Warcbase: Building a Scalable Web Archiving Platform on HBase and Hadoop

Jimmy Lin
University of Maryland
@lintool

Ian Milligan
University of Waterloo
@ianmilligan1

Thursday, June 4, 2015
When does an event become “history”?
When does an event become “history”? ~20-30 years later

The history of the 1960s were written in the 1980s!
“This scandal was brought to you by the digital revolution. That meant we could access all the information we wanted, when we wanted it, anytime, anywhere, and when the story broke in January 1998, it broke online. It was the first time the traditional news was usurped by the Internet for a major news story, a click that reverberated around the world…

…it was before social media, but people could still comment online, email stories, and, of course, email cruel jokes. News sources plastered photos of me all over to sell newspapers, banner ads online, and to keep people tuned to the TV.”

Monica Lewinsky – The Price of Shame
TED Talk, March 2015

So, where are those web pages now?
When does an event become “history”?  
~20-30 years later

Historians are getting ready to write about the 90s…
Right about now!
Can you write a history of the 1990s without the web?

So, where are those web pages now?
And beyond...

70+ web archiving efforts worldwide!
Can you write a history of the 1990s without the web?

So, where are those web pages now?
Great, let’s get to work...
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.)

**Tool builders** (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
What would a web archiving platform built on modern big data infrastructure look like?
Petabox by Internet Archive; NAS, SAN, etc. by others

Scalable storage of archived data
Efficient random access
Scalable processing and analytics
Scalable storage and access of derived data

OpenWayback: monolithic Tomcat application

Some work by the Internet Archive, Common Crawl, and others…

Ad hoc storage in flat text WAT files
Desiderata

HDFS
Scalable storage of archived data
Efficient random access HBase
Scalable processing and analytics Hadoop
Scalable storage and access of derived data HBase

Existing tools aren't adequate!
Open source implementation of the Google File System
Stores data blocks across commodity servers
Scales to 100s of PBs of data
Open source implementation of Google’s framework
Suitable for batch processing on HDFS data

Shuffle and Sort: aggregate values by keys
A collection of tables, each of which represents a sparse, distributed, persistent multidimensional sorted map.

Source: Bonaldo Big Table by Alain Gilles
Warcbase

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

http://warcbase.org/

WARC data

Ingestion

Hadoop MapReduce

Spark

Processing & Analytics
text analysis, link analysis, …

Apache HBase

Applications and Services
Warcbase: here.
Warcbase: here.
Warcbase in a box

✗ cylinder
Warcbase: here.
Portable Warcbase
What’s the big deal?

Historians probably can’t afford Hadoop clusters…
But they can probably afford a Mac Pro or Macbook

How will this change historical scholarship?

Visual graph analysis on longitudinal data, select subsets for further textual analysis
Drill down to examine individual pages… all on your desktop/laptop!
Warcbase: here.
The price of three cocktails
What’s the big deal?

Store every page you’ve ever visited in your pocket!
Throw in search, lightweight analytics, …

What will you do with the web in your pocket?
How will this change how you interact with the web?
So what can you do with Warcbase?