The Cost of a WARC: Analyzing Web Archives in the Cloud

Ryan Deschamps,[1] Samantha Fritz,[1] Jimmy Lin,[2] Ian Milligan,[1] and Nick Ruest[3]
[1] Department of History, University of Waterloo
[2] David R. Cheriton School of Computer Science, University of Waterloo
[3] York University Libraries

ABSTRACT

The value of web archives to support scholarship in the humanities and social sciences is slowly being realized by the increasing availability of scalable tools and platforms. The cost of providing scholarly access is a critical component of developing a long-term sustainability strategy. This paper attempts to answer a straightforward question: How much does it cost to analyze web archives in the cloud? To make this question more concrete, we examine the creation of three derivatives (extraction of collection statistics, full text, and the webgraph) that serve as the starting points of many scholarly inquiries. Our analysis shows that these typical derivatives costs around US$7 per TB using our Archives Unleashed Toolkit. We describe in detail the methodology and assumptions made to arrive at this figure. To our knowledge, we are the first to quantify the economics of scholarly access to web archives, and we believe that this information is valuable for service planning by archives, libraries, and other institutions.

1 INTRODUCTION

Born-digital historical sources have the potential to reshape the humanities and social sciences. In the domain of web archiving, the Internet Archive and other organizations have already crawled and captured hundreds of billions of URLs. They are already becoming crucial for historical research: for example, military historians draw on posts by soldiers, and social historians delve into blogs and social media to examine perspectives of ordinary citizens [1, 4, 6].

Yet while librarians, archivists, and other curators are rapidly collecting content, scholarly access has lagged [3]. Although researchers have already begun to explore the "mechanics" of working with web archives at scale—via analytics toolkits and platforms such as ArchiveSpark [2], Warcbase [3] and its successor, the Archives Unleashed Toolkit—there is, to our knowledge, little work exploring the economics associated with providing scholarly access. It has become widely accepted that the pay-as-you-go model offered by the cloud can be an attractive alternative to the large capital investments necessary to deploy on-premise infrastructure for large-scale data processing. The cloud, of course, is not a panacea, but it should be part of any conversation.

While there has been related work costing out basic data repository functions in the cloud [7], sustainability studies have become increasingly important in today's budget environments. In this paper, we provide realistic cost estimates for scholarly analysis on web archives, supported by a process model derived from our own experiences analyzing over 160 TB of web archives and from organizing multiple in-person "datathons" that have brought together nearly two hundred stakeholders.

This work tries to answer a simple question: How much does it cost to analyze a WARC (the standard container file format of web archives) in the cloud? As the bottom line, we estimate the cost to be roughly US$7 per terabyte for a typical analytics product that would provide the starting point to scholarly inquiry—which we would characterize as quite affordable. In the remainder of this paper, we detail the methodology and assumptions made to arrive at this figure, based on a process model that moves data into the cloud only "on demand". We embarked on this study for our own internal sustainability planning as we strive to build a cloud-based analytics platform to support web archival research, but we believe these insights would also be valuable for libraries, archives, and even individual researchers as the community collectively grapples with the challenges of big data.

2 THE ARCHIVES UNLEASHED TOOLKIT

We begin with an overview of our analytics platform and our process model, as well as assumptions made in our study. Since our focus is on providing scholarly access, we assume that a web archive has already been harvested and is comprised of a number of files in the standard WARC (Web ARChive) file format or its predecessor, the ARC file format. We further assume the existence of a "preservation copy" held in stable, long-term, archival storage.[1] In practice, all of our current content partners use Internet Archive's Archive-It platform, but this need not be the case in general.

Answering the question "How much does it cost to analyze a WARC in the cloud?" first requires us to confront three details:

(1) What do we mean by "analysis"?
(2) What are we performing the analysis with?
(3) What exactly do we mean by the cloud?

The last is the simplest to answer: Our experiments were conducted on the Compute Canada Cloud, an instance of the OpenStack platform, made possible by a research grant. Compute Canada is an organization dedicated to providing researchers across Canada with computing support. Since OpenStack is the most popular open-source platform for managing cloud resources and is deployed by many organizations, our findings should be generalizable. For the purposes of estimating cost, however, we have roughly translated costs onto Amazon Web Services (AWS), currently the most popular commercial cloud provider.

The answers to the first two questions are related: throughout this paper, we assume processing of web archives using our Archives Unleashed Toolkit (AUT). This toolkit, which grew out of our earlier Warcbase project [3], represents a collaboration between computer scientists and historians who engaged in an iterative co-design process to build an analytics framework usable by humanities scholars and social scientists with no formal computer science

[1] We leave aside the question of where the preservation copy should be stored and the associated costs, which is outside the scope of this paper since an organization would face this challenge regardless.

training. AUT is designed as a Scala domain-specific language on top of the Apache Spark open-source data analysis platform, where scholars manipulate large web archives by defining data-parallel transformations over collections of records.

Based on our own experiences and subsequent engagements (more discussion below), we've discovered that scholars are often unsure where to even begin when interrogating a web archive. To provide guidance, we have previously proposed a model for scholarly interactions that begins with a question and proceeds iteratively through four main steps: filter, analyze, aggregate, and visualize [3]. Common analytics tasks, ranging from probing crawl statistics to visualizing web graphs to analyzing frequent mentions of named entities (person names, locations, organizations, etc.) all fit nicely into our proposed model.

One of the ways in which we have validated the effectiveness of the Archives Unleashed Toolkit has been through several in-person "datathons" in North America over the past several years that have brought together librarians, scholars, computer scientists, and other stakeholders [5]. Cumulatively, we have, with the help of our collaborators, engaged nearly two hundred members of the web archiving community. At each event, AUT has been deployed as a tool for hands-on exploration. These have been valuable sessions in teaching us what scholars really want and the barriers to access, and have informed the technical direction of AUT.

One important lesson from our datathons is that while AUT is within the technical reach of our participants, many are more comfortable extracting derivatives from web archives and then bringing those data over to their own laptops for further exploration. These derivatives, typically orders of magnitude smaller than the raw web archives, are then further manipulated using tools the scholars are already familiar with: Python, R, or even Microsoft Excel, exactly along the lines of the filter–analyze–aggregate–visualize workflow we've proposed. One example might be to restrict the analysis to a few domains of interest (filter), identify phrases of interest with a regular expression (analyze), and then count the occurrences of those phrases over time (aggregate) to display in a time series (visualize). In practice, we've discovered that one common role of AUT is to serve as a bridge between scholars' existing tools and large web archives.

We have further discovered that scholars are frequently interested in the same types of derivatives—they are requested so frequently that we have recently begun to pre-generate them as part of our data ingestion pipeline. Thus, we argue that the creation of these derivatives serves as a reasonable proxy for what "analysis" of web archives means. More concretely, in our experiments we used the Archives Unleashed Toolkit to:

- 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);

Our experiments further show that these analytics tasks serve as good representatives because processing time is dominated by the need to scan the entire collection (which can be terabytes).

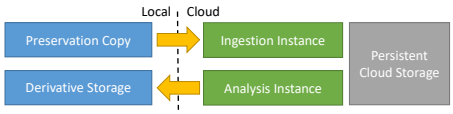

Figure 1: Our process model for providing scholarly access.

3 PROCESS MODEL

Given the background and context provided above, our experiments realize the process model shown in Figure 1. Everything to the right of the dotted line occurs in the cloud (the Compute Canada Cloud in our case). An "Ingestion Instance" virtual machine is used to copy the web archive from the source preservation copy over to persistent cloud storage (in our case, attached volumes backed by the Ceph file system; in the case of AWS, S3).[2] This is necessary because local storage on virtual machine instances is ephemeral and disappears once the instance is shut down.

Once the web archive has been ingested into persistent cloud storage, we start an "Analysis Instance" virtual machine to actually perform the data processing with AUT (i.e., generation of the derivatives discussed in the previous section). Note that although our toolkit is built on Spark, a distributed data platform, for simplicity we decided to run our jobs on individual multi-core machines: This decision is justified as follows: First, Spark is able to take advantage of multiple cores on a single server as well as multiple servers in a cluster, and thus we are still able to exploit data parallelism (albeit "scale up" as opposed to "scale out"). Second, our jobs are not latency sensitive—in the sense that scholars are for the most part willing to submit a job and wait a reasonable amount of time to obtain results—and thus the faster processing times that come with a distributed cluster are not worth the complexity of managing the cluster (e.g., handling startup, configuration, failover, etc.).

We experimented with a variety of virtual machine instance types and settled on a 16 core, 64 GB memory virtual machine. Note that while the Ingestion Instance can be the same as the Analysis Instance, in practice this would not be an effective use of resources since the server would be mostly idle while downloading data. We can allocate a far less powerful instance type for ingestion.

In our process model, the generated derivatives can then be copied over to the scholar's local machine for subsequent analysis or retained alongside the web archive in the persistent cloud storage (or both). However, we specifically discuss the costs associated with storage below.

4 FINDINGS AND DISCUSSION

In the Archives Unleashed Project thus far, we have processed over 160 TB of web archives from our content partners. For this study we focused on 57 collections analyzed in early 2018 from six different Canadian universities, collected using the Archive-It platform. We excluded from analysis nine collections smaller than one gigabyte, as they are too small to benefit from processing by AUT (leaving 48 in total). The largest collection, at 4.3 TB in size, was the Canadian Government Information Collection (from the

[2] As an alternative, an institution might eschew the need for the Ingestion Instance by directly pushing data into persistent cloud storage from a local server, but this is not possible in our case.

Size	Count
≥ 1 GB, < 10 GB	10
≥ 10 GB, < 100 GB	18
≥ 100 GB, < 1 TB	15
≥ 1 TB	5
Total	48

Table 1: Sizes of the collections in our study.

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

Table 2: Processing times per GB in seconds.

University of Alberta); the smallest collection, at 1.2 GB, was the University of Victoria's academic calendar. We believe that this sample is representative of the types of collections we are likely to encounter from Archive-It users. The complete distribution of collection sizes is shown in Table 1; all size figures are given in base 10 and all collection sizes refer to the raw, compressed WARCs.

We have automated the process model described in the previous section, with scripts that start up virtual machine instances to perform the various stages of processing. For data ingestion, we used the data transfer functionalities of WASAPI (Web Archiving Systems API)[3] provided by Archive-It. Our analysis is derived from the execution logs of these scripts.

In Table 2, we show the processing time (in seconds) per GB of source web archive for each derivative as well as the total. The column marked "all" shows analyses for all collections: we further break down results into large collections (larger than 1 TB, denoted "L"), medium collections (between 100 GB and 1 TB, denoted "M"), and small collections (less than 100 GB, denoted "S"). From these results, we make a few observations: Despite the different nature of these derivatives, running times are quite similar because the analytical queries are all dominated by the time to scan the entire collection. Extracting the webgraph is more computationally intensive, but not substantially more so. We see that total processing time for all three derivatives drops as the collection size increases, likely because the startup costs associated with AUT are amortized over longer running times. As expected, there exists a linear correlation between the raw collection size and the total amount of time required to generate all three derivatives: this is shown in Figure 2, where we observe an R^2 value of 0.970.

How large are these derivatives? The answer is shown in Table 3, which reports the sizes of the derivatives per GB raw archive: we report overall statistics as well as statistics broken into large, medium, and small collections (note the different units). These averages hide the fact that actual values vary by collection, depending on the nature of the crawl (e.g., wide multi-site crawls vs. narrow

[3]https://github.com/WASAPI-Community/data-transfer-apis

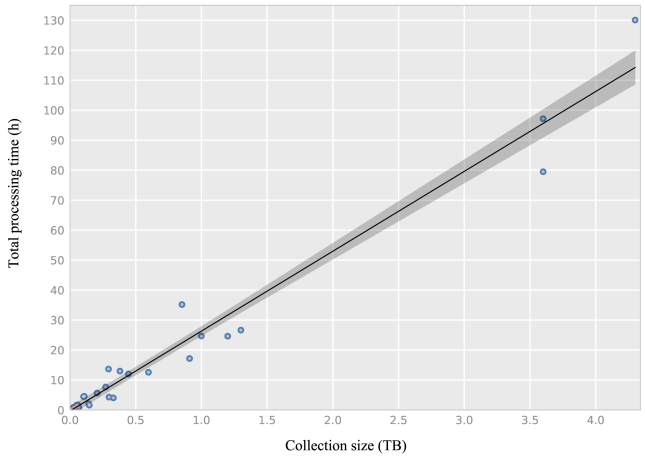

Figure 2: Scatter plot between collection size and total processing time, illustrating a linear relationship.

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

Table 3: Derivative sizes per GB.

deep crawls, text-heavy vs. media-heavy sites, etc.). However, in rough terms, for a typical medium site, domain distribution data is usually less than 1 MB, the raw text is perhaps 10s GB, and the webgraph is 10s MB. These values support our observation that AUT provides a bridge between web archives and scholars' existing tools, since datasets of these sizes are well within the capabilities of modern laptops. Furthermore, the long-term preservation of these derivatives presents no serious challenges: they can be treated as first-class citizens in the scholarly community (e.g., given DOIs and archived in institutional repositories).

Next, our cost analysis is shown in Table 4, organized in the same manner as Table 2, showing the cost in USD per TB of raw web archive on Amazon's EC2 service. Based on available statistics, the instance type used in our experiments on Compute Canada aligns roughly with a `c5.4xlarge` instance, with 16 virtual cores and 68 GB memory, currently costing US$0.68 per hour in the US East (Ohio) region. We assume per-minute billing (i.e., processing times are rounded up to the nearest minute) but do not account for instance startup costs. For consistency, we show cost per TB even for the small collections. These values report an macro-average, i.e., an average across individual collections. Note that our approach for computing these figures leads to inflated costs for small collections because they finish quickly (typically, only a few minutes).

All considered, a "bottom line" figure of US$7 per TB for a typical analytics product is a fair summary of our findings. We argue that further attempts to refine these estimates are not particularly meaningful for two reasons: First, we are mapping between instance types from two different cloud providers, which is imprecise at best. Second, instance costs are constantly changing, and thus an estimate today will likely be inaccurate in a few months. While we

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

Table 4: Processing cost per TB in USD.

hesitate to make more refined cost estimates, we are confident that our figures are in the right ballpark, and this granularity should be sufficient for resource planning purposes.

The above analyses only characterize the costs for generating the derivatives: there are other cost components that need to be quantified as well. Based on our process model, the web archive data need to be staged in from the preservation copy (Figure 1). While bandwidth for inbound data transfers are free in AWS, our workflow requires an Ingestion Instance to be running for the transfer. How much this costs depends on the data transfer speeds that can be achieved, which will vary by network connection, geographic location, and many other factors. Nevertheless, our own experiences provide a data point: under normal circumstances, we can achieve a sustained ingestion rate of around 30 MB/s, which means that even our largest collection, at 4.3 TB, can be copied over in less than two days. As mentioned above, the Ingestion Instance can be a less powerful (hence cheaper) instance. For example, the EC2 `t3.medium` instance costs only US$0.0416 per hour and has sufficient network performance. Thus, the costs associated with data ingestion are relatively small.

The final component of cost is storage. Somewhat simplifying, holding 1 TB of data on AWS S3 costs US$23 per month at present rates. Given this fact, we can optimize for different usage scenarios: to minimize storage costs, for example, we can copy the data into the cloud, generate and capture the derivatives, and then immediately delete the cloud copy of the data. At 30 MB/s with an EC2 `t3.medium` instance, transferring a TB of data effectively costs around US$0.40, which is less than the per-day cost of holding data on S3. These figures suggest that unless the scholar is continuously issuing queries, it makes more economic sense to transfer the web archive into the cloud only when needed and immediately remove it after each analysis. Of course, this assumes that the scholar is tolerant of the data transfer delays. While this is only a back-of-the-envelope calculation and various details need refinement, it is not hard for an institution to conduct such cost/benefit analyses based on the quality of service they wish to provide, balancing processing times with costs. Such analyses would be quite similar to libraries today deciding when to physically move a book to offsite storage. However, a broader and more general finding is that cloud data ingestion and processing is cheap, but cloud storage is expensive. This observation affirms our process model: as long as the preservation copy is secure, organizations can aggressively create and delete "processing copies" on a whim without careful consideration, treating such copies almost like a cache.

To further contextualize processing costs, one useful point of comparison is Google's BigQuery, a fully-managed cloud data warehouse, which offers a similar pricing model at US$5 per TB of data that a query scans. However, several caveats are necessary in order to make this comparison meaningful: BigQuery provides an SQL interface to relational data and cannot directly analyze web archives out of the box. Although in theory it would be possible to build AUT capabilities into the platform, we have not done so. Fundamentally a columnar query engine, BigQuery excels on analytics tasks that involve narrow projections of relational data, as in traditional data warehousing scenarios. Given this limitation, WARCs would be treated as a sequence of plain text records, which is not a use case that BigQuery is optimized for. Finally, BigQuery charges for *uncompressed* bytes read, whereas our figures are reported in terms of raw *compressed* WARCs. While web archives are too heterogeneous to draw a straightforward comparison, from an arbitrary sample we estimate that a compressed WARC is roughly 60% of the uncompressed size. From this simple analysis, AUT appears to be cost-competitive with a commercial service (but of course, our figures do not include a profit margin).

5 CONCLUSIONS

The Archives Unleashed Project is being primarily funded by the Andrew W. Mellon Foundation with three goals: First, to develop an analytics toolkit for web archives—that's AUT. Second, to build a community around scholarly use of web archives—that's the role of our datathons. Finally, to strive towards a sustainable platform for scholarly access to web archives. Our vision for accomplishing this is a platform we call the Archives Unleashed Cloud, and this study provides a step towards this vision. Such an enterprise would be sustainable, without any external assistance, if we are able to recover the costs associated with data processing (with appropriate overhead). 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.

ACKNOWLEDGMENTS

This work was primarily supported by the Andrew W. Mellon Foundation, with additional funding from Start Smart Labs, the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Ontario Ministry of Research and Innovation's Early Researcher Award program. We'd like to thank our content partners and Raymie Stata for comments on an earlier draft.

REFERENCES

[1] N. Brügger. 2018. *The Archived Web. Doing History in the Digital Age.* MIT Press, Cambridge, MA, USA.
[2] H. Holzmann, V. Goel, and A. Anand. 2016. ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. In *JCDL*. 83–92.
[3] J. Lin, I. Milligan, J. Wiebe, and A. Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *J. Comput. Cult. Herit.* 10, 4, Article 22 (July 2017), 30 pages.
[4] I. Milligan. 2016. Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. *Inter. J. of Humanities and Arts Comput.* 10, 1 (March 2016), 78–94.
[5] I. Milligan, N. Casemajor, S. Fritz, J. Lin, N. Ruest, M. Weber, and N. Worby. 2019. Building Community and Tools for Analyzing Web Archives through Datathons. In *JCDL*.
[6] I. Milligan, N. Ruest, and J. Lin. 2016. Content Selection and Curation for Web Archiving: The Gatekeepers vs. The Masses. In *JCDL*. 107–110.
[7] Z. Xie, Y. Chen, J. Speer, and T. Walters. 2016. Evaluating Cost of Cloud Execution in a Data Repository. In *JCDL*. 247–248.