

State of Solr Community

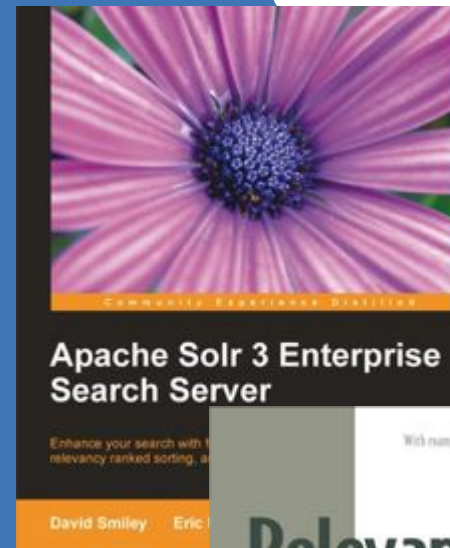
BioSolr Workshop Feb 3, 2016

twitter: @dep4b email: epugh@o19s.com

slides: <http://bit.ly/state-of-solr>

Who Am I?

- Founder and CEO of OpenSource Connections
- First project w/ Solr in 2007
- Co-Author of first book on Solr in 2008
- Focused on Search in 2010, specifically Solr and Lucene
- Interested in data engineering (integration, scaling)
- *Relevant Search* by my colleague Doug in March!



Looking at Solr



Origins of Solr

- Created by Yonik Seeley at CNET in 2004
- Open sourced in 2006, donated to Apache Software Foundation
- January 2007 became a TLP: Top Level Project



But Something Came before..

- Lucene is the core search library
- Doug Cutting wrote it in 1999!
- Joined ASF in 2001



Adding German analysis code contributed by Gerhard Schwarz ...

cutting committed on Sep 25, 2001

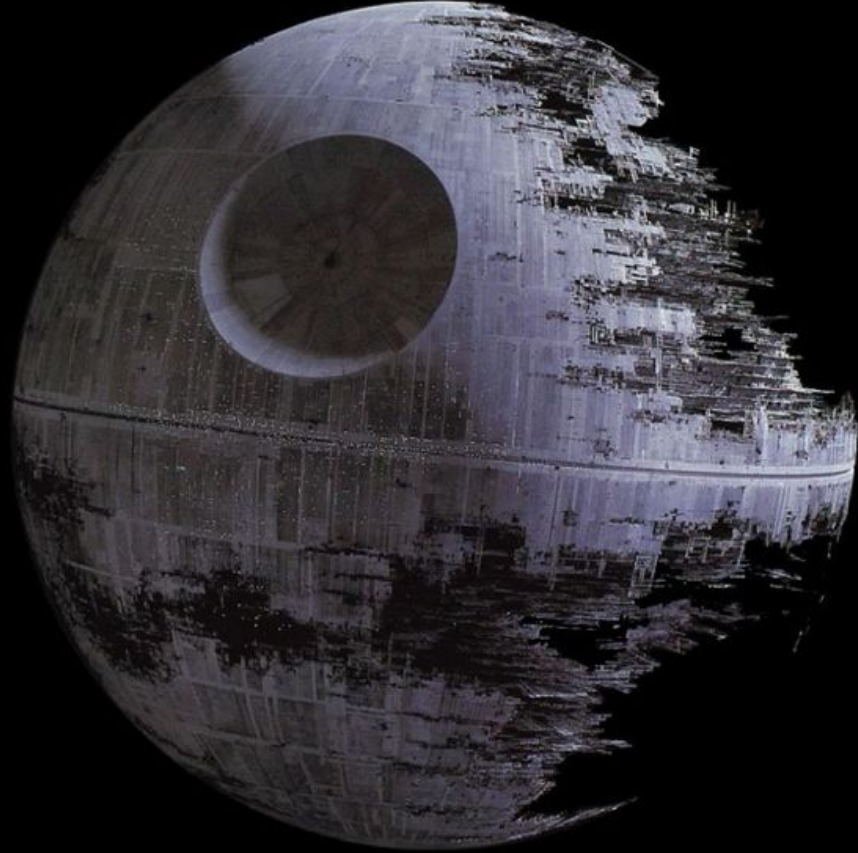


Solr 1.x

- v1.1 Dec 2006 through v1.4 Nov 2009
- Solr become more stable, incremental features added
- Steady growth in adoption
- SOLR-284 for Parsing Rich Document types such as PDF and MS Office
- Solr supports Apache Tika for Content Extraction

Solr 3.x

- Lucene and Solr communities merge development paths, Solr version jumped to match Lucene 3.
- v3.1 Mar 2011 through v3.6.2 Dec 2012
- Range facets, grouping/field collapsing, finite state machine based suggesters and about a million other improvements



OMG: It's Elasticsearch!

Solr 4.x

- v4.0 Alpha July 2012 through v4.10.4 Mar 2015
- SolrCloud is a *huge* effort
- Many features inspired from ES: Schemaless cores, pivot on facets and statistics, simpler startup, nested documents, JSON updates, partial updates.

Today

Release Cycle

Major Version	Minor Version	Defect Version	Dates	Months
1	4	0	12/2006 - 11/2009	36
3	6	2	03/2011 - 12/2012	21
4	7	4	07/2012 - 03/2015	20
5	4	4	02/2015 - current	12..

Solr 5.x

- v5.0 Feb 2012 through v5.4 Dec 2015
- Clean up!
 - v2 API work
 - Slick new AngularJS based Admin Interface
 - Pluggable Authentication/Authorization
 - JSON Query API
- Streaming data
- More analytics features (Hyperloglog, stats field enhancements)

Tomorrow

Solr 6

- In active development
- Solr becomes the pre-eminent IR platform
- Becomes part of the “Big Data” Ecosystem
- Feature set diverges from Elasticsearch
- Not just about “text” anymore

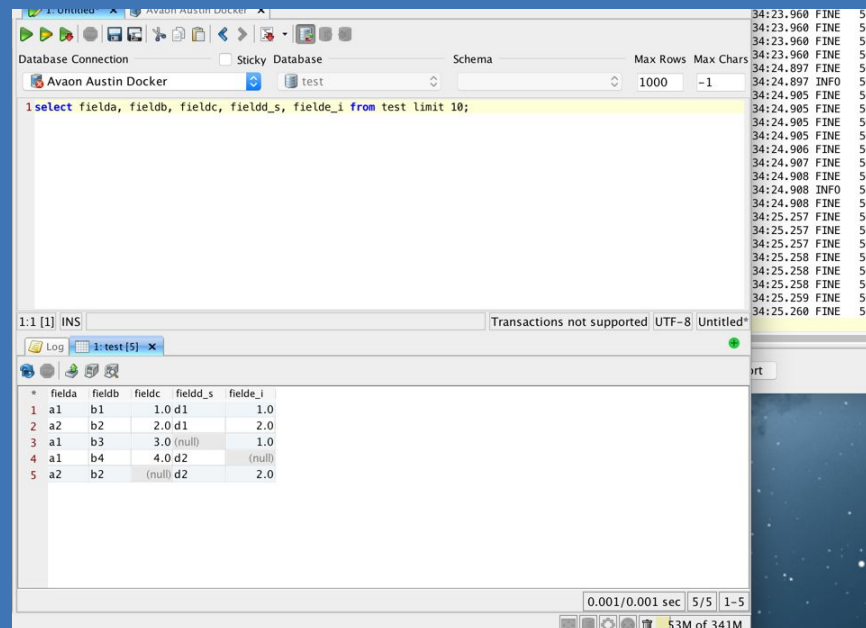
Pre-eminent IR platform

- *Point Values* data structure supports multi dimensional data structures.
- “Learn to Rank” and “Query ReRanking”
- BM25 scoring model replaces 25 year old TF-IDF
- Heatmaps generated right out of Solr!

Relevancy is important again!

“Big Data” Ecosystem

- Spark Integration
- Streaming aggregations API
- SolrJ client now supports JDBC! (Java 8 move)



The screenshot shows a database client window with the following components:

- Database Connection:** Avaon Austin Docker, test
- Schema:** test
- Max Rows:** 1000
- Max Chars:** -1
- Query:** `1 select fielda, fieldb, fieldc, fieldd_s, fielde_i from test limit 10;`
- Results:** A table with 5 rows and 5 columns: fielda, fieldb, fieldc, fieldd_s, and fielde_i.
- Log:** 1:1 [1] INS
- Footer:** 0.001/0.001 sec 5/5 1-5, 53M of 341M

	fielda	fieldb	fieldc	fieldd_s	fielde_i
1	a1	b1	1.0	d1	1.0
2	a2	b2	2.0	d1	2.0
3	a1	b3	3.0	(null)	1.0
4	a1	b4	4.0	d2	(null)
5	a2	b2	(null)	d2	2.0

spark-solr


? ⚙ default ▾

Took 56 seconds.

```
tweets.printSchema()

root
 |-- _indexed_at_tdt: timestamp (nullable = true)
 |-- _version_: long (nullable = true)
 |-- author_s: string (nullable = true)
 |-- id: string (nullable = false)
 |-- provider_s: string (nullable = true)
 |-- type_s: string (nullable = true)
```

FINISHED ▶ ⌵ ⌲ ⌳ ⌴

Took 2 seconds.

```
tweets.registerTempTable("tweets")
```

FINISHED ▶ ⌵ ⌲ ⌳ ⌴

Took 2 seconds.

```
sqlContext.sql("SELECT COUNT(type_s) FROM tweets WHERE type_s='echo']").show()
```

FINISHED ▶ ⌵ ⌲ ⌳ ⌴

```
+-----+
| _c0|
+-----+
| 19684|
+-----+
```

Took 24 seconds.

Feature set Diverges from ES

- No longer concerned by Elasticsearch
- Streaming API is unique to Solr
- Not about just log analytics
- Leverages the “Apache Way” of multiple stakeholders contributing
- Geo related capabilities continue to grow

Not just about “text”

- Used as the backend to many different things (think web content caching and dynamic content use cases)
- Able to execute Graph Queries in Solr
- Build knowledge graphs in Solr



Thank you!

email: epugh@opensourceconnections.com

twitter: [@dep4b](https://twitter.com/dep4b)

slides: bit.ly/state-of-solr