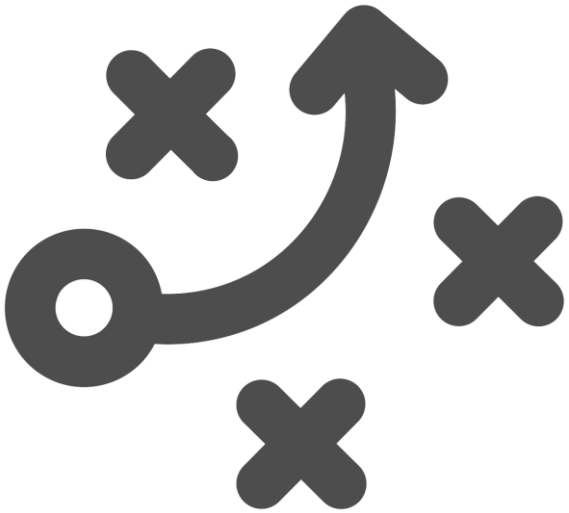
See a little Warclight:
Building an open-source web archive portal with Project Blacklight

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

IIPC WAC 2019 - Zagreb

Plan for The Talk



- Introduction
- Background
- Warclight
- Implementation at scale
- Next steps

Background

One perspective

Standing on Shoulders of Giants

- Analytical Access to the Domain Dark Archive
  - webarchive-discovery
- Big UK Domain Data for the Arts and Humanities
  - Shine
- Ruby on Rails
- Project Blacklight (UVa)
  - Library Catalogues
  - Panama Papers
  - GeoBlacklight
  - Arclight
  - Warclight

R+L = J

(Solr + webarchive-discovery) + Project Blacklight = Warclight

Warclight

What is it?

- Rails Engine
- Tight integration with webarchive-discovery
  - Collection, and Collection-ID
  - Institution
  - Resource name
- Features
  - Field search
  - Faceted search
  - Memento Reply URL check
  - Live web check
  - Thumbnail display
  - Plug-in architecture
    - Advanced Search
    - Range limit

Live Demo

warclight

All Text | mi'kmaq | Search

Start Over | mi'kmaq | ✖

Limit your search

General Content Type

html	274
text	36
other	9
pdf	6

Crawl Year

2013 – 2018 Limit

2013 2014 2015 2016 2017 2018 2019

Public Suffix

ca	188
com	109
org	15
ab.ca	8
gc.ca	3
fm	2

« Previous | **1** - **10** of **325** | Next »

1. Mi'kmaq Archives - APTN NewsAPTN News

Title:	Mi'kmaq Archives - APTN NewsAPTN News
Host:	aptnnews.ca
Crawl Date:	2017-10-13T18:44:09Z
General Content Type:	html
Content Language:	en
Domain:	aptnnews.ca
Collection:	Truth and Reconciliation
Collection Id:	7685
This page links to:	aptn.ca; aptncommunity.com; aptncsr.com; aptnnews.ca; aptnpress.com...

2. Young aboriginal musicians pay tribute to poet Rita Joe at the NAC | CBC News

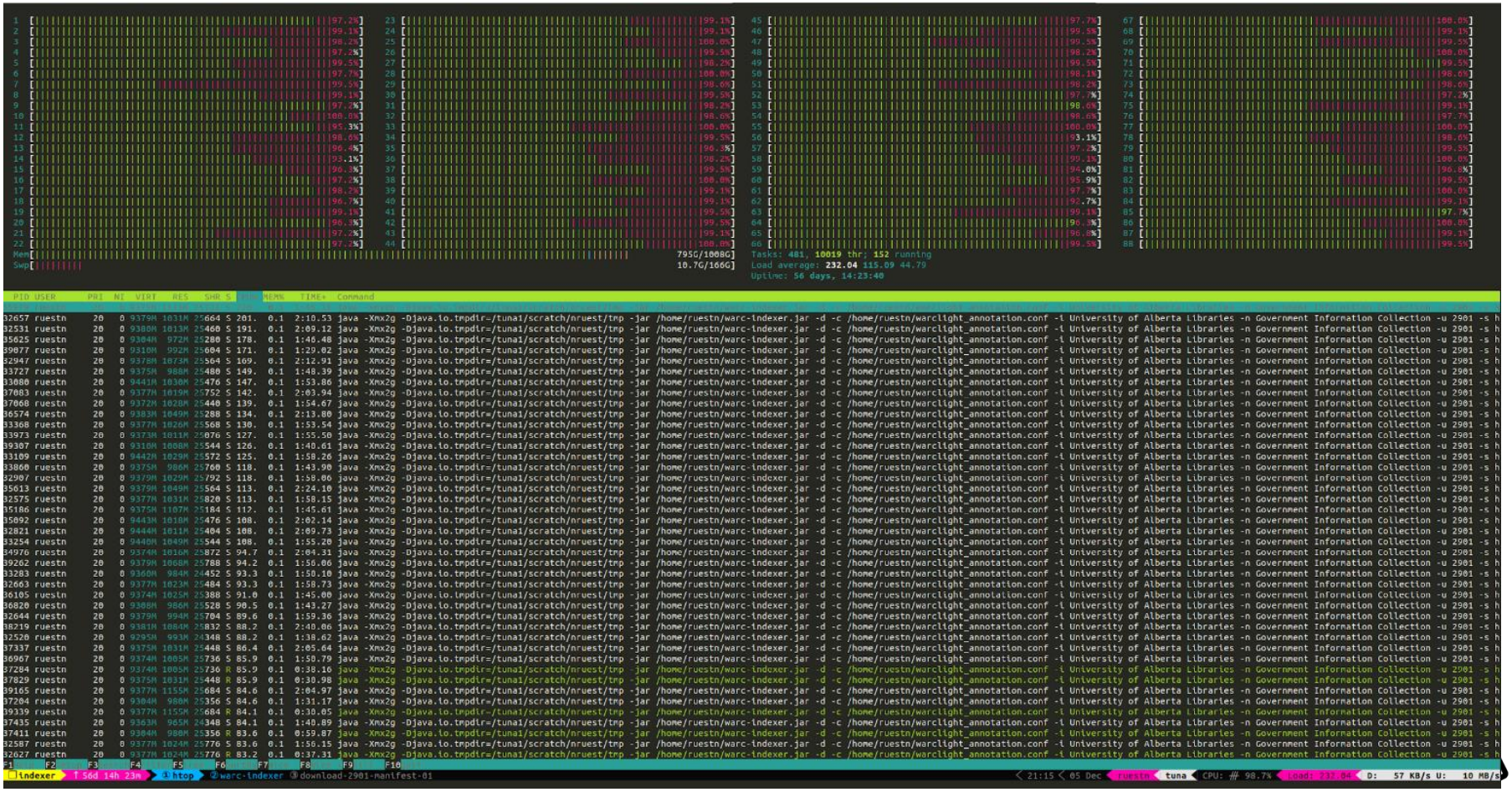
Title:	Young aboriginal musicians pay tribute to poet Rita Joe at the NAC | CBC News
Host:	cbc.ca
Crawl Date:	2018-04-13T18:22:29Z
General Content Type:	html
Content Language:	en
Domain:	cbc.ca
Collection:	Truth and Reconciliation
Collection Id:	7685
This page links to:	cbc.ca; cbckids.ca; cbcshop.ca; facebook.com; instagram.com...

Implementation at Scale

WALK

- 6 Canadian Universities
- Archive-It Partners
- 30T of WARCs
- 1+ Billion Solr docs
- ~1-2 months to download and index

Indexing

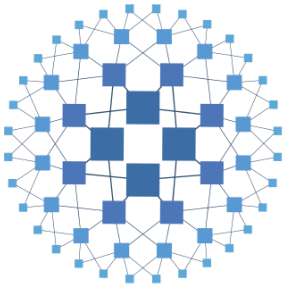

795G/1008G]
10.7G/166G]
Tasks: 481, 10019 thr; 152 running
Load average: 232.04 115.09 44.79
Uptime: 56 days, 14:23:40
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
F1 F2 F3 F4 F5 F6 F7 F8 F9 F10
indexer
56d 14h 23m
htop
warc-indexer
download-2901-manifest-01
21:15
05 Dec
ruestn
tuna
CPU: 98.7%
Load: 232.04
D: 57 KB/s U: 10 MB/s

Warclight apps

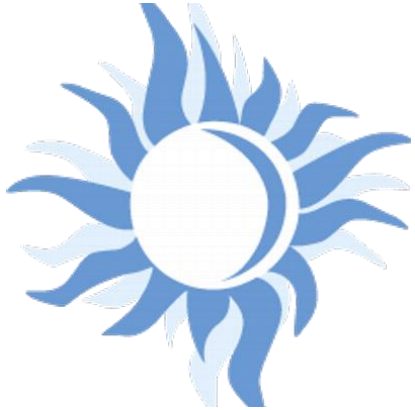
- warclight demo - warclight.archivesunleashed.org (2,775,723)
- University of Toronto - utoronto.archivesunleashed.org (691,803,697)
- University of Alberta - ualberta.archivesunleashed.org (331,924,814)
- University of Victoria - uvic.archivesunleashed.org (18,873,434)
- University of Winnipeg - uwinnipeg.archivesunleashed.org (30,793,376)
- Dalhousie University - dalhousie.archivesunleashed.org (736,023)
- Simon Fraser University - sfu.archivesunleashed.org (787,991)
- GeoCities - geocities.archivesunleashed.org (316,694,661)
- WALK (federated) - walk.archivesunleashed.org (1,077,695,058)

Components

APACHE



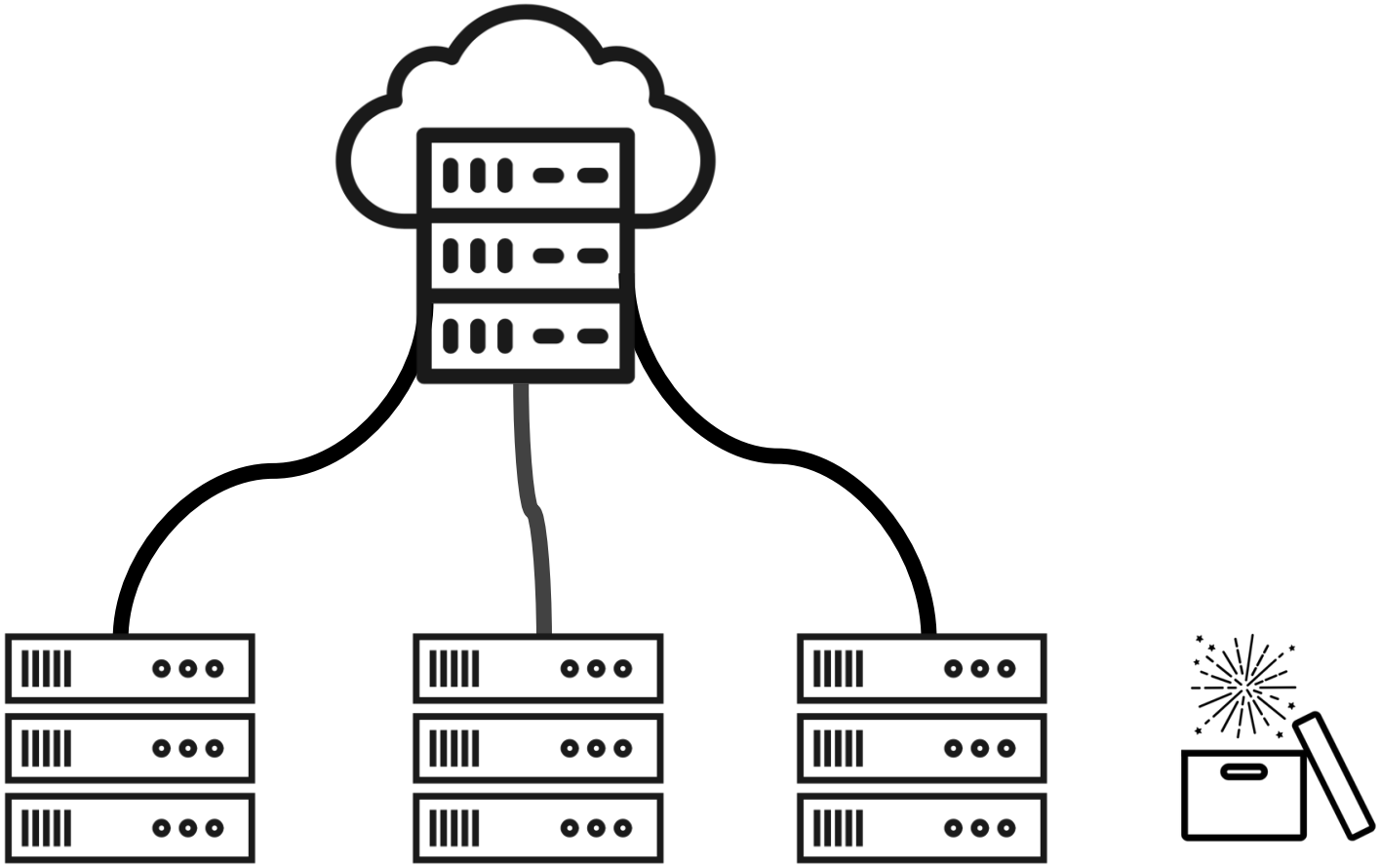
HAPROXY



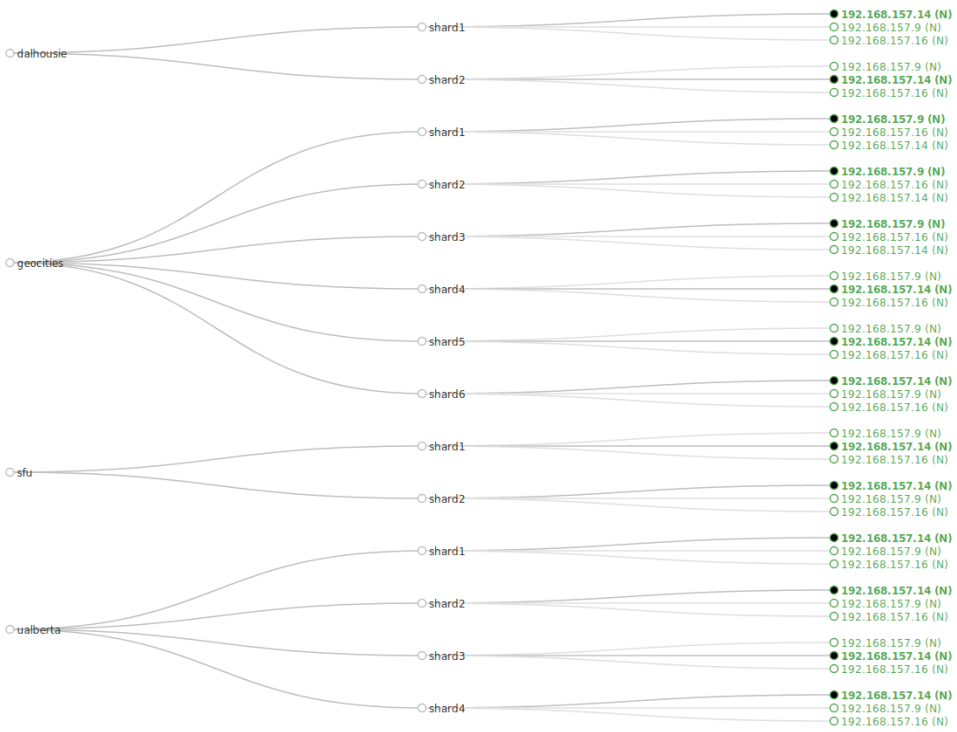
Solr

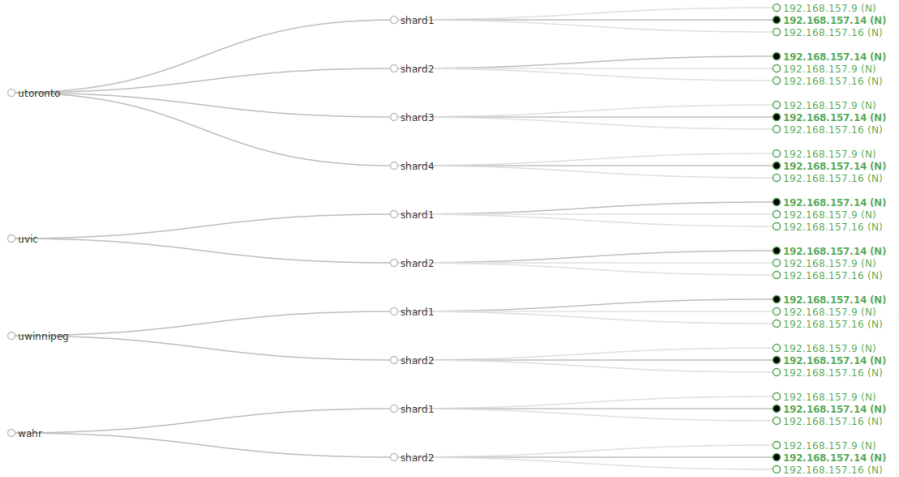
RAILS

System diagram

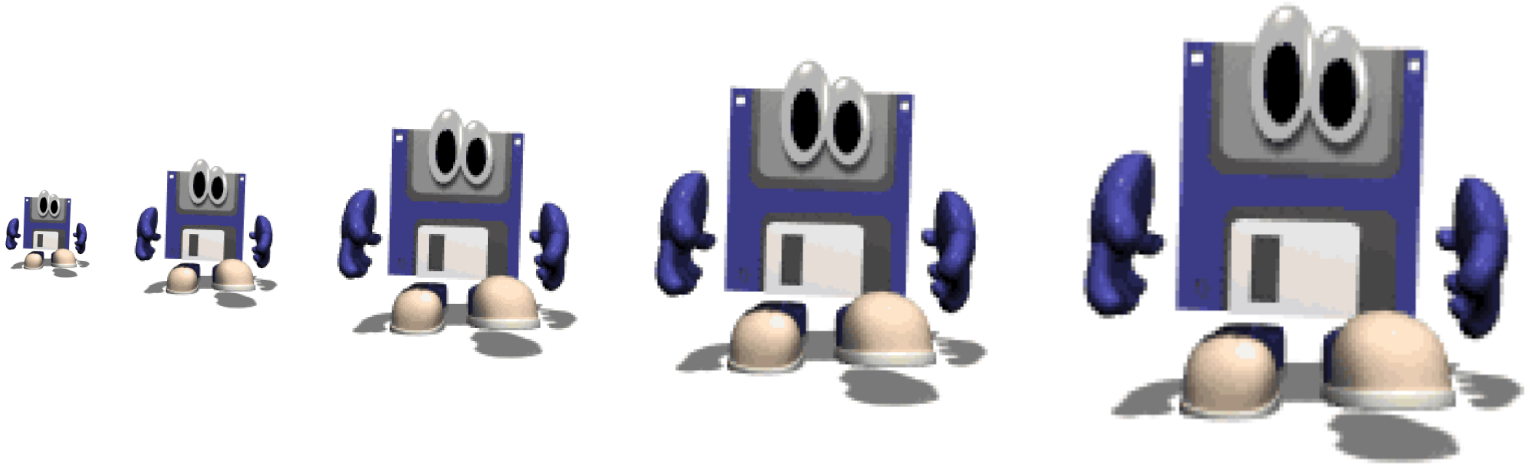


SolrCloud Collections





Implementation



Facet caching:

- https://issues.apache.org/jira/browse/SOLR-13132
- https://github.com/magibney/solr-facet-cache

Solr index:

-
-

Next Steps

Discussion

- Another framework for webarchive-discovery
- Ruby on Rails vs Play Framework
- Parity features with Shine?
  - Concordance
  - Trend search
  - Blacklight plugins
- Growing the community - is there one?
- Reality of time and Archives Unleashed priorities

We look forward to your questions and thoughts.

Thank you!

- Toke Eskildsen, Thomas Egense
- Andy Jackson, Gil Hoggarth
- Chris Beer, Jesse Keck, Justin Coyne (Blacklight Community)

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

Links

- archivesunleashed.org/warclight
- archivesunleashed.org
- cloud.archivesunleashed.org
- github.com/archivesunleashed
- slack.archivesunleashed.org
- news.archivesunleashed.org
- twitter.com/unleasharchives