



Search and browse super powers an introduction to Solr

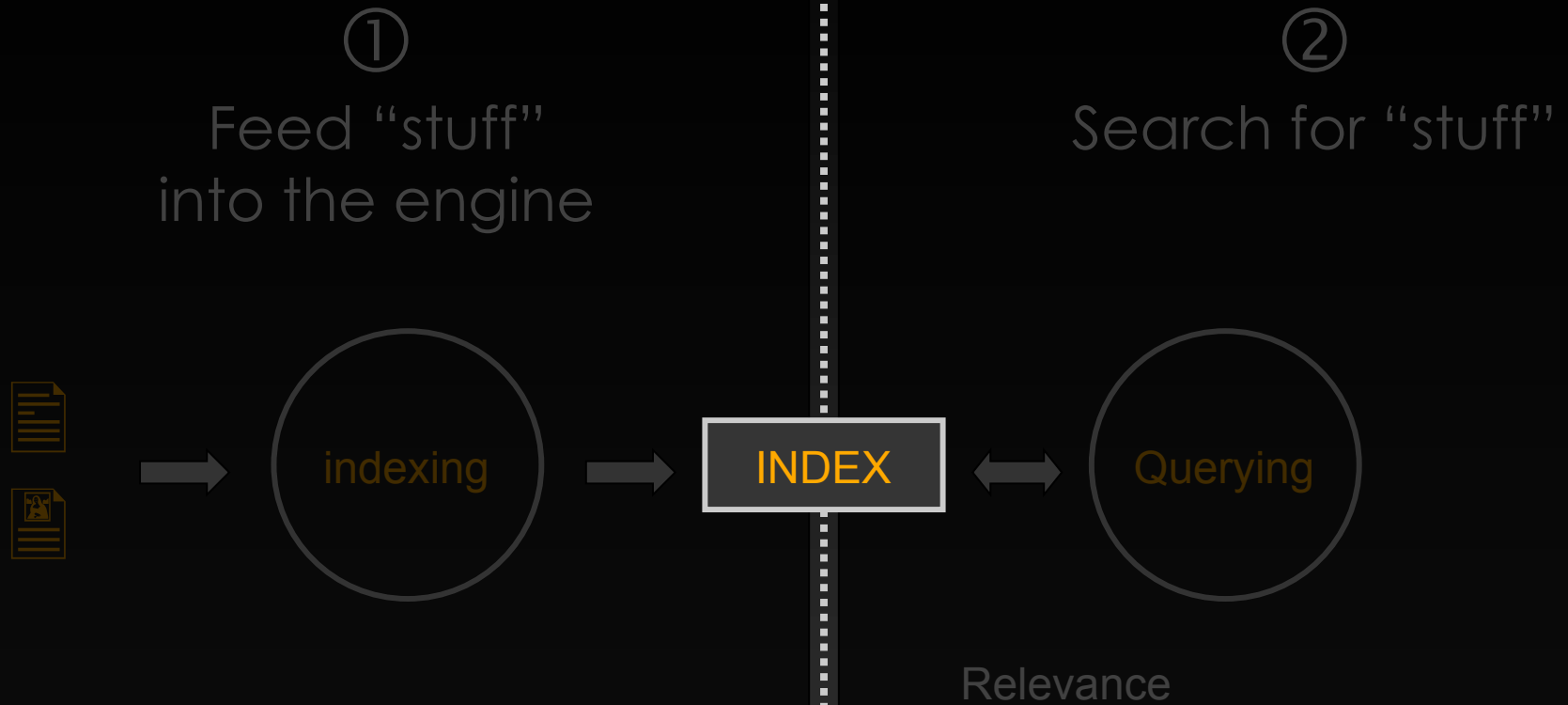
Ben Gaucherin - 11/05/2011

A quick tour of the VM

- The VirtualBox VM is a TurnkeyLinux Core (Ubuntu) setup with both a LAMP stack and a Solr instance running on Tomcat
- The user `root`'s password is `turnkey123`
- You can access a `shell` at: <http://192.168.56.101:12320>
- You can access `WebAdmin` at: <http://192.168.56.101:123221>
- Tomcat is installed in `/usr/share/tomcat6/`
- You can access the `Tomcat Admin` interface at: <http://192.168.56.101:8080/manager/html>
- The Tomcat admin user is `admin` and the password is `turnkey123`
- The Solr home directory which includes both a `conf` and `data` directory is `/etc/solr/`
- The php files are in `/var/www/solr_tut`

Part I – Key concepts for search

The basic model



Inverted index

- A “forward index” establishes a relationship between a document and the terms within it
 - Doc 1 → {“The” “little” “brown” “fox”}
 - Doc 2 → {“Bounce” “little” “brown” “ball”}
- An “inverted index” establishes a relationship between terms and the documents where they can be found
 - “The” → Doc 1
 - “Bounce” → Doc 2
 - “little” → Doc 1, Doc 2
 - “brown” → Doc 1, Doc 2
 - “ball” → Doc 2
 - “fox” → Doc 1

Search Results: *solr*

solr

SEARCH

About 104 results (0.07 seconds)

[Obama names Elena Kagan **solicitor** general | Harvard ...](#)

President-elect Barack Obama has nominated Harvard Law School (HLS) Dean Elena Kagan as **solicitor** general. If confirmed ...

news.harvard.edu/gazette/story/2009/01/obama-names-elena-kagan-solicitor-general/

[Howell Jackson named as prospective acting dean of Harvard ...](#)

... Harvard Law School (HLS), subject to the US Senate's confirmation of Dean Elena Kagan's nomination to serve as US **Solicitor** General, President ...

news.harvard.edu/gazette/story/2009/02/howell-jackson-named-as-prospective-acting-dean-of-harvard-law-school/

[Harvard Gazette: Archibald Cox dies at 92](#)

Professor <i>Emeritus</i> Archibald Cox, former Watergate special prosecutor and **solicitor** general, died Saturday (May 29) in his home in ...

news.harvard.edu/gazette/2004/06.03/15-cox.html

[Guantanamo Bay: Memo to W: Finally we agree on something ...](#)

... On your behalf, the **Solicitor** General of the United States argued that all of Guantánamo, including every square inch on which your base sits, is ...

www.people.fas.harvard.edu/~jldoming/images/jld_guantanamo.pdf

[HPAC PR | Harvard Gazette](#)

... Obama names Elena Kagan **solicitor** general. January 5, 2009. President-elect Barack Obama has nominated Harvard Law ...

news.harvard.edu/gazette/section/hpac-pr/page/9/?gz_page_type=archive

[Colleagues recall Kagan's years at Harvard | Harvard Gazette](#)

... 2009. In 2009, she was confirmed as US **solicitor** general. The ... Obama names Elena Kagan **solicitor** general. President-elect ...

news.harvard.edu/gazette/story/2010/08/senate-approves-kagan-for-high-court/

TOP STORIES

HARVARDgazette

1. **Often, we are what we were**
2. **Harvard serves up its own 'Plate'**
3. **Meditation found to increase brain size**
4. **New approach to traumatic brain injuries**
5. **Fewer drops to drink**
6. **Wandering mind not a happy mind**

TRENDING NOW

HARVARDgazette

1. **Weighing the risks of fracking**
2. **A better view of heart disease**
3. **Fewer drops to drink**
4. **Wandering mind not a happy mind**

Faceted search

- Items in the index have facets
 - For example working with books we could have the following facets: author(s), isbn, subject(s), publication date, publisher, etc.
- When you perform a search, facets provide additional information about the results
- Just about anything can be a facet
- With faceted search: no single taxonomy, no assumption of global relevance, etc.

Cross Registration

Any Cross-Reg Status

[Eligible for cross-registration](#) (319)

[Not eligible for cross-registration](#) (122)

Calendar Period

Any Calendar Period

[Jun to Aug 2011 \(Summer Term\)](#) (21)

[Sep to Dec 2011 \(Fall Term\)](#) (206)

[Jan 2012 \(Winter Session\)](#) (8)

[Jan to May 2012 \(Spring Term\)](#) (197)

Level

Any Credit Level

[Graduate](#) (354)

[Undergraduate](#) (118)

School

All Schools

[Faculty of Arts and Sciences](#) (252)

[Harvard Business School - MBA Program](#) (1)

[Harvard Extension School](#) (77)

[Harvard Graduate School of Design](#) (15)

[Harvard Graduate School of Education](#) (12)


[Harvard Kennedy School](#) (7)


[Harvard Law School](#) (13)

[Harvard Medical School](#) (11)

[Harvard School of Public Health](#) (33)





computer

SEARCH 

Start Over 

441 Courses Found

Sort by: [Course Name](#) | [School](#) | [Department](#) | [Day](#) | [Time](#) | [Relevance](#)

Course Information	Schedule	Relevance							
Abstraction and Design in Computation (CSCI E-250 (23371)) Harvard Extension School Spring 2012 John Gregory Morrisett PhD, Allen B. Cutting Professor of Computer Science, Harvard University More Detail ▼	<table border="1"><tr><td>S</td><td>M</td><td>T</td><td>W</td><td>T</td><td>F</td><td>S</td></tr></table> Online only	S	M	T	W	T	F	S	 0.78
S	M	T	W	T	F	S			
Accumulating Evidence: How to Conduct a Quantitative Research Synthesis That Informs Educational Policy and Practice (Part I) (S032A1) Harvard Graduate School of Education Spring 2012 James S. Kim More Detail ▼	<table border="1"><tr><td>S</td><td>M</td><td>T</td><td>W</td><td>T</td><td>F</td><td>S</td></tr></table> 1:00 p.m. - 4:00 p.m.	S	M	T	W	T	F	S	 0.22
S	M	T	W	T	F	S			
Adv Comp Bio & Bioinformatics (BIO513-01) Harvard School of Public Health Fall 2011 Guocheng Yuan and Winston Hide More Detail ▼	<table border="1"><tr><td>S</td><td>M</td><td>T</td><td>W</td><td>T</td><td>F</td><td>S</td></tr></table> 3:30 p.m. - 5:20 p.m.	S	M	T	W	T	F	S	 0.54
S	M	T	W	T	F	S			
Advanced Epidemiologic Methods (EPI207-01) Harvard School of Public Health Fall 2011 James M. Robins and Miguel Hernan More Detail ▼	<table border="1"><tr><td>S</td><td>M</td><td>T</td><td>W</td><td>T</td><td>F</td><td>S</td></tr></table> Monday 3:30 p.m. - 5:20 p.m.; Wednesday 3:30 p.m. - 5:20 p.m.; Thursday 1:30 p.m. - 3:20 p.m.	S	M	T	W	T	F	S	 0.25
S	M	T	W	T	F	S			

Part II – The Solr application

What is Solr?

- An Apache Software Foundation project
- Extends the capabilities of the Apache Lucene search engine
- Solr is a Java web application distributed as a WAR file – it runs in a servlet container (e.g. Tomcat)



HTTP
Request/Response



Solr
Engine



Solr application tour

Key applications elements

- Using `tomcat_dir/conf/Catalina/localhost/solr.xml` we tell Tomcat where `solr/home` is
- `solr/home` contains a `conf` directory that contains a number of configuration files:
 - `solrconfig.xml`
 - `schema.xml`
 - `spellings.txt`, `stopwords.txt`, `synonyms.txt`, etc.
- And the index will be located here:

```
<dataDir>  
    ${solr.data.dir:/Applications/tomcat/conf/solr_conf/data}  
</dataDir>
```

Part III – Indexing our data

The schema - types

```
<?xml version="1.0" encoding="utf-8" ?>
<schema name="ShelfLife Demo" version="1.0">
<types>
  <fieldType name="text" class="solr.TextField"
    positionIncrementGap="100">
    <analyzer type="index">
      ...
    </analyzer>
    <analyzer type="query">
      ...
    </analyzer>
  </fieldType>
</types>
```

The schema - fields

```
<fields>
  <field name="isbn" type="text" indexed="true" stored="true"
    required="true" />

  <field name="title" type="string" indexed="true" stored="true"
    required="true" />
  <field name="title det" type="text" index="true" stored="true"
    required="true" />
  <copyField source="title" dest="title_det" />

  ...

</fields>

<uniqueKey>isbn</uniqueKey>

<defaultSearchField>title</defaultSearchField>

<solrQueryParser defaultOperator="OR" />
</schema>
```

The data

- Posted to `/update`

```
<?xml version="1.0" encoding="utf-8" ?>
<add>
  <doc>
    <field name='title'>The little prince</field>
    <field name='subject'>space</field>
    <field name='subject'>story</field>
    ...
  </doc>
  <doc>
    ...
  </doc>

</commit/>
</optimize/>
</add>
```

`solr_data_post.php`

Part IV – Solr queries 101

You can talk to Solr directly

- The `/select` request handler
- `q`: the actual query
- `wt`: to specify the format for the results
- `start` and `rows`: to help with pagination through large numbers of results

```
http://192.168.56.101:8080/solr/  
select?q=*:*&wt=xml&start=5&rows=10
```

Adding facets

- We just turn facets on and specify the facet fields and additional configuration as needed

```
http://192.168.56.101:8080/solr/select?q=*:*&wt=xml&  
facet=true&facet.field=author
```

And many more

- Other query parameters (e.g. `hl` for highlighting)
- Other request handlers (see `solrconfig.xml` for a full list)
 - `/admin`
 - `/admin/ping`
 - `/debug/dump`
 - ...

`solr_search.php`

Part V – Faceted search user interface

Some examples...

Subjects:

Games(32) , Games/Puzzles(32) , Games / Gamebooks / Crosswords(31) , Business & Economics(27) , Business
Card Games - Poker(23) , General(23) , Poker(20) , Management - General(19) , Hobbies, Pastimes & Indoor G
Games(13) , Computers(13) , Business & Economics / Management(11) , Gambling(11) , Gambling: theories &
Professional & Programming(9) , Management(9) , Management & management techniques(9) , Card Games - C
Computer Books: General(7) , Reference(7) , Computer Books: Languages(6) , Computers - Languages / Progr
Programming - General(6) , Programming Languages - General(6) , Psychology(6) , Sale Books(6) , United Stat
Textbooks & Study Guides(5) , Conjuring & magic(5) , Entrepreneurship(5) , Gambling - General(5) , Gambling
Industrial management(5) , Magic(5) , Management & Business: General(5) , Religion(5) , Self-Help(5) , Body,
Success(4) , Computers - General Information(4) , Computers / Programming / General(4) , Computers / Progr
business & management(4) , New Age(4) , New Age / Body, Mind & Spirit(4) , Object-oriented programming (C
Politics/International Relations(4) , Popular Science(4) , Reference - General(4) , Reference / General(4) , Secur
Technological innovations(4) , Advice on careers & achieving success(3) , Art(3) , Art & Art Instruction(3) , Body
Leadership(3) , Business & Economics / Reference(3) , Casinos(3) , Computer software(3) , Education(3) , Exec
Games / Magic(3) , General Theory of Computing(3) , Health/Fitness(3) , Hobbies, quizzes & games(3) , Intern
Professional(3) , Management decision making(3) , Mathematical recreations(3) , Mathematics(3) , Motivational
Politics / Current Events(3) , Popular psychology(3) , Production management(3) , Religion / General(3) , Ruby
Growth / General(3) , Science(3) , Science / General(3) , Science: general issues(3) , Social aspects(3) , Sports
General(3) , Accounting(2) , Accounting - General(2) , Alchemy(2) , Alternative medicine(2) , BUSINESS & ECO

Some examples...

Cross Registration

» Any Cross-Reg Status

[Eligible for cross-registration](#) (319)

[Not eligible for cross-registration](#) (122)

Calendar Period

» Any Calendar Period

[Jun to Aug 2011 \(Summer Term\)](#) (21)

[Sep to Dec 2011 \(Fall Term\)](#) (206)

[Jan 2012 \(Winter Session\)](#) (8)

[Jan to May 2012 \(Spring Term\)](#) (197)

Level

» Any Credit Level

[Graduate](#) (354)

[Undergraduate](#) (118)

School

» All Schools

[Faculty of Arts and Sciences](#) (252)

[Harvard Business School - MBA Program](#) (1)

[Harvard Extension School](#) (77)

[Harvard Graduate School of Design](#) (15)

[Harvard Graduate School of Education](#) (12)

[Harvard Kennedy School](#) (7)

[Harvard Law School](#) (13)

[Harvard Medical School](#) (11)

[Harvard School of Public Health](#) (33)

[Harvard Summer School](#) (20)

Days

» Any Day

[Monday](#) (85)

[Tuesday](#) (114)

[Wednesday](#) (96)

[Thursday](#) (105)

[Friday](#) (48)

[Saturday](#) (2)

Start Time

» Any Time

[Morning \(8am-Noon\)](#) (87)

[Afternoon \(Noon-5pm\)](#) (105)

[Evening \(after 5pm\)](#) (64)

Offered

» Any Offered Status

[Not offered](#) (1)

[Offered](#) (440)

Some examples...

Harvard University 2011-2012 Course Catalog

Cross Registration

» Any Cross-Reg Status

[Eligible for cross-registration](#) (78)

[Not eligible for cross-registration](#) (4)

Calendar Period

» Any Calendar Period

» [Sep to Dec 2011 \(Fall Term\)](#)

Level

» Any Credit Level

[Graduate](#) (74)

[Undergraduate](#) (12)

School

SEARCH

Start Over

NARROWED BY

✗ [Sep to Dec 2011 \(Fall Term\)](#)

✗ [Engineering and Applied Sciences](#)

✗ [Faculty of Arts and Sciences](#)

✗ [Fall 2011-2012](#)

82 Courses FoundSort by: [Course Name](#) | [School](#) | [Department](#) | [Day](#) | [Time](#) | [Relevance](#)

Course Information

Schedule

Relevance

[Applied Mathematics 205 - Advanced Scientific Computing: Numerical Methods](#)
(1370)

Faculty of Arts and Sciences Fall 2011-2012

David J. Knezevic

[More Detail](#) ▼

S	M	T	W	T	F	S
---	---	---	---	---	---	---

10:00 a.m. - 11:00 a.m.

0.45

Part VI – Beyond the basics

Beyond the basics

- DisMax –meant to simplify complex queries
- Multi-core – multiple schemas in one instance
- Shards – Distributed search, BIG data split into partitions
- Solr a new kind of database?

A quick recap

- Install Solr
 - Install a servlet container (e.g. Tomcat)
 - Install the Solr WAR file
 - Set solr/home and solr.data.dir
- Define your schema and create/find your data
- Load data into Solr using /update
- Create a user interface that generates Solr queries

- The art of implementing Solr
 - Get it to work with simple data and settings first
 - Then tweak, adjust, refine, iterate

The End

bengaucherin.wordpress.com