage (NAS) 17  
 Newton's law of cooling 340  
 Non Uniform Memory Architecture (NUMA) 17  
 normalized excess entropy (nEE) 404, 412  
 NoSQL 1-2, 7-10, 12-14, 21, 213, 284, 288, 291-292, 313, 417-418, 426

## O

Object Oriented Data Technology 203, 220  
 Ocean Acidification Data Stewardship (OADS) 187  
 Ocean Acidification Program (OAP) 187  
 Ocean Archive System (OAS) 188  
 Online Analytics Processing (OLAP) 17  
 On Line Transaction Processing (OLTP) 416  
 ontological representation 126, 136, 149  
 Open Information Model (OIM) 386  
 Open Source Development Lab (OSDL) 424  
 Open Systems Group (OSG) 420  
 optical character recognition software (OCR) 282  
 Oracle Application Standard Benchmark 423  
 outcoming flow 229-230, 238

## P

parallelization 47-49, 53, 56-60, 66, 71, 90, 283  
 parallel processing techniques 82, 89, 93  
 Performance Impact Factor (PIF) 398  
 power iteration clustering (PIC) 223  
 predicate calculus 118, 128, 146, 149  
 Project Gutenberg 272, 292  
 Project MUSE 180  
 Publishers Association (PA) 100  
 purposeful sampling 100, 103  
 Python 26, 28, 32, 44-46, 206

## R

real-time data warehousing 150-151, 154  
 Reduce phase 43, 46, 322  
 reengineering process 115-116, 118, 145, 149  
 Relational Database Management System (RDBMS) 107  
 relational database system (RDBMS) 284  
 Remote Procedure Calls (RPC) 3  
 Representational State Transfer (REST) 279  
 Research Group (RG) 420, 4255

## S

scale-space image 329-331, 335  
 schemas integration 145, 149  
 Script Workflow Analysis for MultiProcessing (SWAMP) 84  
 Semi-Streaming Index Join (SSIJ) 154  
 semi-structured data 107, 114, 149, 390, 429  
 sensor data analysis 314-315, 318  
 Shared Memory System (SMS) 82, 93  
 single-nucleotide polymorphism (SNP) 349-350  
 Single Program Multiple Data (SPMD) 24, 46  
 site frequency spectrum (SFS) 353-354  
 Solr 197, 199, 205-206, 211, 213-214, 216-218, 221, 277, 286-289  
 space-filling curves (SFC) 298  
 Space Telescope Science Institute (STScI) 182  
 Square Kilometre Array (SKA) 197-198, 202  
 Standard Performance Evaluation Corporation (SPEC) 415-416, 418, 420  
 Storage Area Network (SAN) 14, 17  
 storage area networks (SANs) 417  
 Storage Performance Council (SPC) 416, 418, 421  
 stratified sampling 75, 105-106  
 Streaming Multiprocessors (SMs) 354  
 stream processing 150-151, 159  
 structured data 8, 14, 16, 20, 83, 107, 114, 119, 137, 149, 181, 223, 291, 311, 318, 390, 428-429, 432  
 Structured Query Language (SQL) 3  
 subsequence clustering (SSC) 322  
 suffix array (SA) 366  
 super-band matrix 358-360  
 Symmetric Hash Join (SHJ) 153  
 Symmetric Multi Processing (SMP) 19  
 synchronization 55, 66, 71, 84, 176, 289, 349, 362-363, 367

## ***Index***

syntactical analyzer (SA) 129  
systematic sampling 74-75, 101-102  
system under test (SUT) 428

## **T**

TableMapper 9  
TableReducer 9  
Tabu Search heuristic (TS) 234  
TaskTracker 3  
Tika 197, 199, 205, 207-208, 211, 217-219, 221  
Time-invariant 339  
token 129-130, 276, 284  
TPC's Technology Conference Series (TPCTC) 425  
Transaction Processing Performance Council (TPC)  
415-416, 418

## **V**

V-FASTR 208-210, 213-216, 220-221  
virtual machine 384-385, 393, 395-396, 410, 423  
Virtual Research Environment (VRE) 179, 194  
VMmark® 423

## **W**

warp scheduler 354  
wavecluster algorithm 82, 85-87, 89-90  
Web 2.0 114, 417, 424-425  
Workshop Series on Big Data Benchmarking  
(WBDB) 426-427  
wrap-around problem 332, 334

## **Y**

Yahoo! Cloud Serving Benchmark (YCSB) 13, 424

## **Z**

zipfian distribution 152, 163, 165-166, 173, 309  
Z-ordering (ZO) 308  
Z-regions 298, 303-305, 307