

contents

foreword xv
preface xvii
acknowledgments xix
about this book xxi

PART 1 MEET SOLR.....1

I *Introduction to Solr* 3

1.1 Why do I need a search engine? 4

Managing text-centric data 4

Common search-engine use cases 7

1.2 What is Solr? 9

Information retrieval engine 11 ▪ *Flexible schema management* 13 ▪ *Java web application* 13

Multiple indexes in one server 15 ▪ *Extendable (plugins)* 15

Scalable 15 ▪ *Fault-tolerant* 16

1.3 Why Solr? 17

Solr for the software architect 17 ▪ *Solr for the system administrator* 18 ▪ *Solr for the CEO* 19

- 1.4 Features overview 19
 - User-experience features* 19
 - *Data-modeling features* 21
 - New features in Solr 4* 23
- 1.5 Summary 24

2 **Getting to know Solr** 26

- 2.1 Getting started 27
 - Installing Solr* 27
 - *Starting the Solr example server* 28
 - Understanding Solr home* 32
 - *Indexing the example documents* 33
- 2.2 Searching is what it's all about 34
 - Exploring Solr's query form* 34
 - *What comes back from Solr when you search* 38
 - *Ranked retrieval* 39
 - *Paging and sorting* 40
 - Expanded search features* 41
- 2.3 Tour of the Solr administration console 43
- 2.4 Adapting the example to your needs 45
- 2.5 Summary 46

3 **Key Solr concepts** 48

- 3.1 Searching, matching, and finding content 49
 - What is a document?* 49
 - *The fundamental search problem* 50
 - The inverted index* 53
 - *Terms, phrases, and Boolean logic* 54
 - Finding sets of documents* 56
 - *Phrase queries and term positions* 59
 - *Fuzzy matching* 60
 - *Quick recap* 65
- 3.2 Relevancy 65
 - Default similarity* 65
 - *Term frequency* 67
 - Inverse document frequency* 68
 - *Boosting* 69
 - Normalization factors* 69
- 3.3 Precision and Recall 71
 - Precision* 72
 - *Recall* 73
 - *Striking the right balance* 73
- 3.4 Searching at scale 74
 - The denormalized document* 75
 - *Distributed searching* 77
 - Clusters vs. servers* 78
 - *The limits of Solr* 79
- 3.5 Summary 80

4 **Configuring Solr** 82

- 4.1 Overview of solrconfig.xml 85
 - Common XML data-structure and type elements* 87
 - Applying configuration changes* 87
 - *Miscellaneous settings* 88

- 4.2 Query request handling 90
 - Request-handling overview* 90
 - *Search handler* 93
 - Browse request handler for Solr: an example* 94
 - Extending query processing with search components* 98
- 4.3 Managing searchers 103
 - New searcher overview* 103
 - *Warming a new searcher* 104
- 4.4 Cache management 107
 - Cache fundamentals* 107
 - *Filter cache* 109
 - Query result cache* 112
 - *Document cache* 113
 - Field value cache* 113
- 4.5 Remaining configuration options 114
- 4.6 Summary 114

5 Indexing 116

- 5.1 Example microblog search application 117
 - Representing content for searching* 117
 - Overview of the Solr indexing process* 119
- 5.2 Designing your schema 121
 - Document granularity* 121
 - *Unique key* 122
 - Indexed fields* 123
 - *Stored fields* 123
 - Preview of schema.xml* 124
- 5.3 Defining fields in schema.xml 125
 - Required field attributes* 126
 - *Multivalued fields* 127
 - Dynamic fields* 128
 - *Copy fields* 131
 - *Unique key field* 133
- 5.4 Field types for structured nontext fields 133
 - String fields* 134
 - *Date fields* 135
 - *Numeric fields* 137
 - Advanced field type attributes* 138
- 5.5 Sending documents to Solr for indexing 141
 - Indexing documents using XML or JSON* 141
 - *Using the Solr client library to add documents from Java* 144
 - *Other tools for importing documents into Solr* 146
- 5.6 Update handler 147
 - Committing documents to the index* 148
 - *Transaction log* 151
 - Atomic updates* 152
- 5.7 Index management 155
 - Index storage* 155
 - *Segment merging* 158
- 5.8 Summary 160

6 Text analysis 162

- 6.1 Analyzing microblog text 163
- 6.2 Basic text analysis 167
 - Analyzer* 168 ▪ *Tokenizer* 168 ▪ *Token filter* 169
 - StandardTokenizer* 169 ▪ *Removing stop words with StopFilterFactory* 170 ▪ *LowerCaseFilterFactory—lowercase letters in terms* 171 ▪ *Testing your analysis with Solr's analysis form* 172
- 6.3 Defining a custom field type for microblog text 174
 - Collapsing repeated letters with PatternReplaceCharFilterFactory* 177
 - Preserving hashtags, mentions, and hyphenated terms* 178
 - Removing diacritical marks using ASCIIFoldingFilterFactory* 182
 - Stemming with KStemFilterFactory* 182 ▪ *Injecting synonyms at query time with SynonymFilterFactory* 183 ▪ *Putting it all together* 184
- 6.4 Advanced text analysis 187
 - Advanced field attributes* 187 ▪ *Per-language text analysis* 189
 - Extending text analysis using a Solr plug-in* 190
- 6.5 Summary 194

PART 2 CORE SOLR CAPABILITIES 195

7 Performing queries and handling results 197

- 7.1 The anatomy of a Solr request 198
 - Request handlers* 198 ▪ *Search components* 203
 - Query parsers* 206
- 7.2 Working with query parsers 207
 - Specifying a query parser* 207 ▪ *Local params* 207
- 7.3 Queries and filters 210
 - The fq and q parameters* 210 ▪ *Handling expensive filters* 213
- 7.4 The default query parser (Lucene query parser) 215
 - Lucene query parser syntax* 215
- 7.5 Handling user queries (eDisMax query parser) 222
 - eDisMax query parser overview* 222 ▪ *eDisMax query parameters* 223 ▪ *Searching across multiple fields* 223
 - Boosting queries and phrases* 224 ▪ *Field aliasing* 226
 - User-accessible fields* 227 ▪ *Minimum match* 228
 - eDisMax benefits and drawbacks* 230

7.6 Other useful query parsers 232

Field query parser 232 ▪ *Term and Raw query parsers* 232
Function and Function Range query parsers 233 ▪ *Nested queries and the Nested query parser* 233 ▪ *Boost query parser* 234
Prefix query parser 235 ▪ *Spatial query parsers* 235
Join query parser 236 ▪ *Switch query parser* 236
Surround query parser 236 ▪ *Max Score query parser* 237
Collapsing query parser 238

7.7 Returning results 238

Choosing a response format 238 ▪ *Choosing fields to return* 240
Paging through results 243

7.8 Sorting results 245

Sorting by fields 245 ▪ *Sorting by functions* 247
Fuzzy sorting 247

7.9 Debugging query results 248

Returning debug information 248

7.10 Summary 249

8 *Faceted search* 250

8.1 Navigating your content at a glance 251

8.2 Setting up test data 254

8.3 Field faceting 259

8.4 Query faceting 264

8.5 Range faceting 266

8.6 Filtering upon faceted values 269

Applying filters to your facets 269

Safely filtering on faceted values 273

8.7 Multiselect faceting, keys, and tags 275

Keys 275 ▪ *Tags, excludes, and multiselect faceting* 277

8.8 Beyond the basics 279

8.9 Summary 280

9 *Hit highlighting* 281

9.1 Overview of hit highlighting 282

9.2 How highlighting works 283

Set up a new Solr core for UFO sightings 284 ▪ *Preprocess UFO sightings before indexing* 284 ▪ *Exploring the UFO sightings dataset* 286 ▪ *Hit highlighting out of the box* 288
Nuts and bolts 290 ▪ *Refining highlighter results* 296

- 9.3 Improving performance using FastVectorHighlighter 300
- 9.4 PostingsHighlighter 302
- 9.5 Summary 304

10 Query suggestions 306

- 10.1 Spell-check 307
 - Indexing Wikipedia articles* 307
 - Spell-check example* 309
 - Spell-check search component* 311
- 10.2 Autosuggesting query terms 318
 - Autosuggest request handler* 318
 - Autosuggest search component* 320
- 10.3 Suggesting document field values 321
 - Using n-grams for suggestions* 321
 - N-gram-driven request handler* 323
- 10.4 Suggesting queries based on user activity 324
- 10.5 Summary 329

11 Result grouping/field collapsing 330

- 11.1 Result grouping vs. field collapsing 331
- 11.2 Skipping duplicate documents 332
- 11.3 Returning multiple documents per group 339
- 11.4 Grouping by functions and queries 343
 - Grouping by function* 343
 - Grouping by query* 345
- 11.5 Paging and sorting grouped results 347
- 11.6 Grouping gotchas 348
 - Faceting upon result groups* 349
 - Distributed result grouping* 352
 - Returning a flat list* 352
 - Grouping on multivalued and tokenized fields* 352
 - Grouping performance* 353
- 11.7 Efficient field collapsing with the collapsing query parser 353
- 11.8 Summary 355

12 Taking Solr to production 356

- 12.1 Developing a Solr distribution 357
- 12.2 Deploying Solr 357
 - Building your Solr distribution* 358
 - Embedded Solr* 359

- 12.3 Hardware and server configuration 359
 - RAM and SSDs* 359 ▪ *JVM settings* 360
 - The index shuffle* 361 ▪ *Useful system tricks* 365
- 12.4 Data acquisition strategies 367
- 12.5 Sharding and replication 371
 - Choosing to shard* 371 ▪ *Choosing to replicate* 375
- 12.6 Solr core management 378
- 12.7 Managing clusters of servers 384
 - Load balancers and Solr health check* 384
 - Generic vs. customized configuration* 385
- 12.8 Querying and interacting with Solr 388
 - REST API* 388 ▪ *Available Solr client libraries* 388
 - Using SolrJ from Java* 389
- 12.9 Monitoring Solr's performance 392
 - Solr's Plugins / Stats page* 393 ▪ *Solr cache performance* 396
 - Pulling stats from request handlers and MBeans* 398
 - External monitoring options* 399 ▪ *Solr logs* 400
 - Load testing* 400
- 12.10 Upgrading between Solr versions 401
- 12.11 Summary 402

PART 3 TAKING SOLR TO THE NEXT LEVEL.....403

13 SolrCloud 405

- 13.1 Getting started with SolrCloud 406
 - Starting Solr in cloud mode* 406 ▪ *Motivation behind the SolrCloud architecture* 411
- 13.2 Core concepts 416
 - Collections vs. cores* 416 ▪ *ZooKeeper* 417 ▪ *Choosing the number of shards and replicas* 421 ▪ *Cluster-state management* 422 ▪ *Shard-leader election* 423
 - Important SolrCloud configuration settings* 424
- 13.3 Distributed indexing 427
 - Document shard assignment* 428 ▪ *Adding documents* 429
 - Near real-time search* 431 ▪ *Node recovery process* 433
- 13.4 Distributed search 433
 - Multistage query process* 434
 - Distributed search limitations* 436

- 13.5 Collections API 436
 - Create a collection* 436 ▪ *Collection aliasing* 440
- 13.6 Basic system-administration tasks 442
 - Configuration updates* 443 ▪ *Rolling restart* 443
 - Restarting a failed node* 444 ▪ *Is node X active?* 444
 - Adding a replica* 444 ▪ *Offsite backup* 445
- 13.7 Advanced topics 446
 - Custom hashing* 446 ▪ *Shard splitting* 447
- 13.8 Summary 449

14 Multilingual search 450

- 14.1 Why linguistic analysis matters 451
- 14.2 Stemming vs. lemmatization 452
- 14.3 Stemming in action 454
- 14.4 Handling edge cases 458
 - KeywordMarkerFilterFactory* 459
 - StemmerOverrideFilterFactory* 459
- 14.5 Available language libraries in Solr 460
 - Language-specific analyzer chains* 460
 - Dictionary-based stemming (Hunspell)* 463
- 14.6 Searching content in multiple languages 464
 - Separate field per language* 464 ▪ *Separate index per language* 470 ▪ *Multiple languages in one field* 473
 - Creating a field type to handle multiple languages per field* 474
- 14.7 Language identification 485
 - Update processors for language identification* 486
 - Dynamically assigning detected language analyzers within a field* 494
- 14.8 Summary 499

15 Complex query operations 501

- 15.1 Function queries 502
 - Function syntax* 502 ▪ *Searching on functions* 504
 - Returning functions like fields* 507 ▪ *Sorting on functions* 508
 - Available functions in Solr* 509 ▪ *Implementing a custom function* 515
- 15.2 Geospatial search 521
 - Searching near a single point* 521
 - Advanced geospatial search* 527

- 15.3 Pivot faceting 538
- 15.4 Referencing external data 541
- 15.5 Cross-document and cross-index joins 543
- 15.6 Big data analytics with Solr 546
- 15.7 Summary 547

16 *Mastering relevancy* 548

- 16.1 The impact of relevancy tuning 549
 - 16.2 Debugging the relevancy calculation 550
 - 16.3 Relevancy boosting 556
 - Per-field boosting* 556 ▪ *Per-term boosting* 558
 - Payload boosting* 559 ▪ *Function boosting* 560
 - Term-proximity boosting* 562 ▪ *Elevating the relevancy of important documents* 564
 - 16.4 Pluggable Similarity class implementations 567
 - 16.5 Personalized search and recommendations 569
 - Search vs. recommendations* 570 ▪ *Attribute-based matching* 571 ▪ *Hierarchical matching* 573
 - More Like This* 574 ▪ *Concept-based matching* 579
 - Geographical matching* 585 ▪ *Collaborative filtering* 586
 - Hybrid approaches* 590
 - 16.6 Creating a personalized search experience 591
 - 16.7 Running relevancy experiments 592
 - 16.8 Summary 595
- appendix A Working with the Solr codebase* 596
- appendix B Language-specific field type configurations* 605
- appendix C Useful data import configurations* 610
- index* 616