Apache Solr crash course

Tommaso Teofili

sourcesense

Agenda

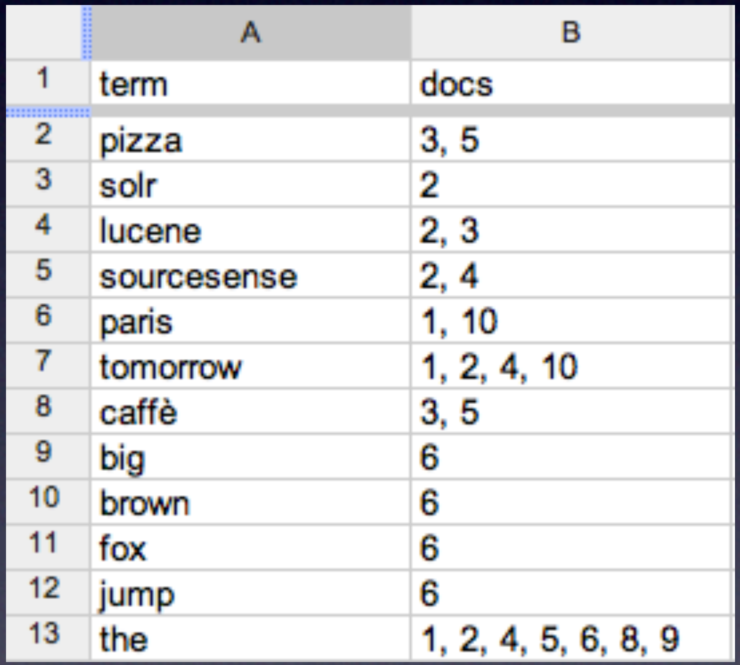
- IR
- Solr
- Tips&tricks
- Case study
- Extras

Information Retrieval

- “Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)” - P. Nayak, Stanford University

Inverted index

- Each document has an id and a list of terms
- For each term t we must store a list of all documents that contain t
- Identify each document by its id



	A	B
1	term	docs
2	pizza	3, 5
3	solr	2
4	lucene	2, 3
5	sourcesense	2, 4
6	paris	1, 10
7	tomorrow	1, 2, 4, 10
8	caffè	3, 5
9	big	6
10	brown	6
11	fox	6
12	jump	6
13	the	1, 2, 4, 5, 6, 8, 9

IR Metrics

- How good is an IR system?
- Precision: Fraction of retrieved docs that are relevant to user’s information need
- Recall: Fraction of relevant docs in collection that are retrieved

Apache Lucene

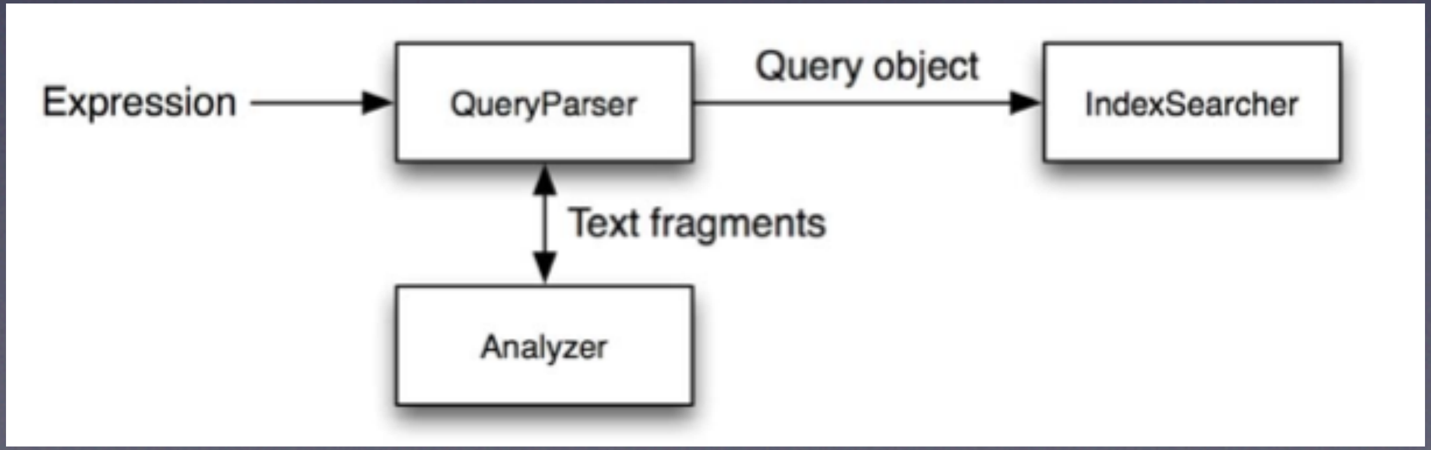
- Information Retrieval library
- Inverted index of documents
- Vector space model
- Advanced search options (synonims, stopwords, similarity, proximity)

Lucene API - indexing

- Lucene indexes are built on a Directory
- Directory can be accessed by IndexReaders and IndexWriters
- IndexSearchers are built on top of Directories and IndexReaders
- IndexWriters can write Documents inside the index
- Documents are made of Fields
- Fields have values
- Directory > IndexReader/Writer > Document > Field

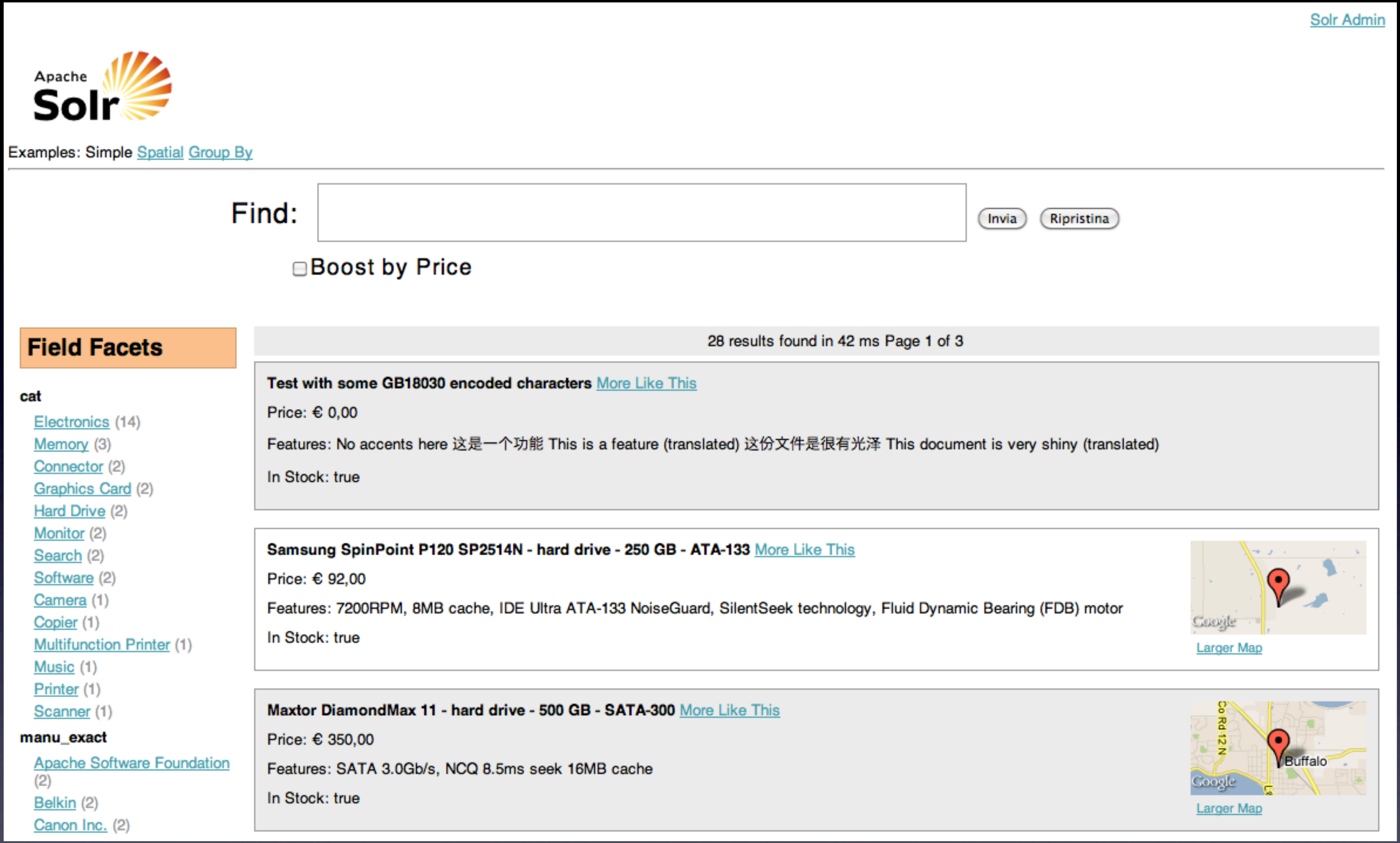
Lucene API - searching

- Open an IndexSearcher on top of an IndexReader over a Directory
- Many query types: TermQuery, MultiTermQuery, BooleanQuery, WildcardQuery, PhraseQuery, PrefixQuery, MultiPhraseQuery, FuzzyQuery, NumericRangeQuery, ...
- Get results from a TopDocs object



Apache Solr

- Ready to use enterprise search server
- REST (and programmatic) API
- Results in XML, JSON, PHP, Ruby, etc...
- Exploit Lucene power
- Scaling capabilities (replication, distributed search, ...)
- Administration interface
- Customizable via plugins



Solr Admin

Apache Solr

Examples: Simple Spatial Group By

Find: Invia Ripristina

□ Boost by Price

Field Facets

cat

- Electronics (14)
- Memory (3)
- Connector (2)
- Graphics Card (2)
- Hard Drive (2)
- Monitor (2)
- Search (2)
- Software (2)
- Camera (1)
- Copier (1)
- Multifunction Printer (1)
- Music (1)
- Printer (1)
- Scanner (1)

manu_exact

- Apache Software Foundation (2)
- Belkin (2)
- Canon Inc. (2)

28 results found in 42 ms Page 1 of 3

Test with some GB18030 encoded characters More Like This

Price: € 0,00

Features: No accents here 这是一个功能 This is a feature (translated) 这份文件是很有光泽 This document is very shiny (translated)

In Stock: true

Samsung SpinPoint P120 SP2514N - hard drive - 250 GB - ATA-133 More Like This

Price: € 92,00

Features: 7200RPM, 8MB cache, IDE Ultra ATA-133 NoiseGuard, SilentSeek technology, Fluid Dynamic Bearing (FDB) motor

In Stock: true

Larger Map

Maxtor DiamondMax 11 - hard drive - 500 GB - SATA-300 More Like This

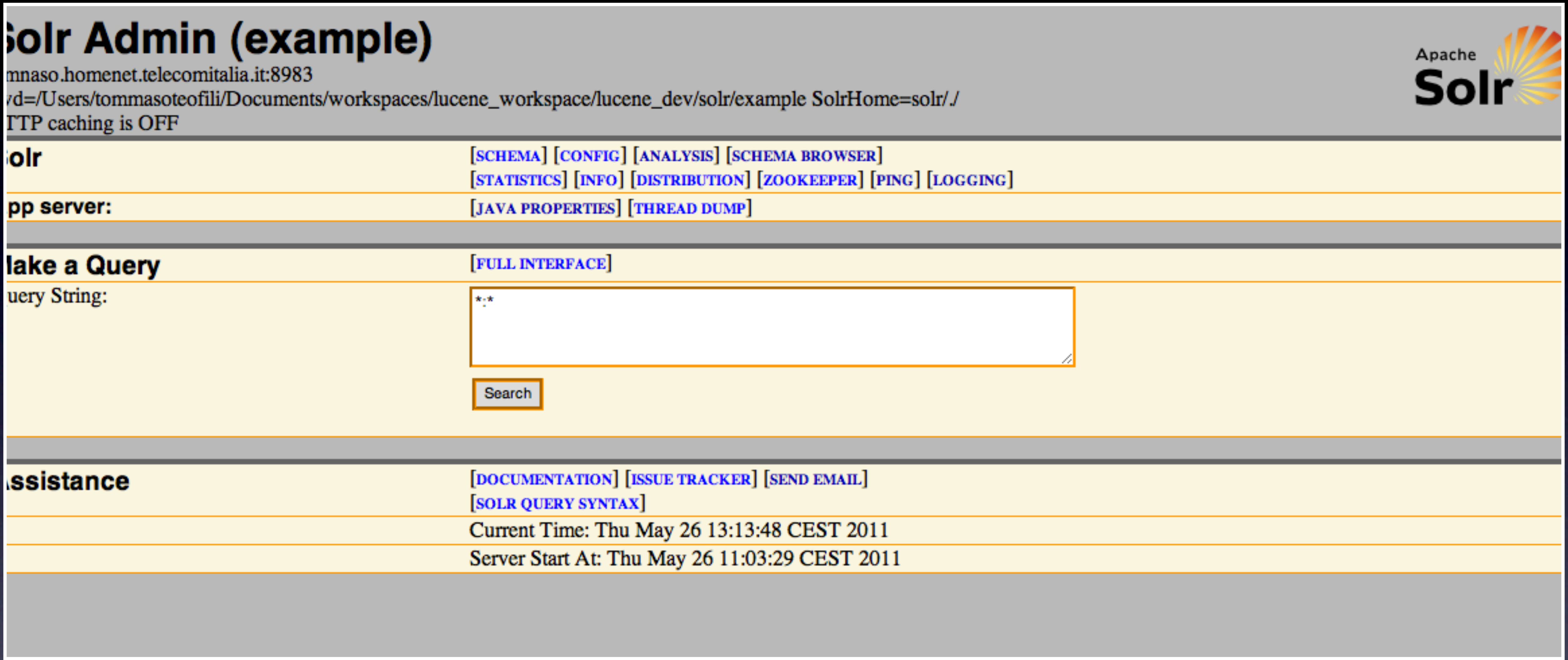
Price: € 350,00

Features: SATA 3.0Gb/s, NCQ 8.5ms seek 16MB cache

In Stock: true

Larger Map

Apache Solr 3.1.0



ṡolr Admin (example)

mnaso.homenet.telecomitalia.it:8983
/d=/Users/tommasoteofili/Documents/workspaces/lucene_workspace/lucene_dev/solr/example SolrHome=solr/./
ITP caching is OFF

;olr	[SCHEMA] [CONFIG] [ANALYSIS] [SCHEMA BROWSER] [STATISTICS] [INFO] [DISTRIBUTION] [ZOOKEEPER] [PING] [LOGGING]
pp server:	[JAVA PROPERTIES] [THREAD DUMP]
lake a Query	[FULL INTERFACE]
uery String:	*:*
	Search
ıssistance	[DOCUMENTATION] [ISSUE TRACKER] [SEND EMAIL] [SOLR QUERY SYNTAX]
	Current Time: Thu May 26 13:13:48 CEST 2011
	Server Start At: Thu May 26 11:03:29 CEST 2011

Apache Solr - admin UI

Solr - project status

- Solr 3.1.0 version released in March 2011
- Lucene/Solr is now a single project
- Huge community
- Backed by Lucid Imagination

Solr basic configuration

- schema.xml
  - contains types definitions for field analysis (field type+tokenizers+filters)
  - contains field definitions
- solrconfig.xml
  - contains the Solr instance configuration

Solr - schema.xml

- Types (with index/query Analyzers)
- Fields with name, type and options
- Unique key
- Dynamic fields
- Copy fields

Solr - content analysis

- define documents' model
- each document consists of fields
- each field
  - has attributes telling Solr how to handle its contents
  - contains free text, keywords, dates, numbers, etc.

Solr - content analysis

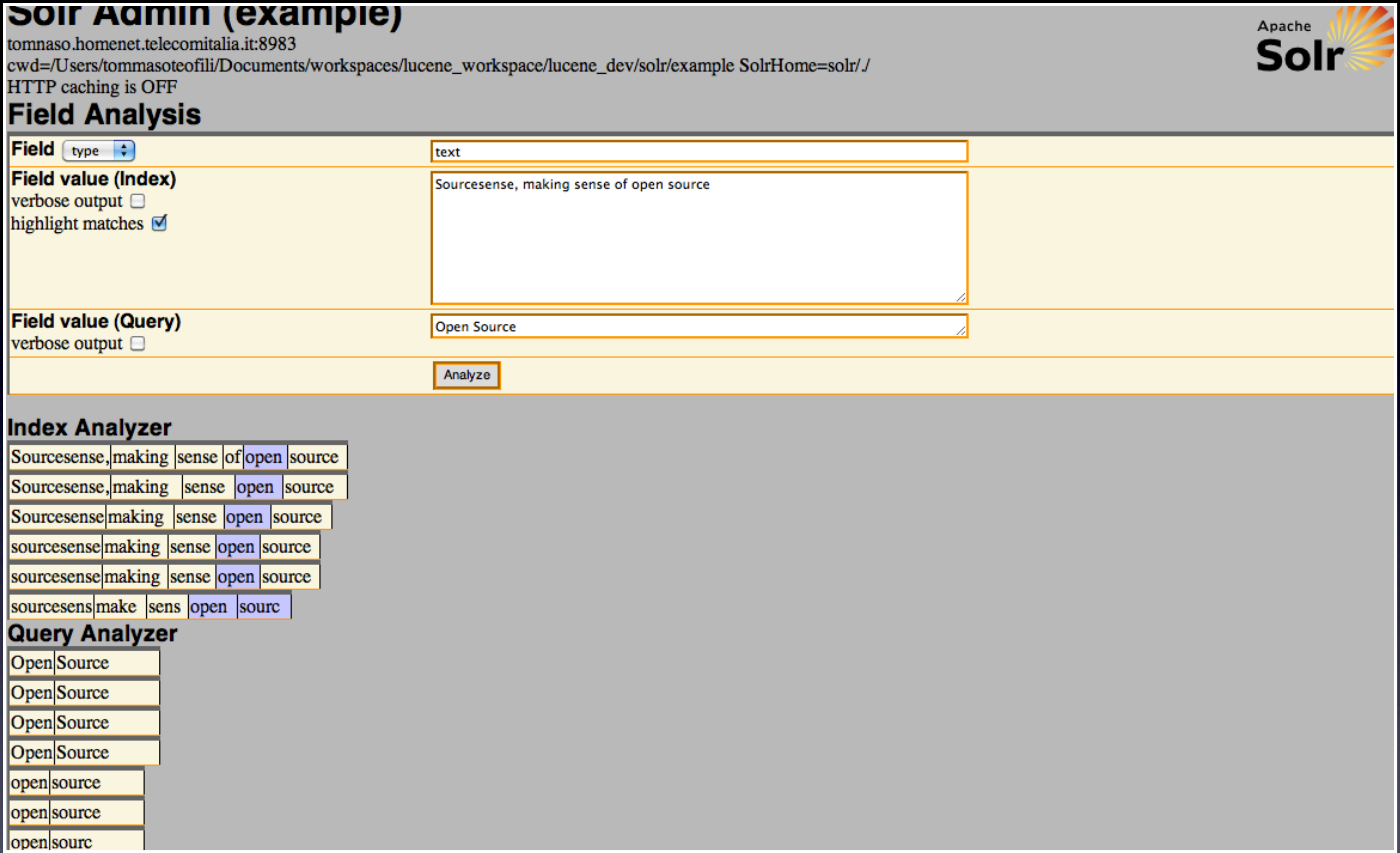
- Analyzer: create tokens using a Tokenizer and, eventually, some filters (TokenFilters)
- Each field can define an Analyzer at ‘query’ time and another at ‘index’ time, or the same in both cases
- Each field can be indexed (searchable), stored (possibly fetched with results), multivalued, required, etc.

Solr - content analysis

- Commonly used tokenizers:
  - WhitespaceTokenizerFactory
  - StandardTokenizerFactory
  - KeywordTokenizerFactory
  - PatternTokenizerFactory
  - HTMLStripWhitespaceTokenizerFactory
  - HTMLStripStandardTokenizerFactory

Solr - content analysis

- Commonly used TokenFilters:
  - SnowballPorterFilterFactory
  - StopFilterFactory
  - LengthFilterFactory
  - LowerCaseFilterFactory
  - WordDelimiterFilterFactory
  - SynonymFilterFactory
  - PatternReplaceFilterFactory
  - ReverseWildcardFilterFactory
  - CharFilterFactories (Mapping,HtmlString)



Solr Admin (example)

tomnaso.homenet.telecomitalia.it:8983

cwd=/Users/tommasoteofili/Documents/workspaces/lucene_workspace/lucene_dev/solr/example SolrHome=solr/./

HTTP caching is OFF

Field Analysis

Field type: text

Field value (Index): Sourcesense, making sense of open source

verbose output □

highlight matches ☑

Field value (Query): Open Source

verbose output □

Index Analyzer

Sourcesense, | making | sense | of | open | source

Sourcesense, | making | sense | open | source

Sourcesense | making | sense | open | source

sourcesense | making | sense | open | source

sourcesense | making | sense | open | source

sourcesens | make | sens | open | sourc

Query Analyzer

Open | Source

Open | Source

Open | Source

Open | Source

open | source

open | source

open | sourc

Debugging analysis

Solr - solrconfig.xml

- Data directory (where Solr will write the Lucene index)
- Caches configuration: documents, query results, filters
- Request handlers definition (search/update handlers)
- Update request processor chains definition
- Event listeners (newSearcher, firstSearcher)
- Fine tuning parameters
- ...

Solr - indexing

- Update requests on index are given with XML commands via HTTP POST
- <*add*> to insert and update
    - *<add> <doc boost="2.5">*
    - *<field name="employeeId">05991</field>*
    - *</doc></add>*
- <*delete*> to remove by unique key or query
    - *<delete><id>05991</id></delete>*
    - *<delete><query>office:Bridgewater</query></delete>*
- <*commit/*> reopen readers on the new index version
- <*optimize/*> optimize index internal structure for faster access

Solr - basic indexing

- REST call - XML/JSON
  - curl ‘http://localhost:8983/solr/update?commit=true’ -H "Content-Type: text/xml" --data-binary '<add><doc><field name="id">testdoc</field></doc></add>'
  - curl 'http://localhost:8983/solr/update/json?commit=true' -H 'Content-type:application/json' -d ' { "add": {"doc": {"id" : "TestDoc1", "title" : "test1"} } }’

Solr - binary files indexing

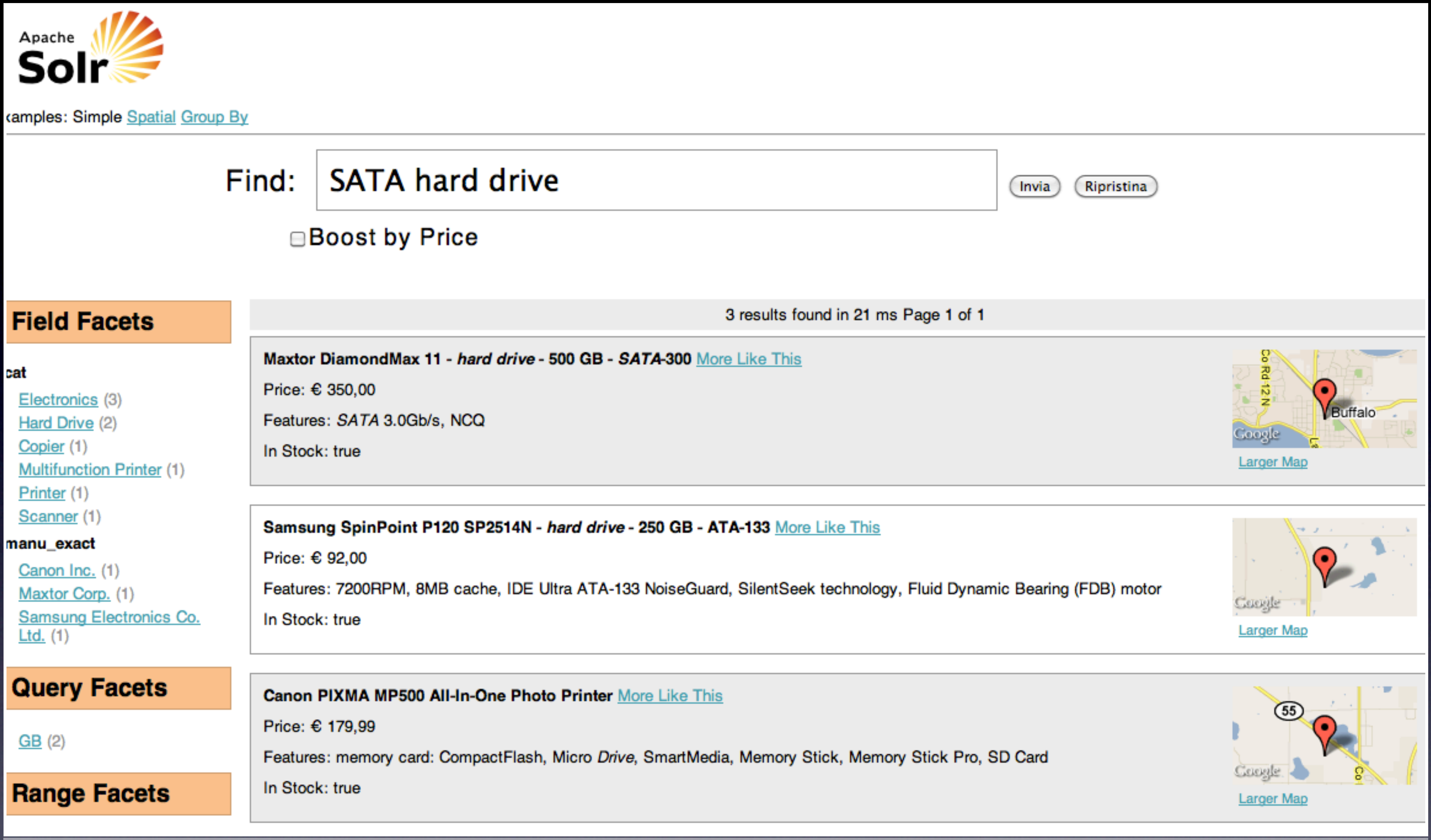
- Many documents are produced in (properietary) binary formats : PDF, RTF, XLS, etc.
- Apache Tika integrated in Solr REST service for indexing such documents
- curl "http://localhost:8983/solr/update/extract?literal.id=doc1&commit=true" -F "myfile=@tutorial.html"

Solr - index analysis

- Luke is a tool for navigating Lucene indexes
- For each field : top terms, distinct terms, terms histogram, etc.
- LukeRequestHandler :
  - http://localhost:8983/solr/admin/luke?wt=xslt&tr=luke.xsl

Solr - data import handler

- DBMS
- FileSystem
- HTTP



Apache Solr

amples: Simple Spatial Group By

Find: SATA hard drive Invia Ripristina

☐ Boost by Price

Field Facets

cat

- Electronics (3)
- Hard Drive (2)
- Copier (1)
- Multifunction Printer (1)
- Printer (1)
- Scanner (1)

manu_exact

- Canon Inc. (1)
- Maxtor Corp. (1)
- Samsung Electronics Co. Ltd. (1)

Query Facets

- GB (2)

Range Facets

3 results found in 21 ms Page 1 of 1

Maxtor DiamondMax 11 - *hard drive* - 500 GB - *SATA*-300 More Like This

Price: € 350,00

Features: *SATA* 3.0Gb/s, NCQ

In Stock: true

Larger Map

Samsung SpinPoint P120 SP2514N - *hard drive* - 250 GB - ATA-133 More Like This

Price: € 92,00

Features: 7200RPM, 8MB cache, IDE Ultra ATA-133 NoiseGuard, SilentSeek technology, Fluid Dynamic Bearing (FDB) motor

In Stock: true

Larger Map

Canon PIXMA MP500 All-In-One Photo Printer More Like This

Price: € 179,99

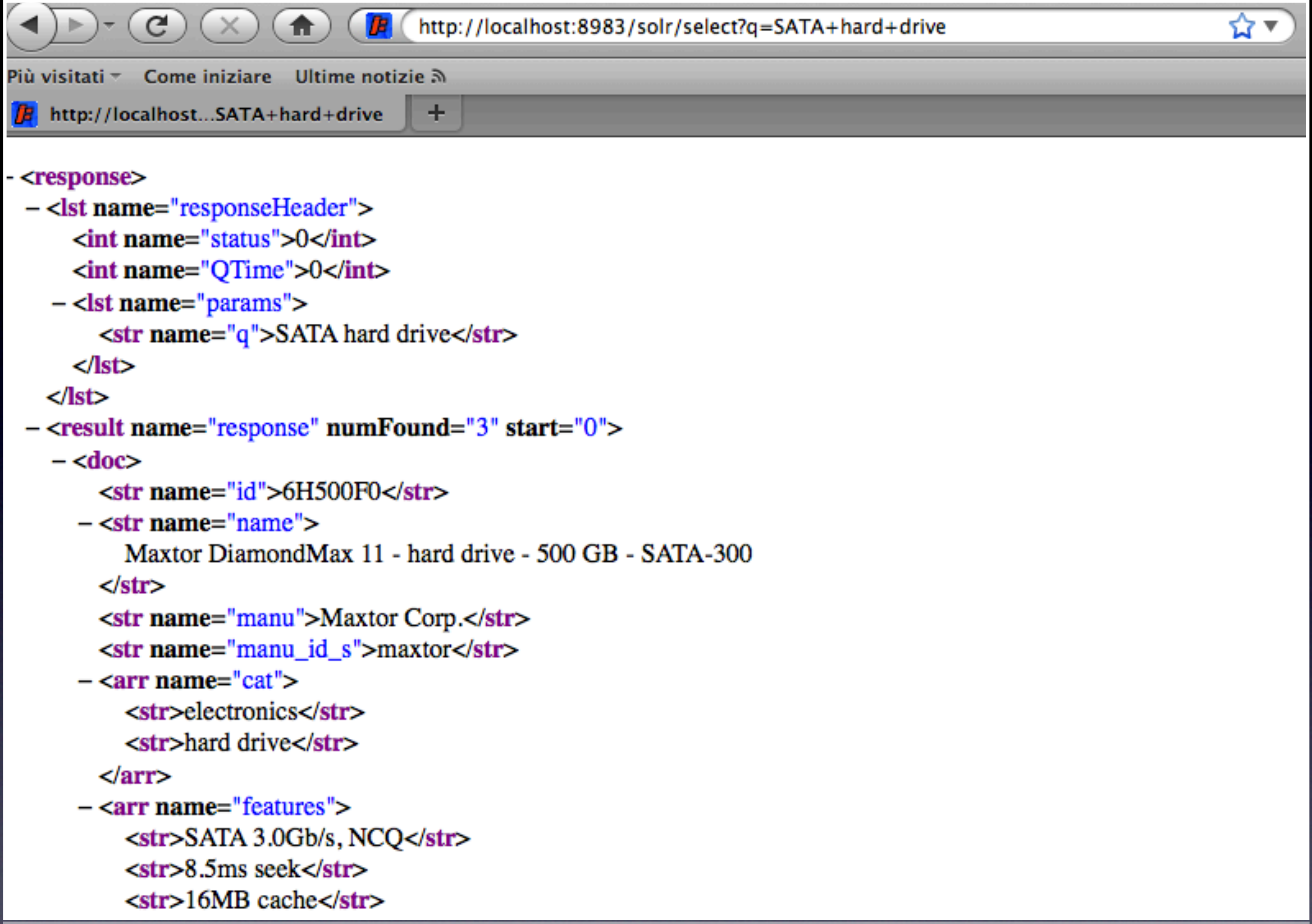
Features: memory card: CompactFlash, Micro *Drive*, SmartMedia, Memory Stick, Memory Stick Pro, SD Card

In Stock: true

Larger Map

Solr - searching

Solr - searching


http://localhost:8983/solr/select?q=SATA+hard+drive
Più visitati
Come iniziare
Ultime notizie
http://localhost...SATA+hard+drive

```
- <response>
  – <lst name="responseHeader">
      <int name="status">0</int>
      <int name="QTime">0</int>
    – <lst name="params">
        <str name="q">SATA hard drive</str>
      </lst>
    </lst>
  – <result name="response" numFound="3" start="0">
    – <doc>
        <str name="id">6H500F0</str>
      – <str name="name">
          Maxtor DiamondMax 11 - hard drive - 500 GB - SATA-300
        </str>
        <str name="manu">Maxtor Corp.</str>
        <str name="manu_id_s">maxtor</str>
      – <arr name="cat">
          <str>electronics</str>
          <str>hard drive</str>
        </arr>
      – <arr name="features">
          <str>SATA 3.0Gb/s, NCQ</str>
          <str>8.5ms seek</str>
          <str>16MB cache</str>
```

Solr - query syntax

- query fields with fieldname:value
- + - AND OR NOT operators
- Range queries on date or numeric fields, ex: timestamp:[* TO NOW]
- Boost terms, ex: people^4 profits
- Fuzzy search, ex: roam~0.6
- Proximity search, ex: “apache solr”~2
- ...

Solr - basic search

- parameters:
  - q: the query
  - start: offset of the first result
  - rows: max no. of results returned
  - fl: comma separated list of fields to return
  - defType: specify the query parser
  - debugQuery: enable query debugging
  - wt: result format (xml, json, php, ruby, javabin, etc)

Solr - query parsers

- Most used:
  - Default Lucene query parser
  - DisMax query parser
  - eDisMax query parser

Solr - highlighting

- can be done on fields with stored=”true”
- returns a snippet containing the higlighted terms for each doc
- enabled with hl=true&hl.fl=fieldname1,fieldname2

Solr - sorting results

- Sorting can be done on the "score" of the document, or on any multiValued="false" indexed="true" field provided that field is either non-tokenized (ie: has no Analyzer) or uses an Analyzer that only produces a single term
- add parameter &sort=score desc, inStock desc, price asc
- can sort on function queries (see later)

Solr - filter queries

- get a subset of the index
- place it in a cache
- run queries for such a “filter” in memory
- add parameter &fq=category:hardware
- if multiple fq parameters the query will be run against the intersection of the specified filters

Solr - facets

- facet by:
  - field value
  - arbitrary queries
  - range
- can facet on fields with indexed=”true”

Solr - function queries

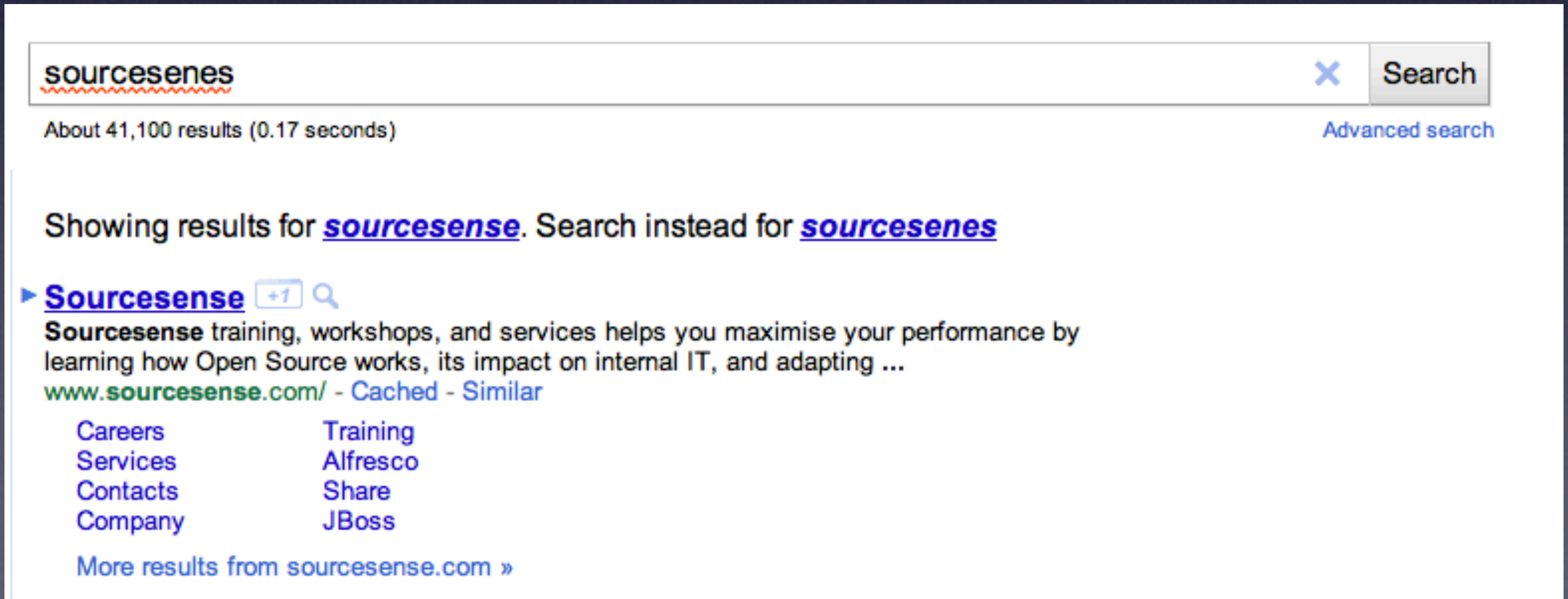
- allow deep customization of ranking :
  - http://localhost:8983/solr/select/?fl=score,id&q=DDR&sort=termfreq(text,memory)%20desc
- functions : sum, sub, product , div ,pow, abs, log, sqrt, map, scale, termfreq, ...

Solr - query elevation

- useful for “marketing”
- configure the top results for a given query regardless of the normal Lucene scoring
- http://localhost:8983/solr/elevate?q=best%20product&enableElevation=true

Solr - spellchecking

- collects suggestions about input query
- eventually correct user query with "suggested" terms



About 41,100 results (0.17 seconds)

Showing results for ***sourcesense***. Search instead for ***sourcesenes***

Sourcesense

Sourcesense training, workshops, and services helps you maximise your performance by learning how Open Source works, its impact on internal IT, and adapting ...

www.**sourcesense**.com/ - Cached - Similar

Careers	Training
Services	Alfresco
Contacts	Share
Company	JBoss

More results from sourcesense.com »

Solr - spellchecking

- build a spellcheck index dynamically
- return suggested results
- http://localhost:8983/solr/spell?q=hell ultrashar&spellcheck=true&spellcheck.collate=true&spellcheck.build=true
- useful to create custom query converters
  - <queryConverter name="queryConverter" class="org.apache.solr.spelling.SpellingQueryConverter"/>

Solr - similarity

- get documents “similar” to a given document or a set of documents
- Vector Space Model
- http://localhost:8983/solr/select?q=apache&mlt=true&mlt.fl=manu,cat&mlt.mindf=1&mlt.mintf=1&fl=id,score

Solr - geospatial search

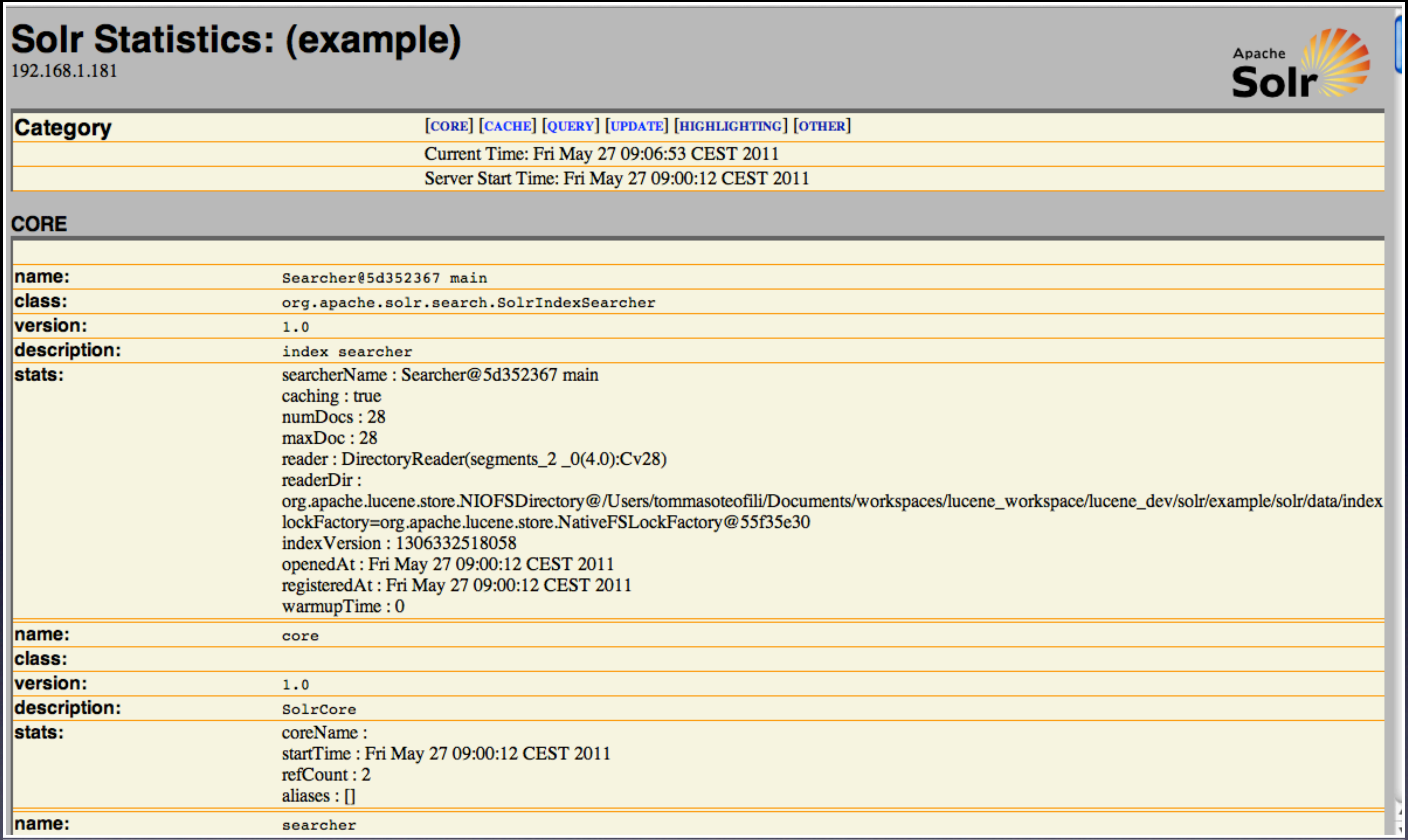
- index location data
- query by spatial concepts and sort by distance
- find all documents with store position at no more than 5km than a specified point
- http://localhost:8983/solr/select?&indent=true&fl=name,store&q=*:*&fq={!geofilt%20sfield=store}&pt=45.15,-93.85&d=5

Solr - field collapsing

- group resulting documents on per field basis
  - http://localhost:8983/solr/select?&indent=true&fl=id,name&q=solr+memory&group=true&group.field=manu_exact
- useful for displaying results in a smart way
- see SOLR-236

Solr - join

- new feature (SOLR-2272)
- many users ask for it
- quite of a paradigm change
- http://localhost:8983/solr/select?q={!join+from=manu_id_s%20to=id}ipod&fl=id,manu_id&debugQuery=true



Solr Statistics: (example)

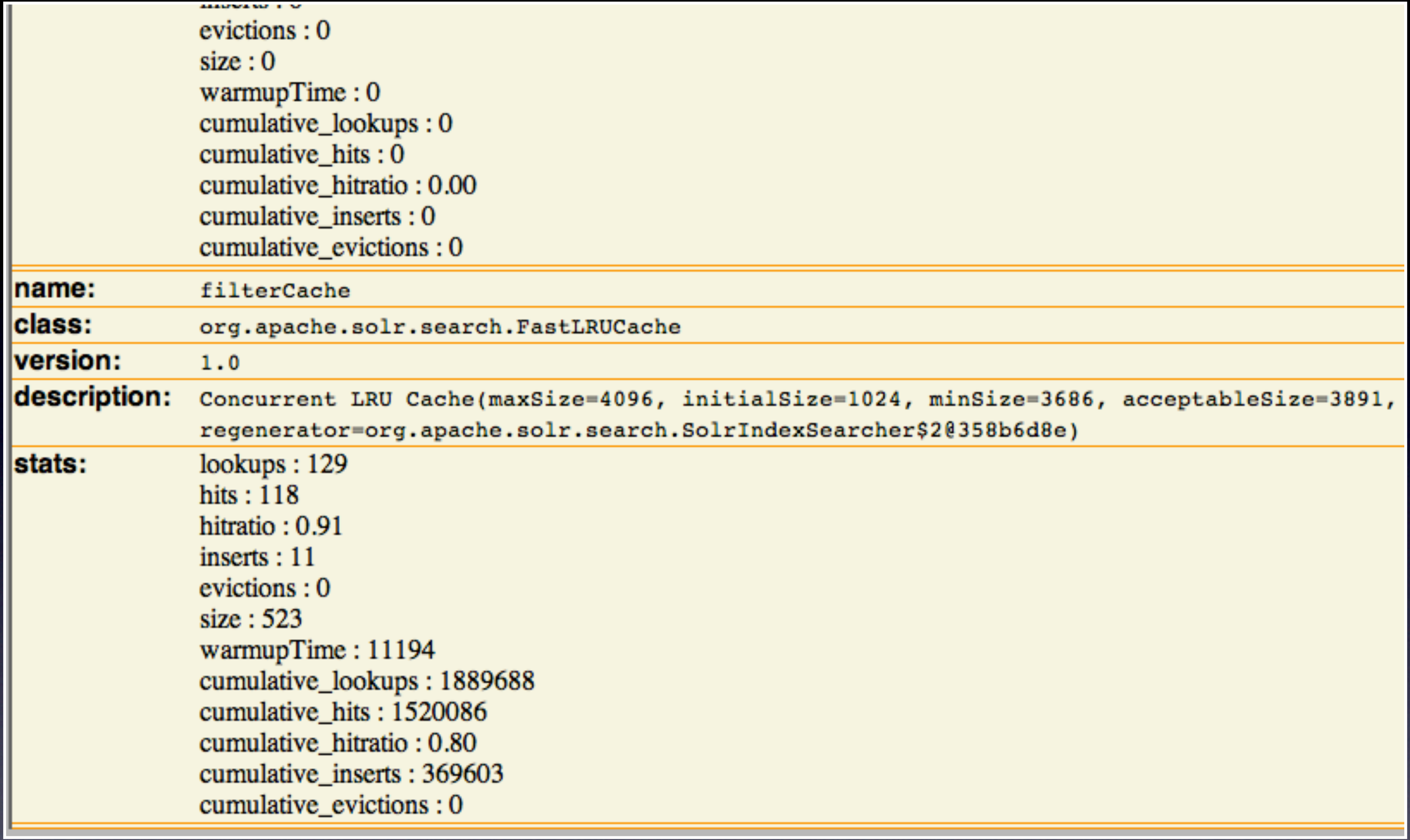
192.168.1.181

Category	[CORE] [CACHE] [QUERY] [UPDATE] [HIGHLIGHTING] [OTHER]
	Current Time: Fri May 27 09:06:53 CEST 2011
	Server Start Time: Fri May 27 09:00:12 CEST 2011

CORE

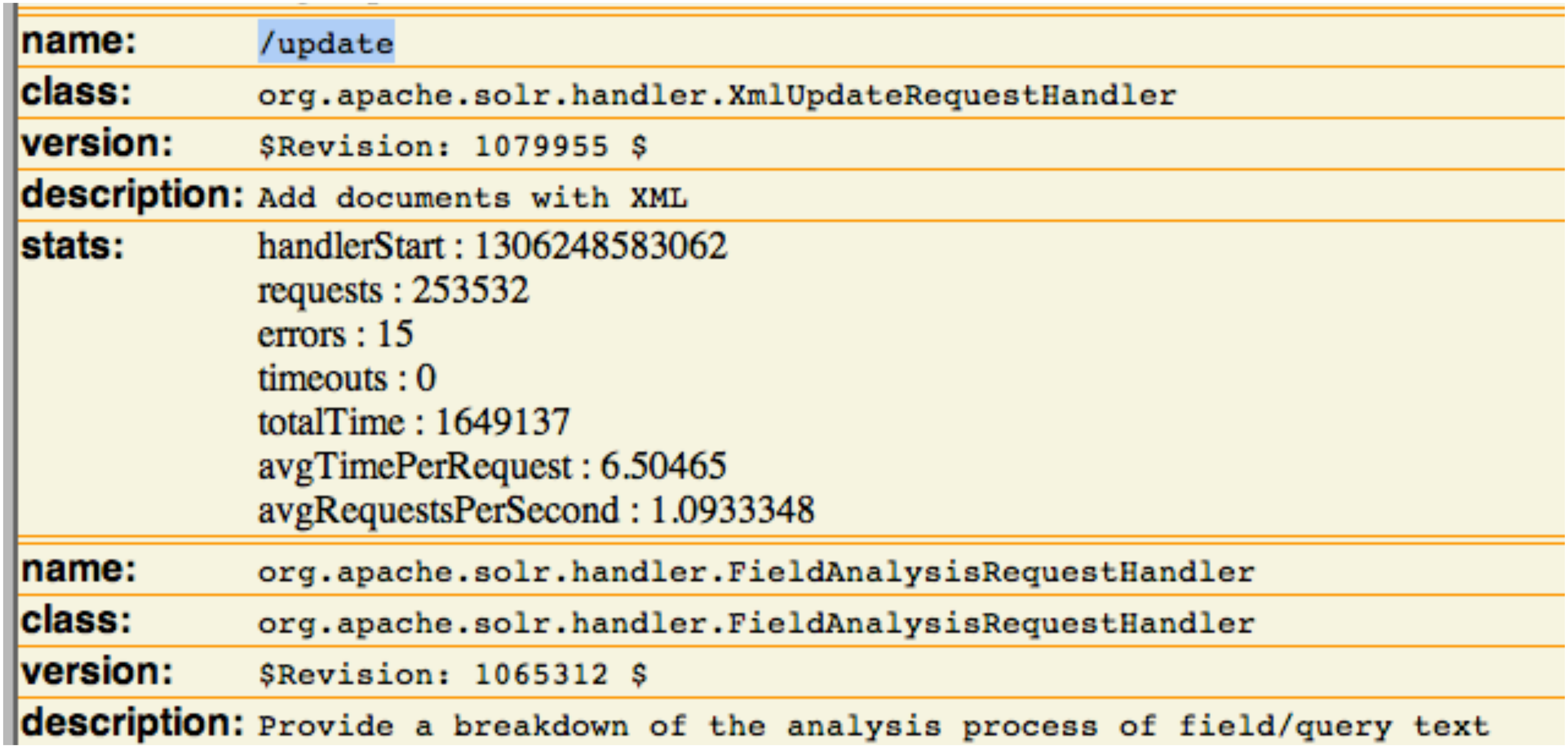
name:	Searcher@5d352367 main
class:	org.apache.solr.search.SolrIndexSearcher
version:	1.0
description:	index searcher
stats:	searcherName : Searcher@5d352367 main caching : true numDocs : 28 maxDoc : 28 reader : DirectoryReader(segments_2 _0(4.0):Cv28) readerDir : org.apache.lucene.store.NIOFSDirectory@/Users/tommasoteofili/Documents/workspaces/lucene_workspace/lucene_dev/solr/example/solr/data/index lockFactory=org.apache.lucene.store.NativeFSLockFactory@55f35e30 indexVersion : 1306332518058 openedAt : Fri May 27 09:00:12 CEST 2011 registeredAt : Fri May 27 09:00:12 CEST 2011 warmupTime : 0
name:	core
class:	
version:	1.0
description:	SolrCore
stats:	coreName : startTime : Fri May 27 09:00:12 CEST 2011 refCount : 2 aliases : []
name:	searcher

Solr statistics



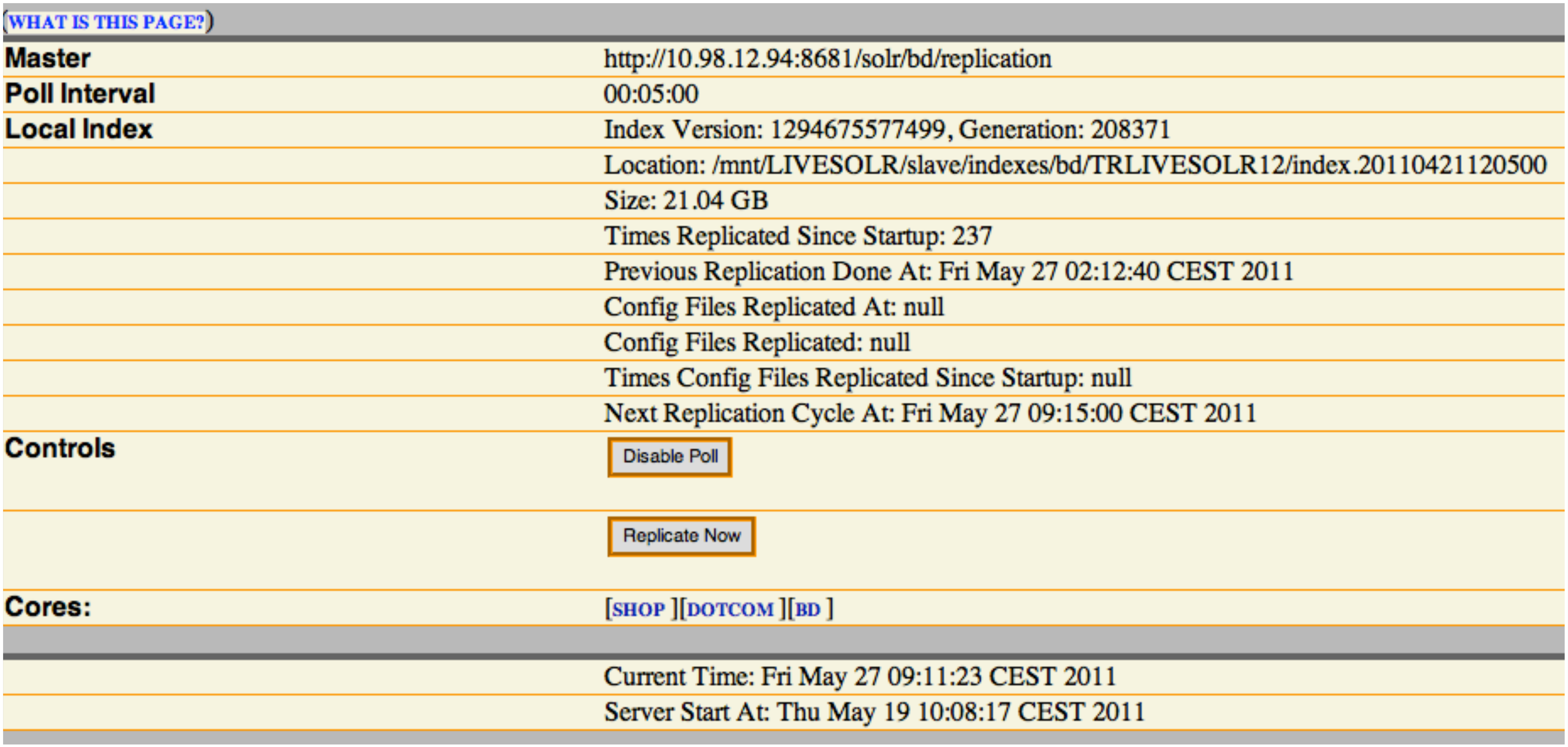
	inserts : 0 evictions : 0 size : 0 warmupTime : 0 cumulative_lookups : 0 cumulative_hits : 0 cumulative_hitratio : 0.00 cumulative_inserts : 0 cumulative_evictions : 0
name:	`filterCache`
class:	`org.apache.solr.search.FastLRUCache`
version:	`1.0`
description:	`Concurrent LRU Cache(maxSize=4096, initialSize=1024, minSize=3686, acceptableSize=3891, regenerator=org.apache.solr.search.SolrIndexSearcher$2@358b6d8e)`
stats:	lookups : 129 hits : 118 hitratio : 0.91 inserts : 11 evictions : 0 size : 523 warmupTime : 11194 cumulative_lookups : 1889688 cumulative_hits : 1520086 cumulative_hitratio : 0.80 cumulative_inserts : 369603 cumulative_evictions : 0

Solr statistics



name:	/update
class:	org.apache.solr.handler.XmlUpdateRequestHandler
version:	$Revision: 1079955 $
description:	Add documents with XML
stats:	handlerStart : 1306248583062 requests : 253532 errors : 15 timeouts : 0 totalTime : 1649137 avgTimePerRequest : 6.50465 avgRequestsPerSecond : 1.0933348
name:	org.apache.solr.handler.FieldAnalysisRequestHandler
class:	org.apache.solr.handler.FieldAnalysisRequestHandler
version:	$Revision: 1065312 $
description:	Provide a breakdown of the analysis process of field/query text

Solr statistics



(WHAT IS THIS PAGE?)

Master	http://10.98.12.94:8681/solr/bd/replication
Poll Interval	00:05:00
Local Index	Index Version: 1294675577499, Generation: 208371
	Location: /mnt/LIVESOLR/slave/indexes/bd/TRLIVESOLR12/index.20110421120500
	Size: 21.04 GB
	Times Replicated Since Startup: 237
	Previous Replication Done At: Fri May 27 02:12:40 CEST 2011
	Config Files Replicated At: null
	Config Files Replicated: null
	Times Config Files Replicated Since Startup: null
	Next Replication Cycle At: Fri May 27 09:15:00 CEST 2011
Controls	Disable Poll
	Replicate Now
Cores:	[SHOP][DOTCOM][BD]
	Current Time: Fri May 27 09:11:23 CEST 2011
	Server Start At: Thu May 19 10:08:17 CEST 2011

Solr replication

Solr Architectures

- Simple
- Multicore
- Replication
- Sharded

Solr - MultiCore

- Define multiple Solr cores inside one only Solr instance
- Each cores maintain its own index
- Unified administration interface
- Runtime commands to create, reload, load, unload, delete, swap cores
- Cores can be thought as ‘collections’
- Allow no downtime while deploying new features/ bugfixes

Solr - Replication

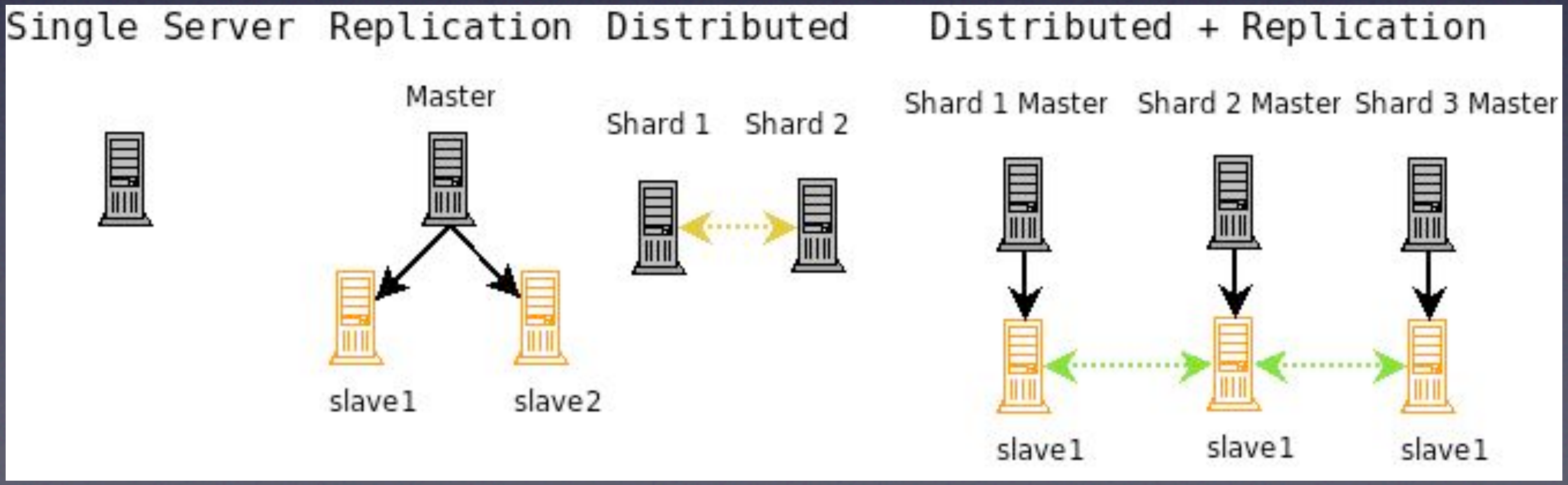
- It’s useful in case of high traffic to replicate a Solr instance and split (with eventually a VIP in front) the search load
- Master has the “original” index
- Slave polls master asking the latest version of index
- If slave has a different version of the index asks the master for the delta (rsync like)
- In the meanwhile indexes remain available
- No impact of indexing on search (almost)

Solr - Shards

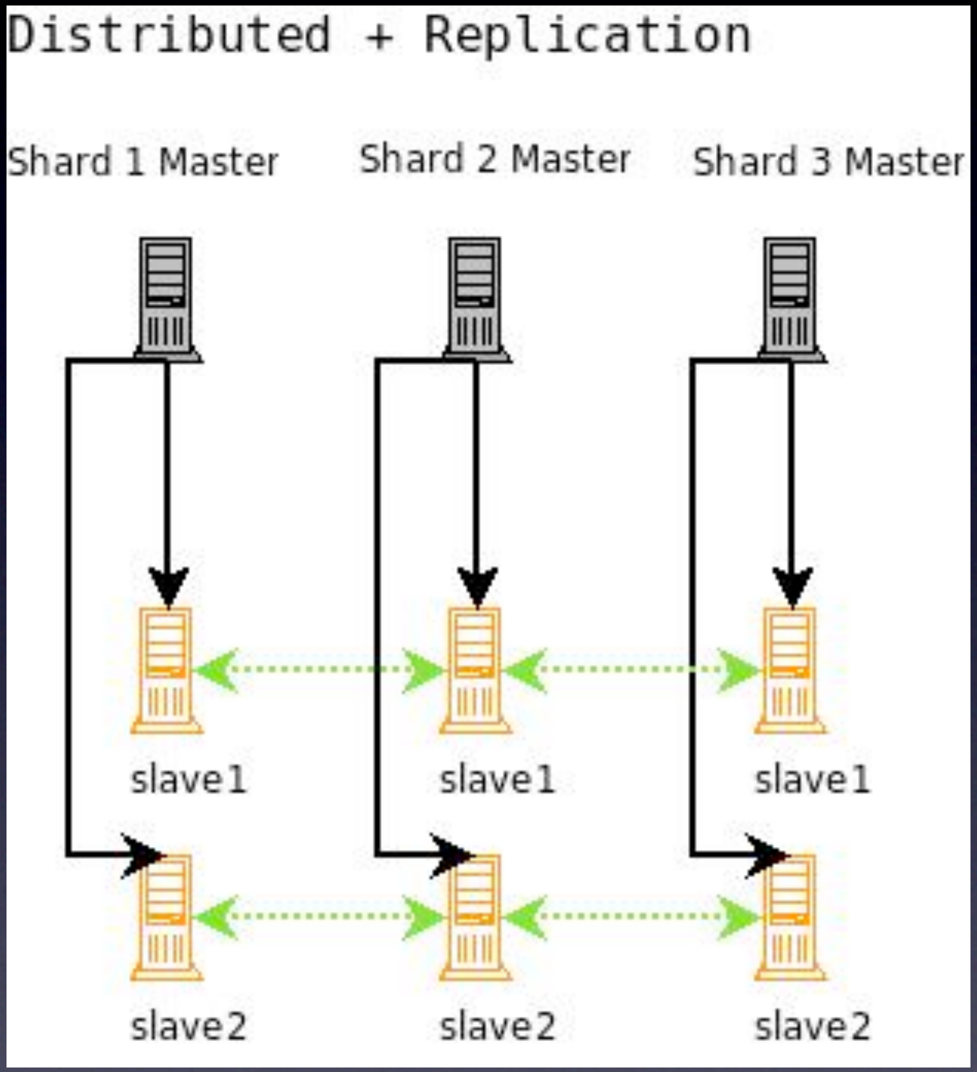
- When an index is too large, in terms of space or memory required, it can be useful to define two or more **shards**
- A shard is a Solr instance and can be searched or indexed independently
- At the same time it’s possible to query all the shards having the result be merged from the sub-results of each shard
- http://localhost:8983/solr/select?shards=localhost:8983/solr,localhost:7574/solr&q=category:information
- Note that the document distribution among indexes is up to the user (or who feeds the indexes)

Solr - Architectures

- When to use each?
- KISS principle
- High query load : replication
- Huge index : shard



Solr - Architectures



- High queries w/ large indexes : shard + replication

Solr - Architectures

- Tips & Tricks:
  - Don’t share indexes between Master and Slaves on distributed file systems (locking)
  - Anyway get rid of distributed file systems (slow)
  - Lucene/Solr is I/O intensive thus behaves better with quick disks
  - Always use MultiCore - hot deploy of changes/bugfixes
  - Replication is network intensive
  - Check replication poll time and indexing rates

Tips&Tricks

- Solr based SE development process
- Plugins
- Performance tuning
- Deploy

Process - t_0 analysis

- Analyze content
- Analyze queries
- Analyze collections
- Pre-existing query/index load (if any)
- Expected query/index load
- Desired throughput/avg response time
- First architecture

Process - n-th iteration

- index 10-15% content
- search stress test (analyze peaks) - use SolrMeter
- quality tests from stakeholders (accuracy, recall)
- eventually add/reconfigure features
- check http://wiki.apache.org/solr/FieldOptionsByUseCase and make sure fields used for faceting/sorting/highlighting/ etc. have proper options
- need to change field types/analysis/options - rebuild the index

Solr - Plugins

- QParserPlugin
- RequestHandler (Search/UpdateHandler)
- UpdateRequestProcessor
- ResponseWriter
- Cache

Performance tuning

- A huge tuning is done in schema.xml
- Configure Solr caches
- Set auto commit where possible
- Play with mergeFactor

Performance tuning

- The number of indexed fields greatly increases memory usage during indexing, segment merge time, optimization times, index size
- Stored fields impact on index size, search time, ...
- set omitNorms=”true” where it makes sense (disabling length normalization and index time boosting)
- set omitTermFreqAndPositions=”true” if no queries on this field using positions or should not influence score

Performance tuning

- FilterCache - unordered document ids for caching filter queries
- QueryResultCache - ordered document ids for caching queries results (caching only the returned docs)
- DocumentCache - stores stored fields (at least <max_results> * <max_concurrent_queries>
- Setup autowarming - keep caches warm after commits

Performance tuning

- Choose correct cache implementation FastLRUCache vs LRUCache
- FastLRUCache has faster gets and slower puts in single threaded operation and thus is generally faster than LRUCache when the hit ratio of the cache is high (> 75%)

Performance tuning

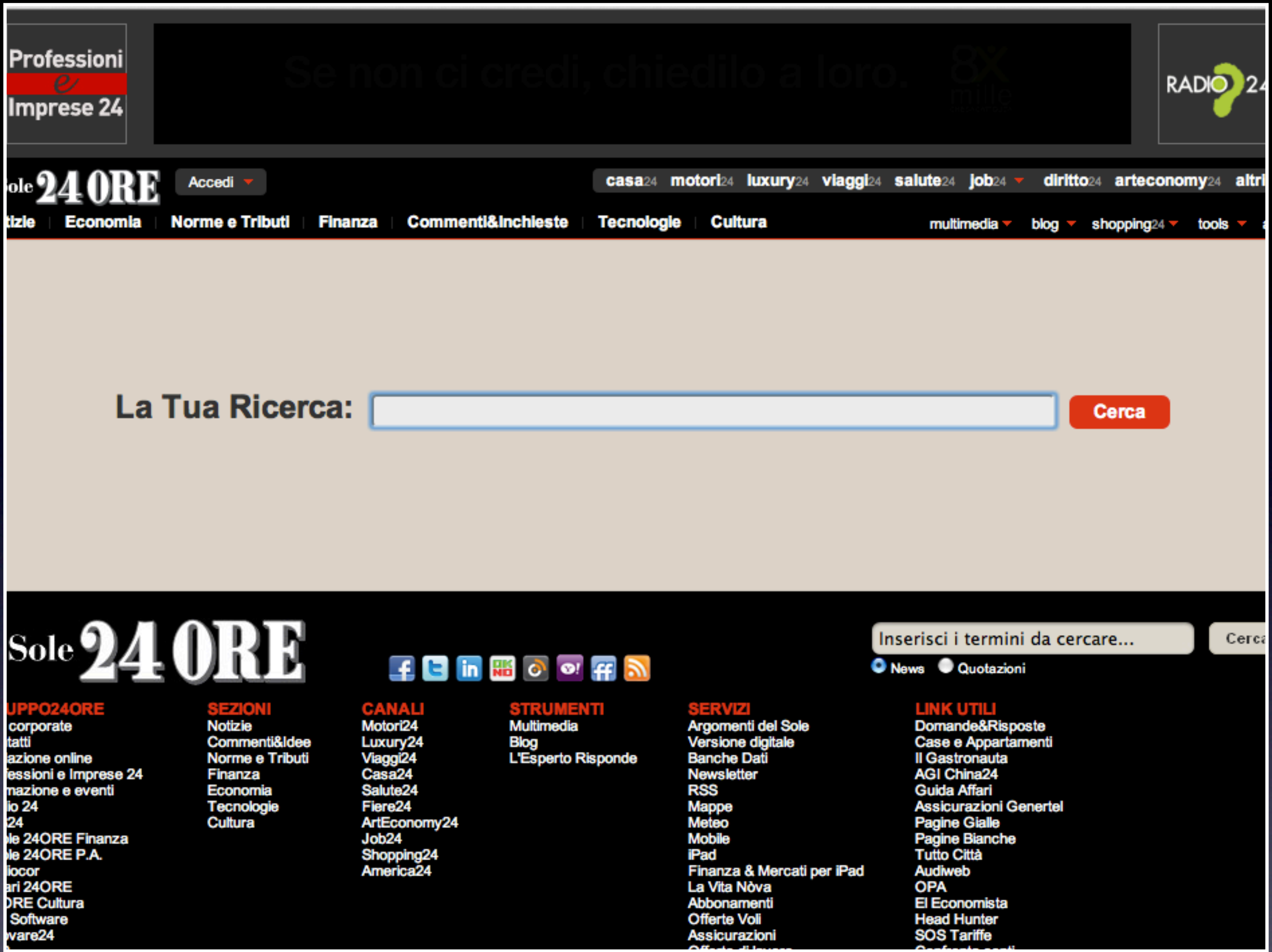
- Explicit warm sorted fields
- Often check cache statistics
- JVM options - don’t let the OS without memory!
- mergeFactor - impacts on the number of index segments created on the disk
  - low mF : smaller number of index files, which speeds up searching but more segment merges slow down indexing
  - high mF : generally improves indexing speed but gets less frequent merges, resulting in a collection with more index files which may slow searching

Performance tuning

- set autocommit where possible, this will avoid close and reopen of IndexReaders everytime a document is indexed - can choose max number of documents and/or time to wait before automatically do the commit
- finally...need to get your hand dirty!

Deploy

- SolrPackager by Simone Tripodi!
- It’s a Maven archetype
- Create standalone/multicore project
- Each project will generate a master and a slave instance
- Define environment dependent properties without having to manage N config files
- ‘mvn -Pdev package’ // will create a Tomcat package for the development environment



ole 24 ORE Accedi

casa24 motori24 luxury24 viaggi24 salute24 job24 diritto24 arteconomy24 altri

tizie | Economia | Norme e Tributi | Finanza | Commenti&Inchieste | Tecnologie | Cultura

multimedia blog shopping24 tools

La Tua Ricerca: Cerca

Sole 24 ORE

Inserisci i termini da cercare... Cerca

◉ News ○ Quotazioni

UPPO24ORE
- corporate
- tatti
- azione online
- essioni e Imprese 24
- mazione e eventi
- io 24
- 24
- le 24ORE Finanza
- le 24ORE P.A.
- iocor
- ari 24ORE
- DRE Cultura
- Software
- vare24

SEZIONI
- Notizie
- Commenti&Idee
- Norme e Tributi
- Finanza
- Economia
- Tecnologie
- Cultura

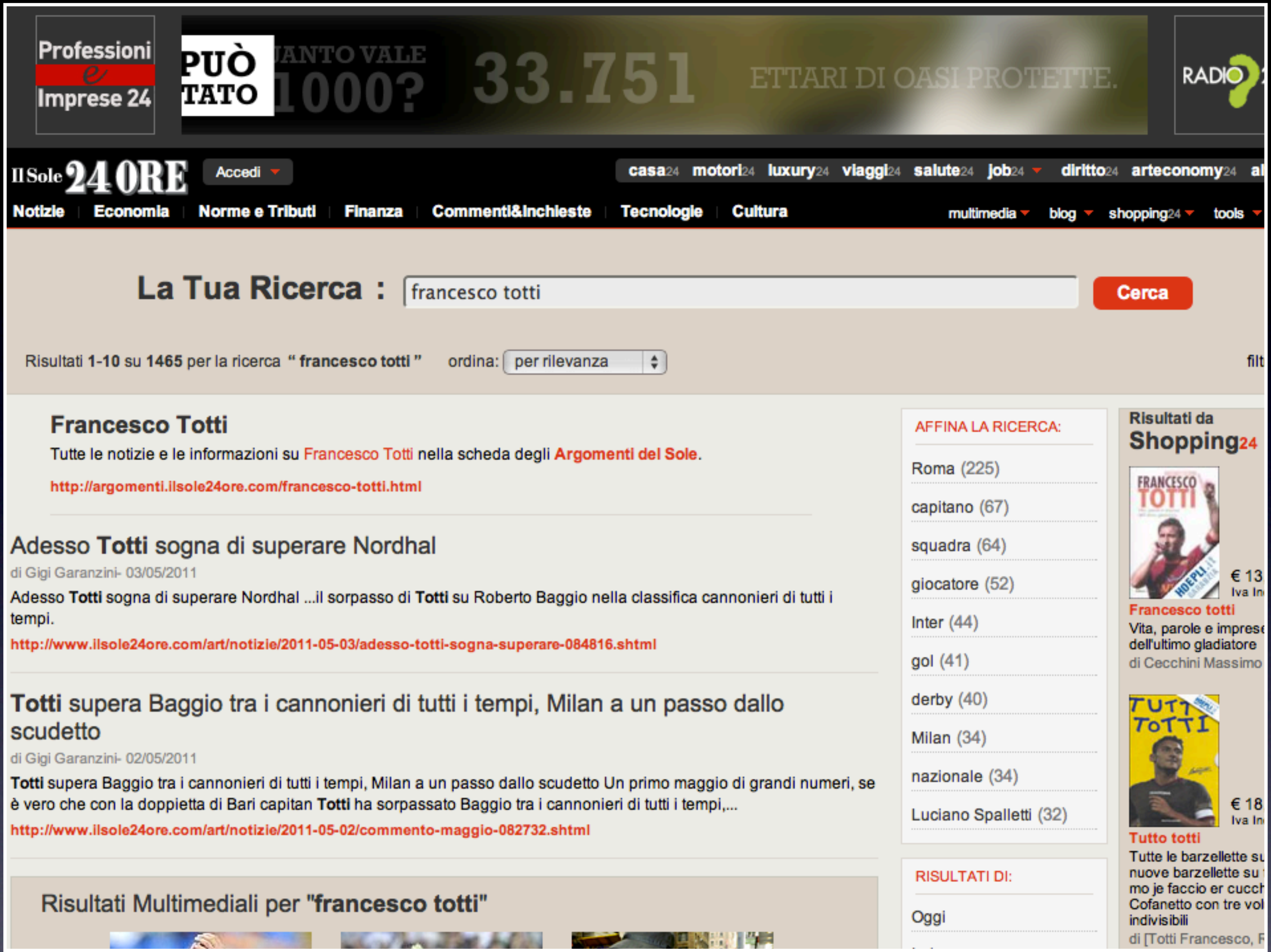
CANALI
- Motori24
- Luxury24
- Viaggi24
- Casa24
- Salute24
- Fiere24
- ArtEconomy24
- Job24
- Shopping24
- America24

STRUMENTI
- Multimedia
- Blog
- L'Esperto Risponde

SERVIZI
- Argomenti del Sole
- Versione digitale
- Banche Dati
- Newsletter
- RSS
- Mappe
- Meteo
- Mobile
- iPad
- Finanza & Mercati per iPad
- La Vita Nòva
- Abbonamenti
- Offerte Voli
- Assicurazioni
- Offerte di lavoro

LINK UTILI
- Domande&Risposte
- Case e Appartamenti
- Il Gastronauta
- AGI China24
- Guida Affari
- Assicurazioni Genertel
- Pagine Gialle
- Pagine Bianche
- Tutto Città
- Audiweb
- OPA
- El Economista
- Head Hunter
- SOS Tariffe
- Confronta conti

Case study



La Tua Ricerca : francesco totti Cerca

Risultati **1-10** su **1465** per la ricerca **" francesco totti "** ordina: per rilevanza filt

Francesco Totti

Tutte le notizie e le informazioni su Francesco Totti nella scheda degli **Argomenti del Sole**.

http://argomenti.ilsole24ore.com/francesco-totti.html

Adesso **Totti** sogna di superare Nordhal

di Gigi Garanzini- 03/05/2011

Adesso **Totti** sogna di superare Nordhal ...il sorpasso di **Totti** su Roberto Baggio nella classifica cannonieri di tutti i tempi.

http://www.ilsole24ore.com/art/notizie/2011-05-03/adesso-totti-sogna-superare-084816.shtml

Totti supera Baggio tra i cannonieri di tutti i tempi, Milan a un passo dallo scudetto

di Gigi Garanzini- 02/05/2011

Totti supera Baggio tra i cannonieri di tutti i tempi, Milan a un passo dallo scudetto Un primo maggio di grandi numeri, se è vero che con la doppietta di Bari capitan **Totti** ha sorpassato Baggio tra i cannonieri di tutti i tempi,...

http://www.ilsole24ore.com/art/notizie/2011-05-02/commento-maggio-082732.shtml

Risultati Multimediali per "**francesco totti**"

AFFINA LA RICERCA:

Roma	(225)
capitano	(67)
squadra	(64)
giocatore	(52)
Inter	(44)
gol	(41)
derby	(40)
Milan	(34)
nazionale	(34)
Luciano Spalletti	(32)

RISULTATI DI:

Oggi

Risultati da **Shopping24**

€ 13 Iva In

Francesco totti

Vita, parole e imprese dell'ultimo gladiatore

di Cecchini Massimo

€ 18 Iva In

Tutto totti

Tutte le barzellette su nuove barzellette su mo je faccio er cucc Cofanetto con tre vol indivisibili

di [Totti Francesco, F

Case study

Case Study

- Architecture analysis
- Plugin development
- Testing and support

Challenges

- Architecture
- Schema design

Challenge

- Architecture
  - 4B docs of ~4k each
  - ~3 req/sec overall
  - 3 collections:
    - |archive| = 3B
    - |2010-2011| = 1M
    - |intranet| = 0.9B

Challenge

- Content analysis
- get the example Solr schema.xml
- optimize the schema in order to enable both stemmed and unstemmed versions of fields: author, title, text, cat
- add omitNorms=”true” where possible
- add a field ‘html_content’ which will contain an HTML text but will be searched as clean text
- all string fields should be lowercased

Extras

- Clustering (Solr-Carrot2)
- Named entity extraction (Solr-UIMA)
- SolrCloud (Solr-Zookeeper)
- ManifoldCF
- Stanbol EntityHub
- Solandra (Solr-Cassandra)

THANKS!