Apache Lucene 5 New Features and Improvements for Apache Solr and Elasticsearch

Uwe Schindler Apache Software Foundation | SD DataSolutions GmbH | PANGAEA

My Background

- Committer and PMC member of Apache Lucene and Solr main focus is on development of Lucene Core.
- Implemented fast numerical search and maintaining the new attribute-based text analysis API. Well known as Generics and Sophisticated Backwards Compatibility Policeman.
- Elasticsearch lover.
- Working as consultant and software architect at SD DataSolutions GmbH in Bremen, Germany.
- Maintaining PANGAEA (Publishing Network for Geoscientific & Environmental Data) where I implemented the portal's geo-spatial retrieval functions with Apache Lucene Core and Elasticsearch.

An Overview

APACHE LUCENE ?

Lucene's data structures



c:\docs\einstein.txt:

The important thing is not to stop questioning.

c:\docs\shakespeare.txt:

c:\docs\einstein.txt:

The important thing is not to stop questioning.

c:\docs\shakespeare.txt:



c:\docs\shakespeare.txt:



c:\docs\einstein.txt: The important thing is stop questioning.

c:\docs\shakespeare.txt:

String comparison slow!

Solution: Inverted index

c:\docs\einstein.txt: The important thing is stop questioning.

c:\docs\shakespeare.txt:

Query: not



c:\docs\shakespeare.txt:

67

153

151

148

82

Dunque, 133

328

Index

Canvas of life turned upside down, 68 "Carbonate of pork," 315 Carracci, the, 147 Casina di Banda, 119 Castelletto, 265 Cavagnago, 76 Cenere, Monte, narcissuses on, 228 Ceres, 161 Cerrea, 133 Chalk, Conté, the Italian for whom this was the one thing needful, 136 Chalk eggs, 43 Chamois, foot of, 283 Change, repudiation of desire for sudden, 186 - importance of, depends on the rate of introduction, 196 — either the circumstances or the sufferer will, 196 Changes, sweeping, to be felt hereafter as vibrations, 60 Cheapissimo, 165 Cheese and the alpi, 289 Cherries, 33, 35, 46 Chestnuts, 118 Chicory and seed onions, weary utterness in, 227 Children, subalpine, 301 — what becomes of the clever, 149 Chinese, the examination-ridden, 151 Chironico, 75 " Chow," 52 Church-going, subalpine, 303 Circulation of people like blood, Ciseri, his picture at Locarno, 271 Civilisation, antiquity of Italian, 124 - stationary, of ants and bees, 195 Class distinction inevitable, 195 Classification only possible through sense of shock, 63

Clergy our English and S Michele

Cocking, Wednesbury, 55, 305 Collects, unsympathetic prical bristling with, III Colleone, Medea, 231 Colma di San Giovanni, 163 Comba di Susa, 119 Comfort as a moral influence, 185 Comic song, the landlord's, 128 Common sense, the safest guide, 108 Consistent, who ever is ? 153 Contradictory principles, there must be a harmonious fusing of 152 Converting things by eating them. 153 Corpses, desiccated, at S. Michele, Cousins, my, the lower animals, (m) Cows fighting in farmyard, 120 Cricco, 125 Cristoforo, S., church of, at Mesocco, 208 - at Castello, 234 Crossing, efficacy of, 152 - unexpected results of, 55 - useless if too wide, 157 Crucifixion, fresco at Fusio, 110 Culture and priggishness, 141 - a mode of concealing weakness 192 Current feeling, the safest much 108 Cutlets, burnt, and the waiter, the Dalpe, 38 Dante a humbug, 156 Darwin, Charles, no place the meeting, 69 Darwin, Erasmus, 23, 131 Dazio, Signor Pietro, of Fusion Death, no man can die to human 277 Deceit a necessary alloy of his 280

Index

Deportment, good technique re-English priests and Italian, 106 sembles, 156 — why introspective, 18 Desire and power, 108 Equilibrium only attainable at Development of power to know the cost of progress, 195 our own likes and dislikes. 22 Eritis, a panic concerning, 204 Devil's Bridge, 23 Eternal punishment, 111, 196 Diatonic scale, and song of birds Eusebius, St., 178 in New Zealand, 232 Evolution and illusion, 43 Dirt, eating a peck of moral, 71 - essence of, consists in not Disgrazia and misfortune, 58 shocking too much, 110 D'Israeli, Isaac, quotations from, Extreme, every, an absurdity, 153 Dissenters all narrow-minded, Faido, 22 Faith, doubt lives in honest, 67 Distribution of plants and animals - more assured in the days of often inexplicable, 135 spiritual Saturnalia, 68 Diversion of mental images, 54 -foundations of our system Doera, fresco at, 145, 221 based on, 107, 277 Dogs, 156, 202, 260, 313 - and reason, 108 Doing, the only mode of learning, — catholic, of protoplasm, 152 - a mode of impudence, 283 Doors, how they open in time, 151 Falsehood turning to truth, 71 Doubt, " There lives more doubt Famine prices at Locarno, 276 in honest faith," 67 Feeling, current, the safest guide, Downs, the South, like Monte * IO8 Generoso, 230 Fertile, rich and poor rarely fertile Draughtsman, first business of a, inter se, 195 Fires, how Italians manage their, Drawing, the old manner of II7 teaching, 150 Fishmonger choosing a bloater, Dream, my, at Lago di Cadagno, Flats and sharps, a maze of meta-Drunkenness and imagination, 46 physical, 23 Fleet Street, beauties of, 19 Duso, Agostino, his fresco at Sta. Flowers, names of, 291 Maria in Calanca, 225 Fossil-soul, 234 Foundations of action lie deeper Earnestness, 142, 192 than reason, 107 Eating, a mode of bigotry, 153 - of a durable system laid on Echo at Graglia, 192 faith, 277 Edelweiss, 291 Francis, St., and Insurance Co.'s Electricity and Alpine roads, 60 plate, 191 Elephant brays a third, 233 Friction, which prevents the un-" Elongated " honey, 293 duly rapid growth of inventions, Embryonic stages, the artist 60

67

153

151

148

82

Dunque, 133

328

Index

Canvas of life turned upside down, 68 "Carbonate of pork," 315 Carracci, the, 147 Casina di Banda, 119 Castelletto, 265 Cavagnago, 76 Cenere, Monte, narcissuses on, 228 Ceres, 161 Cerrea, 133 Chalk, Conté, the Italian for whom this was the one thing needful, 136 Chalk eggs, 43 Chamois, foot of, 283 Change, repudiation of desire for sudden, 186 - importance of, depends on the rate of introduction, 196 — either the circumstances or the sufferer will, 196 Changes, sweeping, to be felt hereafter as vibrations, 60 Cheapissimo, 165 Cheese and the alpi, 289 Cherries, 33, 35, 46 Chestnuts, 118 Chicory and seed onions, weary utterness in, 227 Children, subalpine, 301 — what becomes of the clever, 149 Chinese, the examination-ridden, 151 Chironico, 75 " Chow," 52 Church-going, subalpine, 303 Circulation of people like blood, Ciseri, his picture at Locarno, 271 Civilisation, antiquity of Italian, 124 - stationary, of ants and bees, 195 Class distinction inevitable, 195 Classification only possible through sense of shock, 63

Clergy our English and S Michele

280

Cocking, Wednesbury, 55, 305 Collects, unsympathetic prical bristling with, III Colleone, Medea, 231 Colma di San Giovanni, 163 Comba di Susa, 119 Comfort as a moral influence, 185 Comic song, the landlord's, 128 Common sense, the safest guide, 108 Consistent, who ever is ? 153 Contradictory principles, there must be a harmonious fusing of 152 Converting things by eating them 153 Corpses, desiccated, at S. Michele, Cousins, my, the lower animals, the Cows fighting in farmyard, 120 Cricco, 125 Cristoforo, S., church of, at Mesocco, 208 - at Castello, 234 Crossing, efficacy of, 152 - unexpected results of, 55 - useless if too wide, 157 Crucifixion, fresco at Fusio, 110 Culture and priggishness, 141 - a mode of concealing weakness 192 Current feeling, the safest much 108 Cutlets, burnt, and the waiter, the Dalpe, 38 Dante a humbug, 156 Darwin, Charles, no place the meeting, 69 Darwin, Erasmus, 23, 131 Dazio, Signor Pietro, of Fusion Death, no man can die to himself 277 Deceit a necessary alloy of his

Index

Deportment, good technique re-English priests and Italian, 106 sembles, 156 — why introspective, 18 Desire and power, 108 Equilibrium only attainable at Development of power to know the cost of progress, 195 our own likes and dislikes. 22 Eritis, a panic concerning, 204 Devil's Bridge, 23 Eternal punishment, 111, 196 Diatonic scale, and song of birds Eusebius, St., 178 in New Zealand, 232 Evolution and illusion, 43 Dirt, eating a peck of moral, 71 - essence of, consists in not Disgrazia and misfortune, 58 shocking too much, 110 D'Israeli, Isaac, quotations from, Extreme, every, an absurdity, 153 Dissenters all narrow-minded, Faido, 22 Faith, doubt lives in honest, 67 Distribution of plants and animals - more assured in the days of often inexplicable, 135 spiritual Saturnalia, 68 Diversion of mental images, 54 -foundations of our system Doera, fresco at, 145, 221 based on, 107, 277 Dogs, 156, 202, 260, 313 - and reason, 108 Doing, the only mode of learning, — catholic, of protoplasm, 152 - a mode of impudence, 283 Doors, how they open in time, 151 Falsehood turning to truth, 71 Doubt, " There lives more doubt Famine prices at Locarno, 276 in honest faith," 67 Feeling, current, the safest guide, Downs, the South, like Monte * IO8 Generoso, 230 Fertile, rich and poor rarely fertile Draughtsman, first business of a, inter se, 195 Fires, how Italians manage their, Drawing, the old manner of II7 teaching, 150 Fishmonger choosing a bloater, Dream, my, at Lago di Cadagno, Flats and sharps, a maze of meta-Drunkenness and imagination, 46 physical, 23 Fleet Street, beauties of, 19 Duso, Agostino, his fresco at Sta. Flowers, names of, 291 Maria in Calanca, 225 Fossil-soul, 234 Foundations of action lie deeper Earnestness, 142, 192 than reason, 107 Eating, a mode of bigotry, 153 - of a durable system laid on Echo at Graglia, 192 faith, 277 Edelweiss, 291 Francis, St., and Insurance Co.'s Electricity and Alpine roads, 60 plate, 191 Elephant brays a third, 233 Friction, which prevents the un-" Elongated " honey, 293 duly rapid growth of inventions, Embryonic stages, the artist 60

328

Index

Canvas of life turned upside down, 68 "Carbonate of pork," 315 Carracci, the, 147 Casina di Banda, 119 Castelletto, 265 Cavagnago, 76 Cenere, Monte, narcissuses on, 228 Ceres, 161 Cerrea, 133 Chalk, Conté, the Italian for whom this was the one thing needful, 136 Chalk eggs, 43 Chamois, foot of, 283 Change, repudiation of desire for sudden, 186 - importance of, depends on the rate of introduction, 196 — either the circumstances or the sufferer will, 196 Changes, sweeping, to be felt hereafter as vibrations, 60 Cheapissimo, 165 Cheese and the alpi, 289 Cherries, 33, 35, 46 Chestnuts, 118 Chicory and seed onions, weary utterness in, 227 Children, subalpine, 301 — what becomes of the clever, 149 Chinese, the examination-ridden, 151 Chironico, 75 " Chow," 52 Church-going, subalpine, 303 Circulation of people like blood, Ciseri, his picture at Locarno, 271 Civilisation, antiquity of Italian, 124 - stationary, of ants and bees, 195 Class distinction inevitable, 195 Classification only possible through sense of shock, 63

Clergy our English and S Michele

280

Cocking, Wednesbury, 55, 305 Collects, unsympathetic prical bristling with, III Colleone, Medea, 231 Colma di San Giovanni, 163 Comba di Susa, 119 Comfort as a moral influence, 185 Comic song, the landlord's, 128 Common sense, the safest guide, 108 Consistent, who ever is ? 153 Contradictory principles, there must be a harmonious fusing of 152 Converting things by eating them. 153 Corpses, desiccated, at S. Michele, Cousins, my, the lower animals, (m) Cows fighting in farmyard, 120 Cricco, 125 Cristoforo, S., church of, at Mesocco, 208 - at Castello, 234 Crossing, efficacy of, 152 - unexpected results of, 55 - useless if too wide, 157 Crucifixion, fresco at Fusio, 110 Culture and priggishness, 141 - a mode of concealing weakness 192 Current feeling, the safest much 108 Cutlets, burnt, and the waiter, the Dalpe, 38 Dante a humbug, 156 Darwin, Charles, no place the meeting, 69 Darwin, Erasmus, 23, 131 Dazio, Signor Pietro, of Fusion Death, no man can die to human 277 Deceit a necessary alloy of his

sembles, 156 Desire and power, 108 Development of power to know our own likes and dislikes. 22 Devil's Bridge, 23 Diatonic scale, and song of birds in New Zealand, 232 Dirt, eating a peck of moral, 71 Disgrazia and misfortune, 58 D'Israeli, Isaac, quotations from, 67 Dissenters all narrow-minded, 153 Distribution of plants and animals often inexplicable, 135 Diversion of mental images, 54 Doera, fresco at, 145, 221 Dogs, 156, 202, 260, 313 Doing, the only mode of learning, 151 Doors, how they open in time, 151 Doubt, " There lives more doubt in honest faith," 67 Downs, the South, like Monte Generoso, 230 Draughtsman, first business of a, 148 Drawing, the old manner of teaching, 150 Dream, my, at Lago di Cadagno, 82 Drunkenness and imagination, 46 Dunque, 133 Duso, Agostino, his fresco at Sta. Maria in Calanca, 225 Earnestness, 142, 192 Eating, a mode of bigotry, 153 Echo at Graglia, 192 Edelweiss, 291 Electricity and Alpine roads, 60

Elephant brays a third, 233 " Elongated " honey, 293 Embryonic stages, the artist

Index

Deportment, good technique re-English priests and Italian, 106 — why introspective, 18 Equilibrium only attainable at the cost of progress, 195 Eritis, a panic concerning, 204 Eternal punishment, 111, 196 Eusebius, St., 178 Evolution and illusion, 43 - essence of, consists in not shocking too much, 110 Extreme, every, an absurdity, 153 Faido, 22 Faith, doubt lives in honest, 67 - more assured in the days of spiritual Saturnalia, 68 -foundations of our system based on, 107, 277 - and reason, 108 — catholic, of protoplasm, 152 - a mode of impudence, 283 Falsehood turning to truth, 71 Famine prices at Locarno, 276 Feeling, current, the safest guide, * IO8 Fertile, rich and poor rarely fertile inter se, 195 Fires, how Italians manage their, II7 Fishmonger choosing a bloater, Flats and sharps, a maze of metaphysical, 23 Fleet Street, beauties of, 19 Flowers, names of, 291 Fossil-soul, 234 Foundations of action lie deeper than reason, 107 - of a durable system laid on faith, 277 Francis, St., and Insurance Co.'s plate, 191 Friction, which prevents the unduly rapid growth of inventions, 60

67

153

151

148

82

Dunque, 133

328

Index

Canvas of life turned upside down, 68 "Carbonate of pork," 315 Carracci, the, 147 Casina di Banda, 119 Castelletto, 265 Cavagnago, 76 Cenere, Monte, narcissuses on, 228 Ceres, 161 Cerrea, 133 Chalk, Conté, the Italian for whom this was the one thing needful, 136 Chalk eggs, 43 Chamois, foot of, 283 Change, repudiation of desire for sudden, 186 - importance of, depends on the rate of introduction, 196 — either the circumstances or the sufferer will, 196 Changes, sweeping, to be felt hereafter as vibrations, 60 Cheapissimo, 165 Cheese and the alpi, 289 Cherries, 33, 35, 46 Chestnuts, 118 Chicory and seed onions, weary utterness in, 227 Children, subalpine, 301 — what becomes of the clever, 149 Chinese, the examination-ridden, 151 Chironico, 75 " Chow," 52 Church-going, subalpine, 303 Circulation of people like blood, Ciseri, his picture at Locarno, 271 Civilisation, antiquity of Italian, 124 - stationary, of ants and bees, 195 Class distinction inevitable, 195 Classification only possible through sense of shock, 63

Clergy our English and S Michele

Cocking, Wednesbury, 55, 305 Collects, unsympathetic prical bristling with, III Colleone, Medea, 231 Colma di San Giovanni, 163 Comba di Susa, 119 Comfort as a moral influence, 185 Comic song, the landlord's, 128 Common sense, the safest guide, 108 Consistent, who ever is ? 153 Contradictory principles, there must be a harmonious fusing of 152 Converting things by eating them. 153 Corpses, desiccated, at S. Michele, Cousins, my, the lower animals, (m) Cows fighting in farmyard, 120 Cricco, 125 Cristoforo, S., church of, at Mesocco, 208 - at Castello, 234 Crossing, efficacy of, 152 - unexpected results of, 55 - useless if too wide, 157 Crucifixion, fresco at Fusio, 110 Culture and priggishness, 141 - a mode of concealing weakness 192 Current feeling, the safest much 108 Cutlets, burnt, and the waiter, the Dalpe, 38 Dante a humbug, 156 Darwin, Charles, no place the meeting, 69 Darwin, Erasmus, 23, 131 Dazio, Signor Pietro, of Fusion Death, no man can die to human 277 Deceit a necessary alloy of his 280

Index

Deportment, good technique re-English priests and Italian, 106 sembles, 156 — why introspective, 18 Desire and power, 108 Equilibrium only attainable at Development of power to know the cost of progress, 195 our own likes and dislikes. 22 Eritis, a panic concerning, 204 Devil's Bridge, 23 Eternal punishment, 111, 196 Diatonic scale, and song of birds Eusebius, St., 178 in New Zealand, 232 Evolution and illusion, 43 Dirt, eating a peck of moral, 71 - essence of, consists in not Disgrazia and misfortune, 58 shocking too much, 110 D'Israeli, Isaac, quotations from, Extreme, every, an absurdity, 153 Dissenters all narrow-minded, Faido, 22 Faith, doubt lives in honest, 67 Distribution of plants and animals - more assured in the days of often inexplicable, 135 spiritual Saturnalia, 68 Diversion of mental images, 54 -foundations of our system Doera, fresco at, 145, 221 based on, 107, 277 Dogs, 156, 202, 260, 313 - and reason, 108 Doing, the only mode of learning, — catholic, of protoplasm, 152 - a mode of impudence, 283 Doors, how they open in time, 151 Falsehood turning to truth, 71 Doubt, " There lives more doubt Famine prices at Locarno, 276 in honest faith," 67 Feeling, current, the safest guide, Downs, the South, like Monte * IO8 Generoso, 230 Fertile, rich and poor rarely fertile Draughtsman, first business of a, inter se, 195 Fires, how Italians manage their, Drawing, the old manner of II7 teaching, 150 Fishmonger choosing a bloater, Dream, my, at Lago di Cadagno, Flats and sharps, a maze of meta-Drunkenness and imagination, 46 physical, 23 Fleet Street, beauties of, 19 Duso, Agostino, his fresco at Sta. Flowers, names of, 291 Maria in Calanca, 225 Fossil-soul, 234 Foundations of action lie deeper Earnestness, 142, 192 than reason, 107 Eating, a mode of bigotry, 153 - of a durable system laid on Echo at Graglia, 192 faith, 277 Edelweiss, 291 Francis, St., and Insurance Co.'s Electricity and Alpine roads, 60 plate, 191 Elephant brays a third, 233 Friction, which prevents the un-" Elongated " honey, 293 duly rapid growth of inventions, Embryonic stages, the artist 60





Query: not



Query: not



Query: not



Query: not



Information Retrieval Model

Lucene is based on a combination of two well known Information Retrieval models:

- Vector Space Model scoring and relevance
- Boolean Model narrowing down the documents to score

$$\cos \Theta = \frac{v_1 \cdot v_2}{|v_1| \cdot |v_2|} \qquad w_{t,d} = tf_t \cdot idf_{t,d}$$

Term-Freque
$$idf_{t,d} = \log \frac{|D|}{|t \in d|} document d.$$

Inverse Document Frequency $(idf) \rightarrow$ the relation between the number of documents in the corpus and the number of documents containing term t (global parameter).

Indexing with Lucene

- Fast: over 200 GB/hour
- Incremental and "near-realtime"
- Multi-threaded
- Beyond full-text: numbers, dates, binary,...
- Customize what is indexed ("analysis")
- Customize index format ("codecs")

History

ON THE WAY TO LUCENE 5...

- Lucene started > 10 years ago
 - Lucene's VINT format is old and not as friendly as new compression algorithms to CPU's optimizers (exists since Lucene 1.0)

- Lucene started > 10 years ago
 - Lucene's VINT format is old and not as friendly as new compression algorithms to CPU's optimizers (exists since Lucene 1.0)
- It's hard to add additional statistics for scoring to the index
 - IR researchers don't use Lucene to try out new algorithms

- Lucene started > 10 years ago
 - Lucene's VINT format is old and not as friendly as new compression algorithms to CPU's optimizers (exists since Lucene 1.0)
- It's hard to add additional statistics for scoring to the index
 - IR researchers don't use Lucene to try out new algorithms
- Small changes to index format are often huge patches covering tons of files

• Major release in October 2012

- Major release in October 2012
- New index engine:
 - -Codec support (pluggable via SPI)
 - DocValues fields

- Major release in October 2012
- New index engine:
 - Codec support (pluggable via SPI)
 DocValues fields
- New relevancy models: not only TF/IDF !

-e.g., **BM25**

- Major release in October 2012
- New index engine:
 - Codec support (pluggable via SPI)
 DocValues fields
- New relevancy models: not only TF/IDF !
 - -e.g., **BM25**
- FSAs / FSTs everywhere

Complete overhaul of all APIs



Complete overhaul of all APIs

- Terms got byte[]
- Low level terms enumerations and postings enumerations refactored
- Query API internals (scorer, weight)
- Analyzers: new module, package structure changed (pluggable via SPI)
- IndexReader => AtomicReader, CompositeReader

- Every Lucene 4 release got new features!
 - API glitches!!!

- Burden of maintaining the old stuff:
 - old index formats
 - especially support for
 Lucene 3.x indexes



On-going Disasters

• Not only problems with bugs in Java runtimes



On-going Disasters

Not only problems with bugs in Java runtimes

– Story could fill another talk! 🙂



On-going Disasters

- Not only problems with bugs in Java runtimes
 Story could fill another talk! ⁽ⁱ⁾
- Major problems with old index formats:
 Lucene 3 had a completely different index format
 - without codec support (missing headers,...)
On-going Disasters

- Not only problems with bugs in Java runtimes
 Story could fill another talk! ⁽ⁱ⁾
- Major problems with old index formats:
 Lucene 3 had a completely different index format
 - without codec support (missing headers,...)

Lot's of hacks!

Chronology

- Lucene 4.2.0: Lucene deletes entire index if exception is thrown due do too many open files with OpenMode.CREATE_OR_APPEND (LUCENE-4870)
- Lucene 4.9.0: Closing NRT reader after upgrading from 3.x index can cause index corruption *(LUCENE-5907)*
- Lucene 4.10.0: Index version numbers caused CorruptIndexException (LUCENE-5934)

Apache Lucene 5

A lot new features!

Apache Lucene 5

A lot new features!

 But not so many as you would expect for major release!

Apache Lucene 5

A lot new features!

- But not so many as you would expect for major release!
- Some more than in previous minor 4.x releases...

Lucene 5: "Anti-Feature"

Removal of Lucene 3 index support!



Lucene 5: "Anti-Feature"

Removal of Lucene 3 index support!

- Get rid of old index segments: IndexUpgrader in latest Lucene 4 release helps!
- Elasticsearch has automatic index upgrader already implemented / Solr users have to manually do this



Lucene 5: New data safety features

Lucene 5: New data safety features

- Checksums in all index files
 - Checksums are validated on each merge!
 - Can easily be validated during Solr's / Elasticsearch's replication!

Lucene 5: New data safety features

- Checksums in all index files
 - Checksums are validated on each merge!
 - Can easily be validated during Solr's / Elasticsearch's replication!
- Unique per segment ID
 - ensures that the reader really sees the segment mentioned in the commit
 - prevents bugs caused by failures in replication (e.g., duplicate segment file names)

Lucene 5: New index safety features

Cutover to NIO.2 (Java 7, JSR 203)

atomic rename to publish commit
fsync() on index directory

Java 7 support

• Introduced in Lucene 4.8

– Could have been Lucene 5 already 😇

• Why?

- EOL of Java 6, but still bugs that affected Lucene
- Java 8 released
- use of new features for <u>index safety</u>!

- Try-With-Resources
 - Nice, but we had it already implemented: IOUtils.closeWhileHandlingExceptions()

- Try-With-Resources
 - Nice, but we had it already implemented: IOUtils.closeWhileHandlingExceptions()
- Some syntactic sugar ☺

- Try-With-Resources
 - Nice, but we had it already implemented: IOUtils.closeWhileHandlingExceptions()
- Some syntactic sugar \odot
- Partial implementation of NIO.2 for FSDirectory
 - (allows to delete open files on Windows!)

- Try-With-Resources
 - Nice, but we had it already implemented: IOUtils.closeWhileHandlingExceptions()
- Some syntactic sugar $\textcircled{\odot}$
- Partial implementation of NIO.2 for FSDirectory
 - (allows to delete open files on Windows!)
- MethodHandle / ClassValue for Tokenization API's internals
 - Huge speedup for dynamic instantiation of token Attributes, especially in Java 8!

Java 7u55+ has no serious bugs anymore

(still a no-go for G1GC with Lucene)

• Complete overhaul of Lucene I/O APIs

• Complete overhaul of Lucene I/O APIs

• java.io.File* => forbidden-apis *)

*) https://code.google.com/p/forbidden-apis/

• Complete overhaul of Lucene I/O APIs

• java.io.File* => forbidden-apis *)

- Atomic rename to publish commit
 - no more segments.gen
 - fsync() on directory metadata

*) https://code.google.com/p/forbidden-apis/

No more index corruption because of broken Exception handling:

- Exceptions now have a clear meaning, you can rely on
- NIO.2 APIs now throw useful exceptions
- before that, File.rename() / delete()
 could do nothing at all!

- Don't use Future.cancel(true) !!!
 - Never interrupt searching threads, it kills your IndexReader!
 - Alternative: org.apache.lucene.store.RAFDirectory (RAF = RandomAccessFile, only available in "misc" module)

- Don't use Future.cancel(true) !!!
 - Never interrupt searching threads, it kills your IndexReader!
 - Alternative: org.apache.lucene.store.RAFDirectory (RAF = RandomAccessFile, only available in "misc" module)
- All other file I/O is now channel based (or MMap)
 - If cancelled throws ClosedByInterruptException
 - also SimpleFSDirectory !

- Don't use Future.cancel(true) !!!
 - Never interrupt searching threads, it kills your IndexReader!
 - Alternative: org.apache.lucene.store.RAFDirectory (RAF = RandomAccessFile, only available in "misc" module)
- All other file I/O is now channel based (or MMap)
 - If cancelled throws ClosedByInterruptException
 - also SimpleFSDirectory!
- Use Paths.get() while opening DirectoryReader / IndexWriter
 - Alternative: use File.toPath()

Lucene 5.0: Overhaul of Codec API

- Pull APIs throughout Codec components
 E.g., PostingsFormat
- Norms are now handled separate codec component

Lucene 5.0: Index merging

Lucene 5.0: Index merging

- Linux: Detection if index is on SSD
 - Better default merging settings
 - Other operating systems assume spinning disks (no change)

Lucene 5.0: Index merging

- Linux: Detection if index is on SSD
 - Better default merging settings
 - Other operating systems assume spinning disks (no change)
- Merge Scheduler: Auto Throttling
 - Automatically controls I/O rates based on indexing/merging rate
 - Stalling under high load is more unlikely!

Lucene 5.0: Reduced Heap Usage

- Query Filters uses new bit set types
- CachingWrapperFilter replacement:
 - New, highly configureable filter cache
 - Tracks filter's frequency of use
 - Simplifies code in Apache Solr and Elasticsearch
- Merging uses much less heap

Lucene 5.0: Reduced Heap Usage

- Query Filters uses new bit set types
- CachingWrapperFilter replacement:
 - New, highly configureable filter cache
 - Tracks filter's frequency of use
 - Simplifies code in Apache Solr and Elasticsearch
- Merging uses much less heap
- Most classes now implement Accountable
 - Allows to query heap usage
 - Nice "tree view" on heap usage of index components

Lucene 5.0: Reduced Heap Usage

- Query Filters uses new bit set types
- CachingWrapperFilter replacement:
 - New, highly configureable filter cache
 - Tracks filter's frequency of use
 - Simplifies code in Apache Solr and Elasticsearch
- Merging uses much less heap

```
_cz(5.0.0):C8330469: 28MB
postings [...]: 5.2MB
...
field 'latitude' [...]: 678.5KB
term index [FST(nodes=6679, ...)]: 678.3KB
```

Lucene 5.0: CustomAnalyzer

- Freely configurable Analyzer
- Based on SPI framework for Tokenizers, TokenFilters and CharFilters
- Similar to Apache Solr's schema.xml:
 - Generic names of components (like Elasticsearch)
 - Same config options like Apache Solr
- Builder API

Lucene 5.0: CustomAnalyzer

• Freely configurable Analyzer

```
Analyzer ana =
CustomAnalyzer.builder(Paths.get("/path/to/config"))
.withTokenizer("standard")
.addTokenFilter("standard")
.addTokenFilter("lowercase")
.addTokenFilter("stop",
    "ignoreCase", "false",
    "words", "stopwords.txt",
    "format", "wordset")
.build();
```

Die, FieldCache,... die, die, die!

- FieldCache is gone from Lucene Core
- Use DocValues fields and APIs!
Die, FieldCache,... die, die, die!

- FieldCache is gone from Lucene Core
- Use DocValues fields and APIs!

- Not completely gone:
 - UninvertingReader in misc/module emulates
 DocValues by uninverting index
 - UninvertingReader allows to merge to a new index, automatically adding DocValues!



Apache Solr 5.0

New release bundled with Lucene 5.0 release

Improved fault tolerance

Solr 5.0: No Webapp anymore!

- Solr ships as server software
 - like MySQL, PostgreSQL,...
 - or Elasticsesarch ☺
- Start/Stop scripts for SysVinit
- JVM tuning by default
- Scripts to create collections
- No "official" WAR file anymore
 - Maven
 - Download distribution

Caly E A. Na Mahanna anumana

C:\Windows\system32\cmd.exe

C:\Users\Uwe Schindler\Projects\lucene\lucene_solr_5_0\solr\bin>dir Datenträger in Laufwerk C: ist SSD Volumeseriennummer: D8D1-55DC

Verzeichnis von C:\Users\Uwe Schindler\Projects\lucene\lucene_solr_5_0\solr\bin

29.01.2015	17:19	<dir></dir>				
29.01.2015	17:19	<dir></dir>				
14.01.2015	15:43	<dir></dir>		init.d		
24.01.2015	13:43	8	.563	install_solr_s	ervice	.sh
14.01.2015	15:43	1	.285	oom_solr.sh		
24.01.2015	13:43	7	.589	post		
29.01.2015	17:19	56	.174	solr		
29.01.2015	17:19	46	.225	solr.cmd		
18.01.2015	13:19	4	.143	solr.in.cmd		
18.01.2015	13:19	4	.527	solr.in.sh		
	7 Da	tei(en),	1	128.506 Bytes		
	3 Ve	rzeichnis(se), 19	52.308.031.488	Bytes	frei

C:\Users\Uwe Schindler\Projects\lucene\lucene_solr_5_0\solr\bin>solr.cmd start -p 4711 Backing up C:\Users\Uwe Schindler\Projects\lucene\lucene_solr_5_0\solr\server\logs\solr.log 1 Datei(en) verschoben.

Backing up C:\Users\Uwe Schindler\Projects\lucene\lucene_solr_5_0\solr\server\logs\solr_gc.log 1 Datei(en) verschoben.

Starting Solr on port 4711 from C:\Users\Uwe Schindler\Projects\lucene\lucene_solr_5_0\solr\server

Direct your Web browser to http://localhost:4711/solr to visit the Solr Admin UI

C:\Users\Uwe Schindler\Projects\lucene\lucene_solr_5_0\solr\bin>solr.cmd stop -all Stopping Solr process 10844 running on port 4711

Gewartet wird Ø Sekunden. Weiter mit beliebiger Taste...

C:\Users\Uwe Schindler\Projects\lucene\lucene_solr_5_0\solr\bin>

Solr 5.0: Distributed IDF

Support for distributed *Inverse Document Frequency:*

- Makes use of caching of IDF from other nodes
- Several caching implementations

Solr 5.0: Distributed IDF

Support for distributed *Inverse Document Frequency:*

- Makes use of caching of IDF from other nodes
- Several caching implementations

Should only be used if exact scoring is really needed

• If documents are not well (randomly) distributed

Solr 5.0: Config API

- Makes parameters of RequestHandlers configurable
- Allows to change RequestHandlers
- Upload of plugin JARs

Solr 5.0: Other features

- Bandwidth control for index replication
- BLOBs API
- SolrJ improvements:
 - Rename SolrServer to SolrClient
 - Support of Collections API
- Split Clusterstate
 - Scales better for hundreds of nodes

THANK YOU!

Questions?

Contact

Uwe Schindler

uschindler@apache.org http://www.thetaphi.de @thetaph1



SD DataSolutions GmbH Wätjenstr. 49 28213 Bremen, Germany +49 421 40889785-0 http://www.sd-datasolutions.de

