Content: Schemas, Documents and Indexing

Xavier Morera
@xmorera
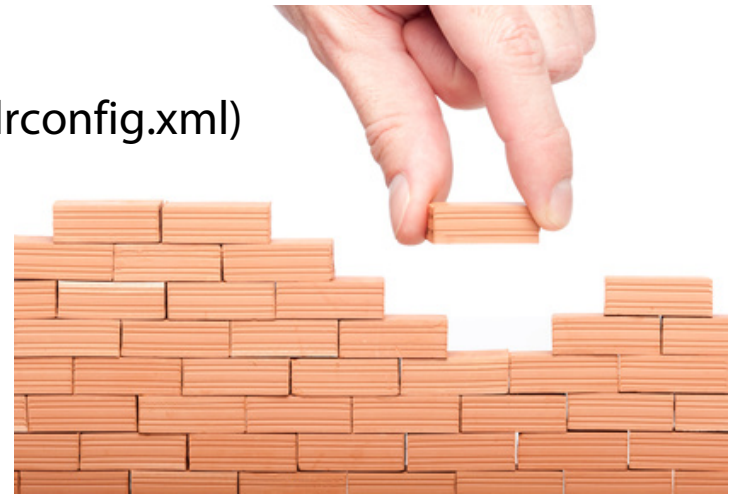www.searchtechnologies.com

pluralsight
hardcore dev and IT training

So You Have a Running Solr…

- **Big deal! Now what?**

- **Documents, Fields and Schema Design**
  - ☐ Learn a few concepts and definitions
  - ☐ Configure your instance (Schema.xml & Solrconfig.xml)

- **Start building our sample**

- **Add a few documents (indexing!)**

- **Run a query or two**

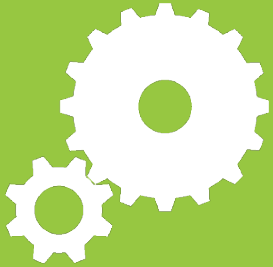- **And then play with data through many interesting demos!**

Solr Content, Schemas & Documents

Concepts & Definitions

Schema.xml Solrconfig.xml

Indexing

Playing with Data

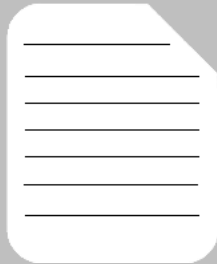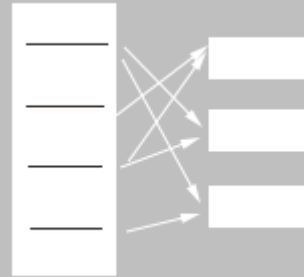# Solr Content, Schemas & Documents

Concepts & Definitions

Schema.xml Solrconfig.xml

Indexing

Playing with Data

Concepts and Definitions

- **Let's start with a few definitions and concepts**

- **Review**
  - Documents
  - Fields
  - Schema
  - Solrconfig

doc·u·ment

noun
/ˈdäkyəmənt/

field

/fēld/ ◀))

noun

sche·ma

/ˈskēmə/ ◀))

noun *technical*

Solr Document

Basic unit of information

Set of data that describes something

Made up of fields

Solr Document

Solr Document

```
{
  "courseid": "scrum-development-jira-agile",
  "coursetitle": "Scrum Development with Jira & JIRA Agile from Xavier",
  "everything": [
    "Scrum Development with Jira & JIRA Agile from Xavier",
    "Having a great idea is just the start, implementation and execution are key. Improve your chances of success by usin
  ],
  "durationinseconds": 5834,
  "releasedate": "2014-01-14T00:00:00Z",
  "description": "Having a great idea is just the start, implementation and execution are key. Improve your chances of su
  "assessmentstatus": "Live",
  "iscourseretired": "no",
  "_version_": 1464487979512758300
}
```

Solr Field

Fields contain different types of data

Flexible definition

Tell Solr how to interpret it with Field Type

Solr Field

- **Field: specific piece of information within a document**
  - i.e. name of a Pluralsight course

- **Copy Field: data may be interpreted in more than one way. Copy data into different fields and field types.**

- **Dynamic Fields: index fields that are not explicitly defined in the schema**

- **Each field can have many different attributes**

Solr Field

Schema.xml Field Attributes

```
<field name="description" type="text_general" indexed="true" stored="true"/>
```

- **Name: field name**

- **Type: field type as declared in <types> section**

- **Indexed: true if indexed (searchable | sortable)**

- **Stored: true if retrievable**

- **Compressed: [false] stored using gzip compression**
  - only if field type is compressable
  - only TextField and StrField are from standard types

- **Multivalued: true if multiple values per document**

Schema.xml Field Attributes

- **Required: instruct Solr to reject if missing**

- **DocValues: If true, the value of the field will be put in a column-oriented DocValues structure**

- **OmitNorms: true omits the norms associated with this field**

- **TermVectors: [false] true to store the term vector for a given field**
  - With MLT store for best performance

- **TermPositions: Store position information with the term vector**
  - Increases storage costs

- **TermOffsets: Store offset information with the term vector**
  - Increases storage costs

- **Default: value if none providing on adding document**

Schema.xml Field Types

- **List is long…**

BCDIntField, BCDLongField, BCDStrField, BinaryField, BoolField,
ByteField, CollationField, CurrencyField, DateField, DoubleField,
ExternalFileField, EnumField, FloatField, ICUCollationField, IntField,
LatLonType, LongField, PointType, PreAnalyzedField, RandomSortField,
ShortField, SortableDoubleField, SortableFloatField, SortableIntField,
SortableLongField, SpatialRecursivePrefixTreeFieldType, StrField,
TextField, TrieDateField, TrieDoubleField, TrieField, TrieFloatField,
TrieIntField, TrieLongField, UUIDField, …

Included in org.apache.solr.schema

Properties by Use Case

Use Case	indexed	stored	multiValued	omitNorms	termVectors	termPositions
search within field	TRUE					
retrieve contents		TRUE				
use as unique key	TRUE		FALSE			
sort on field	TRUE		FALSE	TRUE		
use field boosts				FALSE		
document boosts affect searches within field				FALSE		
highlighting	true	TRUE				TRUE
faceting	TRUE					
add multiple values, maintaining order			TRUE			
field length affects doc score				FALSE		
MoreLikeThis					TRUE	

FieldOptionsByUseCase in Apache Wiki
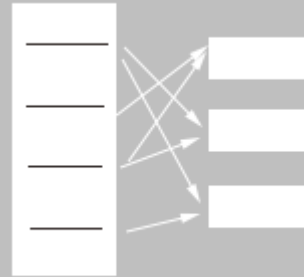
Solr Content, Schemas & Documents
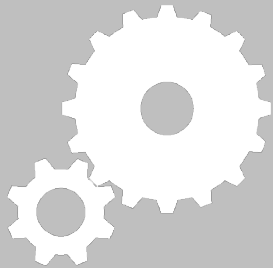
| Concepts & Definitions | Schema.xml Solrconfig.xml | Indexing | Playing with Data |

Schema.xml

Contains details about document fields and how to treat when documents *added to* or *queried from* the index

Schema.xml

- **1ˢᵗ file configured (usually)**

- **Declare**
  - Fields, Field Types & Attributes
  - Primary key (unique) field
  - Which fields are required
  - How to index and search

- **Quick note → Schemaless mode**
  - Index data without manually modifying schema
  - Not covered here

Schema.xml at a Glance

<schema>

    <types>

    <fields>

    <uniqueKey>

    <defaultSearchField>

    <solrQueryParser defaultOperator>

    <copyField>

</schema>

Schema.xml Sins

- **Keep schema.xml simple and organized**

- **Just changing the field names from default configuration**

- **Over planning and "everything but the kitchen sink"**

- **Attributes and fields you don't need**

- **Not upgrading your Solr**

Pluralsight's Schema.xml Courses Demo

01 schema

Pluralsight's Schema.xml Courses Demo

- **Create new pscollection core**
    - Clone *collection1* in *\pluralsight-solr\example\solr\collection1*
    - Change name in *core.properties*
    - Load core using Admin UI (or on Solr start)

- **Modify schema.xml**
    - Add fields

```
<field name="courseid" type="string" indexed="true" stored="true" required="true" multiValued="false" />
<field name="coursetitle" type="string" indexed="true" stored="true" multiValued="false"/>
<field name="durationinseconds" type="int" indexed="true" stored="true" />
<field name="releasedate" type="date" indexed="true" stored="true"/>
<field name="description" type="text_general" indexed="true" stored="true"/>
<field name="assessmentstatus" type="string" indexed="true" stored="true"/>
<field name="iscourseretired" type="string" indexed="true" stored="true"/>
<field name="tag" type="string" multiValued="true" indexed="true" stored="true"/>
```

SolrConfig.xml

Contains most of the parameters to configure the Solr core

Solrconfig.xml

- **XML statements that set configuration values**
  - □ Index configuration
  - □ Query
  - □ Caching
  - □ Event listeners
  - □ Request Handler
  - □ Request Dispatcher
  - □ Highlighter plugin configuration
  - □ Admin GUI section

Solrconfig.xml RequestHandler

- **Defines logic executed for any request**
  - Multiple can be created to define a scope, sorting, etc…

- **Give your users something interesting**
  - Love at first sight?
  - i.e. Newly added? Best discount? Trending topic?

- **Or change number of results**

```xml
<requestHandler name="/select" class="solr.SearchHandler">
  <!-- default values for query parameters can be specified, these
       will be overridden by parameters in the request
   -->
  <lst name="defaults">
    <str name="echoParams">explicit</str>
    <int name="rows">1</int>
    <str name="df">text</str>
  </lst>
```

Solrconfig.xml Sins

- **Not doing minimalism**

- **Not always caching**

- **Over warming**

- **Too many handlers**

- **Not changing the default configuration**

- **Not upgrading your Solr**

Pluralsight's Solrconfig.xml Courses Demo

Pluralsight's Schema.xml Courses Demo

- **Open Solrconfig.xml**

- **Look for the /select RequestHandler**

```xml
<requestHandler name="/select" class="solr.SearchHandler">
  <!-- default values for query parameters can be specified, these
       will be overridden by parameters in the request
   -->
  <lst name="defaults">
    <str name="echoParams">explicit</str>
    <int name="rows">10</int>
    <str name="df">text</str>
  </lst>
```

- **Modify ResponseWriter and default facets**

- **Load some data**

- **Query, reload core , query**

Solr Content, Schemas & Documents
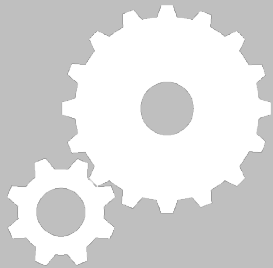
Concepts & Definitions
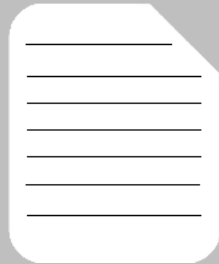
Schema.xml Solrconfig.xml

Indexing

Playing with Data

Indexing

Adding content to a Solr index

If necessary, modifying it

And making it searchable

Indexing – Many Ways

- **Post.jar tool**

- **Solr Cell framework built on Apache Tika for binary files**
    - PDF, Word, …

- **Uploading XML files sending via HTTP requests**

- **Database via DataImportHandler**

- **Custom Java application via Solr's Java Client API**

- **SolrNet for those .Net fans! (Oh yeah SolrJ too!)**

- **Other tools, for example Search Technologies' ASPIRE PostToSolr**

Solr Content, Schemas & Documents
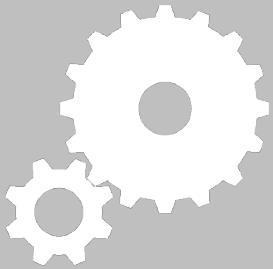
Concepts & Definitions
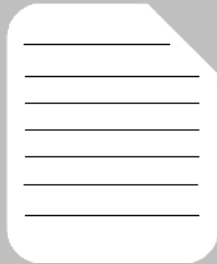
Schema.xml
Solrconfig.xml

Indexing

D Playing with Data

Let's Play with Some Data

- **Update Documents using CSV**

- **Updating Documents Admin UI**

- **Deleting Documents**

- **Index with Solr XML format**

- **cURL & Fiddler**

- **DataImportHandler**

- **Documents with Apache Tika**

- **Web pages with Apache Nutch**

- **Index for our Pluralsight Demo**

IMPROVE YOUR SKILLS !

Updating Documents using CSV

- **Updating documents is easy**

- **Unique key per document and re index**

- **Test with course authors**

- **Modify the info and re post.jar**

```
C:\pluralsight-solr\example\exampledocs>java -Durl=http://localhost:8983/solr/psdemo/update  -Dt
csv"
SimplePostTool version 1.5
Posting files to base url http://localhost:8983/solr/psdemo/update using content-type text/csv..
POSTing file Two sample courses with author test.csv
1 files indexed.
COMMITting Solr index changes to http://localhost:8983/solr/psdemo/update..
Time spent: 0:00:00.080
```

Updating Data via Admin UI

- **Playing around with the Admin UI: partial update vs full update**
  - {"courseid":"scrum-development-jira-agile","coursetitle":{"set":"Small update"}}
  - {"courseid":"scrum-development-jira-agile","coursetitle":"Nothing"}

Document(s)

```
{"courseid":"scrum-development-jira-agile","coursetitle":{"set":"Small
update"}}
```

Commit Within

```
1000
```

Overwrite

```
true
```

Boost

```
1.0
```

Submit Document

Status: success
Response:

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 15
  }
}
```

Deleting Data

- **Several ways of deleting data**

- **By query**
  - http://localhost:8983/solr/psdemo/update?stream.body=<delete><query>courseid:scrum-development-jira-agile</query></delete>&commit=true

- **All items**
  - <delete><query>*:*</query></delete>

- **SolrNet, SolrJ, …**
  - Solr.deleteByQuery("*:*")

Solr's XML Format

- **Solr's own XML format**

- **Specific and verbose yet easy to read**

- **Deals better with multivalued/optional fields, complex strings and real life requirements**

java -Dauto -Durl="http://localhost:8983/solr/psdemo/update" -jar post.jar solr-example-course.xml

```xml
<add>
    <doc>
        <field name="courseid">getting-started-enterprise-search-apache-solr</field>
        <field name="coursetitle">Getting Started with Enterprise Search using Apache Solr</field>
        <field name="durationinseconds">8849</field>
        <field name="releasedate">2014-05-14T00:00:00Z</field>
        <field name="description">Search is one of the most misunderstood functionalities in the IT industry.
        <field name="assessmentstatus">Live</field>
        <field name="iscourseretired">no</field>
    </doc>
</add>
```

cURL Is Your Friend

- **Tool for transferring data using various protocols**

- **Simple and easy to use, post information, including files!**

- **Let's add, update and delete a document**

- curl http://localhost:8983/solr/psdemo/update?commit=true -H "Content-Type: text/xml" --data-binary "<add><doc><field name=\"courseid\">getting-started-enterprise-search-apache-solr</field><field name=\"coursetitle\">Getting Started with Enterprise Search using Apache Solr</field></doc></add>"

```
C:\Users\xavier>curl http://localhost:8983/solr/psdemo/update?commit=true -H "Content-Type:
<?xml version="1.0" encoding="UTF-8"?>
<response>
<lst name="responseHeader"><int name="status">0</int><int name="QTime">16</int></lst>
</response>
```

Fiddler Is Your Pal

- **Debugging proxy that logs all HTTP traffic in your computer**

- **Use it to inspect, reissue and compose requests**

1	200	HTTP	localhost:8983	/solr/psdemo/select?q=apache&wt=json&indent=true&...	361	text/plain;charse...	chrome...
16	200	HTTP	localhost:8983	/solr/psdemo/select?q=apache&wt=json&indent=true&...	560	text/plain;charse...	chrome...
17	200	HTTP	localhost:8983	/solr/tpl/documents.html	4,570	text/html	chrome...
18	200	HTTP	localhost:8983	/solr/tpl/query.html?_=1397508390721	13,119	text/html	chrome...
19	200	HTTP	localhost:8983	/solr/psdemo/select?q=*%3A*&wt=json&indent=true&...	557	text/plain;charse...	chrome...
21	200	HTTP	localhost:8983	/solr/	5,887	text/html;charset...	chrome...
22	200	HTTP	localhost:8983	/solr/admin/cores?wt=json&indexInfo=false&_=139750...	997	application/json; ...	chrome...
23	200	HTTP	localhost:8983	/solr/admin/info/system?wt=json&_=1397508399012	2,778	application/json; ...	chrome...
24	200	HTTP	localhost:8983	/solr/tpl/documents.html	4,570	text/html	chrome...
32	200	HTTP	localhost:8983	/solr/psdemo/update?wt=json	53	text/plain;charse...	chrome...
33	200	HTTP	localhost:8983	/solr/psdemo/select?q=*%3A*&wt=json&indent=true&...	514	text/plain;charse...	chrome...
41	200	HTTP	localhost:8983	/solr/psdemo/update?wt=json	53	text/plain;charse...	fiddler:...

Indexing Files with Apache Tika

- **Content analysis toolkit**
  - Detect and extract metadata
  - Formerly subproject of Lucene

- **Use it to upload a binary document and extract information from it**

- **Tokenization and meta data extraction require much more work**

- **And let's use our friend cURL**

curl "http://localhost:8983/solr/pstika/update/extract?literal.id=doc1&commit=true" -F "myfile=@apache-solr-enteprise-search-m4-content-schemas-documents.pptx"

Indexing Web Pages with Apache Nutch

- **Download nutch, unzip and move to c:\pluralsight-nutch**
  - Bin vs src?

- **Install Cygwin (functionality similar to Linux on Windows)**

- **Set NUTCH_JAVA_HOME to the same location as JAVA_HOME**
  - Recommend DOS style paths

- **Open nutch-site.xml in C:\pluralsight-nutch\conf**

- **Create seed.txt within urls folder and add list of web sites**
  - http://www.pluralsight.com

Indexing Web Pages with Apache Nutch

- **Create new psnutch collection**

- **Add Nutch required fields in schema.xml**

  ```
  <field name="digest" type="text" stored="true" indexed="true"/>
  <field name="boost" type="text" stored="true" indexed="true"/>
  <field name="segment" type="text" stored="true" indexed="true"/>
  <field name="host" type="text" stored="true" indexed="true"/>
  <field name="site" type="text" stored="true" indexed="true"/>
  <field name="content" type="text" stored="true" indexed="true"/>
  <field name="tstamp" type="text" stored="true" indexed="false"/>
  <field name="url" type="string" stored="true" indexed="true"/>
  <field name="anchor" type="text" stored="true" indexed="false"
  multiValued="true"/>
  ```

- **Set unique key to url but leave id**

Indexing with DataImportHandler

- **DIH is a Solr contrib that provides a configuration driven way to import data into Solr**

- **Build documents by aggregating from multiple columns and tables**

- **Both full builds and incremental**

- **Databases, XML, feeds, emails, …**

- **Scheduling allowed**

- **Very powerful and decently scalable**
  - Homework: investigate indexing Wikipedia with DIH

Indexing with DataImportHandler

- **http://localhost:8983/solr/rss/dataimport?command=full-import**

- **Slashdot feed: http://rss.slashdot.org/Slashdot/slashdot**

```
"docs": [
  {
    "title": "A 2560x1440 VR Headset That's Mobile",
    "description": "New submitter oldmildog writes: \"GameFace Labs may very well be the furthest along in the quest to cre
    "link": "http://rss.slashdot.org/~r/Slashdot/slashdot/~3/i99Wss6siQc/story01.htm",
    "creator": "Soulskill"
  },
  {
    "title": "Zombie Plants Help To Spread Bacterial Pathogen",
    "description": "bmahersciwriter writes: \"We've all heard stories about how parasites can 'zombify' organisms, getting
    "slash-comments": 1,
    "link": "http://rss.slashdot.org/~r/Slashdot/slashdot/~3/Cg7EU6hoobo/story01.htm",
    "creator": "Soulskill"
  },
  {
    "title": "Snowden: NSA Spied On Human Rights Workers",
    "description": "Hugh Pickens DOT Com writes: \"The Guardian reports that according to Edward Snowden, the NSA has spied
    "slash-comments": 2,
    "link": "http://rss.slashdot.org/~r/Slashdot/slashdot/~3/LSQKL7_9TWU/story01.htm",
    "creator": "Soulskill"
  },
```

Index All Documents for Pluralsight Demo

- **Let's just fill up our index for future demos**

- **Same process as we did for loading the 2 documents**
  - Just use a the CSV with all the courses
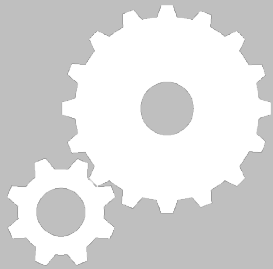
```json
{
  "responseHeader": {
    "status": 0,
    "QTime": 16,
    "params": {
      "indent": "true",
      "q": "*:*",
      "_": "1398139129685",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 934,
    "start": 0,
    "docs": [
      {
        "courseid": "getting-started-enterprise-search-apache-solr",
```

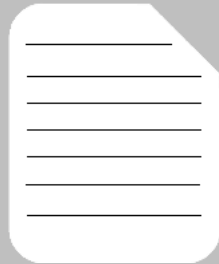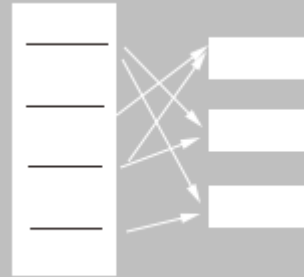# Solr Content, Schemas & Documents

Concepts & Definitions

Schema.xml Solrconfig.xml

Indexing

Playing with Data

Takeaway

- **Document, schema, field, attribute and type**

- **Schema.xml & Solrconfig.xml**

- **Index**
  - CSV, XML, Json
  - Post tool, Fiddler, cURL
  - Tika, Nutch, DataImportHandler

- **Played around with many demos**

- **Prepared data for our next module**