Warcbase
Building a Scalable Platform on HBase and Hadoop
Part Two: Historian Use Case

Jimmy Lin
University of Maryland
College Park, MD

Ian Milligan
University of Waterloo
Waterloo, ON Canada
Why should a historian care?

The sheer amount of social, cultural, and political information generated every day presents new opportunities for historians.
Could one even study the 1990s and beyond without web archives?
No.

Historians need to do this now, or we’re going to be left behind.
Nightmare Scenario

• Wayback Machine won’t be enough. We won’t use that.

• Historians rely uncritically on date-ordered keyword search results, putting them at mercy of search algorithms they do not understand;

• Historians are completely left out of post-1996 research, letting everybody else do the work (a la Culturomics project/Nature magazine article);

• Our profession gets left behind…
Unlocking an Archive-It Collection

• Archive-It has **amazing collections** of social, cultural, political, and economic records generated by **everyday people, leaders, businesses, academics, and beyond.**

• Stories waiting to be hold.

• The data is there, but the problem is **access.**
Example Dataset

- Archive-It Collection 227, Canadian Political Parties and Political Interest Groups (University of Toronto)

- October 2005 - Present

- All major and minor political parties, as well as organized political interest groups (Council of Canadians, Coalition to Oppose the Arms Trade, Assembly of First Nations, etc.)

- Started by now-retired librarian, hard to get details on seed list
Two Main Approaches

- Warcbase
  - Link extraction and analytics
  - Full-text extraction and analytics
- Full-text faceted search
  - UK Web Archive’s **Shine** solr front end
Using Warcbase to analyze links and full-text
Basic Link Statistics

- Count number of pages per domain
- Count number of links for each crawl so they can be normalized (very important)
- Run on command line using relatively simple pig scripts
Example Script (counting number of links for each crawl)

register 'target/warcbase-0.1.0-SNAPSHOT-fatjar.jar';

DEFINE ArcLoader org.warcbase.pig.ArcLoader();
DEFINE ExtractLinks org.warcbase.pig.piggybank.ExtractLinks();

raw = load '/shared/collections/CanadianPoliticalParties/arc/' using ArcLoader as
  (url: chararray, date: chararray, mime: chararray, content: bytearray);

a = filter raw by mime == 'text/html' and date is not null;
b = foreach a generate SUBSTRING(date, 0, 6) as date, url, FLATTEN(ExtractLinks((chararray) content, url));
c = group b by $0;
d = foreach c generate group, COUNT(b);
Social Media Appearances -
Twitter

(20080611220246, http://creativecommons.org/, twitter)
(20080930221618, http://www.ndp.ca/home, twitter)
(20080930221618, http://www.ndp.ca/home, twitter)
(20080930221638, http://www.liberal.ca/default_e.aspx, twitter)
(20080930221641, http://www.liberal.ca/story_15081_e.aspx, twitter)
(20080930221714, http://www.liberal.ca/video_e.aspx, twitter)
(20080930221903, http://www.ndp.ca/page/5246, twitter)
(20080930222124, http://www.ndp.ca/bloggingtools, twitter)
(20080930223240, http://www.liberal.ca/depth_e.aspx, twitter)
(20080930223258, http://www.liberal.ca/enews_e.aspx, twitter)
(20080930223315, http://www.liberal.ca/glance_e.aspx, twitter)
(20080930223320, http://www.liberal.ca/story_15073_e.aspx, twitter)
(20080930223323, http://www.liberal.ca/gallery_e.aspx, twitter)
Social Media Appearances - Facebook

(20070418135140, http://www.liberal.ca/glance_e.aspx, facebook)
(20070418151727, http://www.equalvoice.ca/youth/, facebook)
(20070418151843, http://www.equalvoice.ca/youth/Bios.htm, facebook)
(20070418153832, http://greenparty.ca/fr/node/1280, facebook)
(20070518134656, http://www.liberal.ca/glance_e.aspx, facebook)
(20070518134918, http://www.liberal.ca/conversation_e.aspx, facebook)
(20070518134918, http://www.liberal.ca/conversation_e.aspx, facebook)
(20070518134941, http://www.ndp.ca/page/4733, facebook)
Link Analysis

- Extracting links by domain (tab-separated values):

200810 conservative.ca digg.com 2325
200810 conservative.ca facebook.com 2325
200810 conservative.ca mycampaign.conservative.ca 7902

[..]

200902 liberal.ca ctv.ca 16
200902 liberal.ca del.icio.us 1118
200902 liberal.ca digg.com 1118
December 2006
Stephane Dion Elected Leader of Party
April 2008
Fundraising with the Victory Fund/
Fonds de la Victoire
July 2008

The Green Shift Announced!
December 2008

Election campaign Ends; Attacking Harper on Anti-American Grounds (bushharper)
Other Cases

- Extracting all links to the mainstream media, or thinktanks, or other political parties
2005 Canadian Federal Election
register 'target/warcbase-0.1.0-SNAPSHOT-fatjar.jar';

DEFINE ArcLoader org.warcbase.pig.ArcLoader();
DEFINE ExtractRawText org.warcbase.pig.piggybank.ExtractRawText();
DEFINE ExtractTopLevelDomain org.warcbase.pig.piggybank.ExtractTopLevelDomain();

raw = load '/shared/collections/CanadianPoliticalParties/arc/' using ArcLoader as 
  (url: chararray, date: chararray, mime: chararray, content: bytearray);

a = filter raw by mime == 'text/html' and date is not null;
b = foreach a generate SUBSTRING(date, 0, 6) as date,
   REPLACE(ExtractTopLevelDomain(url), '^\s*www\.', '') as url, content;
c = filter b by url == 'greenparty.ca';
d = foreach c generate date, url, ExtractRawText((chararray) content) as text;

store d into 'cpp.text-greenparty';
Text Analysis

- Now have circumscribed corpus for specified query (i.e. liberal.ca, or ndp.ca, or conservative.ca)

- Can now use standard text analysis tools, etc. to extract meaning
  - LDA (topic modeling)
  - NER (named entity recognition)
October 2005

62476 Stephen Harper
30234 Michael Chong
30109 Gwynne Dyer
28011 ami Entrez
26238 Paul Martin
22303 Harper
November 2008

3188 Stéphane Dion
2557 Stephen Harper
2471 Stephen Harper Laureen
2410 Dion
2356 Harper
Shine

- UK Web Archive’s Shine (https://github.com/ukwa/shine)

- Indexing as bottleneck

- ~ 250GB of WARC takes ~ 5 days on a single machine

- Hadoop indexer available if data in HFDS

- ~ 90GB index size
Examples

DEMO GODS

PLEASE LET THIS DEMO WORK
Shine

- **Advantages**: accessible to the general public, easy to use, interactive trend diagram allows digging down for context, can move down to level of document itself.

- **Disadvantage**: keyword searching requires you know what to look for; random sampling misleading when tens of thousands of records; etc.

- Doesn’t take advantage of what makes web sources so powerful: hyperlinks
Building connections between Warcbase and Shine
Conclusions &
Thanks

Jimmy Lin
University of Maryland
College Park, MD

Ian Milligan
University of Waterloo
Waterloo, ON Canada