INTERNET ARCHIVE

Full Text Search of Web Archive Collections

IWAW2005

Michael Stack

Internet Archive

stack@ archive.org

INTERNET ARCHIVE

Internet Archive

- IA is a 501(c)(3)non-profit
- Mission is to build a public Internet digital library.
  - Movies, Books, Web, etc.
  - "*Universal Access to All Human Knowledge*"
    - This talk is about developing *access* to Web Collections

INTERNET ARCHIVE

IA Web Collection

- IA Web Archive Collection currently:
  - ~600TB of data
  - ~60 Billion URLs (Google ~8 Billion).
- Bulk of Web Collection donated by Alexa Internet
  - Private company, crawling since 1996.
- 2-month rolling snapshots
  - Recent: 4.7 billion URIs, 180 million websites, ~44TB.
- Crawling software
  - Sophisticated
  - Weighted towards popular sites
  - Proprietary: We only receive the data

INTERNET ARCHIVE

IA Web Collection Access

- Currently, only public access to Web Collection is via the...

INTERNET ARCHIVE
WayBackMachine

Enter Web Address: http://www.duboce.net | All | Take Me Back | Adv. Search | Compare Archive Pages

Searched for http://www.duboce.net — **14** Results

Note some duplicates are not shown. See all.
* denotes when site was updated.

Search Results for Jan 01, 1996 - Sep 14, 2005

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	0 pages	0 pages	0 pages	3 pages	0 pages	3 pages	6 pages	1 pages
					May 18, 2001 *		Feb 10, 2003 *	May 12, 2004	Jan 04, 2005
					Jul 22, 2001		Apr 22, 2003 *	May 22, 2004	
					Nov 27, 2001		Jun 12, 2003 *	Sep 21, 2004 *	
								Sep 22, 2004 *	
								Nov 14, 2004 *	
								Dec 22, 2004 *	

Public Access Limitation

- But Wayback Machine has major shortcomings:
  - You must know the exact URL of the page you want to browse beforehand.
    - Can use URLs from published sources.
    - Can use search engines to first find URLs.
    - Does not allow for serendipitous exploration of past.
    - Particularly frustrates users accustomed to Google-style full text query box.
  - IA Wayback Machine cannot be distributed.
- Need full text search of Web Archive Collections!

Not Just IA...

- Others also need to be able to full text search Web Archive Collections. E.g..
  - Members of the International Internet Preservation Consortium (IIPC), twelve National Libraries (& IA).
    - Libraries developing own WACs.
  - Users of HERITRIX the IA's open source, extensible, archival-quality, web crawler.

INTERNET ARCHIVE

Searching Web Archive Collections

- Has much in common searching any large collection of Web resources.
  - Parse common types.
    - text/*, application/pdf, application/msword, etc.
  - Return quality search results.
  - Scales, Performant, Cheap, Customizable, etc.
- But also challenges unique to Web Archive Collections (WACs).

INTERNET ARCHIVE

What is a Web Archive Collection?

- Web Archive Collections (WACs) are typically an aggregate of multiple, related, focused Web crawls. Examples:
  - The IA has been making a Hurricane Katrina WAC crawling Katrina sites on a 3-day period.
  - The IA ran 34 weekly crawls of sites pertaining to the US 2004 Presidental election.

WAC Attributes

- Key attributes of WACs:
  - Tend to be large. E.g.:
    - 34 weekly US 2004 election crawls 140million.
    - 60 Billion URLs of the IA Web collection is a WAC
  - A single URL may appear multiple times.
    - 34 instances of http://www.whitehouse.gov in US 2004 election WAC.
  - No help from Web displaying found pages.
    - Click on search result takes you where? 1999? 2001?
    - A Wayback-like *viewer* application or Search Engine Cache required – rendors pages from past in browser

Nutch as Search platform

- Nutch, *lucene.apache.org/nutch*, selected as search engine platform on which to develop WAC search.
  - "Nutch is a complete open-source Web search engine package that aims to index the World Wide Web as effectively as commercial search services"
  - Technically, Nutch provides basic search engine capability, is extensible, aims to be cost-effective, and is demonstrated capable of indexing up to 100 million documents.
    - Convincing development story for how to scale up to billions (More on this later).
  - Just as importantly, policy-wise, Nutch project aligns with mission of IA/IIPC:
    - Transparent alternative to commercial web search engines.
      - Signature feature is ability to *explain* search result rankings.
        - » See 09/17 FT excerpt from John Battelle 'The Search'. Horror stories of how the inscrutable alogrithm can change a small business's fortune overnight.

Nutch Overview

- Popular open source project.
- Java.
- Builds on Lucene search lib, adding: crawler, a link-graph database, parsers for common types, etc.
- Customizable at parse-time, index-time, and query-time via *plugins*.
  - Add/ Amend query terms.
  - Add/ Amend parsers.

Nutch Overview: Indexing 1

- Runs in stepped, batch mode.
  - Should a step fail (machine crash or operator misconfiguration), just redo.
- First step, "segment" the work
  - No single step overwhelms as collection grows.
    - In current implementation, not the case (Will revisit later below).
  - Can distribute “segments” across machines.
- Custom DB maintains state.
  - Between steps and across segments.
  - Computes link structure.

Nutch Overview: Indexing 2

- Steps:
  1. Ask Nutch DB to generate URLs to fetch.
  2. Fetch and Parse the downloaded pages.
  3. Update Nutch DB, run analysis on DB content, update segments.
  4. Index parsed text + in-link anchor text (From Nutch DB).
  5. Optionally remove duplicates (By URL + content MD5).
  6. Optionally merge all segment indices.
- In current Nutch (<= v0.7) implementation:
  - Steps 2 and 4 may be distributed/run in parallel.
  - All other steps require either single process exclusive access to Nutch DB or single process exclusive access to all segment data.
  - A step must complete before the next can begin.

Nutch Overview: Querying

- Start the Nutch search Web application.
  - Run multiple to distribute query processing.
    - Distributes by remotely invoking queries against all query cluster participants.
      - Each query cluster participant is responsible for some subset of all “segments”.
- Queries return ranked Google-like results.
  - Support for basic query terms: *url*, *site*, *etc.*

INTERNET ARCHIVE

Adapting Nutch to WAC Search

- WAC search needs to support 2 distinct modes of operations.
  1. As Google-like search engine.
     - No duplicate URL pollution in results.
  2. Allows study of how pages change over time
     - "*return all versions of* www.whitehouse.gov *crawled in 1999 sorted by crawl date*".
       - This is what IA Wayback Machine does.
       - IA Wayback Machine cannot do "...and contains terms 'Hilary' and 'Clinton'"
     - Also add support for IA WM-like WAC viewer application.

Adapting Nutch: Mode 1

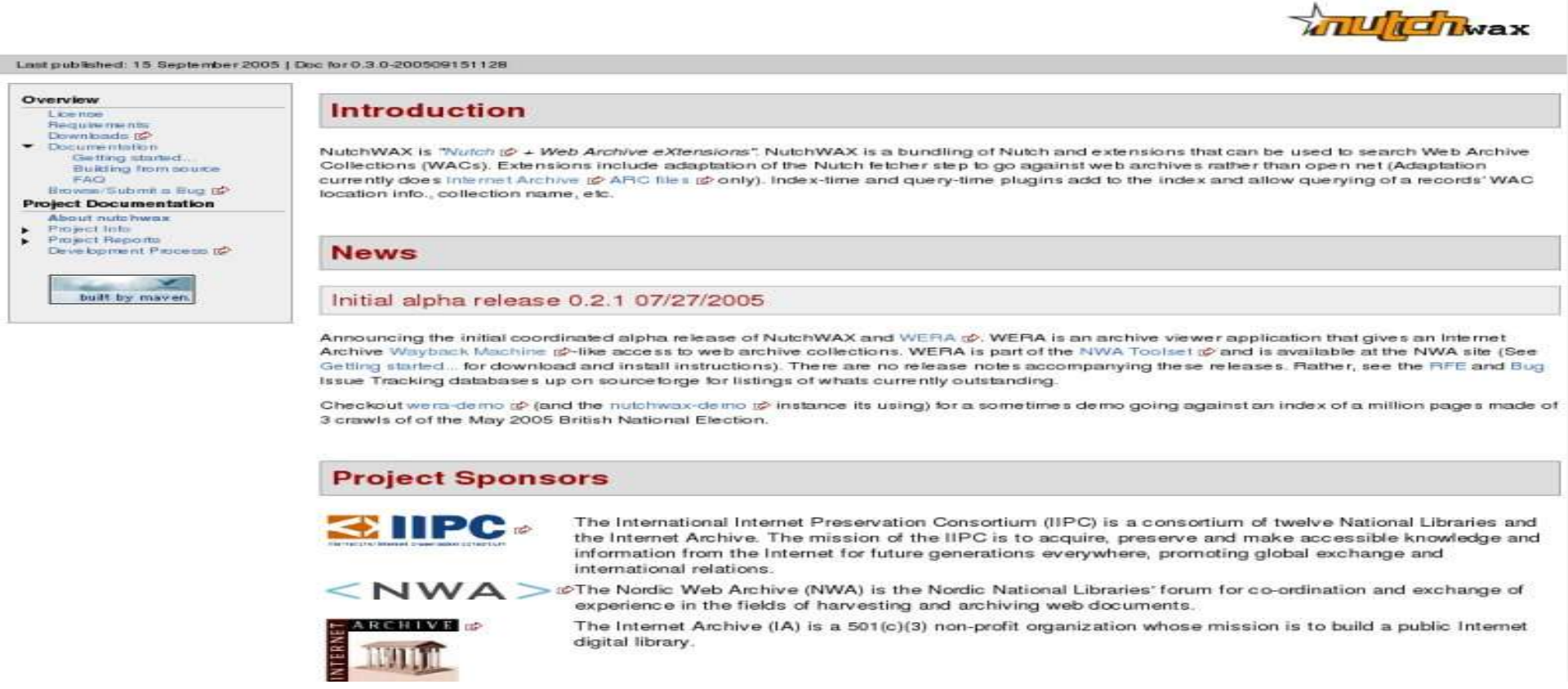
- Nutch fetcher step recast to pull content from a WAC repository rather than from the live Web.
  - WAC content already exists, previously harvested by other means.
    - At IA harvested content is stored in ARC files.
    - ARC-to-segment tool feeds ARCs to Nutch parsers and segment content writers.
      - Adaptation for formats other IA ARC, trivial.
- Upon completion, using Nutch indices purged of exact duplicates, possible to deploy basic WAC search using IA Wayback Machine as WAC viewer.

Adapting Nutch: Mode 2

- Added following to support WAC *viewer* and wayback-like querying:
  - Added support for explicit IA 14-digit – YYYYDDMMHHSS -- date and date range querying.
    - Replaced Nutch native date support.
  - Added ARC location information to search result: *collection*, *arcname*, and *arcoffset*.
  - Added default parser. Used when no parser match found (e.g. types w/o text). Adds meta-info on each resource.
    - Allows stylesheets, images, etc., to be found by WAC *viewer* drawing pages from past.
  - Native Nutch modified to support sort on arbitrary fields: *sort*, *reverse*.
  - Native Nutch modified to support removal of duplicates at query time (rather than at index time): *hitsPerDup*, *dedupField*.
  - Native Nutch modified to return results as XML (A9 OpenSearch RSS).
- Upon completion, both modes of operation possible using same non-deduplicated index. Example queries:
  - *hilary clinton site:www.whitehouse.gov date:1999-2000 sort:date*
  - *levees collection:katrina date:200508-200509 sort:date*

Nutchwax

- All Nutch WAC plugin extensions, documentation, and scripts are open source, hosted on Sourceforge under the *Nutchwax* -- Nutch with Web Archive eXtensions – project: http://archive-access/projects/nutch/



Web Images Groups News Froogle Local more »

Google

thomas jefferson site:gov Search

Web Results 1 - 8 of about 1,340,000 for thomas jefferson site:gov. (0.04 seconds)

Biography of Thomas Jefferson
Biography of **Thomas Jefferson**, the third President of the United States (1801-1809).
www.whitehouse.gov/history/presidents/tj3.html - 35k - Cached - Similar pages

The Thomas Jefferson Papers - (American Memory from the Library of ...
The **Thomas Jefferson** Papers consist of 27000 documents and is the largest collection of original **Jefferson** documents in the world.
lcweb2.loc.gov/ammem/collections/**jefferson**_papers/ - 9k - Cached - Similar pages

Thomas Jefferson Memorial (National Park Service)
Official National Park Service site. Includes information about the monument and its history.
www.nps.gov/thje/ - 26k - Cached - Similar pages

Thomas Jefferson Memorial Home Page
Site about **Jefferson** from the National Park service. Include details on monuments and memorials.
www.nps.gov/thje/home.htm - 9k - Cached - Similar pages
[More results from www.nps.gov]

Thomas Jefferson
Biographical sketch summarizes **Jefferson's** accomplishments.
sc94.ameslab.gov/TOUR/tj**efferson**.html - 17k - Sep 7, 2005 - Cached - Similar pages

Thomas Jefferson
Thomas Jefferson, third president of the United States. **Thomas Jefferson**, third president of the ... **Thomas Jefferson's** drawing of a macaroni machine. ...
www.americaslibrary.gov/cgi-bin/page.cgi/aa/**jefferson** - 12k - Sep 6, 2005 - Cached - Similar pages

Thomas Jefferson (Library of Congress Exhibition)
This exhibition focuses on the extraordinary written legacy of **Thomas Jefferson**—founding father, farmer, architect, inventor, slaveholder, book collector, ...
www.loc.gov/exhibits/**jefferson**/ - 5k - Cached - Similar pages

JEFFERSON, Thomas - Biographical Information
JEFFERSON, Thomas, (1743 - 1826). **JEFFERSON, Thomas**, (father-in-law of **Thomas** Mann Randolph and John Wayles Eppes), a Delegate from Virginia, ...
bioguide.congress.gov/scripts/biodisplay.pl?index=j000069 - 4k - Cached - Similar pages

INTERNET ARCHIVE

Web Search

thomas jefferson Search

Search took 0.708 seconds. Hits **1-8** (out of about 64,161 total matching pages):

THOMAS -- U.S. Congress on the Internet
... Word/Phrase Quick Links : House | House Clerk | House Directory | Senate | Senate Directory | GPO LINKS LEGISLATION CONGRESSIONAL RECORD COMMITTEE INFORMATION About **THOMAS THOMAS** FAQ Congress & Legislative Agencies How Congress Makes Laws: House | Senate Résumés of Congressional Activity Days in**THOMAS** -- U.S. Congress on the Internet The Library of Congress Congress Now : House Floor This ...
http://thomas.loc.gov/ - 2004-10-14 20:58:29 - other versions - explain

Thomas Jefferson Memorial (National Park Service)
Thomas Jefferson Memorial (National Park Service) **Thomas Jefferson** Memorial Visitor Center Open All Year View all Facilities » Fee Information View all Fees » Maps and Brochures » Printable ... the separation between church and state, and in education available to all. **Thomas Jefferson** struck a chord for human liberty 200 years ago that resounds through the decades ... ...
http://www.nps.gov/thje/index.htm - 2004-11-02 10:21:25 - other versions - more from www.nps.gov - explain

Thomas Jefferson
... several years. The Most Important Thing He Ever Wrote Fire on the Capitol Good books about **Thomas Jefferson** Choose another Leader or Statesman **Thomas Jefferson** Fire on the Capitol "**Jefferson** and the Library of Congress" Jeffersoni Macaroni "**Thomas Jefferson** At Home" The Most Important Thing He Ever Wrote "The Declaration of Independence" Library**Thomas Jefferson** U.S. Presidents ...
http://www.americaslibrary.gov/cgi-bin/page.cgi/aa/jefferson - 2004-10-16 10:55:38 - other versions - explain

NOAA Ship Thomas Jefferson
... General Specifications Deck Equipment Electronic Equipment Engineering All Ship Specifications All Ship Specifications in PDF format (182 kb) **THOMAS JEFFERSON**'s E-mail address is: Noaa.Ship.**Thomas Jefferson**@noaa.gov **THOMAS JEFFERSON**'s Mail address is: NOAA Ship **THOMAS JEFFERSON** Marine Operations Center, Atlantic 439 York Street Norfolk, VA 23510-1145 **THOMAS JEFFERSON**'s Telephone Numbers • Return to Marine Operations Home Page • Inquires and Comments • URL: http://www.moc ... ...
http://www.moc.noaa.gov/tj/index.html - 2004-10-14 22:46:07 - other versions - explain

Thomas Jefferson (Library of Congress Exhibition)
... the revolution in individual rights in America and the world. Learn more about **Jefferson**: **Thomas Jefferson** Papers - **Jefferson** Time Line - Read More About It Online Survey - Exhibits Home - Library of Congress Home Library of Congress ... Declaration of Independence Establishing a Federal Republic The West - A Revolutionary World Legacy - **Jefferson**'s Library T his exhibition focuses on the extraordinary legacy of **Thomas** ...
http://www.loc.gov/exhibits/jefferson/ - 2004-11-02 15:36:46 - other versions - explain

Thomas Jefferson
Thomas Jefferson T H E W H I T E H O U S E **Thomas Jefferson** Help Site Map Text Only **Thomas Jefferson** Third President 1801-1809 [Martha Wayles Skelton **Jefferson**] Fast Fact: **Thomas Jefferson** gained the immense Louisiana Territory for the infant Republic. Biography: In the thick ... Kids | White House History White House Tours | Help | Text Only Privacy Statement 1789 - 1850 Washington - Taylor George Washington John Adams **Thomas** ...
http://clinton4.nara.gov/WH/glimpse/presidents/html/tj3.html - 2004-10-30 14:17:36 - other versions - explain

Thomas Jefferson
Thomas Jefferson Thomas Jefferson Picture of **Thomas Jefferson Thomas Jefferson** Third President 1801-1809 [Martha Wayles Skelton **Jefferson**] Fast Fact: **Thomas Jefferson** gained the immense Louisiana Territory for the infant Republic. Biography: In the thick ... WH/glimpse/presidents/html/presidents.html /WH/glimpse/presidents/html/jm4.html 1789 - 1850 Washington - Taylor George Washington John Adams **Thomas Jefferson** ...
http://clinton5.nara.gov/textonly/WH/glimpse/presidents/html/tj3.html - 2004-11-02 08:32:26 - other versions - explain

Biography of Thomas Jefferson
... Union Resources Historical Association Presidential Libraries Military Air Force One Camp David Marine One Home > History & Tours > Past Presidents > **Thomas Jefferson Thomas Jefferson** In the thick of party conflict in 1800, **Thomas Jefferson** wrote in a private letter, "I have sworn upon the altar of God ... Biography of **Thomas** ...
http://www.whitehouse.gov/history/presidents/tj3.html - 2004-11-08 13:59:57 - other versions - explain

next page

nutch powered by

RSS

Running WAC Search: Indexing Stats

- Indexing Machine Profile
  - Single processor 2.80GHz Pentium 4s with 1GB of RAM and 4x400GB IDE disks running Debian GNU/Linux.
  - Indexing, CPU-bound with light I/O loading.
  - RAM sufficient (no swapping).
  - All source ARC data NFS mounted.
- Only documents of type text/* or application/* and HTTP status code 200 were indexed.

Indexing Stats: Small Collection

- Collection
  - Three crawls.
  - Indexing steps run in series on one machine using single disk.
  - 206 100MB ARC files, 37.2GB of uncompressed data.
  - 1.07 million documents indexed.
- Indexing
  - 40.3 hours to complete.
    - 1/3rd segmenting/parsing, 1/3rd indexing, 1/3rd all other steps.
  - Merged index size was 3% the size of src.
    - Index plus cleaned-up segments occupied 16% src.
    - Index plus uncleaned segments made up 40% src.

Indexing Stats: Medium Collection

- Collection
  - 1054 ARCs, 147.2GB of uncompressed data.
  - 4.1 million documents indexed.
  - Two machines to do the segmenting step.
  - Subsequent steps all run in series on a single machine using a single disk.
- Indexing
  - 99 hours of processing time
    - Or 86.4 hours of elapsed time because segmenting split.
  - Merged index size was 5.2GB, 4% source.
  - Index plus cleaned-up segment data 13.5% source.
  - Index plus uncleaned segments 22% source.

Observations 1

- Indexing big collections is a long-running manual process.
  - Requires manual intervention at each step moving process along.
  - Attention compounds the more distributed the indexing.
  - An early indexing of 85 million took approx. a week over 4 machines.
    - Steps restarted as disks overfilled.
    - Little science applied so load suboptimally distributed.
    - Synchronizations waiting on laggard processes.
  - Current toolset, vigilant operator, a week of time, 4 to 5 machines with lots of disk, indexing 100 million doc. WACs is practical limit (200 million documents...perhaps...if segments and indices).
- Automated means of efficiently distributing indexing needed!
  - But some indexing steps are currently single process.
  - And as collection grows, so grows central Nutch DB of page and link content. Eventually larger than any available single disk.

Observations 2

- Inclusion of in-link anchor-text indexing improves search result quality.
  - Without, results rich in query terms but *wrong*.
- Distributed Nutch query clustering works well.
  - At least for low rates of access: ¼ query per second.
    - Search access-rates are expected lower for WACs than live-Web search.
  - But caches to speed querying will turn problematic.
    - Nutch (Lucene) query implementation uses one byte per document per field indexed.
    - Additions made to support query-time deduplication and sorting share cache of each search result's document URL. Such a cache of (Java) UTF-16 Java strings gets large fast.
- Collections of a few million plus hosted on single disk/single machine, show distinct lag drawing search results.
  - Distribute segments over disks and machines.
- Robust, performant, (Java) doc-type parsers are sparse.
  - Especially for proprietary types: application/pdf, msword, etc.
  - External call out to XPDF for application/pdf (Still slow relative text/*).

Future 1

- Nutch project moving onto distributed file system.
  - Nutch Distributed File System (NDFS).
    - "...software for storing very large stream-oriented files over a set of commodity computers. Files are replicated across machines for safety, and load is balanced fairly across the machine set"
    - Java implementation of subset of Google File System (GFS).
    - Nutch DB distributed, segments distributed.
- How to evenly distribute work across cluster?
  - MapReduce!
    - "a platform on which to build scalable computing"
    - Another Google innovation.
    - Cast cluster task in MapReduce mold -- think Python *map* function followed by *reduce* function -- then the MapReduce platform will manage distribution of task across cluster in fault-tolerant way.
    - Java implementation of Google MapReduce underway in Nutch project (Doug Cutting, Mike Cafarella).

Future 2

- Viewer applications
  - NWA WERA viewer using Nutchwax engine (October 2005).
  - Open source Wayback Machine.
- IA moving Web collections to Petabox platform.
  - Racks of low power, high storage density, inexpensive, rack-mounted, “shipping container friendly” computers.
    - Open source design.
    - Previously, racks of commodity PCs.

End

- Future WAC Search development.
  - Immediate: Nutchwax 0.4 release based on Nutch 0.7 coordinated with release of the WERA viewer (Lots of bug fixes, October 2005).
  - Beyond: Scale beyond 100 million limit to WACs of 1 Billion and beyond.
    - Exploitation of Nutch project NDFS/MapReduce platform atop Petabox.

Thank you