# ElasticSearch in Production

lessons learned

Anne Veling, ApacheCon EU, November 6, 2012

# agenda

- Introduction
- ElasticSearch
- Udini
- Upcoming Tool
- Lessons Learned

elasticsearch.

Udini
by ProQuest

# introduction

- Anne Veling, @anneveling

beyondtrees

- Self-employed contractor
  - Software Architect
  - Agile process management
  - Performance optimization
  - Lucene/SOLR/ElasticSearch implementations & training

# ElasticSearch

- Apache Lucene
- Started in 2010 by Shay Banon
- Open Source – Apache License
- A company was formed in 2012: ElasticSearch
  - Training, support and development
- Careful feature development
  - vs. build because you can

@kimchy

elasticsearch.

# ElasticSearch

- Scalable
  - Distributed, Node Discovery
  - Automatic sharding
  - Query distribution
- RESTful, HTTP API
  - With API wrappers for Ruby, Java, Scala, …
  - JSON in, JSON out
- Document Model
  - Maps book.author.lastName
  - “schemaless” -> field type recognition
  - Keeps source, keeps ‘version’ number

# ElasticSearch

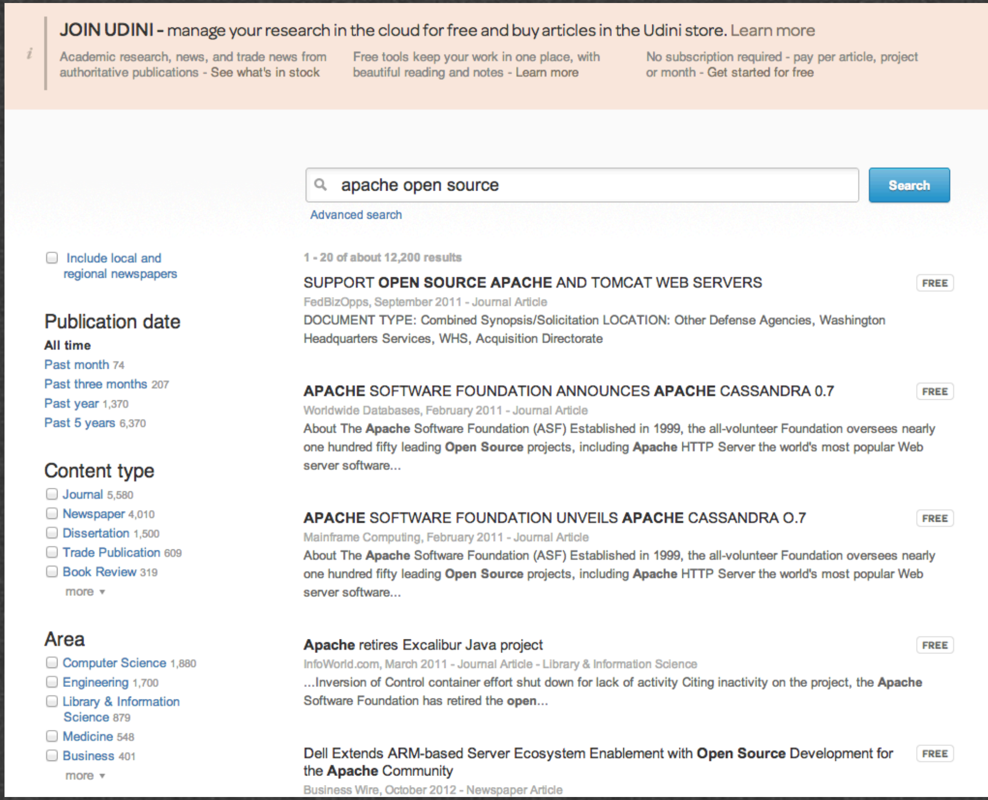
- Integrated faceting
  - With statistical aggregates (sum/avg/…) for free
- Field types and analyzers
  - String, numerics, geo, attachment, …
  - Arrays, subdocuments, nested documents
- Integrated sharding
  - Routing and alias
  - Cross-index searching / multi-document type

# udini.proquest.com

- ProQuest
- The World's Article Store
- Stack
  - Amazon EC2
  - Scala with Unfiltered
  - MongoDB, ElasticSearch

Udini

by ProQuest



**JOIN UDINI** - manage your research in the cloud for free and buy articles in the Udini store. Learn more

Academic research, news, and trade news from authoritative publications - See what's in stock

Free tools keep your work in one place, with beautiful reading and notes - Learn more

No subscription required - pay per article, project or month - Get started for free

apache open source

Advanced search

□ Include local and regional newspapers

Publication date

**All time**

Past month 74

Past three months 207

Past year 1,370

Past 5 years 6,370

Content type

□ Journal 5,580

□ Newspaper 4,010

□ Dissertation 1,500

□ Trade Publication 609

□ Book Review 319

more ▾

Area

□ Computer Science 1,880

□ Engineering 1,700

□ Library & Information Science 879

□ Medicine 548

□ Business 401

more ▾

1 - 20 of about 12,200 results

SUPPORT **OPEN SOURCE APACHE** AND TOMCAT WEB SERVERS — FREE

FedBizOpps, September 2011 - Journal Article

DOCUMENT TYPE: Combined Synopsis/Solicitation LOCATION: Other Defense Agencies, Washington Headquarters Services, WHS, Acquisition Directorate

**APACHE** SOFTWARE FOUNDATION ANNOUNCES **APACHE** CASSANDRA 0.7 — FREE

Worldwide Databases, February 2011 - Journal Article

About The **Apache** Software Foundation (ASF) Established in 1999, the all-volunteer Foundation oversees nearly one hundred fifty leading **Open Source** projects, including **Apache** HTTP Server the world's most popular Web server software...

**APACHE** SOFTWARE FOUNDATION UNVEILS **APACHE** CASSANDRA 0.7 — FREE

Mainframe Computing, February 2011 - Journal Article

About The **Apache** Software Foundation (ASF) Established in 1999, the all-volunteer Foundation oversees nearly one hundred fifty leading **Open Source** projects, including **Apache** HTTP Server the world's most popular Web server software...

**Apache** retires Excalibur Java project — FREE

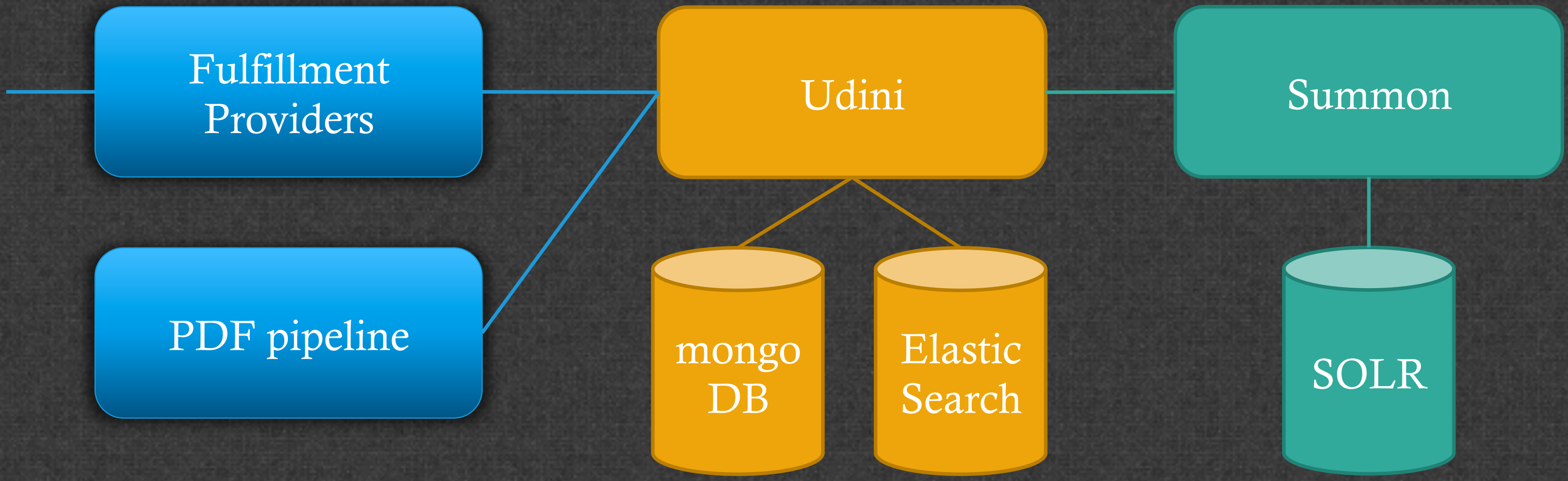
InfoWorld.com, March 2011 - Journal Article - Library & Information Science

...Inversion of Control container effort shut down for lack of activity Citing inactivity on the project, the **Apache** Software Foundation has retired the **open**...

Dell Extends ARM-based Server Ecosystem Enablement with **Open Source** Development for the **Apache** Community — FREE

Business Wire, October 2012 - Newspaper Article

# architecture



# SOLR at Udini

- Connecting to Summon API
  - 700M SOLR Cluster
- In Udini, we serve a subset of 160M full text articles
  - Including fulfillment mechanisms
  - PDF and HTML5 viewing and annotation

# ElasticSearch at Udini

- Local index to search your articles
- Many small user libraries, searching only locally
  - User-id as sharding key
  - Include key in all queries

# Exciting new product

- Developing for ProQuest
- Exciting new research tool for scientific researchers
- Creating a large ElasticSearch index for journal article canonicalization
- Currently in private beta, launching in the coming months

# Lessons Learned

- Very fast indexing
- Bulk indexing ftw
  - Set up without replicas (replicas = 0, not 1)
  - Play with bulk size
  - Simple write to disk and CURL it in, is very fast
  - 1M records in 40s

```
for f in ${BATCH_DIR}/batch-*.json
do
  echo "about to index $f"
  curl --silent --show-error --request POST
             --data-binary @$f localhost:9200/_bulk > /dev/null
  echo
done
```

# Lessons learned

- Schema(less)?
- Automatic field type recognition
  - Can miss types
  - Strict about types #duh
- Mapping of subfields (doc.title vs doc.publication.title)
  - Version dependent
- In reality
  - Schema still needed
  - Mapping changes still non trivial

# Lessons learned

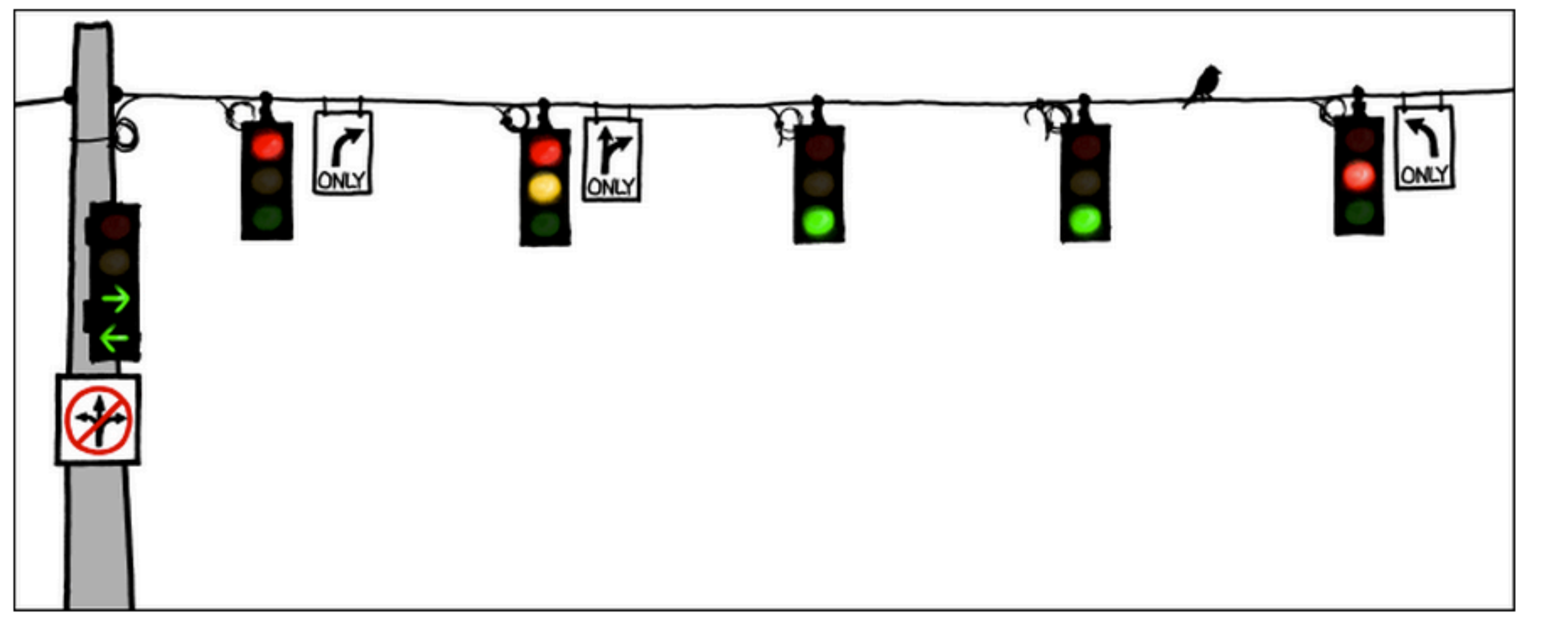
- Learn to trust ElasticSearch
  - Analyzers: do not pretokenize queries yourself…
- Difference between “term” and “text” type queries
  - tokenized or not
- ElasticSearch probably already does what you want it to do
  - Search for it
  - Try it

# Lessons learned

- Issues with automated testing and node discovery/startup
- Start/stop hundreds of times during Jenkins test jobs or development boxes
  - Takes time
  - Locally sometimes picks up previous versions
- Memory issues: ElasticSearch manages a large part of its memory outside of the heap
  - Do not simply increase -Xmx

# Lessons learned

- New tools every month
- waitForYellowStatus
- Aliases, routing allow for clever control



# API

- ElasticSearch is new, connection libraries still in infancy, documentation growing
- Issues using the Java API in Scala
- Happy with Scalastic now
  - synchronous
  - asynchronous
  - bulk prepare

https://github.com/bsadeh/scalastic

# #nodb

- ElasticSearch used as a full nosql datastore?
- Using “version” and optimistic locking scheme
- Could replace MongoDb in our setup

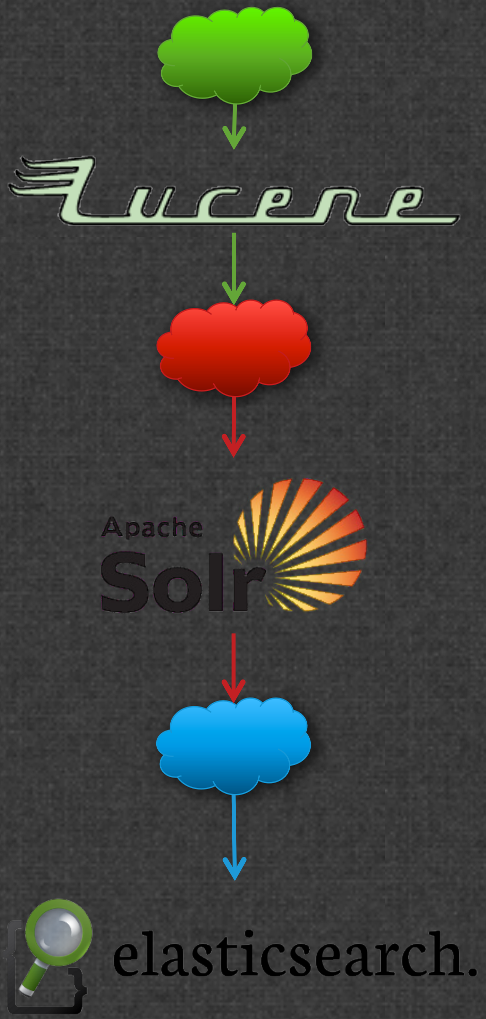
- ElasticSearch is actually a store optimized for getting stuff out, not for getting stuff in
  - With free faceting
  - Who needs multi-table transactions anyway?

# SOLR vs ElasticSearch

- SOLR
  - Well-known, many tools, extensions
  - Feels clunky to configure
  - Manual document to lucene mapping
  - Replication and indexing in a cluster non-trivial
- Old school ;-)

- ElasticSearch
  - New kid on the block
  - Very easy to configure
  - Handles document to lucene mapping
  - Horizontally scalable
    - Easy replication
    - But: shard key
- New school

# search evolution



- Custom indexers

- Inverted index
- Segment merges

- Custom analyzers
- Faceting

- Configuration of analyzers
- Faceting, Geospatial

- Document mapping
- Sub-document queries
- Replication

- JSON document input
- Faceting, complex queries just work

# conclusions

- ElasticSearch benefits
  - Easy to setup
  - Very clever architecture
- Drawbacks
  - Very new software, tool support limited
    - But lots of movement
  - Change sharding in a full index non-trivial
- ElasticSearch
  - Clever architecture, fast, stable
  - Does exactly what you need

elasticsearch.

# thank you

```
Are you still using Solr?
Come on, it’s 2012 already ;-)
```

anne@beyondtrees.com

@anneveling