Cost of a WARC
Analyzing Web Archives in the Cloud

Ryan Deschamps, Samantha Fritz, Jimmy Lin, **Ian Milligan**, and Nick Ruest

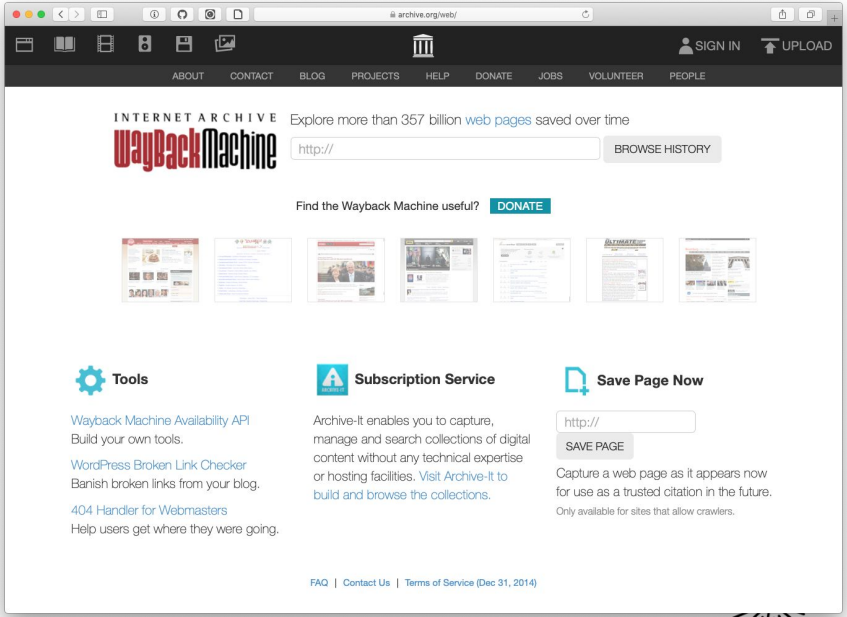
JCDL 2019 - Urbana-Champaign, IL

Why do we care about web archives?

Born-digital sources have the potential to reshape research in the humanities and social sciences;

Research access has lagged (beyond Wayback Machine, analysis ecosystem is mostly command-line-based tools)

As we plan for research access, we need to understand the economics associated with providing this sort of access



How much does it cost to analyze a WARC (the standard container file format of web archives) in the cloud?

US$7 per TB

The TL;DR

What do we mean by the “Cloud”?

We conduct our work on the **Compute Canada Cloud**, which is an **OpenStack** instance supported by a research grant.

As OpenStack is a popular open-source cloud platform, our findings should be generalizable.

We translated all of our compute time into Amazon Web Services costs as it is the most popular commercial provider.

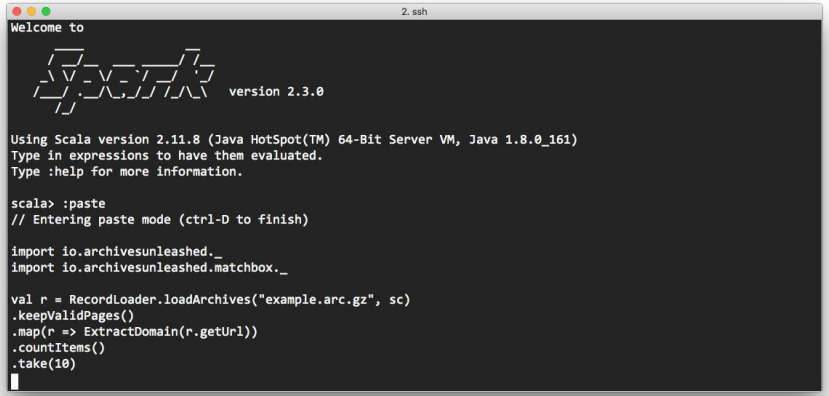
compute canada | calcul canada

aws

What are we performing “analysis” with?

Analysis using the **Archives Unleashed Toolkit** or **AUT**

AUT is a Scala domain-specific language on top of the Apache Spark platform

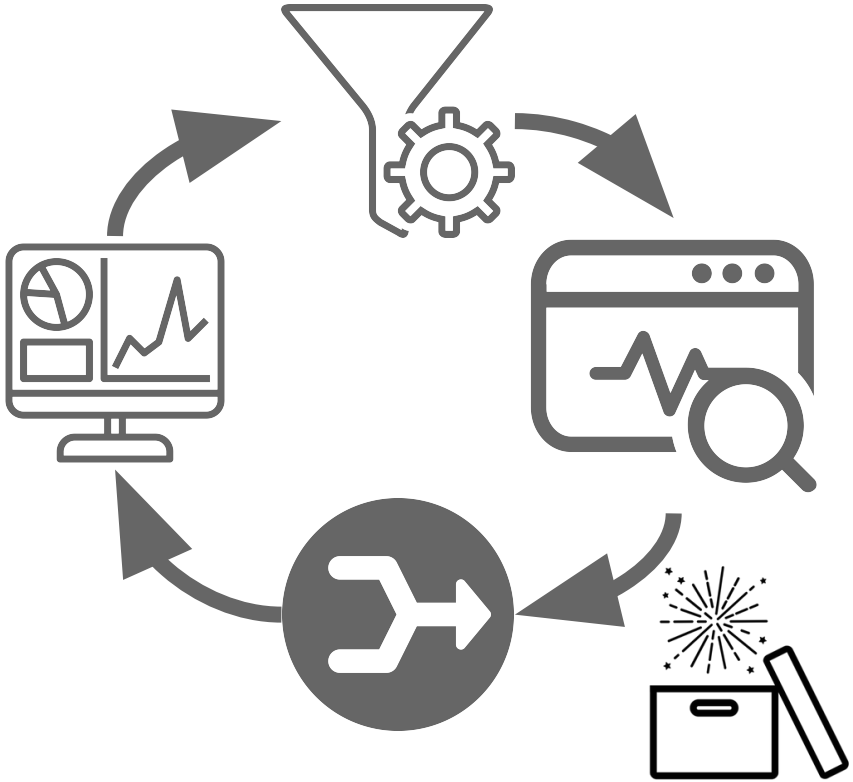


What do we mean by “Analysis”?

The **F**ilter - **A**nalyze - **A**ggregate - **V**isualize (FAAV) Cycle

Common analytics task: crawl statistics to visualizing web graphs to exploring text at scale

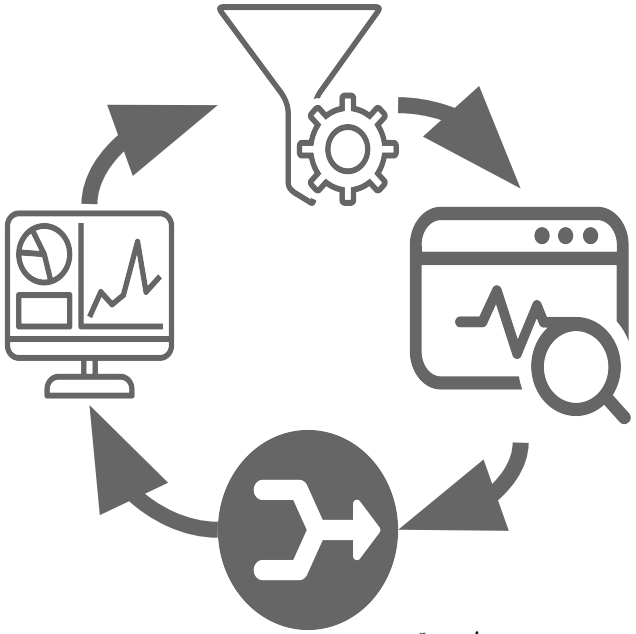
Informed by extensive hands-on collaboration



2. ssh

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.0
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_161)
Type in expressions to have them evaluated.
Type :help for more information.

scala> :paste
// Entering paste mode (ctrl-D to finish)

import io.archivesunleashed._
import io.archivesunleashed.matchbox._

val r = RecordLoader.loadArchives("example.arc.gz", sc)
.keepValidPages()
.map(r => ExtractDomain(r.getUrl))
.countItems()
.take(10)
```

+



ANALYSIS AT SCALE

What do we mean by “Analysis”?

Extract all URLs to compute the frequency of domains appearing in a given collection (domain distribution);

Extract all plain text from all pages, along with metadata such as crawl date, domain name, and URL (full text); and

Extract all hyperlinks to create a domain-to-domain network graph (webgraph);

On to the experiment!

The Experiment

We decided to use a **16 core, 64GB memory virtual machine**

Powerful, but struck the balance between expensive and power

Why not a cluster?

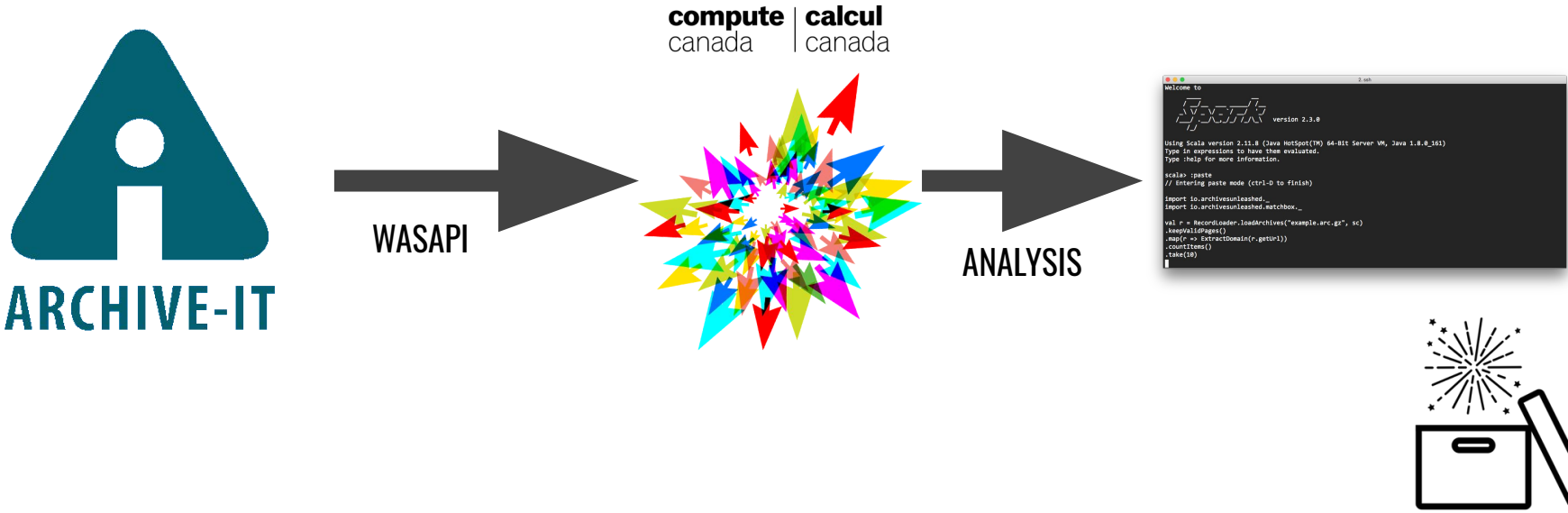
The Experiment

Analysis based on analyzing the cost of processing **48 Archive-It collections** from six Canadian universities (Toronto, Victoria, Simon Fraser University, Manitoba, Dalhousie, and Winnipeg).

A variety of **sizes** – smallest at 1.2GB was Victoria's academic calendar; largest at 4.3TB was Canadian Government Information Collection

Size	Count
$\geq$ 1 GB, $<$ 10 GB	10
$\geq$ 10 GB, $<$ 100 GB	18
$\geq$ 100 GB, $<$ 1 TB	15
$\geq$ 1 TB	5
Total	48

The Experiment (Workflow)



Findings

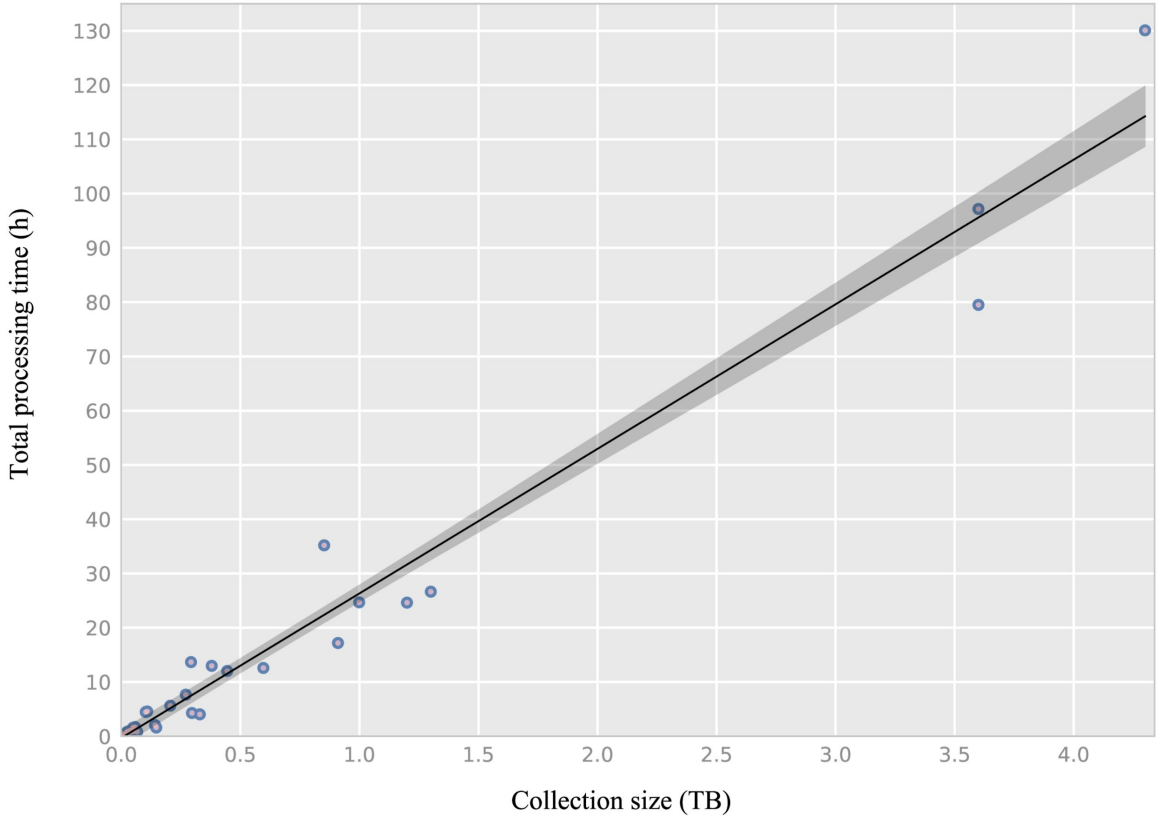
We then took all the times for each job (**Domain**, **Full Text**, **Webgraph**) and found processing time per GB in seconds.

Webgraph is most computationally intensive, but not too much so.

Processing times drop as size increases, as startup costs are amortized.

Derivative	all	L	M	S
domain distribution	32	25	27	36
full text	34	28	35	34
webgraph	36	34	36	36
total	102	87	98	106

Figure: Processing times per GB in seconds



Scatter plot between collection size and total processing time, **illustrating a linear relationship**

Findings

Derivative files are **much smaller**

Researcher can usually work with these derivative files on their own systems in a way they could not work with their WARCs

Derivative	all	L	M	S
domain distribution (KB)	0.95	0.51	0.98	1.01
full text (MB)	78.5	97.6	102.1	62.4
webgraph (KB)	76.9	85.8	122.6	50.9

Figure: Derivative sizes per GB

So we know the times to compute these derivatives.. Show me the money!

Derivative	all	L	M	S
domain distribution	$6.51	$4.67	$5.05	$7.63
full text	$6.73	$5.24	$6.65	$7.04
webgraph	$7.19	$6.46	$6.82	$7.52
total	$20.43	$16.37	$18.52	$22.19

Processing cost per TB in US $

Cost of a WARC

C5.4xlarge (16 core, 68 GB memory) is $0.68/hour in US East (Ohio)

The previous results show a **macro-average**

The bottom line: US$7/TB for a typical analytics operation such as generating **domain frequency** reports, extracting **full text of a collection**, or extracting the link-to-link **webgraph** of hyperlinks.

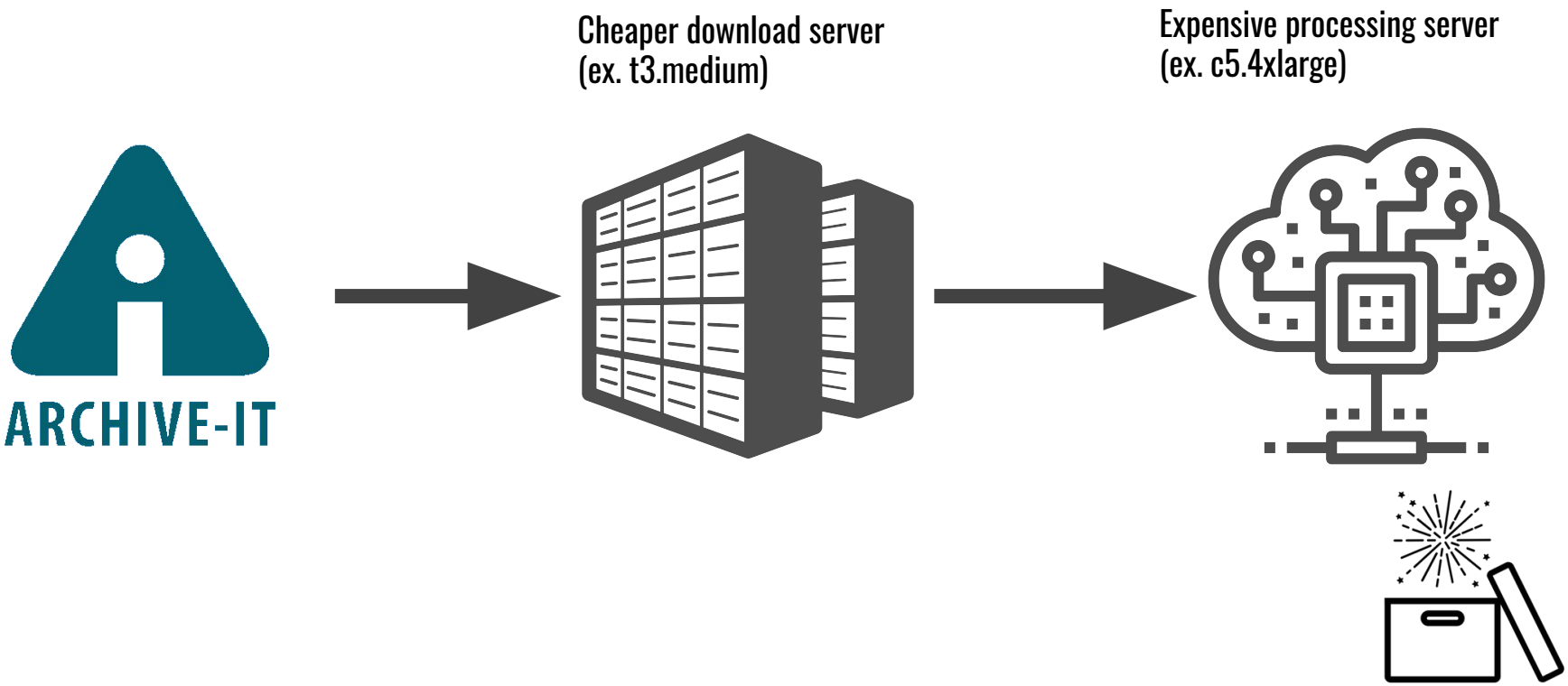
Cost of a WARC

This is **cost-competitive**

Google BigQuery costs US$5 per TB – *but* is SQL based and prices on uncompressed size whereas our calculations were on compressed WARCs (which are roughly 60% the size of uncompressed WARCs)

Archives Unleashed is price competitive with commercial services, albeit without any profit margin.

Proposed Workflow

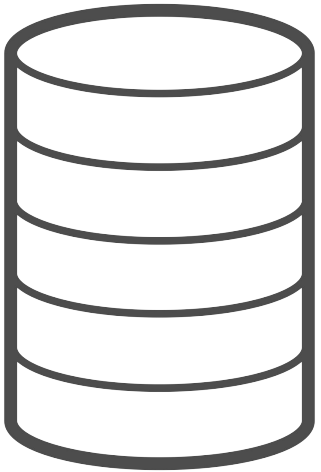


Limitations: Storage

We did not include **storage** in this discussion. 1TB of data costs US$23 per month. Our preferred workflow would be to transfer WARCs, analyze, and then delete them quickly.

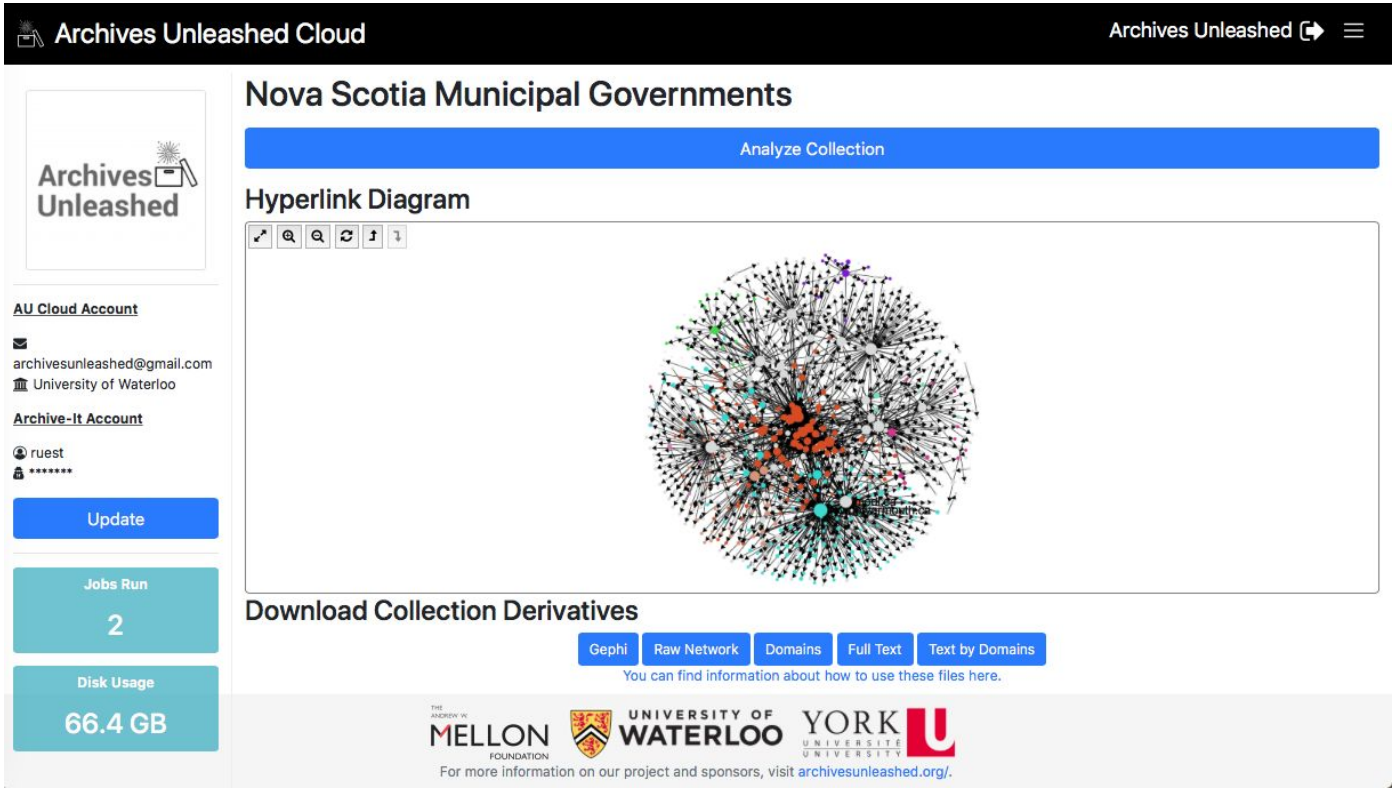
At 30 MB/s data transfer speed, transferring a TB costs US$0.40; less than the per-day cost of S3 data storage

As long as the preservation copy is secure, the “processing copy” can be created and deleted on a whim



Conclusions

Conclusions



We share the beginnings of an economic analysis and believe the costs to be quite affordable; whether institutions or individual scholars find these costs palatable remains to be seen.

US$7 per TB

The TL;DR

Thanks to our supporters!

THE ANDREW W. MELLON FOUNDATION

START SMART LABS

YORK UNIVERSITÉ UNIVERSITY

compute canada | calcul canada

UNIVERSITY OF WATERLOO

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada