‘20th Century Find’ Data Storage & Processing via S3 & EC2

Kris Carpenter Negulescu, Director

The Internet Archive, Web Group

Agenda

- Introductions
- Case Study: 20th Century Find
- AWS: Lessons & Recommendations
- Q & A

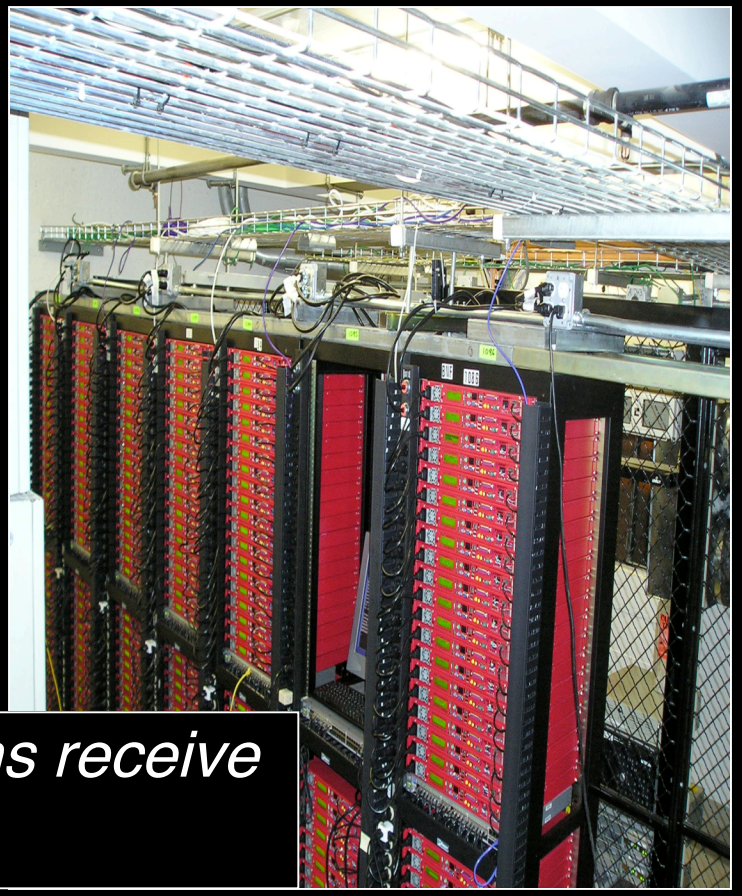

Headquarters in Presidio
2km from Golden Gate Bridge
(DR in downtown SF)

INTERNET ARCHIVE

What Is the Internet Archive?

A digital library of over 3.5 petabytes of information

- Web Pages (100+bil URIs/~1.9 PBs)
- Educational Courseware
- Films & Videos
- Music & Spoken Word
- Books & Texts
- Software
- Images



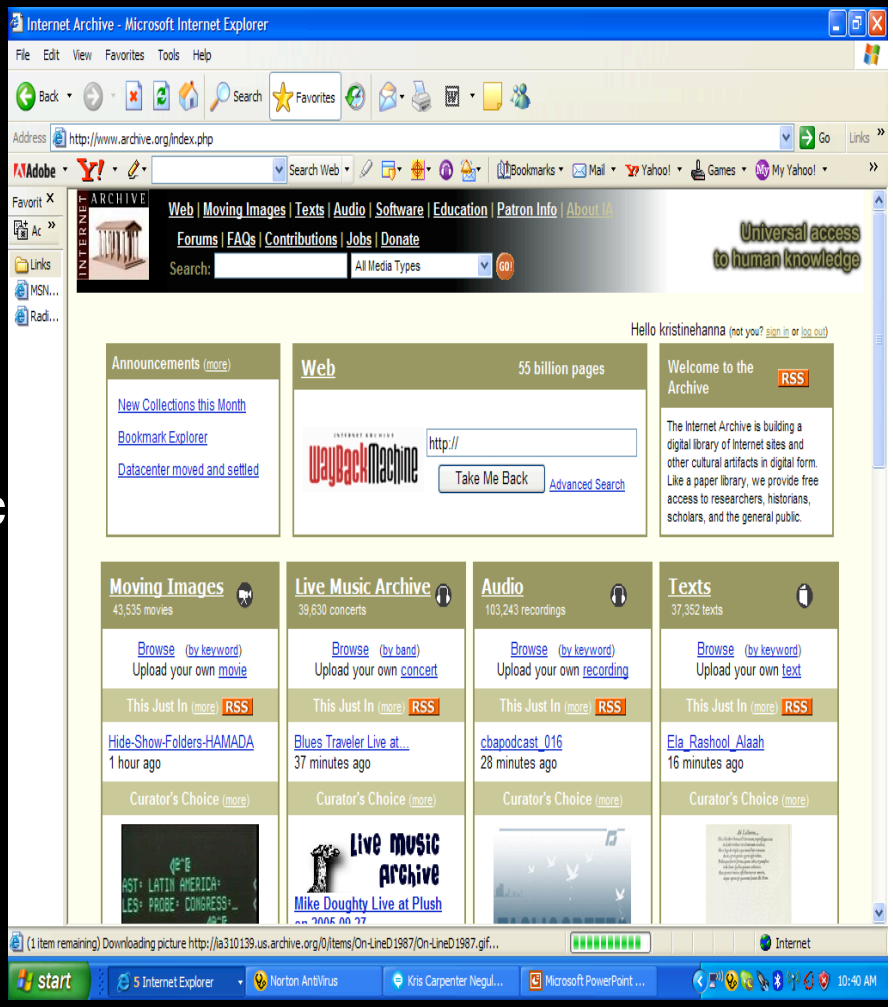
The archive's combined collections receive ~1.5 mil downloads a day

INTERNET ARCHIVE

How Content is Collected?

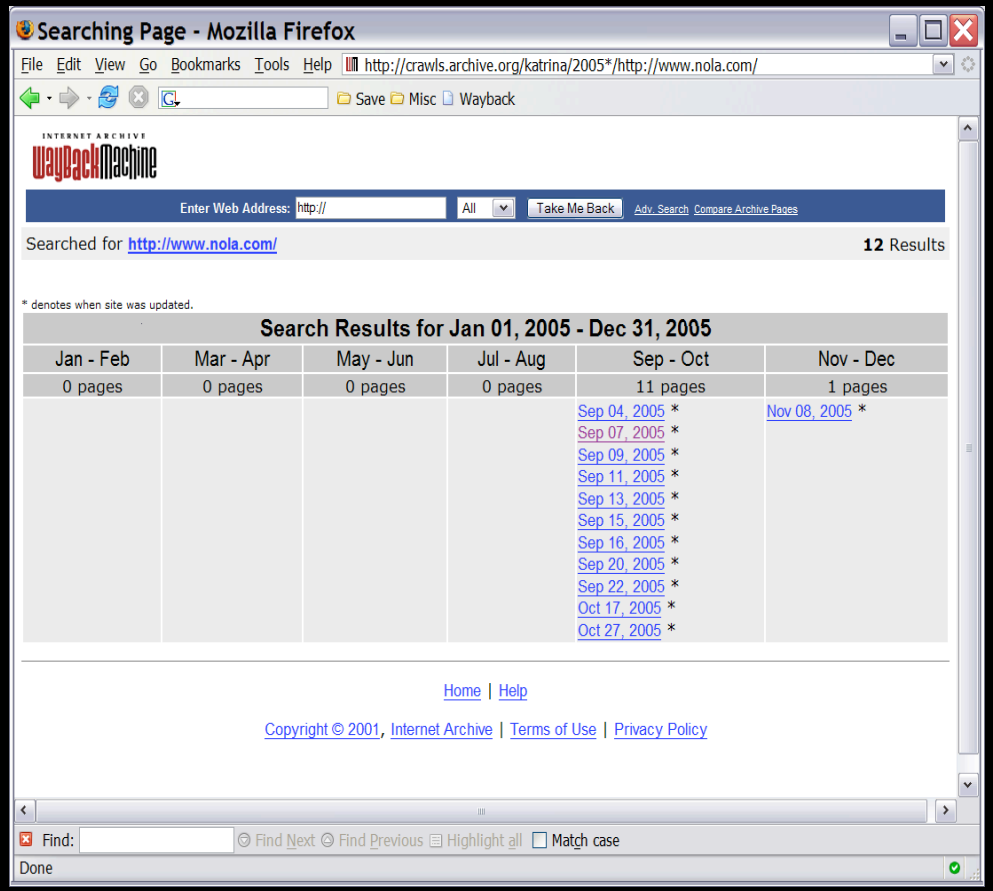
- Ongoing harvests of www by Alexa Internet
- Web harvests by the Internet Archive's Web Group
- Scanning & digitization of public domain texts, stills, moving images, audio clips, etc
- Contributions from the community.

Entire collection accessible for free to the public at www.archive.org



INTERNET ARCHIVE

Public Access to Collections



The Wayback Machine

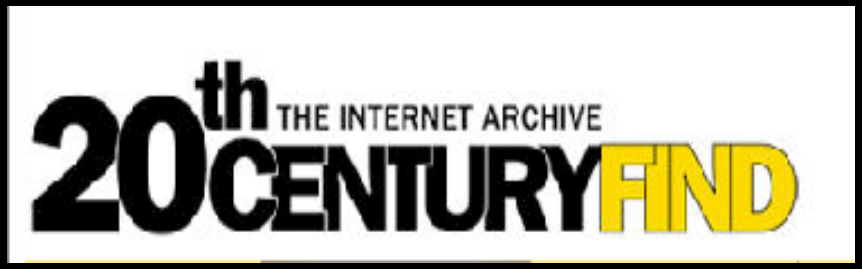
- Applied to web.archive.org
  - ~10 million hits per day
  - Avg of 100-200 hits per second
  - ~100K lookups-by-URL (ie, display dated list of captures) per day
  - ~4 million retrieval requests (URL+exact date) per day (>50/second)

Full-text-search

- Lucene/SOLR for books, audio, video, images, software, etc.
- Not yet available for the web archive

The “20th Century Find”

A Case Study in Progress



Sponsor: Library of Congress/NDIIPP

Contributors: Yahoo! Open Source Team

Project Goals:

- To identify and preserve Web content harvested from 1996-2000 (“The 20th Century Web”)
- To make this historic collection more accessible to the public by adding full-text search.

No web search of this scale has ever been built using open-source tools, nor over a public archive of Web content.

20th Century Find Data

~4.8 billion unique Web captures and ~22TBs data archived at IA and replicated to the Bibliotheca Alexandrina, Egypt and Amazon S3.

	HTML	ALL	
1996	20,475,783	32,168,637	
1997	214,802,568	318,294,397	
1998	195,422,208	245,919,083	
1999	757,339,164	847,540,628	
2000	(2.5B (est))	3,328,040,041	
	==========	==========	
	(est) 3,700,000,000	4,771,962,786	~22TBs

Tools & Services In Use

NutchWAX: tools for full text search of archival web content.

http://archive-access.sourceforge.net/projects/nutch/

NutchWax is an extension of other open source projects:

- **Nutch (v0.9):** open source web search engine software

  http://lucene.apache.org/nutch/

- **Hadoop (v0.10):** a framework for running applications on large clusters of commodity hardware.

  http://lucene.apache.org/hadoop/

Tools & Services In Use

Amazon S3: Amazon S3 provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web.

- Simple HTTP operations for all actions
- Lots of open source code available in many languages to support projects
- Arbitrarily large associative array
  - Each array is called a bucket
  - Text keys
  - Binary values up to 5 GBs in size
  - Iterate in sorted order

Tools & Services In Use

Amazon EC2 (beta): Amazon Elastic Compute Cloud (EC2) is a web service that provides resizable compute capacity you specify and utilize for whatever period of time you need it.

- Uniform hardware profile
- Customizable software images
- Simple command line tools for provisioning and network configuration
- Incredibly rapid provisioning and release of resources
- Flexible network primitives

EC2 Node Options

- Default (small) instance ($0.10/cpu hr)
  - 32 bit platform, moderate IO performance
  - Roughly equivalent to 1.7Ghz Xeon CPU, 1.75GB of RAM, 160GB of local disk, & 250Mb/s of network bandwidth
  - In use for "20th Century Find"
- Large instance ($0.40/cpu hr)
- Extra Large instance ($0.80/cpu hr)

‘20th Century Find’ Indexes

Indexing began in October 2006

1996: indexed via a cluster of 20 EC2 nodes in ~36 hrs.

1997: indexed via a cluster of 100+ EC2 nodes

1998: indexed via a cluster of 300+ EC2 nodes

1999 (which is 3x the size of any earlier year) was attempted in Sept 07 using a cluster of ~270 EC2 nodes. We expected indexing to complete in days but project was halted due to lack of consistent cpu/IO across nodes.

‘20th Century Find’ Service

Alpha launched in April 2007

http://20thcf.archive.org:8080/

- 1996-1998 indexed using NutchWAX, Hadoop
- Hosted on AWS EC2.
  - A front-end machine distributes the query across 21 backend nodes each of which is carrying a piece of the index.
- Deployed index is 1.35TB in size
  - No compression was applied to this experimental alpha
  - 95 559 ARCs, ~600 mil documents

Needed Enhancements

- Improved performance and stability of Nutch
- Enhanced application deployment architecture
  - E.g. index compression, load balancing, efficient exclusion management architecture, etc.
- Ability to handle multiple instances of a page
- Improved ranking of results
- Mechanisms for applying link based weighting w/dimension of time
- UI that exploits the dimension of time and distinguishes the experience from live web searches

A public beta of 20th Century Find is targeted for later this year

Why Amazon Web Services?

- "Convenient dedicated peak capacity" for big index jobs
- 'Pay as We Go' based on what we need/use
  - Helps avoid unnecessary hardware acquisition
- Simple to provision storage and compute nodes
- Committed to support us even though only in "beta"
- Ideal for indexing Web pages & for providing redundant, offsite storage & reliable hosting for the alpha service
- Great platform for experimentation, iteration
- Geographically disperse from IA DR

A Cost Effective* Solution

We budgeted a total investment of only $20k to:

- Index ~2 billion web captures
- Create mirrored, offsite backups of the original files and text indexes
- Store this data for a year

*True, if able to provision and go. Fees can add up fast if not vigilant.

AWS: Lessons to Date

October 2006 - present

What is Working Well

- Programmatic interfaces for both S3 & EC2
- Level of technical support and operational troubleshooting contributed by AWS teams
- Data reliability and availability (S3)
- Ease of data back-up and management (S3)
- Fee structure; Free traffic between S3 and EC2
- Speed of provisioning of capacity
- Uniformity of nodes (S3, not as true for EC2 as of late)

Challenges/Limitations - S3

October 2006 – June 2007

- Available bandwidth was varied, i.e. fairly unpredictable (both external and internal to EC2), perhaps under provisioned?
- No real specific guarantees for data preservation (no SLA, no transparency of choices/policies)
  - In absence of stated policies, must rely on history
- Some problems were due to popularity (we prefer this over under-funding, lack of strategic priority, or pure incompetence)

Challenges/Limitations - S3

Fall 2007

- Available bandwidth has been consistent, i.e. fairly predictable (both external and internal to EC2). E.g. ~4 hrs to move 7.5 TBs of data into EC2 for indexing. Last winter it took weeks.
- There are still no real specific guarantees for data preservation (no SLA, no transparency of choices/policies)

Challenges/Limitations EC2 Beta

October 2006 - June 2007

- Last winter over 55% of EC2 machines allocated to an indexing job crashed when using our image at scale (50+ nodes)
  - Started out better than our own internal rate of failure but was higher during peak use
- Location of S3 nodes relative to EC2 instances was a significant factor for large scale data processing
- Often took too long to assemble and remove files from EC2 (internal from S3 and external)

Challenges/Limitations EC2 Beta

July 2007 - present

- Experienced delays in provisioning new nodes due to AWS infrastructure changes and spikes in demand in late summer
- EC2 machines allocated to indexing jobs performed well with our image at scale (250+ nodes) through early Fall
- No longer takes too long to assemble and remove files from EC2 (internal from S3 and external)
- More recently hit IO and cpu constraints on basic nodes

Consider Using AWS When…

S3

- Need cost-effective secondary or tertiary back-up for primary data
- Need to have a multi-provider preservation plan with geographically disperse storage solutions

EC2 beta

- Have extreme variation in compute capacity needs (i.e. spikes in demand broken by relatively long stretches of minimal usage requirements)
- Can allocate resources to R&D

Consider Using AWS When…

Or in general…

- Have limited capital for purchasing hardware and/or services and/or available capital is spread out over a period of many months/years
- Need to get up and running quickly
- Want/Need to experiment with multiple solutions/partners

New Initiatives: Crawling

- Planning to experiment with Heritrix/AWS
  - Beginning Jan ‘08
- Members of the Heritrix user community already actively crawling via AWS
  - E.g. Page-Store.com (Paul Pedersen, Principal)
  - Running 80-90 EC2 instances for web wide, site based crawling and post processing of harvested data
  - Storing data, post processing, in S3
  - Avg Cost/month: ~$10k

Thank You!

Kris Carpenter Negulescu, Director
The Internet Archive, Web Group
kcarpenter@archive.org