Free Search: Lucene & Nutch

Doug Cutting
<doug@nutch.org>

Lucene is...

- A mature Apache open-source project;
- Java library for text indexing and search;
  - Not an application;
- A large community of contributors;
- The search technology behind a lot of web sites & applications (ZOË, JIRA, Lookout, Furl, etc.)
- http://jakarta.apache.org/lucene/
- A book out this summer!

Nutch is...

- A young open-source project;
- Web search application software;
- Two part-time paid developers;
- A growing number of contributors;
  - paid and un-paid.
- Behind a growing number of sites.

Nutch isn't...

- A business;
  - But is a non-profit legal entity to own copyright;
  - No employees.
- A search site;
  - But want to power lots of search sites;
  - From domain-specific, to whole-web.
- A research project.
  - But want to be platform for research.

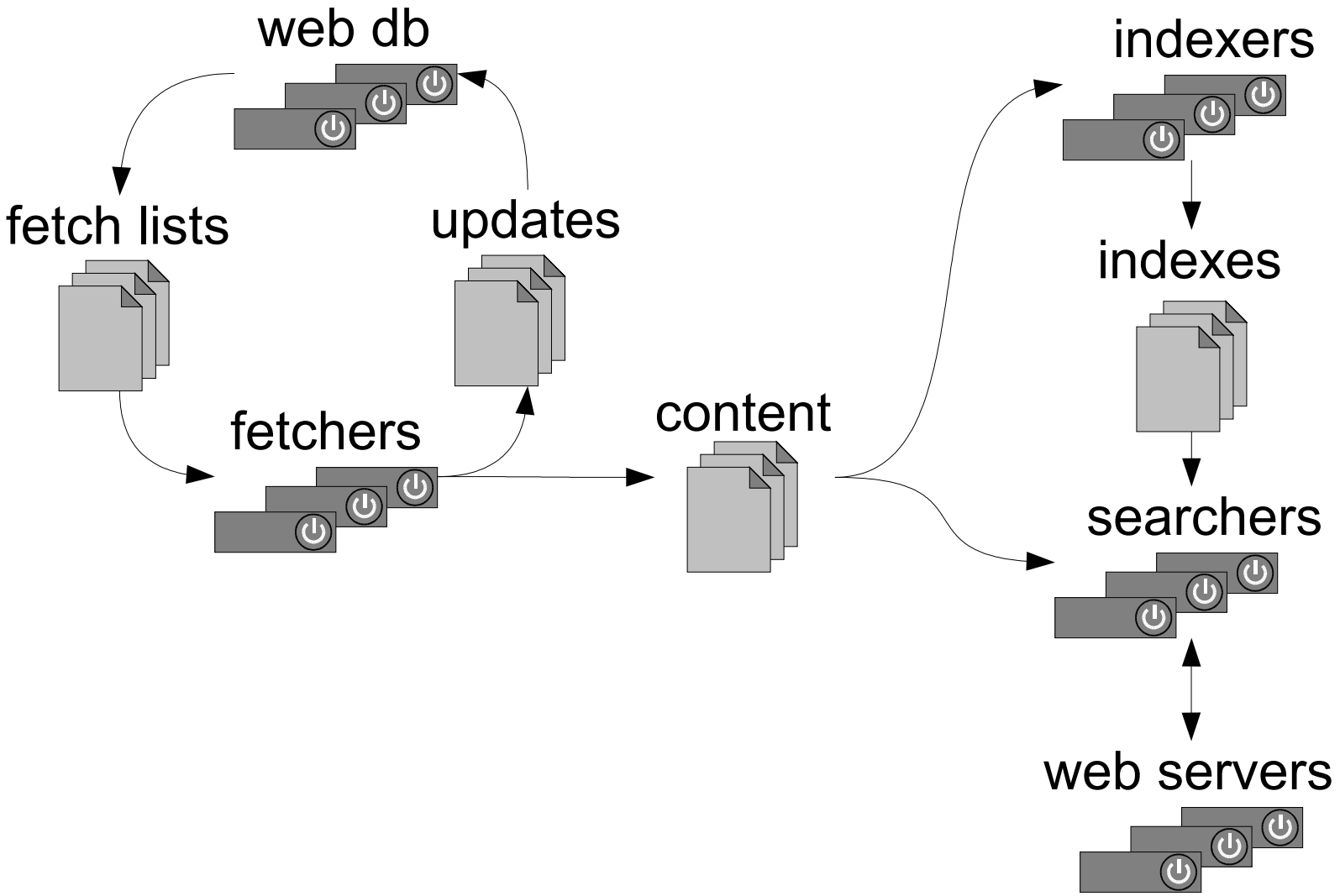
Nutch's Civil Goals

- Increase transparency of web search.
  - search is essential to internet navigation
  - yet algorithms are secret
  - small number of providers
- An open-source implementation can help:
  - enable more providers (free as in beer)
  - enable transparency (free as in freedom)

Nutch Technical Goals

- Scale to entire web
  - pages on millions of different servers
  - billions of pages
  - complete crawl takes weeks
  - very noisy
- Support high traffic
  - thousands of searches per second
- State-of-the-art search quality

Nutch Architecture



Web Database

- Page Database
  - Used for fetch scheduling.
- Link Database
  - Represents full link graph.
  - Stores anchor text associated with each link.
  - Used for:
    - Link analysis;
    - Anchor text indexing.
- This is not an RDBMS application!

Scalability

- Scales up:
  - multiple simultaneous fetches
    (100+ pages/second / CPU, ~10M / day)
  - parallel, distributed db update
    (100M pages @ 100 pages/second / CPU)
  - distributed search
    (2-20M pages, 1-40 searches/second / CPU)
- Scales down:
  - single box can easily handle 1M+ page intranet

Preliminary Evaluation at OSU: Nutch versus a Google Appliance

- For OSU's top-25 queries:
  - 9 queries nutch and google were both perfect: 10/10
  - 2 queries nutch was slightly better
  - 2 queries google was slightly better than nutch
  - 1 query google was much better: 10 to 6
  - 1 query google was much better: 10 to 6
  - 1 query both scored 5
  - Google Appliance had a slight overall advantage.

Demonstrations

- http://labs.yahoo.com/demo/nutch/
- http://www.mozdex.com/search.html
- http://www.objectssearch.com/en/search.html
- http://kodiak.cs.cornell.edu:8080/en/search.html
- http://devjr.cws.oregonstate.edu:8080/
- http://umkreisfinder.eventax.de/umkreisfinder.php

Current Status

- Re-architecting for easy extensibiltity:
  - Protocols (FTP, File, SQL, etc.)
  - Formats (Word, PDF, etc.)
  - Metadata indexing (location, license, pricing, etc.)
  - New query operators (site:, metadata, etc.)
- Working with Creative Commons
  - to develop search engine of CC-licensed content

http://www.nutch.org/

doug@nutch.org

Thanks to http://www.media-style.com/ for the Nutch logo & design.