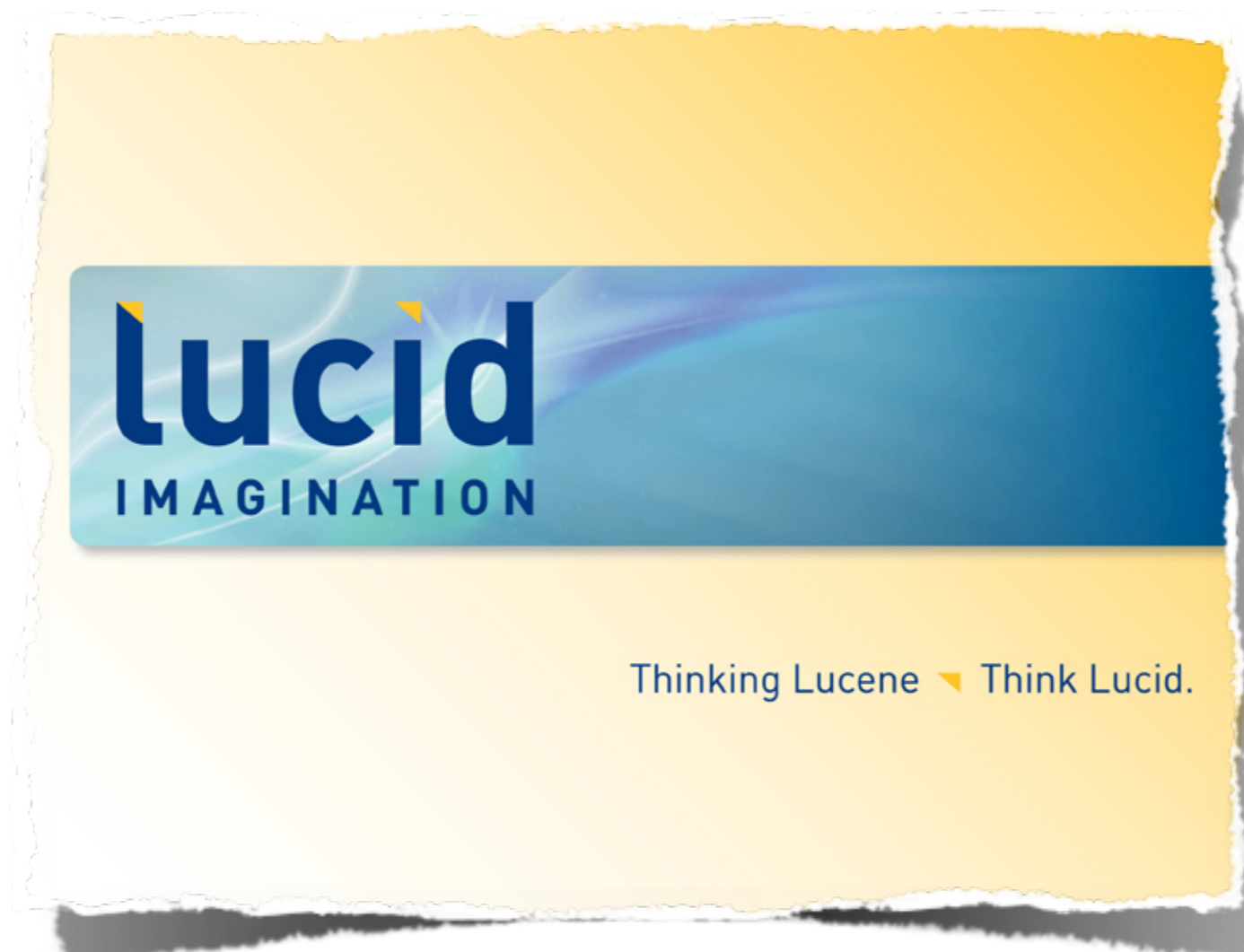


Apache Solr 1.4

Faster, Easier, and More Versatile than Ever



About Erik

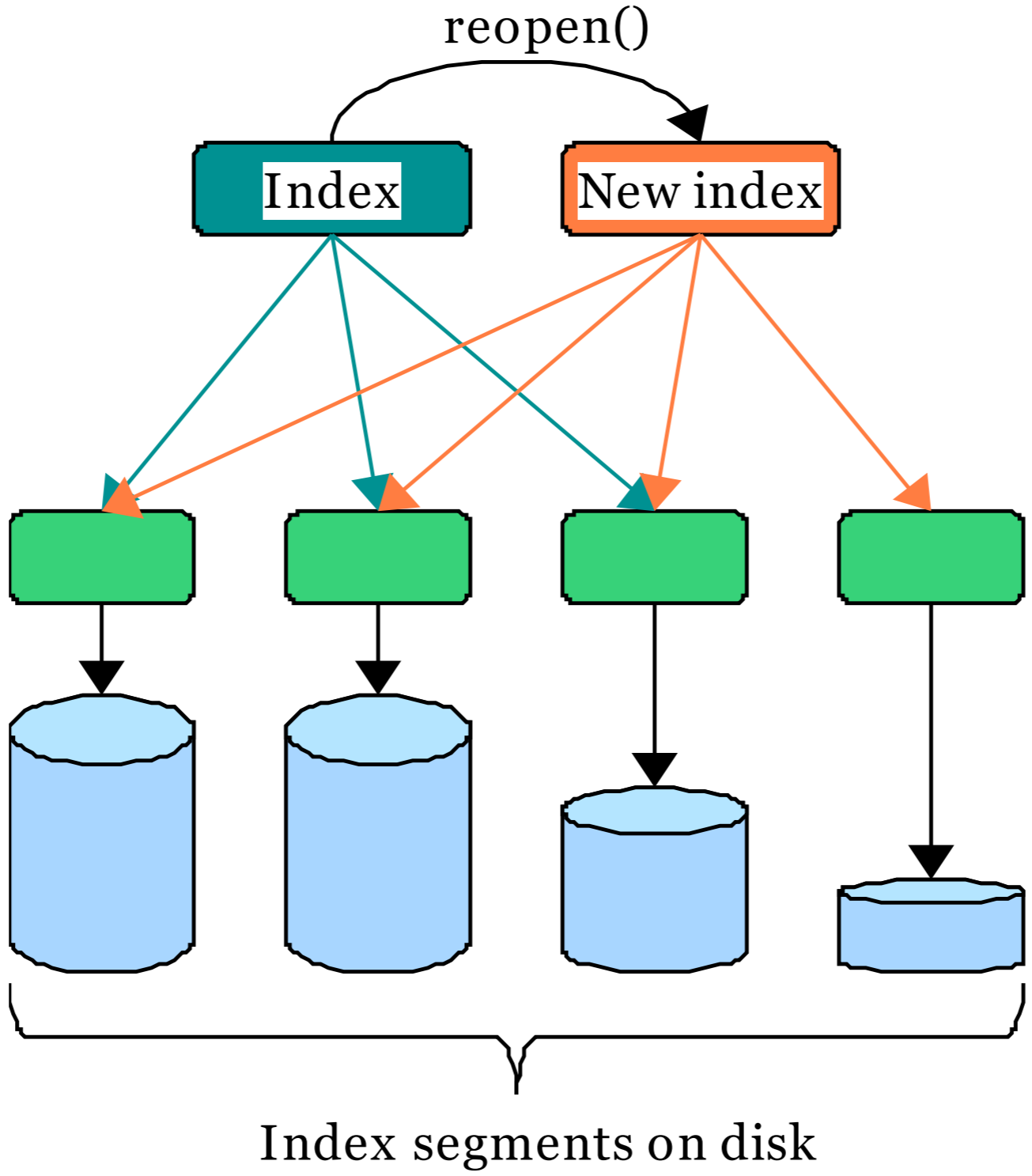
- Member of Technical Staff, Lucid Imagination
- Co-author "Lucene in Action"
- Frequent speaker at industry conferences
- Committer: Lucene and Solr
- Apache Lucene PMC member

Lucene

Lucene 2.9

- `IndexReader#reopen()`
- Faster filter performance, by 300% in some cases
- Per-segment FieldCache
- Reusable token streams
- Faster numeric/date range queries, thanks to trie
- and tons more, see Lucene 2.9's [CHANGES.txt](#)

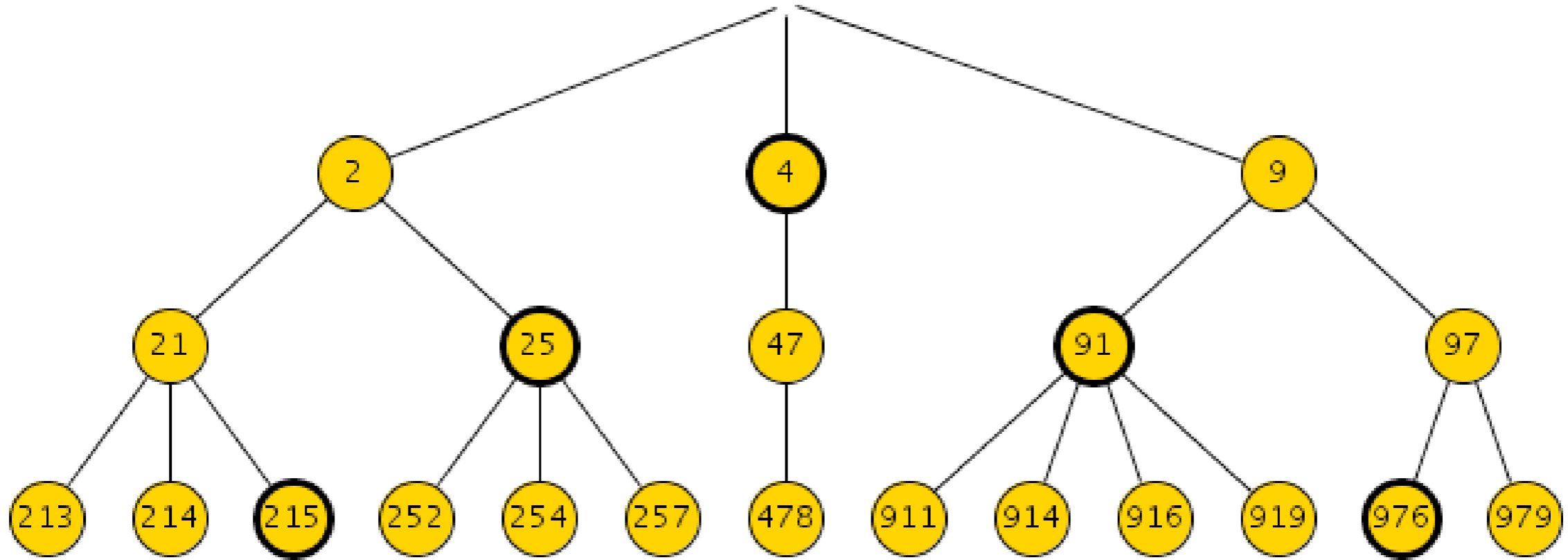
reopen()



Trie fields

- Trie* fields index multiple precision steps
- Works for numerics & dates
- Result: up to 40x faster than standard range queries
- Configurable precision step

Trie Example



Trie-Range Prefix Tree Encoding and relevant nodes for range [215 TO 977]

Solr

New Logo!



Performance Improvements

- Caching
- Concurrent file access
- Per-segment index updates
- Faceting
- DocSet generation, avoids scoring
- Streaming updates for Solrj

Caching

- LRU cache now based on `ConcurrentHashMap`
- Reads are lockless, writes are partitioned
- Minimizes overhead of synchronization
- Improves `filterCache`, `queryCache`, and `documentCache`
- Anecdotal evidence: can double query throughput in some circumstances

Concurrent file access

- Defaults to NIOFSDirectory on non-Windows platforms for better concurrency
- Pluggable DirectoryProvider

Per-segment caching

- FieldCache per segment
- Improves searching, sorting, and function queries

Faceting performance

- Major performance improvements on multi-valued fields!
- <http://yonik.wordpress.com/2008/11/25/solr-faceted-search-performance-improvements/>

Solrj

- Binary updates (bye bye XML)
- StreamingUpdateSolrServer
 - Streams multiple documents over multiple connections
 - Simple test went from 231 docs/sec to 25000 docs/sec!
- LBHttpSolrServer
 - load balancing / failover

Feature Improvements

- Rich document indexing
- DataImportHandler enhancements
- Smoother replication
- More choices for logging
- Multi-select faceting
- Speedier range queries
- Duplicate detection
- New request handler components

Solr Cell

- Content Extraction Library L_?_
- Richly extracts text from Microsoft Office, PDF, HTML, and many other formats
- Powered by Apache Tika
- <http://wiki.apache.org/solr/ExtractingRequestHandler>

DataImportHandler

- SQL deltaImportQuery
- abort, skip, continue options
- FieldReader
 - example: can read XML from CLOB
- HTML strip transformer
- Event listener API (import start/end)
- ContentStreamDataSource
- MailEntityProcessor
- LineEntityProcessor, FileListEntityProcessor

Replication

- Solr 1.3: rsync-based
- Solr 1.4: Java-based, request handler
- Copy config files, and optionally rename
- Basic authentication support
- Custom Solr index deletion policy enables deletion of commit points on various criteria such as number of commits, age of commit point and optimized status (pluggable)

Replication Configuration

Master

```
<requestHandler name="/replication" class="solr.ReplicationHandler">  
  <lst name="master">  
    <str name="replicateAfter">commit</str>  
    <str name="confFiles">schema.xml, stopwords.txt</str>  
  </lst>  
</requestHandler>
```

Slave

```
<requestHandler name="/replication" class="solr.ReplicationHandler">  
  <lst name="slave">  
    <str name="masterUrl">  
      http://masterhostname:8983/solr/replication  
    </str>  
    <str name="pollInterval">00:00:60</str>  
  </lst>  
</requestHandler>
```

New faceting features

- multi-select
- tag a filter query (fq)
- exclude a tagged filter from facet counts
- key - for labeling response structure

Options

[Clear all facets](#)

PROJECT

[clear projects](#)

- Lucene (502)
- Solr (4456)
- Nutch (85)
- Tika (15)
- Mahout (70)
- Droids (8)
- PyLucene (4)
- Lucene.Net (16)
- Lucy (1)
- Open Relevance Project (2)

SOURCE

[clear sources](#)

- Lucid (3)
- Wiki (48)
- Apache Lucene Web (3)
- Email (4548)
 - user (3464)
 - dev (1084)
- Issues (353)

solr 1.4

[Start new search](#)

Search Results for solr 1.4

Found 4,957 results in 0.026 seconds. Displaying page 1 of 496, sorted by relevancy

[\[WIKI\] Solr1.4](#)

2009-09-16 00:07

Solr 1.4 Solr 1.4 has not been released. CURRENT GOAL IS TO START THE SOLR 1.4 RELEASE PROCESS APPROXIMATELY ONE TO TWO WEEKS AFTER LUCENE BEEN RELEASED. If you found some documentation it was there to alert you to the fact that it described a feature that is expected to be included in the Solr 1.4 eventually

<http://wiki.apache.org/solr/Solr1.4>

[\[WIKI\] Search Component](#)

2009-08-21 22:20

/handler/component/StatsComponent.html Statistics] <!-- ["Solr1.4"] * debug - [http://lucene.apache.org/solr/api/org/apache/solr/handler/component/DebugComponent.html Debug] <----- These examples are enabled by default, but can be overridden by

<http://wiki.apache.org/solr/SearchComponent>

[\[WIKI\] Solr Logging](#)

2008-12-17 14:35

. If you don't know much about JDK containers or servlet containers and want a quick recipe for modifying the logging settings for Solr setup, see LoggingInDefaultJettySetup. Solr 1.4 and Above Starting with Solr 1.4, the Solr Code base compiles against the "SLF

Deduplication

- Detect duplicates during indexing and handle them
- Adds a signature field to the document (could be uniqueKey)
- Exact (hash on certain fields) or Fuzzy duplicate detection
- <http://wiki.apache.org/solr/Deduplication>

commit/rollback

- `commitWithin`
- `rollback`

Analysis

- CharFilter
- DoubleMetaphone
- PersianAnalyzer,
ArabicAnalyzer,
SmartChineseAnalyzer
- WordDelimiterFilter
 - splitOnNumerics
 - protected words support
 - stemEnglishPossessive
- avoid splitting or dropping international non-letter characters such as non spacing marks.
- PositionFilter support
- SnowballPorterFilterFactory - protected words support
- HTMLStripCharFilter
- CommonGrams

omitTermFreqAndPositions

- Omits number of terms in that specific field & list of positions
- Saves time and index space for non-text fields

Trie Numeric Field Types

```
<fieldType name="tint" class="solr.TrieIntField"  
  precisionStep="8" omitNorms="true"  
  positionIncrementGap="0"/>
```

```
<fieldType name="tfloat" class="solr.TrieFloatField"  
  precisionStep="8" omitNorms="true"  
  positionIncrementGap="0"/>
```

```
<fieldType name="tlong" class="solr.TrieLongField"  
  precisionStep="8" omitNorms="true"  
  positionIncrementGap="0"/>
```

```
<fieldType name="tdouble" class="solr.TrieDoubleField"  
  precisionStep="8" omitNorms="true"  
  positionIncrementGap="0"/>
```

Trie Date Field Types

```
<fieldType name="date" class="solr.TrieDateField"  
  omitNorms="true" precisionStep="0"  
  positionIncrementGap="0"/>
```

```
<fieldType name="tdate" class="solr.TrieDateField"  
  omitNorms="true" precisionStep="6"  
  positionIncrementGap="0"/>
```

Wildcard handling

- ReversedWildcardFilterFactory
- Index-time reversal, query-time handling
- Uses special marker for end
- Example:
 - Document: `<field name="text_rev">Solr</field>`
 - Indexed: `#rlos`
 - Query: `*olr -> #rlo*`
 - # used here to denote marker character

Function queries

- milliseconds: ms()
- subtraction: sub()
- `{!frange l=6 u=9}sqrt(sum(a,b))`
- <http://yonik.wordpress.com/2009/07/06/ranges-over-functions-in-solr-1-4/>

Stats Component

- min, max, sum, sumOfSquares, count, missing, mean, stddev
- numeric fields only
- `http://localhost:8983/solr/select?q=*&stats=true&stats.field=price&stats.field=popularity&rows=0&indent=true`

Terms Component

- Return indexed terms+docfreq in a field, use for auto-suggest, etc
- `http://localhost:8983/solr/terms?terms.fl=name&terms.lower=a&terms.sort=index`

Term Vector Component

- Returns term info per document (tf, positions)
- `http://localhost:8983/solr/select/?q=%3A*&version=2.2&start=0&rows=10&indent=on&qt=tvrh&tv=true&tv.tf=true&tv.df=true&tv.positions&tv.offsets=true`

Field & Document Request Handlers

- Provides similar capabilities as Solr's admin analysis tool, but with response in Solr's flexible formats

Clustering

- Implemented with Carrot2
- Uses Carrot2 to dynamically cluster the top N search results
- Like dynamically discovered facets
- <http://wiki.apache.org/solr/ClusteringComponent>

Example Clustering Output

```
<lst name="cluster">
  <lst name="labels">
    <str name="label">Car Power Adapter</str>
  </lst>
  <lst name="docs">
    <str name="doc">F8V7067-APL-KIT</str>
    <str name="doc">IW-02</str>
  </lst>
</lst>
<lst name="cluster">
  <lst name="labels">
    <str name="label">Display</str>
  </lst>
  <lst name="docs">
    <str name="doc">MA147LL/A</str>
    <str name="doc">VA902B</str>
  </lst>
</lst>
<lst name="cluster">
  <lst name="labels">
    <str name="label">Hard Drive</str>
  </lst>
  <lst name="docs">
    <str name="doc">SP2514N</str>
    <str name="doc">6H500F0</str>
  </lst>
</lst>
<lst name="cluster">
  <lst name="labels">
    <str name="label">Retail</str>
  </lst>
  <lst name="docs">
    <str name="doc">TWINX2048-3200PRO</str>
```

Distributed Search

- `facet.sort=lex`
- timeout support: `shard-socket-timeout` & `shard-connection-timeout`

VelocityResponseWriter

- celeritas: swiftness, speed (Latin), origin of the symbol "c" for the speed of light
- solritas: Velocity template rendering of Solr responses
- Useful for rapid prototyping and more

AJAX-Solr

- Newest Solr/AJAX library
- Improves upon the old SolrJS library that was to be in Solr 1.4
- <http://github.com/evolvingweb/AJAX-Solr/>

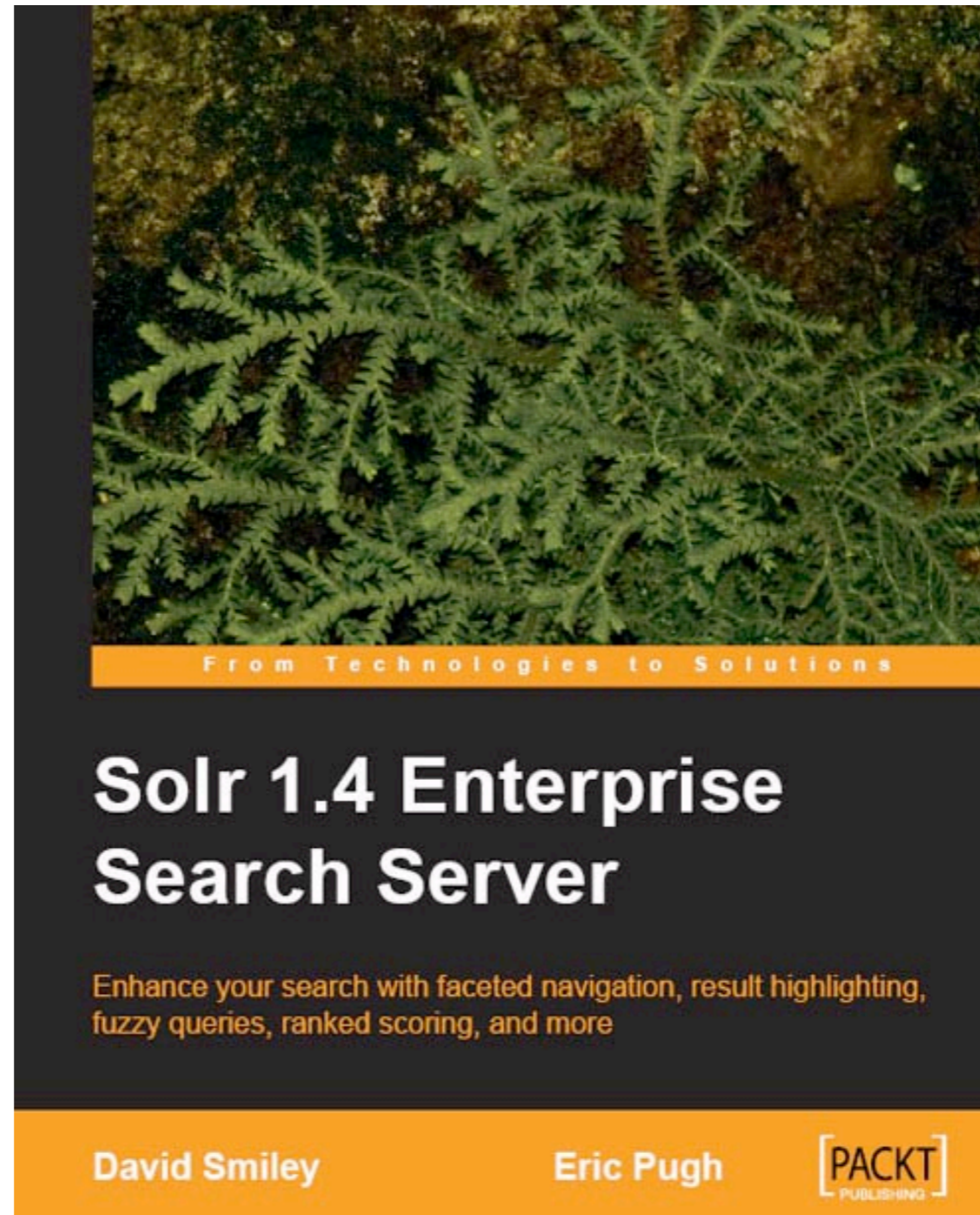
Odds & Ends

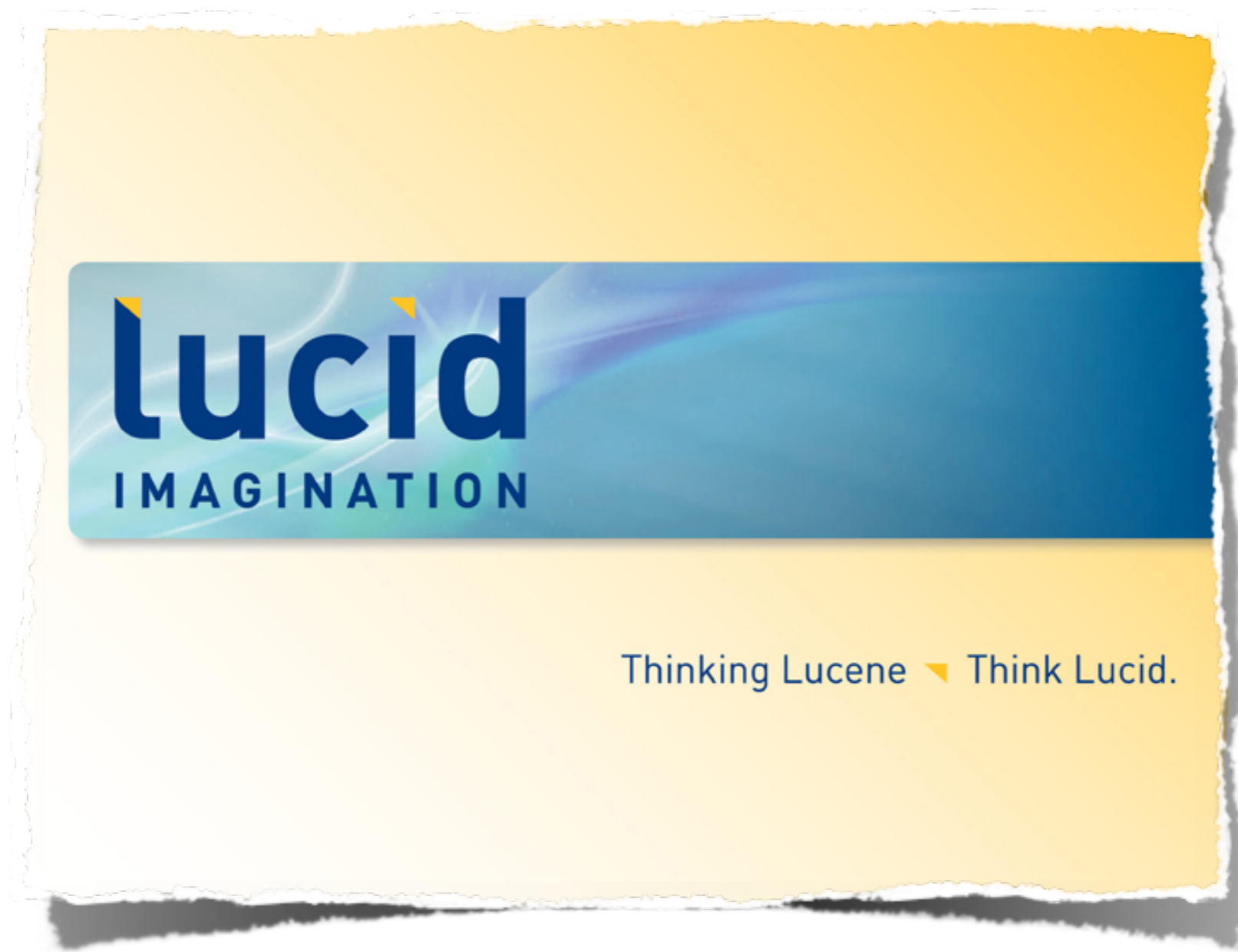
- omitHeader
- logging switched to SLF4J
- spellcheck rebuild on optimize, if configured
- maxChars on copyField
- Nested query support in function query parser
- expungeDeletes
- XInclude support
- multicore merge
- binary field (Base64)
- Highlighter:
 - field globbing:
hl.fl=*_text
 - now supports range/wildcard/fuzzy/prefix
 - FieldCache stats exposed to stats.jsp and JMX
 - plugins: enable flag

Upgrading from 1.3

- `omitTermFreqAndPositions`
- set version from 1.1 to 1.2 in `schema.xml`
- default query parser syntax no longer supports `;sort` options, use `&sort=` instead, or switch `defType=luenePlusSort`
- Potential analysis differences when using `WordDelimiterFilterFactory` (SOLR-1078)
- Reindexing not required, but can't hurt.

Book





lucidimagination.com