Getting Started With Enterprise Search using Apache Solr

Why Solr and Enterprise Search?

Xavier Morera
@xmorera
www.searchtechnologies.com

pluralsight
hardcore dev and IT training

Why Watch This Course?

- **To learn how to implement search using Solr**
- **Search is an incredibly useful feature**
  - People take it for granted
  - Unless missing or poorly implemented

- **While there is a lot information about search engines & Solr**
  - IMHO not simple enough to get people started and scattered
  - Although Solr's Wiki is very complete
- **And the commercial search engines can be hugely expensive**
  - Solr #1 choice (again IMHO and a few thousand more…)
- **And I have a promise for you…**

I promise…

That in the next couple of hours I will teach you to build things that might take you weeks to learn on your own and we will create together a search experience that could cost thousands of dollars to build

- *Not bad for something that comes as part of a Pluralsight subscription, right?*
- *Albeit not a fully advanced and complete app, a great start!*

What to Expect: Full Training Agenda

- M1: Why Solr and Enterprise Search?
- M2: Architecture of an Enterprise Search Application
- M3: Solr Configuration
- M4: Content: Schemas, Solrconfig and Indexing
- M5: Searching & Relevance
- M6: Making it all Work: Put a UI on it!
- M7: Takeaway

Ready to Watch and Learn Solr?

Who's with me?

Who's NOT with me

Agenda

Why (Enterprise) Search	Why Solr? Famous Sites	Other Search Engines
History of Search	Functions of a Search Engine	Features & Scalability

Why Solr and Enterprise Search?

Why (Enterprise) Search	Why Solr? Famous Sites	Other Search Engines
History of Search	Functions of a Search Engine	Features & Scalability

Why (Enterprise) Search?

Definition

- **search (sûrch)*v.* searched, search·ing, search·es, searcha·ble *adj.*, searcher *n.***
  - *v.tr.***1.** To make a thorough examination of; look over carefully in order to find something;
  - **2.** To make a careful examination or investigation of; probe:
  - *v.intr.*To conduct a thorough investigation; seek:

enterprise search

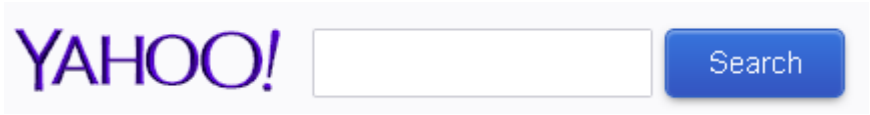
> **“practice of generating content and making it searchable to a defined audience out of multiple enterprise-type data sources like databases or CMS”**

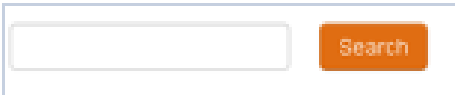
Why (Enterprise) Search?

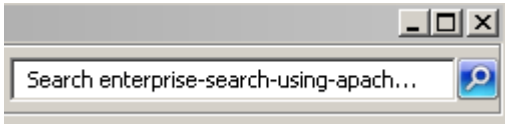
- **People *need to find* things (and fast)**
- **People *expect* to find things**
- **People crave *simplicity***

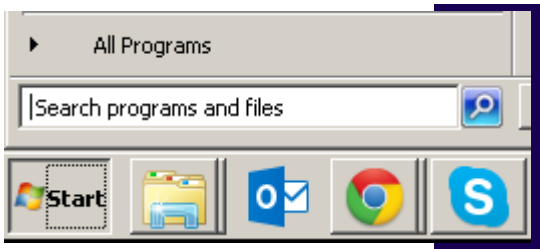
- **Internet search has an interesting side effect**
  - Expect search “Everywhere!”
  - Billions of people trained in search
  - And BOY are we ready to use it!
    - ~5,922,000,000 daily searches in 2013
- **Where else do we see search applications?**

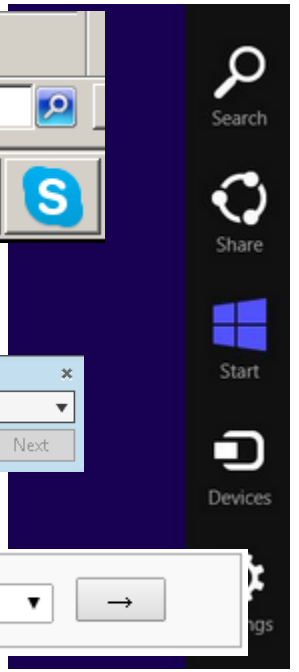
Search is Everywhere!



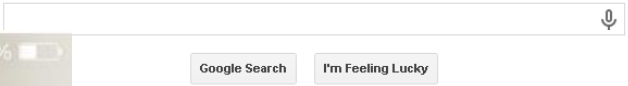


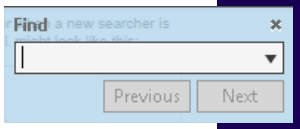


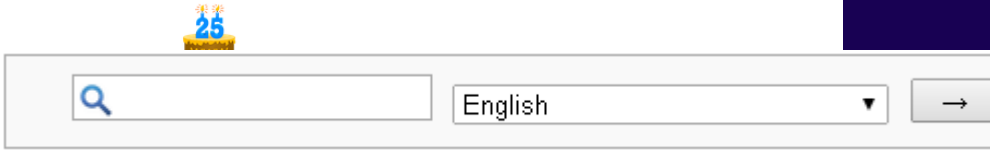




Google







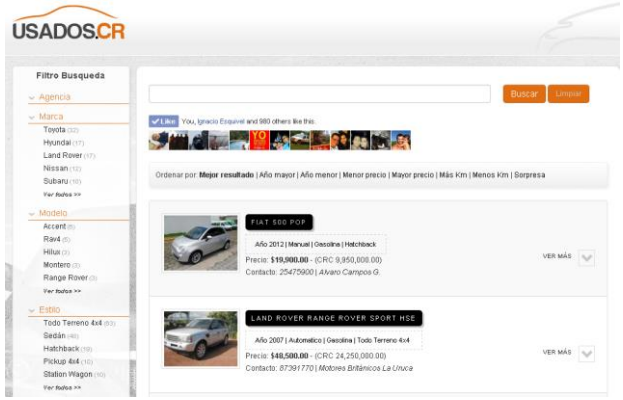
Why (Enterprise) Search?

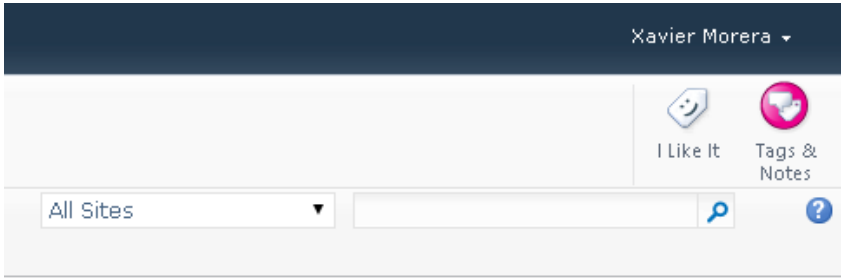
Search makes your life easy

- Personal Search (desktop, cell phone, tablet, tv, …)
- Enterprise Search: Extranet & Intranet
- Retailers: B2C & B2B
- eDiscovery/Compliance
- Recruiting
- Intelligence Analysis

Why (Enterprise) Search

- Two types: outside and inside the firewall
- Outside → Make money!
- Inside → Save money!





Why Solr and Enterprise Search?

Why (Enterprise) Search	Why Solr? Famous Sites	Other Search Engines
History of Search	Functions of a Search Engine	Features & Scalability

Why Solr?

- **Search used to be**
  - Not for the faint of heart
  - With a thin wallet
- **Solr changed it all**
  - Open source
  - Fast and sophisticated text search
  - Highly extensible
  - Highly scalable
  - Dynamic content
  - Great query speed (properly scaled)
  - Many more…

Why Solr?

"Solr has an active development community, both individuals and companies, who contribute new features and bug fixes"

Search Engines

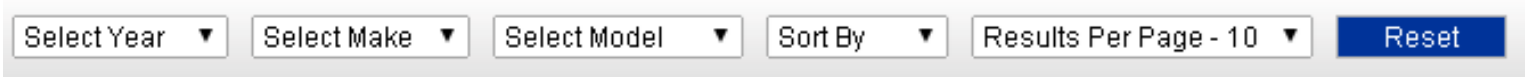
- **Search engines are a totally different animal**
- **You will fall in love with what you can do with them**
- **Or absolutely hate it if tackled head on without proper resources**
- **Efficient and fast because it searches an inverted index**
  - Instead of searching through the text

or

“I have Full Text Search in my SQL Server!”

- **Yeah right…**
- **Check this link**
  - http://wiki.apache.org/solr/WhyUseSolr
  - Or Google “why use solr site:apache.org”
  - Solr vs. Relational Database

(Solr) vs. (Relational Database)



Hint: simplicity is the ultimate sophistication

Public Solr Sites

Netflix
Yellow Pages
Usados.cr
GPO
Sears
Whitehouse.gov
Instagram
Zappos
Comcast

eHarmony
Jobreez
Immonet
Chegg
The Guardian
FCC.gov
Comcast / xfinity
Jobuzu
Openindex

Buy.com
AOL Music
AOL NFL Sports
AOL Recipes
AOL Real Estate
AOL Autos
AOL Travel
AOL StyleList
News.com

Why Solr and Enterprise Search?

Why (Enterprise) Search	Why Solr? Famous Sites	Other Search Engines
History of Search	Functions of a Search Engine	Features & Scalability

Other Search Engines

SharePoint

fast

Autonomy

ATTIVIO

SINEQUA

LucidWorks

amazon CloudSearch

ENDECA BY ORACLE

Rocket Folio/NXT

exalead

MarkLogic

cloudera

Content Analyst
Find | Organize | Discover

Vivísimo

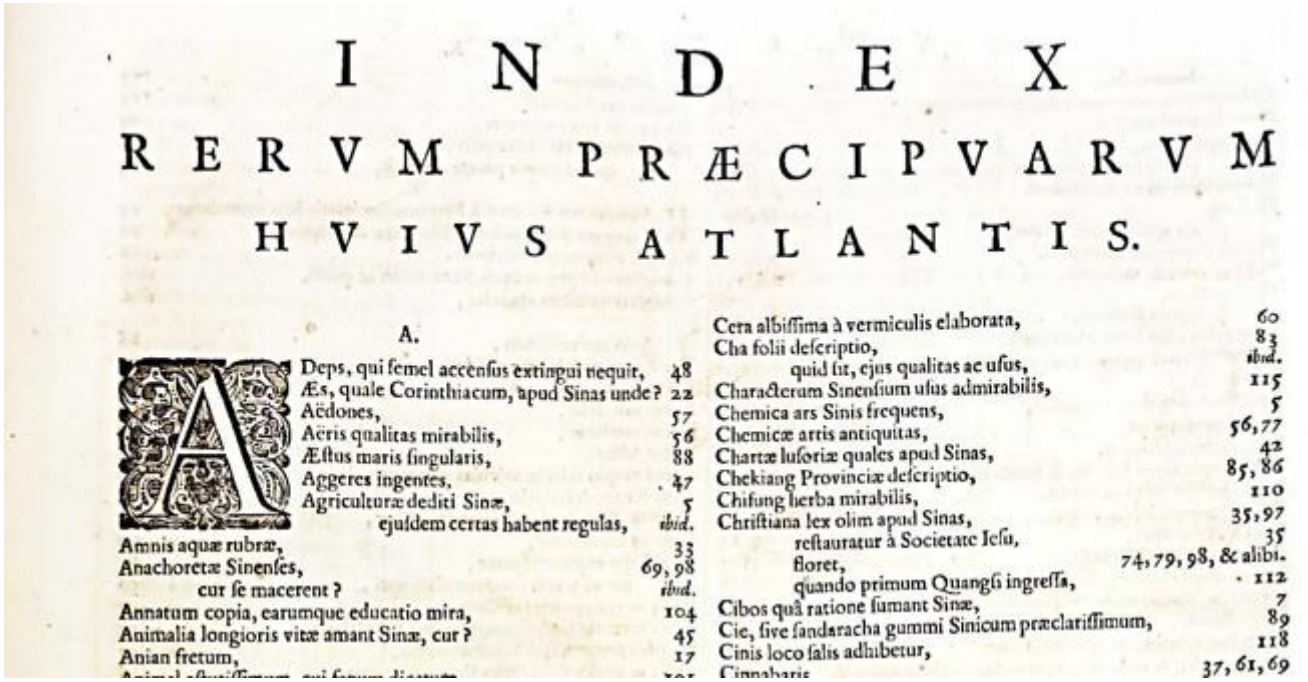
HERITRIX

Why Solr and Enterprise Search?

Why (Enterprise) Search	Why Solr? Famous Sites	Other Search Engines
History of Search	Functions of a Search Engine	Features & Scalability

History of Search

- **Indexing has been around for quite a long time**
  - As early as 1500's



INDEX
RERVM PRÆCIPVARVM
HVIVS ATLANTIS.

A.

A Brief History of Search - Vision

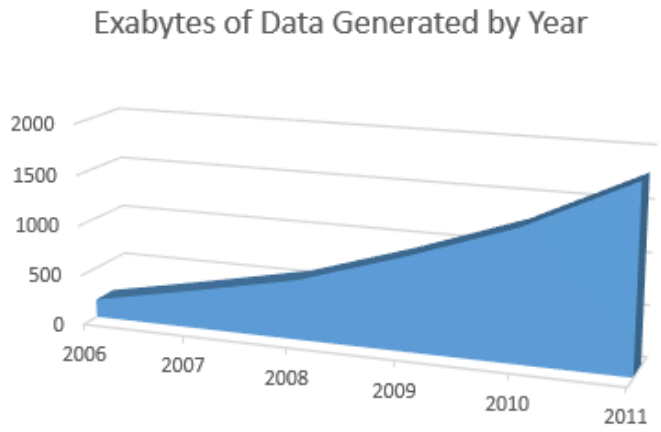
- **Long before the internet existed (1945)**
  - Vannevar Bush head of the U.S. Office of Scientific Research and Development (OSRD) during World War II said:
    - "The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships".
    - "Our ineptitude in getting at the record is largely caused by the artificiality of the systems of indexing it"

A Brief History of Search – Achieving the Vision

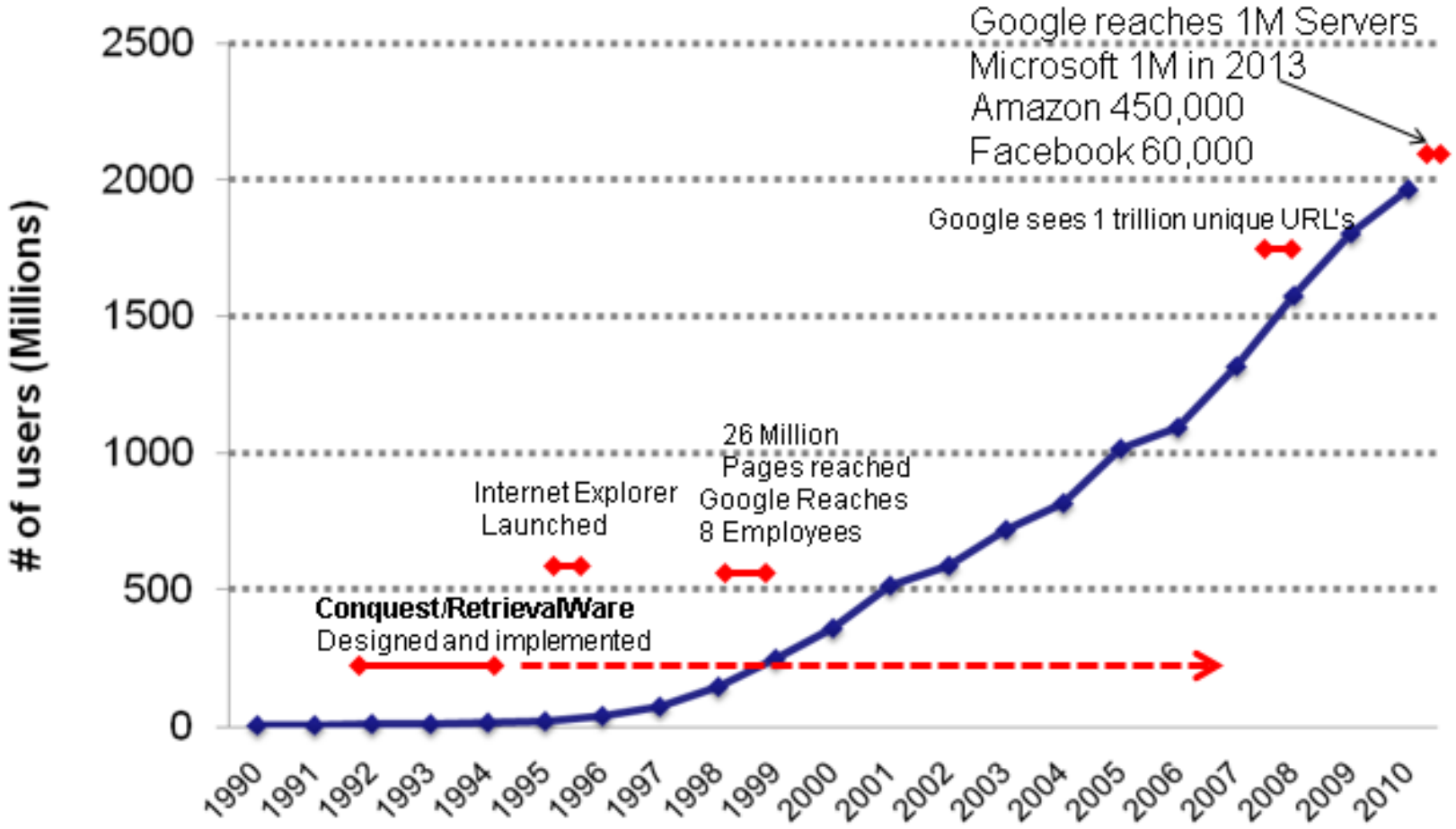
- **First 'modern' indexer of text**
  - Professor Gerard Salton (Harvard / Cornell)
  - *Automatic Information Organization and Retrieval*, (1968)
  - System for the Mechanical Analysis and Retrieval of Text
    - Invented Vector Space Model
      - A way of indexing, finding similar content
    - Invented TF/IDF
      - Relevance ranking
  - A Theory of Indexing (1975)

Proliferation of Data – Can We Keep Up?

- **Google's Eric Schmidt**
  - Between the Dawn of Civilization and late 2003.
    - 5 Exabytes of data created globally (5 Million Tb)
  - In 2010 we were generating
    - 5 Exabytes every two days!
- **IDC**
  - Global Data (2020) will likely exceed
    - 35 Zettabytes
    - 3,500 Exabytes
    - 35 Billion Terabytes



Proliferation of Data and Millions of Users!



Why Solr and Enterprise Search?

Why (Enterprise) Search	Why Solr? Famous Sites	Other Search Engines
History of Search	Functions of a Search Engine	Features & Scalability

Functions of a Search Engine

- Indexing: Organise “all” content for quick access
- Query Parsing : Understand what the user is looking for
- Search & Browse: Quickly find the information of interest
- Security: Filter out documents the user is not allowed to see
- Relevance Ranking: Order the information in useful ways

Search Engine… Be Helpful!

- Be Helpful: Show Snippets of info on hits, i.e. highlighting
- Be Helpful: Offer Query Suggestions
- Be Helpful: Offer Spelling Corrections
- Be Helpful: Offer refinement options to delve further, i.e. facets
- Be Helpful: Never suggest ‘dead ends’

Why Solr and Enterprise Search?

Why (Enterprise) Search	Why Solr? Famous Sites	Other Search Engines
History of Search	Functions of a Search Engine	Features & Scalability

Solr Features

- Advanced Full-Text Search Capabilities
- Optimized for High Volume Web Traffic
- Standards Based Open Interfaces - XML, JSON and HTTP
- Comprehensive HTML Administration Interfaces
- Server statistics exposed over JMX for monitoring
- Near Real-time indexing
- Flexible and Adaptable with XML configuration

Solr Features

- Extensible Plugin Architecture
- A Real Data Schema, with Numeric Types, Dynamic Fields, Unique Keys
- Faceted Search and Filtering
- Highly Configurable and User Extensible Caching
- Performance Optimizations
- Multiple search indices
- And more…

Scalability via SolrCloud

- Linearly scalable, auto index replication, auto failover and recovery with no single point of failure
- Centralized Apache ZooKeeper based configuration
- Automated distributed indexing/sharding

- Near Real-Time indexing with immediate push-based replication (also support for slower pull-based replication)
- Transaction log ensures no updates are lost even if the documents are not yet indexed to disk
- Automated query failover, index leader election and recovery in case of failure

Agenda

Why (Enterprise) Search	Why Solr? Famous Sites	Other Search Engines
History of Search	Functions of a Search Engine	Features & Scalability

Takeaway: We Learned…

- Why Enterprise Search is important
- Why Solr
- Solr is not the only choice, but IMHO (and many) the best
- About the History of Search
- The functions of a Search Engine
- The features of a Search Engine
- And that scaling is done via SolrCloud, which uses Zookeper