Infrastructure for Supporting Exploration and Discovery in Web Archives

Jimmy Lin
University of Maryland
@lintool

Working with Internet Archives for Research
Wednesday, June 18, 2014
From the Ivory Tower...
... to building sh*t that works


I worked on...
- data products to surface relevant content to users
- analytics infrastructure to support data science
circa ~2010

~150 people total
~60 Hadoop nodes
~6 people use analytics stack daily

circa ~2012

~1400 people total
10s of Ks of Hadoop nodes, multiple DCs
10s of PBs total Hadoop DW capacity
~100 TB ingest daily
dozens of teams use Hadoop daily
10s of Ks of Hadoop jobs daily
And back!

Web archives are an important part of our cultural heritage... 

... but they’re underused
Why?

Restrictive use regimes?
But I don’t think that’s all…
Users can’t do much with current web archives

Hard to develop tools for non-existent needs

Source: http://www.flickr.com/photos/cheryne/8417457803/
We need *deep* collaborations between:

**Users** (e.g., archivists, journalists, historians, digital humanists, etc.)

**Technologists** (me and my colleagues)

Goal: tools to support exploration and discovery in web archives

Beyond browsing…
Beyond searching…

Source: http://waterloocyclingclub.ca/wp-content/uploads/2013/05/Help-Wanted-Sign.jpg
Scalable storage of archived data
Efficient random access
Scalable processing and analytics
Scalable storage and access of derived data

HDFS
HBase
Hadoop MapReduce

(Existing work by IA, IIPC, Common Crawl, etc.) (Not much previous work here…)
Warcbase

An open-source platform for managing web archives built on Hadoop and HBase

http://warcbase.org/

A collection of tables, each of which represents a sparse, distributed, persistent multidimensional sorted map

(Open-source implementation of Google’s Bigtable, the system behind Gmail, Google Maps, etc.)
Warbase Application Lifecycle

WARC/ARC

Ingestion

Processing & Analytics

Applications and Services

Apache HBase

Hadoop MapReduce
Processing and Analytics

text analysis, webgraph analysis, …

HDFS files → Hadoop MapReduce → HDFS files

HBase rows → HBase rows

Derived data stored alongside raw source!
Scalability?
We got 99 problems but scalability ain’t one…
– Jay-Z

Scalability of Warcbase limited by Hadoop/HBase
Applications are lightweight clients
Demo Applications

WARC/ARC → Ingestion → Hadoop Map Reduce → Processing & Analytics → Apache HBase → Applications and Services
Sample dataset: crawl of the 108th U.S. Congress

Monthly snapshots, January 2003 to January 2005
1.15 TB gzipped ARC files
29 million captures of 7.8 million unique URLs
23.8 million captures are HTML files

Hadoop/HBase cluster:
16 nodes, dual quad-core processors, 3 × 2TB disks each
**Warcbase Browser**  = Wayback Machine

Implementation
- Lightweight HBase client embedded in Jetty app
- Reuses Wayback code for rendering

Key features
- Offloads storage management to HBase
- Transparently scales out with HBase
Dear Friend,

Thank you for visiting my on-line office! I appreciate your interest in the issues before the United States Senate.

Please let me hear from you about your views on the issues that matter to your family and your community.

Sincerely,

[Signature]

Senator Clinton's Thanksgiving Trip to Afghanistan and Iraq

[Image of Senator Clinton and Senator Reed with U.S. Marines outside of the U.S. Embassy in Kabul, Afghanistan]
GOP Education Leaders Unveil College Cost Central Website to Seek Input from Parents & Students

Washington, D.C. Providing a new resource for parents, students, and taxpayers troubled by dramatically increasing higher education prices, Republicans on the U.S. House Education & the Workforce Committee today announced the launch of the College Cost Central website. Read more.

Click here for the College Cost Central website

House GOP Education Committee Leaders Release Report on College Cost Crisis

WASHINGTON, D.C. On September 4, 2003 the U.S. House Education & the Workforce Committee Chairman John Boehner (R-OH) and 21st Century Competitiveness Subcommittee Chairman Howard P. "Buck" McKeon (R-CA) introduced a congressional report declaring that the nation's higher education system is in crisis as a result of exploding cost increases that threaten to put college out of reach for low and middle income students and families. Read more.
Topic Model Explorer

Implementation
LDA on each temporal slice
Adaptation of Termite visualization
Webgraph Explorer

Implementation

Link extraction with Hadoop, site-level aggregation
Computation of standard graph statistics
d3 interactive visualization
Name: Craig, Larry
Party: Republican
State: Idaho
Homepage: http://craig.senate.gov/

Outgoing Links:
- Simpson, Mike
- Cramer, Bud
- Hunter, Duncan
- Ensign, John
- Cunningham, Randy "Duke"
- Manzullo, Don
- Levin, Carl
- Rogers, Mike
- Walden, Greg
- Lincoln, Blanche
- Hatch, Orrin G.
- Landrieu, Mary L.
- Lowey, Nita M.
- Davis, Jo Ann
Homepage: http://waysandmeans.house.gov/

Outgoing Links:
- Lewis, John
- Herger, Wally
- Sandlin, Max
- McCrery, Jim
- Becerra, Xavier
- Doggett, Lloyd
- Cantor, Eric
- Portman, Rob
- Hulshof, Kenny
- Kleczyka, Jerry
- Weller, Jerry
- United States. Congress. Joint Committee on Taxation.
- Shaw, E. Clay
- Thomas, Bill
- Stark, Fortney Pete
- Crane, Phil
Goal: tools to support exploration and discovery

We need *deep* collaborations between:

**Users** (e.g., archivists, journalisms, historians, digital humanists, etc.)

**Technologists** (me and my colleagues)

Warcbase is just the first step…