Indexing big data with Tika, Solr, and map-reduce

Scott Fisher, Erik Hetzner

California Digital Library

8 February 2012
Outline

- Introduction
- Tika
- Pig
- Solr
- Done!
Web Archiving Service

- Service provided by the California Digital Library
- Fee-based
- Archiving web sites, as selected by curators
43 public archives
18 partners
58k crawls, 35k viewable by public
7535 sites
600 million URLs
40+ TB
Vital statistics

- 43 public archives
- 18 partners
- 58k crawls, 35k viewable by public
- 7535 sites
- 600 million URLs
- 40+ TB
Introduction

Vital statistics

- 43 public archives
- 18 partners
- 58k crawls, 35k viewable by public
- 7535 sites
- 600 million URLs
- 40+ TB
43 public archives
18 partners
58k crawls, 35k viewable by public
7535 sites
600 million URLs
40+ TB
Vital statistics

- 43 public archives
- 18 partners
- 58k crawls, 35k viewable by public
- 7535 sites
- 600 million URLs
- 40+ TB
Vital statistics

- 43 public archives
- 18 partners
- 58k crawls, 35k viewable by public
- 7535 sites
- 600 million URLs
- 40+ TB
Introduction

Tools

Open source and rails UI for crawl management and display of many focused web crawls. Heritrix - NutchWAX - Wayback

The Web has revolutionized our access to information, but Web publications are fragile, and ready access to Web resources cannot be taken for granted. The Web Archiving Service enables librarians and scholars to meet that challenge.

Search Archives
Subject experts have been capturing and preserving critical web sites to ensure that you have lasting access to web content. Public archives can be browsed or searched and provide persistent links to archived documents. Access to published archives is open to all. Click on an archive name on the right to search and view archived materials.

Learn more

Build Archives
With a WAS account you can capture, analyze and archive web sites and documents. Archives can be published or kept for private study. The WAS curator tools are easy to use, fully hosted, and allow collaborative collection building. CDL provides training and guidance for WAS curators.

Learn more
Nutch search

- Using Nutch for full text indexing
- Nutch is slowing down...
- Nutchwax (nutch + web archiving) is no longer supported
- Nutch search is no longer default with Nutch itself
- Deduplicating content requires a more sophisticated index.
Nutch search

- Using Nutch for full text indexing
- Nutch is slowing down...
- Nutchwax (nutch + web archiving) is no longer supported
- Nutch search is no longer default with Nutch itself
- Deduplicating content requires a more sophisticated index.
Nutch search

- Using Nutch for full text indexing
- Nutch is slowing down...
- Nutchwax (nutch + web archiving) is no longer supported
- Nutch search is no longer default with Nutch itself
- Deduplicating content requires a more sophisticated index.
Nutch search

- Using Nutch for full text indexing
- Nutch is slowing down...
- Nutchwax (nutch + web archiving) is no longer supported
- Nutch search is no longer default with Nutch itself

Deduplicating content requires a more sophisticated index.
Nutch search

- Using Nutch for full text indexing
- Nutch is slowing down...
- Nutchwax (nutch + web archiving) is no longer supported
- Nutch search is no longer default with Nutch itself
- Deduplicating content requires a more sophisticated index.
The web can contain anything.
Mostly HTML, but PDFs are very important.
Not to mention Office
The web can contain anything.

 Mostly HTML, but PDFs are very important.

 Not to mention Office
The web can contain anything.
 Mostly HTML, but PDFs are very important.
 Not to mention Office
Tika

- Apache software project
- Java
- Wraps parsers for different file types in a uniform interface.
- Parses most common file types.
- Use the same code to parse different types.
Tika

- Apache software project
- Java
  - Wraps parsers for different file types in a uniform interface.
  - Parses most common file types.
  - Use the same code to parse different types.
Apache software project
Java
Wraps parsers for different file types in a uniform interface.

- Parses most common file types.
- Use the same code to parse different types.
Apache software project

Java

Wraps parsers for different file types in a uniform interface.

Parses most common file types.

Use the same code to parse different types.
Tika

- Apache software project
- Java
- Wraps parsers for different file types in a uniform interface.
- Parses most common file types.
- Use the same code to parse different types.
Tika difficulties

- Some files are slow to parse.
- Some files blow up your memory.
- Some file parses never return.
Tika difficulties

- Some files are slow to parse.
- Some files blow up your memory.
- Some file parses never return.
Tika difficulties

- Some files are slow to parse.
- Some files blow up your memory.
- Some file parses never return.
Tika solutions

- Don’t parse files that are too big (e.g. > 2 MB)
- Fork and monitor process from the outside (Hadoop comes in handy)
Tika solutions

- Don’t parse files that are too big (e.g. > 2 MB)
- Fork and monitor process from the outside (Hadoop comes in handy)
What is Pig?

- Platform for data analysis from Apache.
- Based on Hadoop.
  - fault tolerant
  - distributed processing
- Can be used for ad-hoc analysis, without writing Java code.
- Embraced by the Internet Archive.
What is Pig?

- Platform for data analysis from Apache.
- Based on Hadoop.
  - fault tolerant
  - distributed processing
- Can be used for ad-hoc analysis, without writing Java code.
- Embraced by the Internet Archive.
What is Pig?

- Platform for data analysis from Apache.
- Based on Hadoop.
  - Fault tolerant
  - Distributed processing
- Can be used for ad-hoc analysis, without writing Java code.
- Embraced by the Internet Archive.
What is Pig?

- Platform for data analysis from Apache.
- Based on Hadoop.
  - fault tolerant
  - distributed processing
- Can be used for ad-hoc analysis, without writing Java code.
- Embraced by the Internet Archive.
What is Pig?

- Platform for data analysis from Apache.
- Based on Hadoop.
  - fault tolerant
  - distributed processing
- Can be used for ad-hoc analysis, without writing Java code.
- Embraced by the Internet Archive.
What is Pig?

- Platform for data analysis from Apache.
- Based on Hadoop.
  - fault tolerant
  - distributed processing
- Can be used for ad-hoc analysis, without writing Java code.
- Embraced by the Internet Archive.
Pig example

Data = LOAD 'arclist' USING org.cdlib.was.weari.pig.ArchiveURLParserLoader();
STORE Data INTO 'outputdir.json' USING org.cdlib.was.weari.pig.JsonParsedArchiveRecordStore;
Parse once!

Parsing takes forever. Do it once, store the results.

Storing raw text is cheap, compared to all those PDFs, HTML, etc.
Lessons learned

Parse once!

Parsing takes forever. Do it once, store the results.

Storing raw text is cheap, compared to all those PDFs, HTML, etc.
Parse once!

Parsing takes forever. Do it once, store the results.

Storing raw text is cheap, compared to all those PDFs, HTML, etc.
Lessons learned

Distribute from the start

- Use hadoop, pig, or another system to distribute your computing.
- Don’t use an ad-hoc solution. Take the time up front to distribute things.
Lessons learned

Distribute from the start

- Use hadoop, pig, or another system to distribute your computing.
- Don’t use an ad-hoc solution. Take the time up front to distribute things.
Faceting

California Government Sites Sample Archive

Search

Arnold Schwarzenegger in full text

Limit search to: from: 11/22/2010 to: 02/01/2012

Web site:
- California Housing Finance Agency
- California Institute for Regenerative Medicine
- California Integrated Waste Management Board
- California Law Revision Commission
- California Maritime Academy
- California Office of Health Information Integrity
- California Science Center
- California State Assembly
- California State Association of Counties (CSAC)
- California State Library
- California State Parks

Topic:
- Culture
- Economy
- Education
- Elections & politics
- Environment
- Governor
- Health
- Labor
- Law & judicial
- "Really long tags are a favorite please enjoy"
- Science
- Social services

Media type:
- HTML
- Image
- PDF
- Office
- Compressed
- Audio
- Video

You can narrow your search by date range, to particular web sites, to sites on certain topics, or to particular file types, such as PDF files.

See search tips for further information.
### California Government Sites Sample Archive

**Search**

- **Title:** asm_weekly_X1_20110616_3935
- **Captured:** 06/23/11 01:52 AM
- **URL:** [192.234.213.35/clerkarchive/session/awh061811_01x.pdf](http://192.234.213.35/clerkarchive/session/awh061811_01x.pdf)
- **Abstract:** of the Constitution of the State of California, Governor **Arnold Schwarzenegger**... by Governor **Arnold Schwarzenegger** Convened December 6, 2010. *Superseded*... **ARNOLD SCHWARZENEGGER**, Governor of the State of California, in accordance

- **Title:** History document
- **Captured:** 06/23/11 01:52 AM
- **URL:** [192.234.213.35/clerkarchive/session/awh111209xxxxxx.pdf](http://192.234.213.35/clerkarchive/session/awh111209xxxxxx.pdf)
- **Abstract:** therefore, I, **ARNOLD SCHWARZENEGGER**, Governor of the State of California

---

**Refine Your Results**

- **web site:** California State Assembly [8] (remove)
- **site topic:** Law & Judicial [8] (remove)
- **media type:** Pdf [8] (remove)
- **date:** from: 01/22/2010 to: 02/02/2012

---

**Faceting 2**

Scott Fisher, Erik Hetzner (CDL) Indexing big data 8 February 2012 16 / 19
Mime types

```
SELECT mime_type FROM crawl_mime_type
WHERE mime_type LIKE '%word%';
```
Solr XML

```xml
<str name="institution">CDLQA</str>
<arr name="job">
  <str>00022578</str>
</arr>
<str name="mediatypedet">application/pdf</str>
<str name="mediatypegroupdet">pdf</str>
<str name="mediatypesup">application/pdf</str>
<str name="project">CDLQA_ag_64</str>
<str name="site">192.234.213.35</str>
<arr name="specification">
  <str>spec:0000001x8</str>
</arr>
<arr name="tag">
  <str>Law & Judicial</str>
</arr>
```
Be careful when you try to parse at a bunch of files you downloaded from the web.

Parse and store.

Distribute up front.

Build a test index first.

http://webarchives.cdlib.org/

scott.fisher@ucop.edu, erik.hetzner@ucop.edu