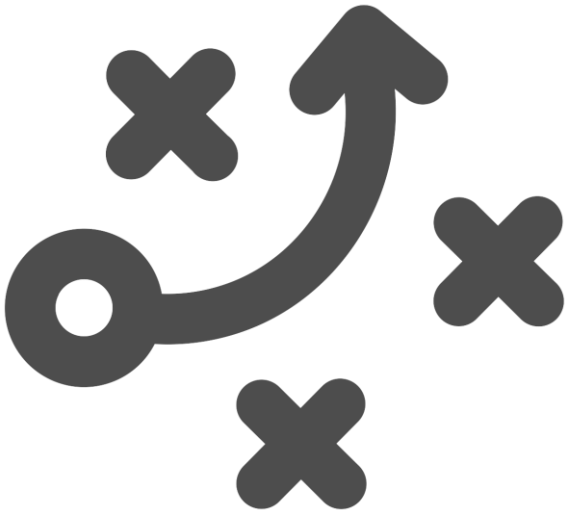
Project Sustainability and Research Platforms:

The Archives Unleashed Project

Nick Ruest (York University)
Ian Milligan (University of Waterloo)

IIPC WAC 2019 - Zagreb

Plan for The Talk



- Introduction
- Background
- What we do?
- How much does it cost?
- Discussion

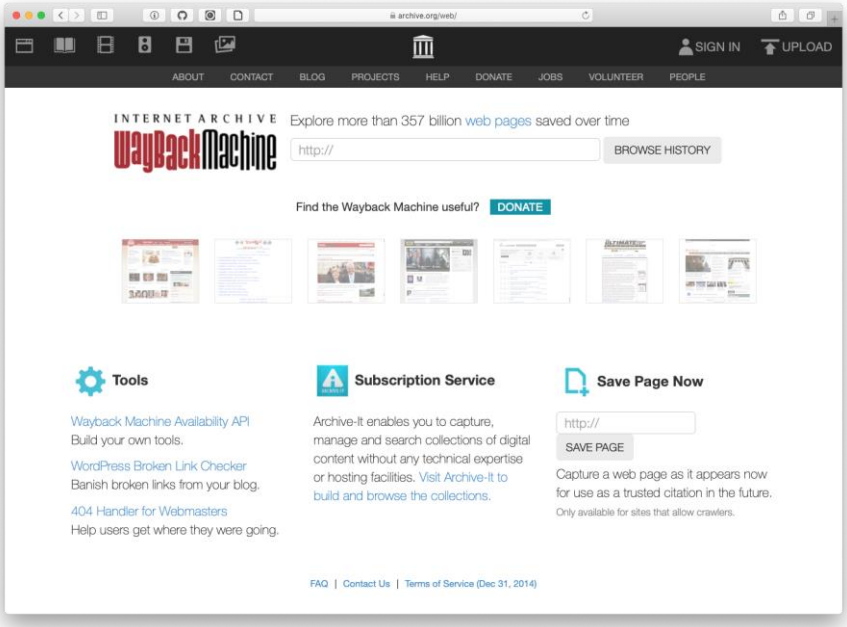
Background

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



Why do we care about web archives?

What do we do?

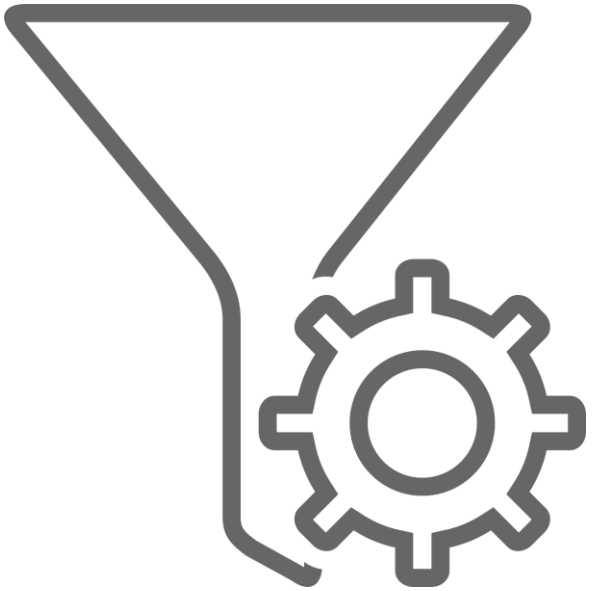
Archives Unleashed Toolkit

- An open-source platform for analyzing web archives with Apache Spark;
- Scalable
  - Can work on a powerful cluster
  - Can work on a single-node server
  - Can work on a laptop (on MacOS, Linux, or on Windows with a Linux VM)
  - Can work on a Raspberry Pi for all your personal web archiving analysis needs 😂

Using the Toolkit is based on the **F**ilter-**A**nalyze-**A**ggregate-**V**isualize (FAAV) Cycle

Filter

- Filter down content
  - Focus on a particular range of crawl dates;
  - Focus on a particular domain;
  - Content-based filter (“global warming”) or those who link to a given site
- Can be nested - i.e. pages from 2012 from liberal.ca that link to conservative.ca and contain the phrase “Keystone XL”



Analyze

- After filtering, want to perform analysis – extracting information of interest.
- Such as:
  - Links and associated anchor text?
  - Tagging or extracting named entities?
  - Sentiment analysis.
  - Topic modeling.

Aggregate

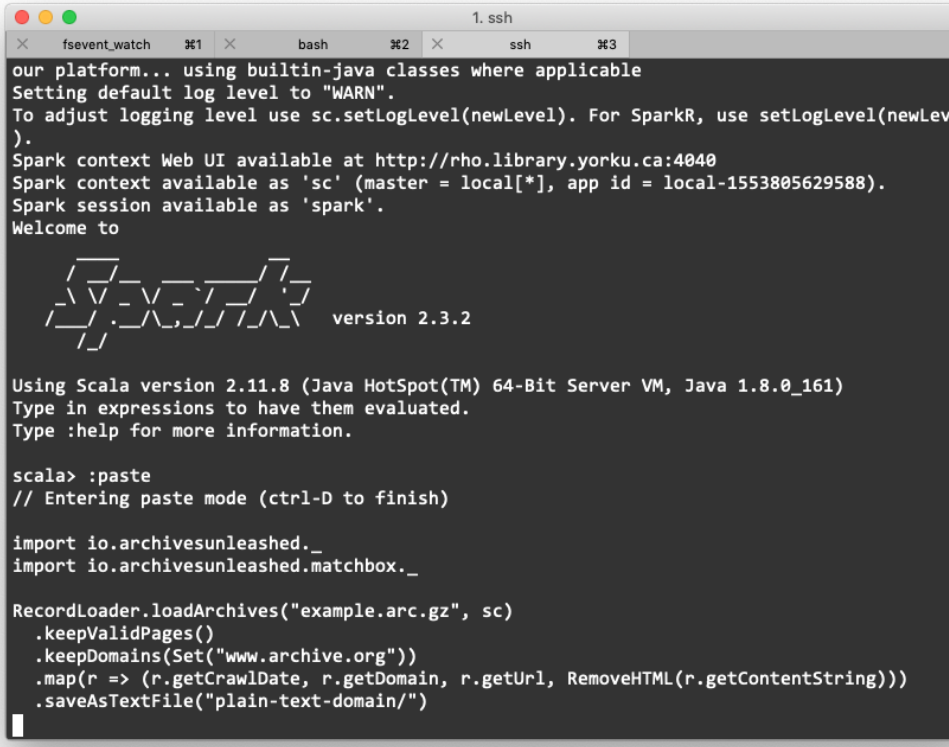
- Summarize the output of the analysis from the previous step.
  - Counting
    - How many times is Jack Layton or Barack Obama mentioned?
    - How many links are there from one domain to another?
- Finding maximum (page with most incoming links?)
- Average (average sentiment about “Barack Obama” or “Donald Trump”)

Visualize

- Output data as a visualization
  - Tables of results
  - External applications (i.e. GEXF files for Gephi)

Great!
So why doesn’t everybody use the Toolkit?!?!

Our Cutting Edge Interface



In other words...

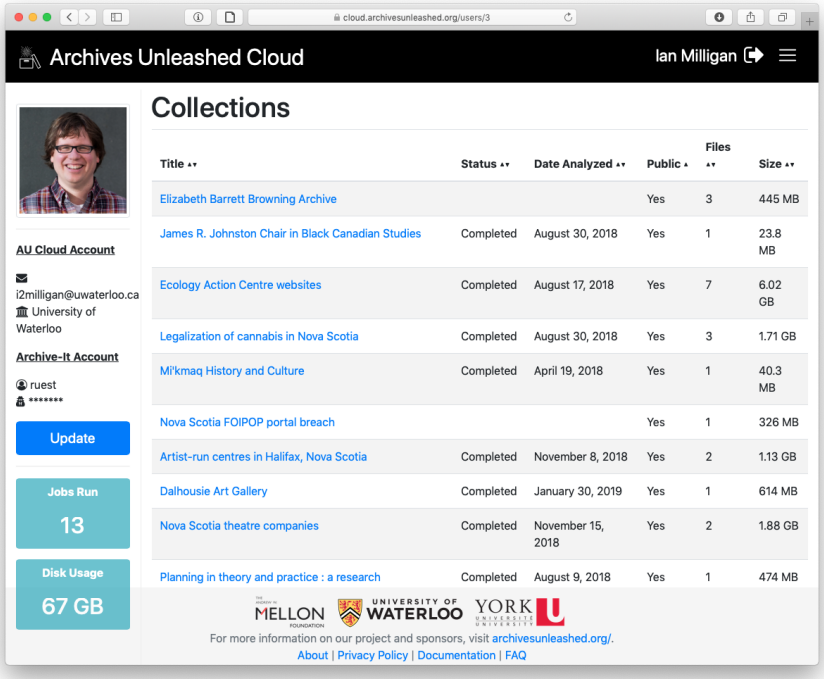
We have a wonderful platform that takes WARC files and converts them into formats that are familiar to digital humanists, computational social scientists, systems librarians, digital archivists, and beyond..

.. but you basically need to be a developer to run the simplest of commands (despite ample documentation and outreach... the command line interface is a bridge too far).

Enter the Archives Unleashed Cloud

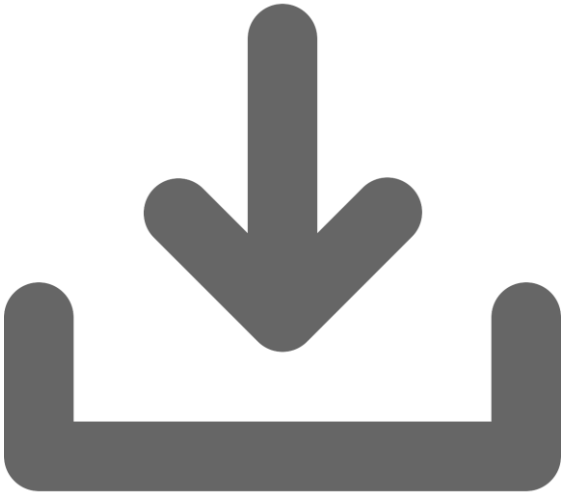
Archives Unleashed Cloud

- A web-based front end for working with the Archives Unleashed Toolkit;
- Runs on our central servers or you can run one yourself;
- Uses WASAPI – Web Archives Systems API – to transfer data
- Generates a basic set of research derivatives for scholars to work with

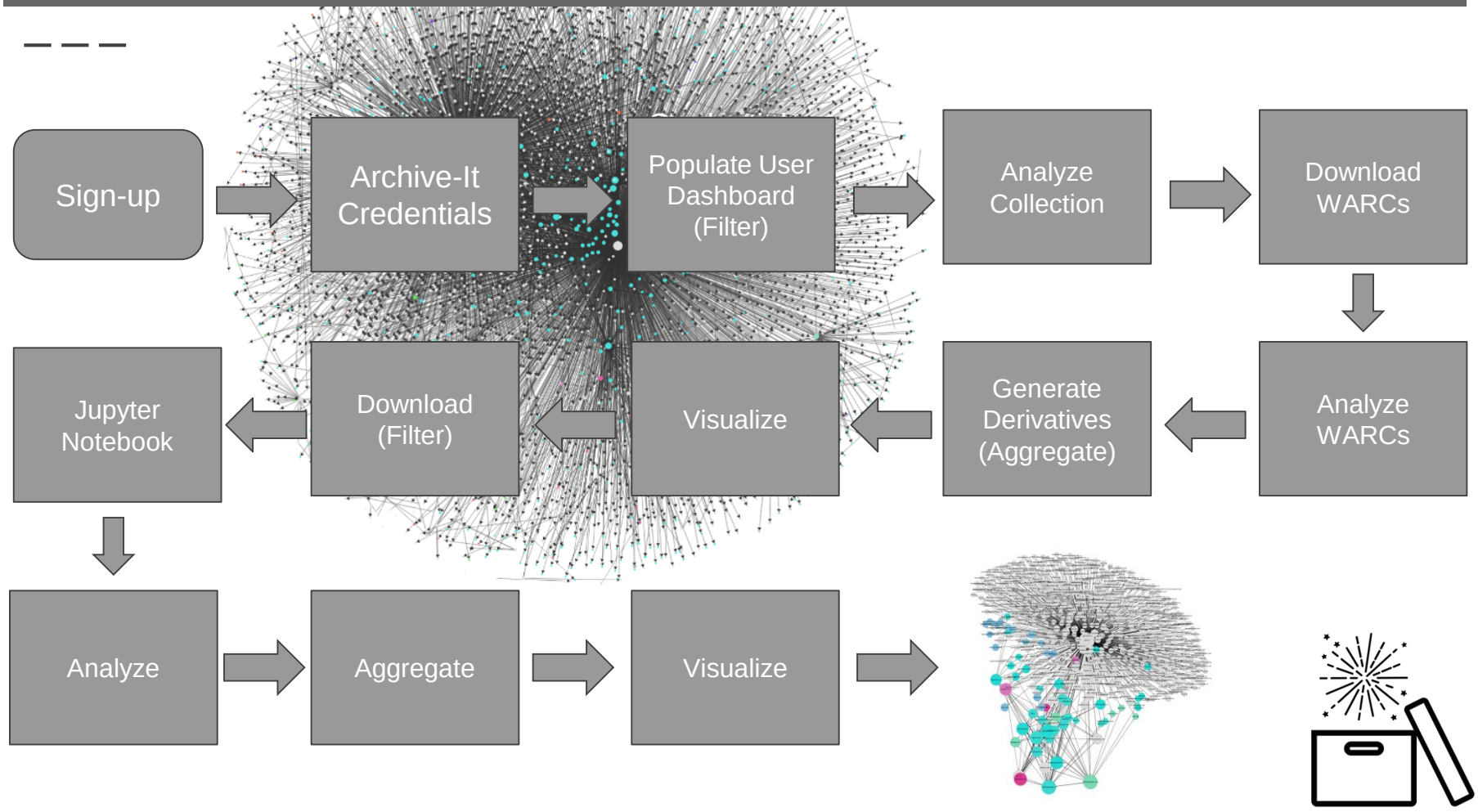


Archives Unleashed Cloud

- Download options for each collection
  - Full text of a web archive;
  - Full text of the top-ten most popular domains in a web archive;
  - Network diagram with characteristics pre-computed (Gephi);
  - Raw network diagram (origin/destination/weight);
  - Domain frequency statistics



How it works


Sign-up
Archive-It Credentials
Populate User Dashboard (Filter)
Analyze Collection
Download WARCs
Analyze WARCs
Generate Derivatives (Aggregate)
Visualize
Download (Filter)
Jupyter Notebook
Analyze
Aggregate
Visualize

Archives Unleashed Cloud


COOL.

Cool.
How much does it cost?
(...to process WARC files in the “CLOUD”)

US$7 per TB

The TL;DR

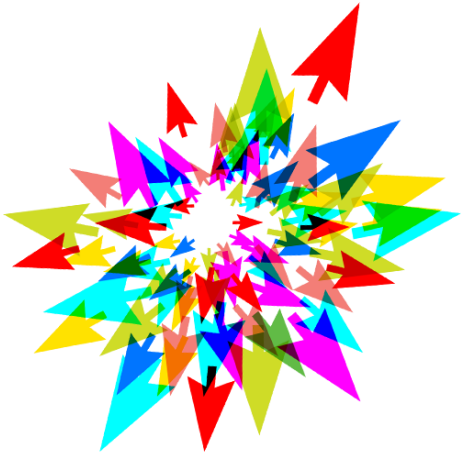
What do we mean by the “Cloud”?

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

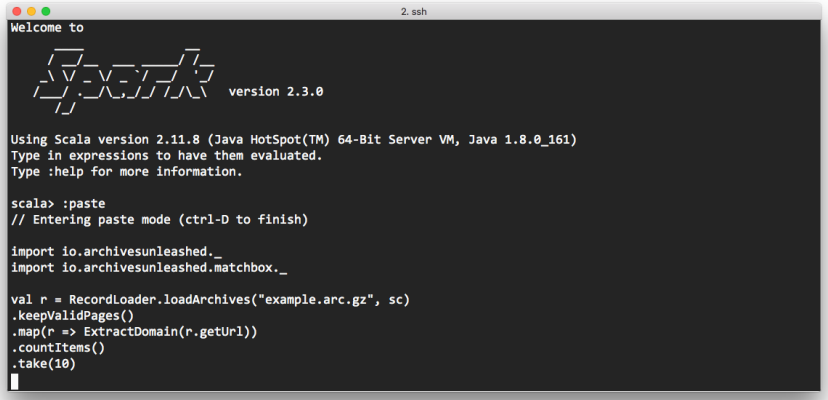


aws

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

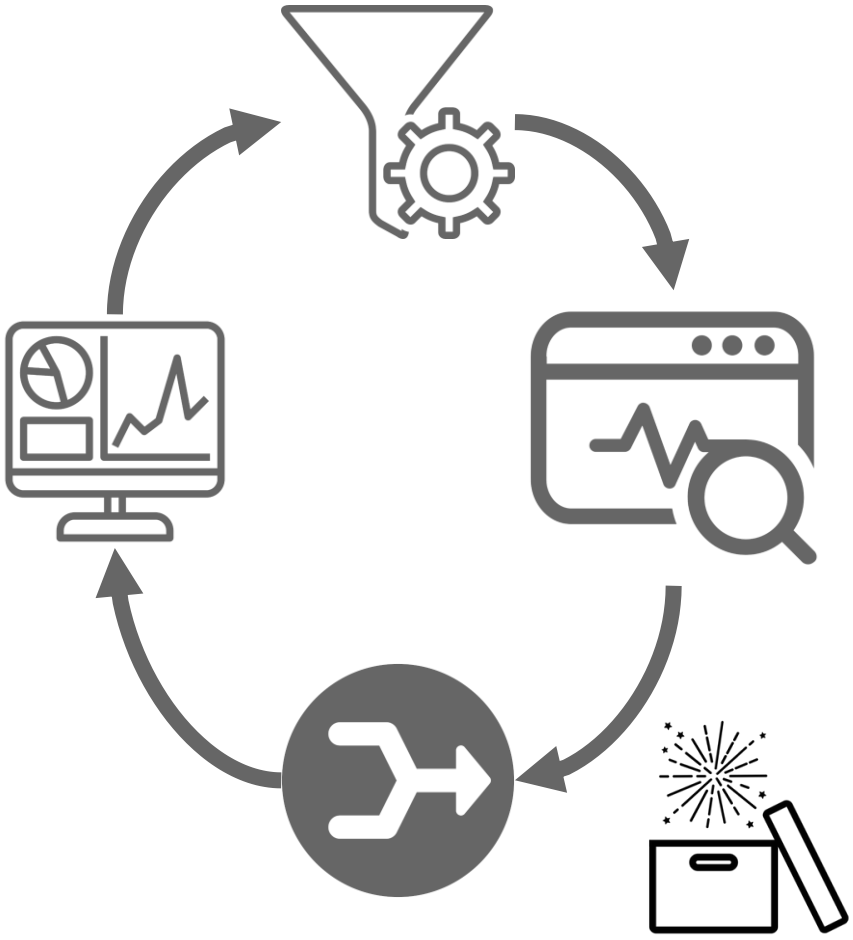


What do we mean by “Analysis”?

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



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);

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?

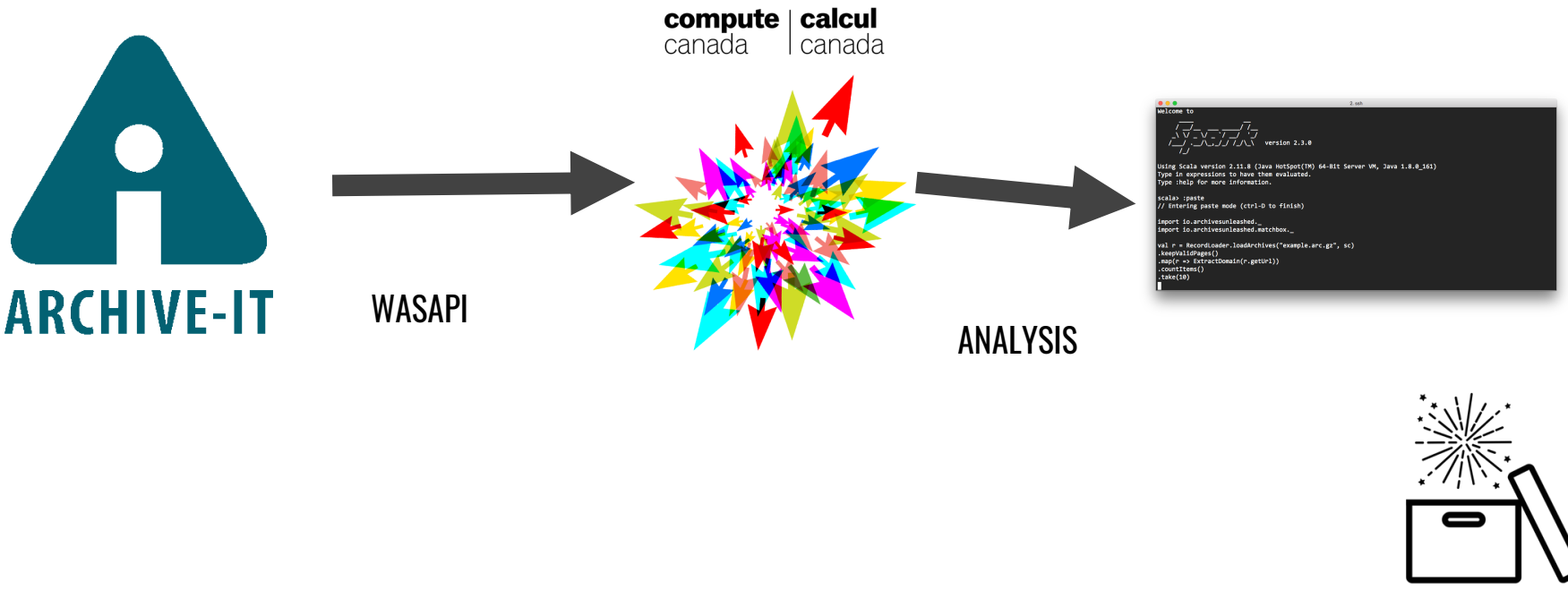
The Experiment

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

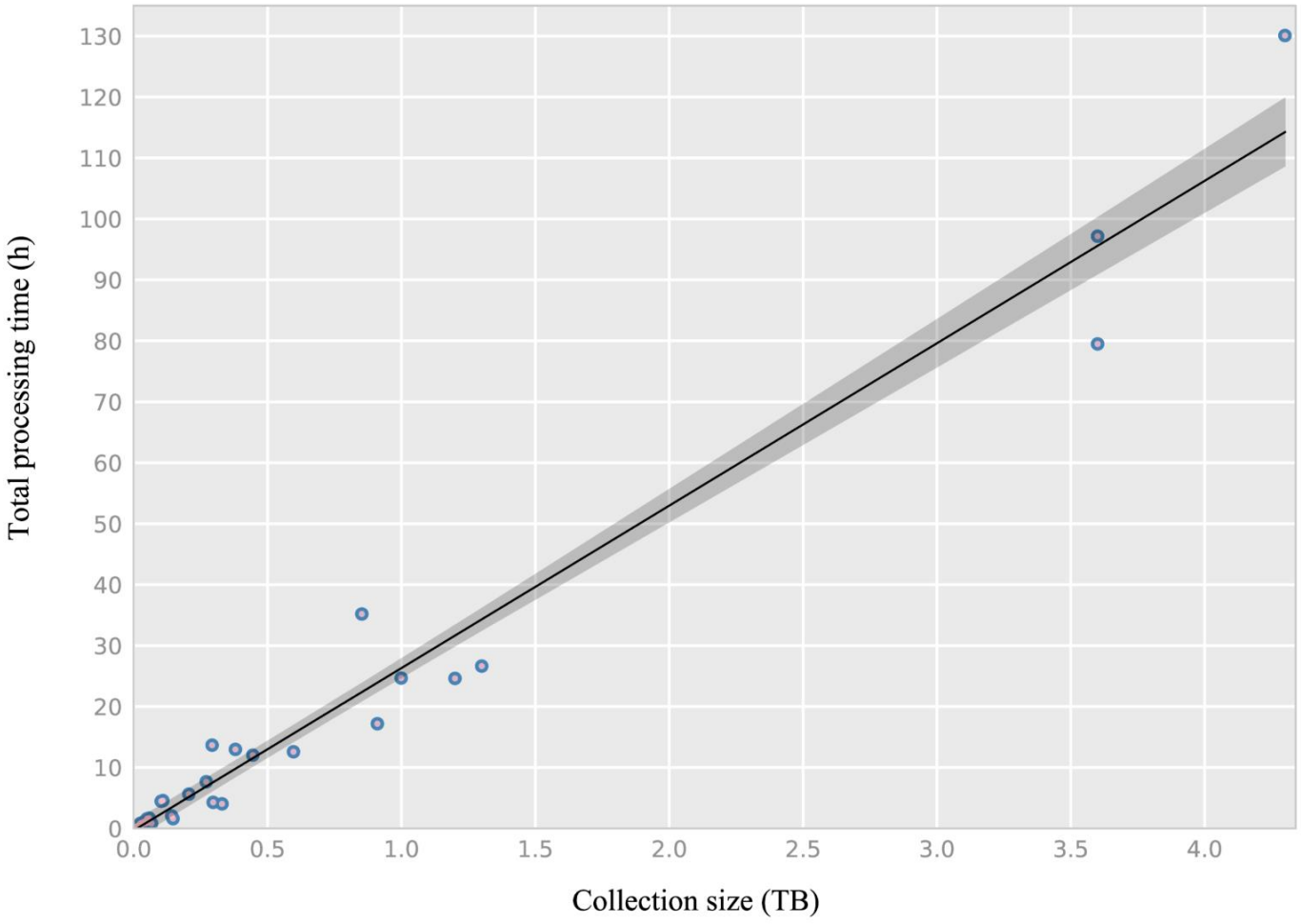
We then took all the times for each job (**Domain**, **Full Text**, **Webgraph**) and found processing time per GB in seconds.

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.

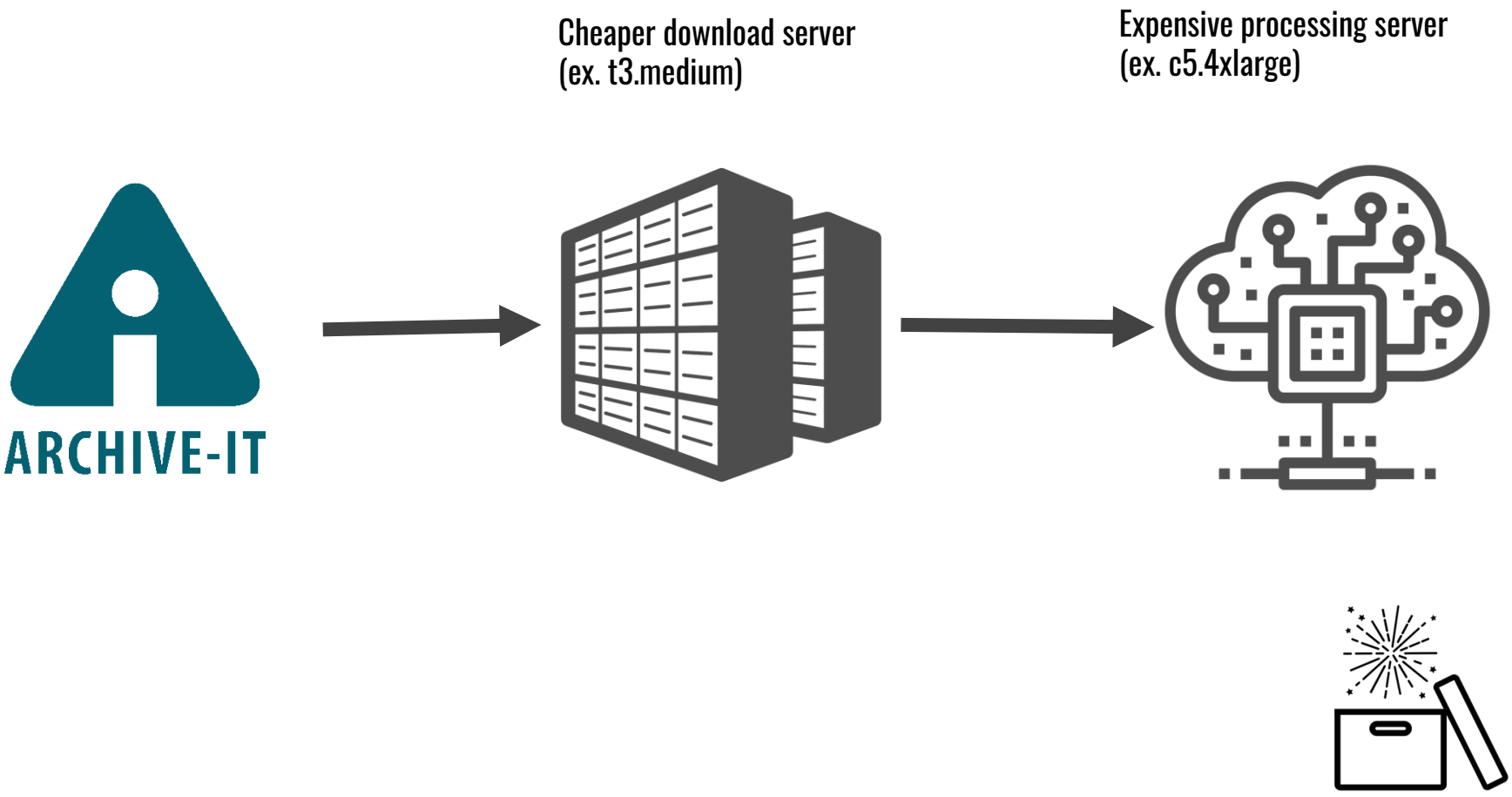
Cost of a WARC

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

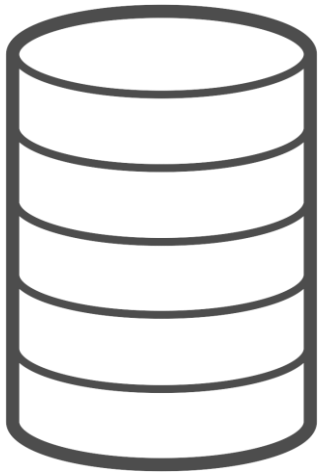


Limitations: Storage

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



Limitations: People

Discussion/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.

GIVE ME STATS

- 4,865 jobs run
- 187T analysed
- 12,987h, 8m, 18s (540+ days)
- 590h, 31m, 49s (24 days)
- 164 users
- 930 collections, 1,235,263 files
- Dataset citations?

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