OPEN SOURCE SEARCH WITH APACHE LUCENE/SOLR

CHRISTINE POERSCHKE, BLOOMBERG

WHAT IS APACHE?

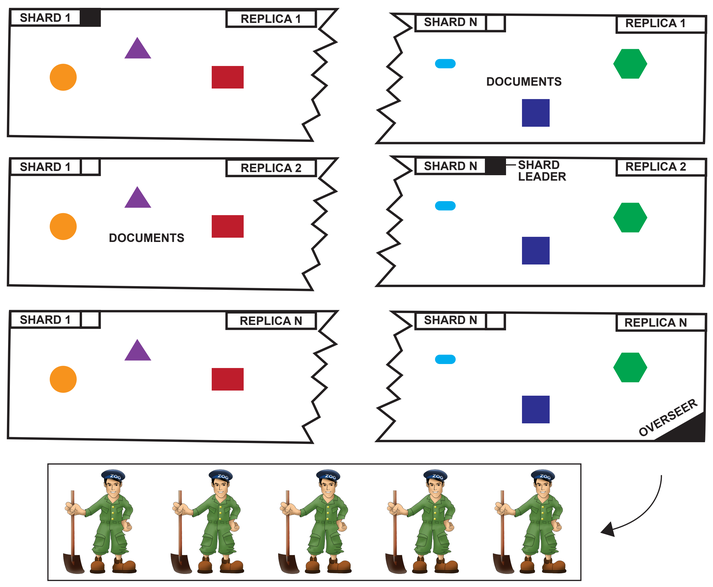
- Apache Software Foundation
- Established 1999 as a charitable organization
- Mission: to provide (free) software for the public good
- 350+ open source projects
- Meritocracy, collaboration, project independence
- Apache License used by many open source projects (including non-Apache projects)

WHAT ARE LUCENE AND SOLR?

- Apache Lucene is a search engine library written in Java.
- Apache Solr is a search platform, built on top of the Lucene library.

A PICTURE OF A SOLR CLOUD

- A Solr Cloud consists of multiple Solr instances on different machines, either virtual machines or 'bare metal'. The overall state of the cloud (the cluster state) is kept in Apache ZooKeeper™.
- A Collection in the Cloud is divided into Shards with each shard hosting a subset of Documents. Shards are replicated within the Cloud so that there is more than one copy of each shard's data. One of the replicas of a given shard will be elected to act as Shard Leader.
- The Cloud scales via replication and sharding: to handle more search queries, add more replicas on more machines; to have capacity for more documents, split your collection into more shards and house the extra shards on extra machines.



NEWS SEARCH AT BLOOMBERG

325K+ Subscribers **1 Million Stories**
9 Million Searches PER DAY **PUBLISHED EACH DAY**

INDEX OF 500 MILLION STORIES

500 Stories PER SECOND

Available for Search in ~100ms

180ms RESPONSE TIME

More. Better. Faster.

Alerts in 100ms

1.5 MILLION SAVED SEARCHES

OUR CHALLENGES

Load, latency, stability

- Very spikey flow of news and user load
- 'Always on', no downtime, 24/7/365

Indexing and searching

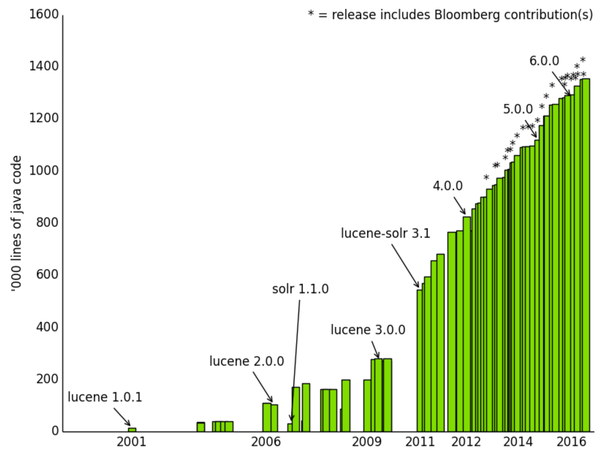
- News stories and searches in 40+ languages
- Content from 125K+ sources, from tweets to 100+ page-long research reports
- Complex user-specific privileging/ACL model
- Arbitrarily complex boolean queries
- Ticker lists e.g. "I want news on all the companies of interest to me."

Custom components

- Indexing, parsing, searching, postfiltering

CONTRIBUTING BACK

- 15+ Solr/Lucene contributors including 3 committers (I am one of them)
- Fixes or new features in 15+ releases, such as
  - o Lucene's SortingMergePolicy and EarlyTerminatingSortingCollector configurable in Solr (SOLR-5730 and SOLR-8621)
  - o Streaming Expressions (SOLR-7377)
  - o Learning To Rank integration into Solr (SOLR-8542)



COMMUNITY

- User mailing list: 3500+ subscribers, ~250 emails per week
- Dev mailing list: 800+ subscribers, ~100 emails per day
- Widely used in all sorts of sites and applications, from A(pple) to Z(appos.com)
- Many, sometimes competing, companies and individuals
- Talks and meetup groups in cities around the world
- New people welcome

DEVELOPMENT

- Task and issue tracking via Apache JIRA
- Source code in Apache Git apache repository
- Building and dependency management via Apache Ant and Apache Ivy
- Multi-platform build, continuous integration and other services hosted by Apache

RESOURCES

- lucene.apache.org/solr/quickstart.html
- Apache Solr Reference Guide
- Solr Community Wiki
- Books and blogs

EXPECTATIONS

- It's impossible to follow and understand everything
- Community of professionals who volunteer their time are enthusiastic, but busy (just like you)
- No one expects big patches; big code patches are rare, just start somewhere …
  - o community.apache.org
  - o wiki.apache.org/lucene-java/HowToContribute
  - o wiki.apache.org/solr/HowToContribute

LET'S BE REAL

Intellectual property

- Understand your employer's open source policies and guidelines
- Respect open source projects' licenses

Balance

- Work
- Open source work
- Life in general (I keep honeybees in my spare time)

INSPIRATION

- Everyone is unique, everyone's contributions are unique, find or create your place and role
- techatbloomberg.com/post-topic/open-source

SCAN HERE TO DOWNLOAD THIS POSTER.