EXPLORING THE FEASIBILITY OF INDEXING CAMPAIGN STORE IN ELASTICSEARCH

EMILY MC NETT
University of Wisconsin-Stout
Mentors: Nathan Hook, Eric Nienhouse, Jason Cunning

NCAR

SIParCS Summer Internships in Parallel Computational Science BOULDER · CO

AUGUST 2, 2023

University of Wisconsin-Stout
Wisconsin's Polytechnic University

NSF

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Overview

- Background
  - Campaign Store
  - The Problem
  - The First Step (Towards a Solution)
- Implementation
  - Experimentation
  - Challenges
- Findings
  - Previous Availability
  - New Availability
- Conclusion
  - Future Work & Objectives

Campaign Store

What is it?

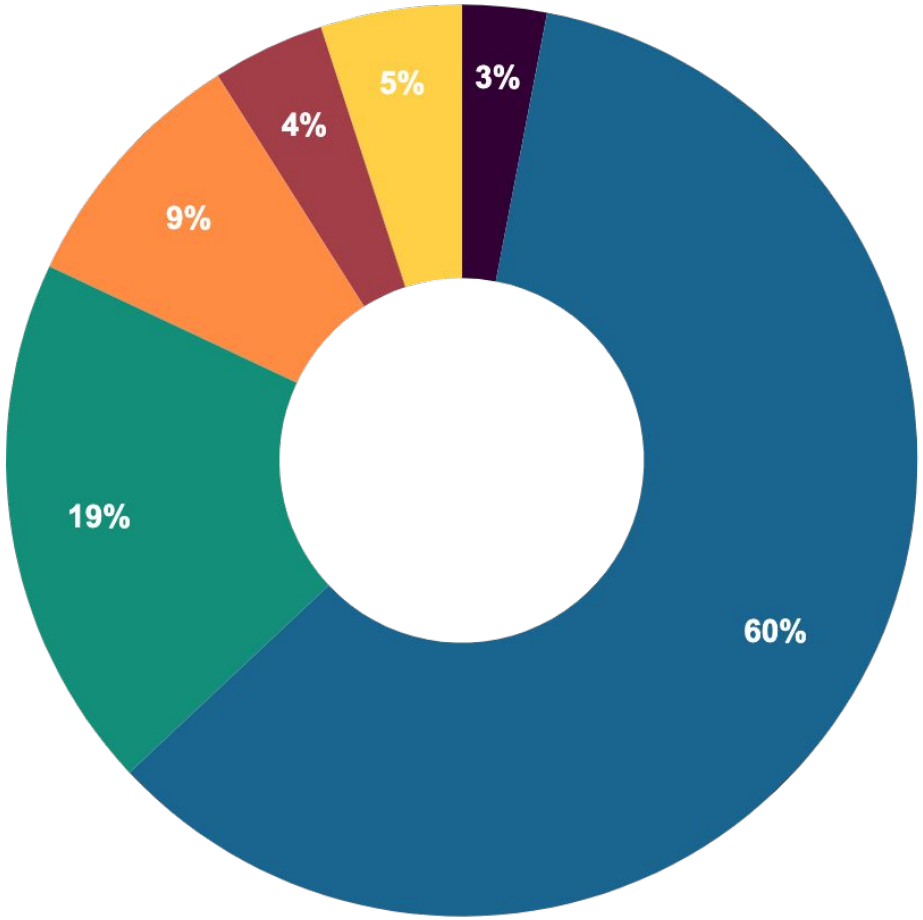
NCAR Campaign Stor[ag]e is a resource for medium-term storage of project data, typically for three to five years, by NCAR labs and universities that have project allocations.[1]

What is known about it?

Monitored for security and storage purposes . . .

[1] Smith, B. (2023, January 12). *Campaign Storage File System*. ARC NCAR. https://arc.ucar.edu/knowledge_base/70549621

The Problem



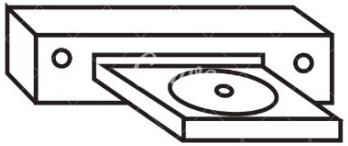
19% collection data, 60% cleaning and organizing data, 21% left to continue scientific work

- Building training sets
- Cleaning and organizing data
- Collecting data sets
- Mining data for patterns
- Refining algorithms
- Other

2016 Crowdflower survey of Data Scientists

The Problem

6,352,950
DVD Movies



12 Human Brains at Memory Capacity[1]

(2019 MacBook Pros)
933090
Laptops

4.7 years to download with U.S. average speed[2]

243.47 years of binge watching 90 minute 4k movies

29,858,865,529,466,000
Bytes

80 Tebibytes — The Entire Internet in 1997

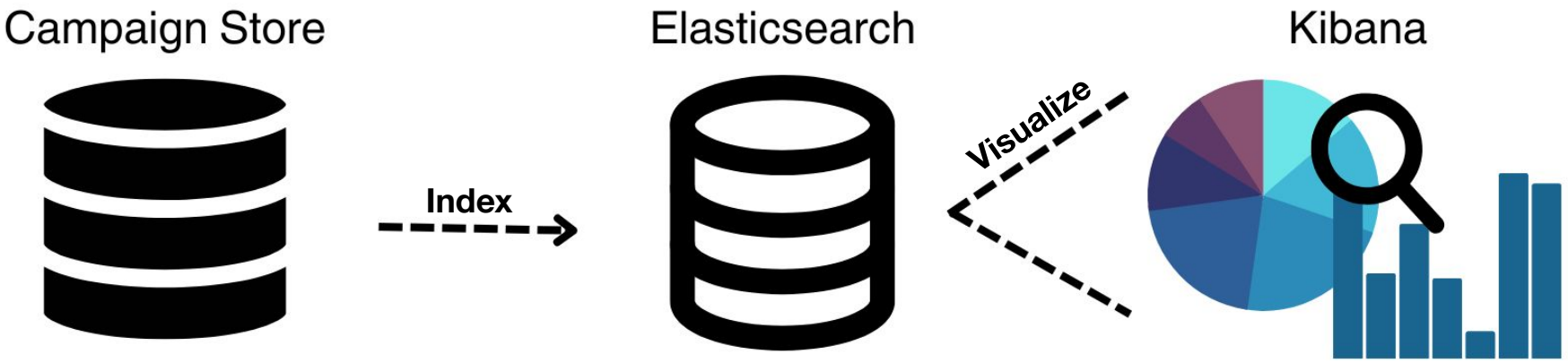
26.52 Pebibytes — **Desired Campaign Store Index**

84.17 Pebibytes — Current Campaign Store Size

[1] Reber, P. (2010, May 1). *What is the memory capacity of the human brain?*. Scientific American. https://www.scientificamerican.com/article/what-is-the-memory-capacity/
[2] Tachus Community. (2023, July 14). *Tachus blog: Internet speeds: USA vs. the rest of the world*. RSS. https://www.tachus.com/post/internet-speeds-usa-vs-the-rest-of-the-world

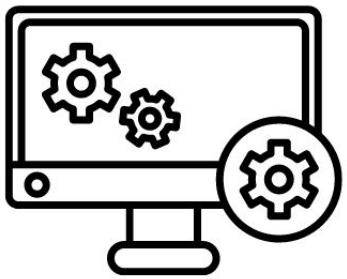
The First Step (Towards a Solution)

Is it feasible?



Experimentation

File Walker



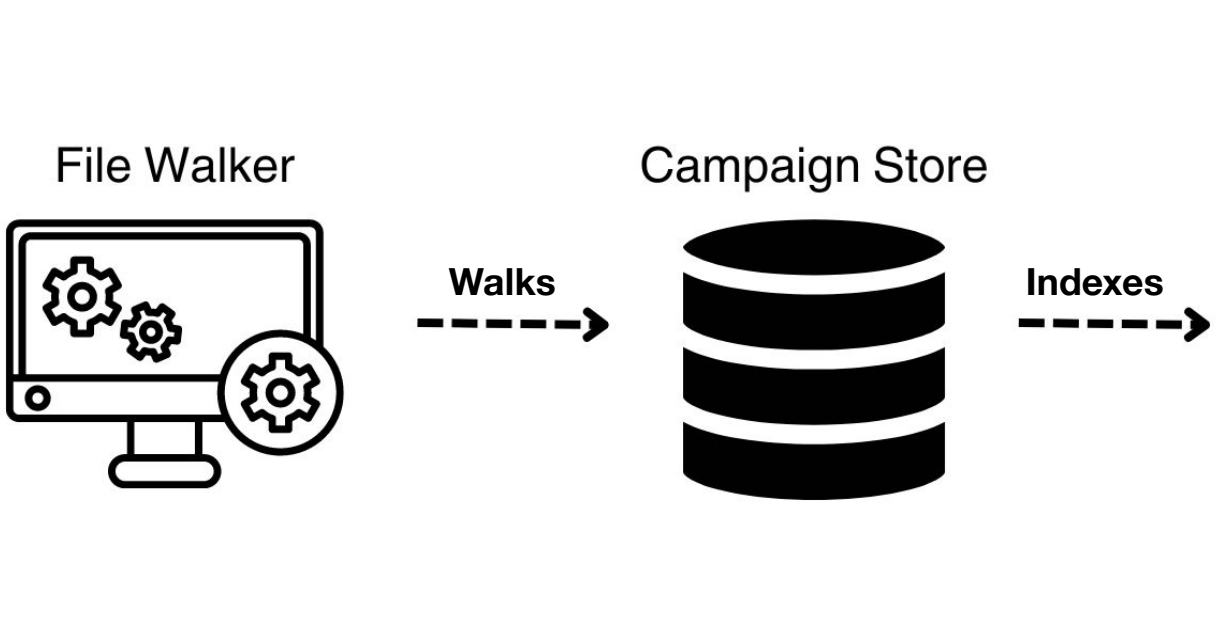
Walks

Campaign Store

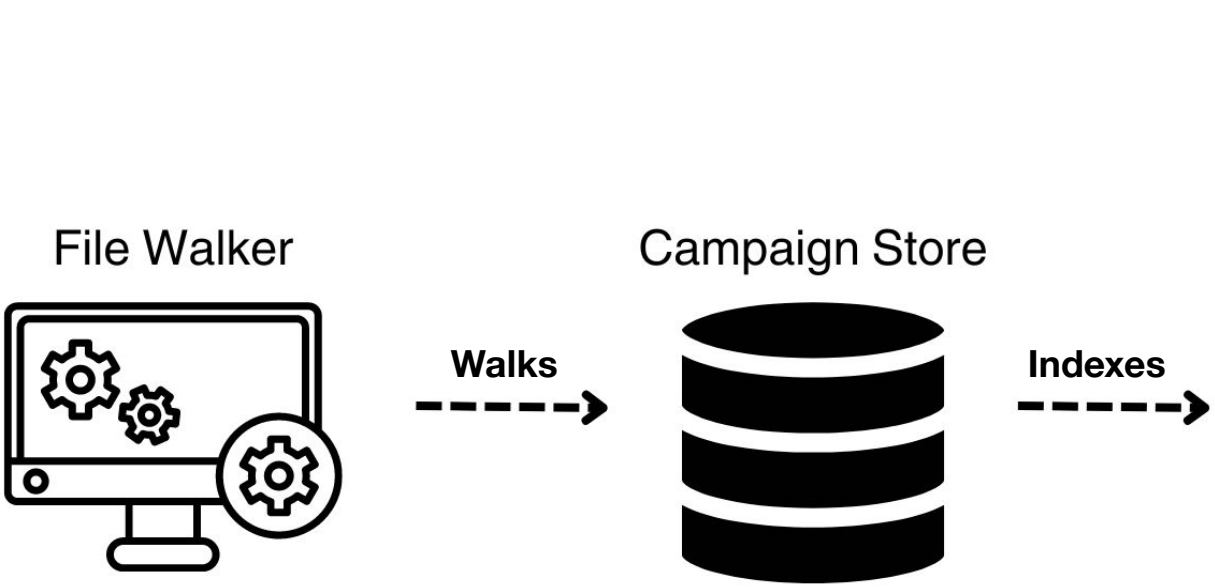
Displays

User Interface

Experimentation



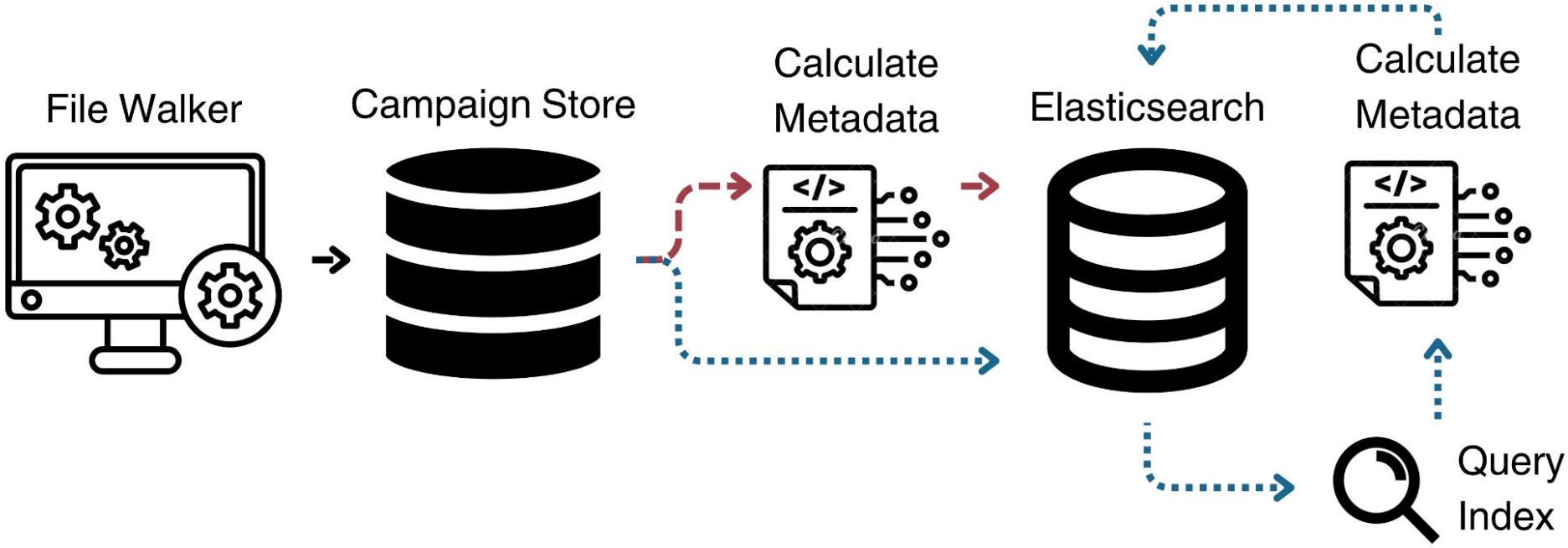
Experimentation



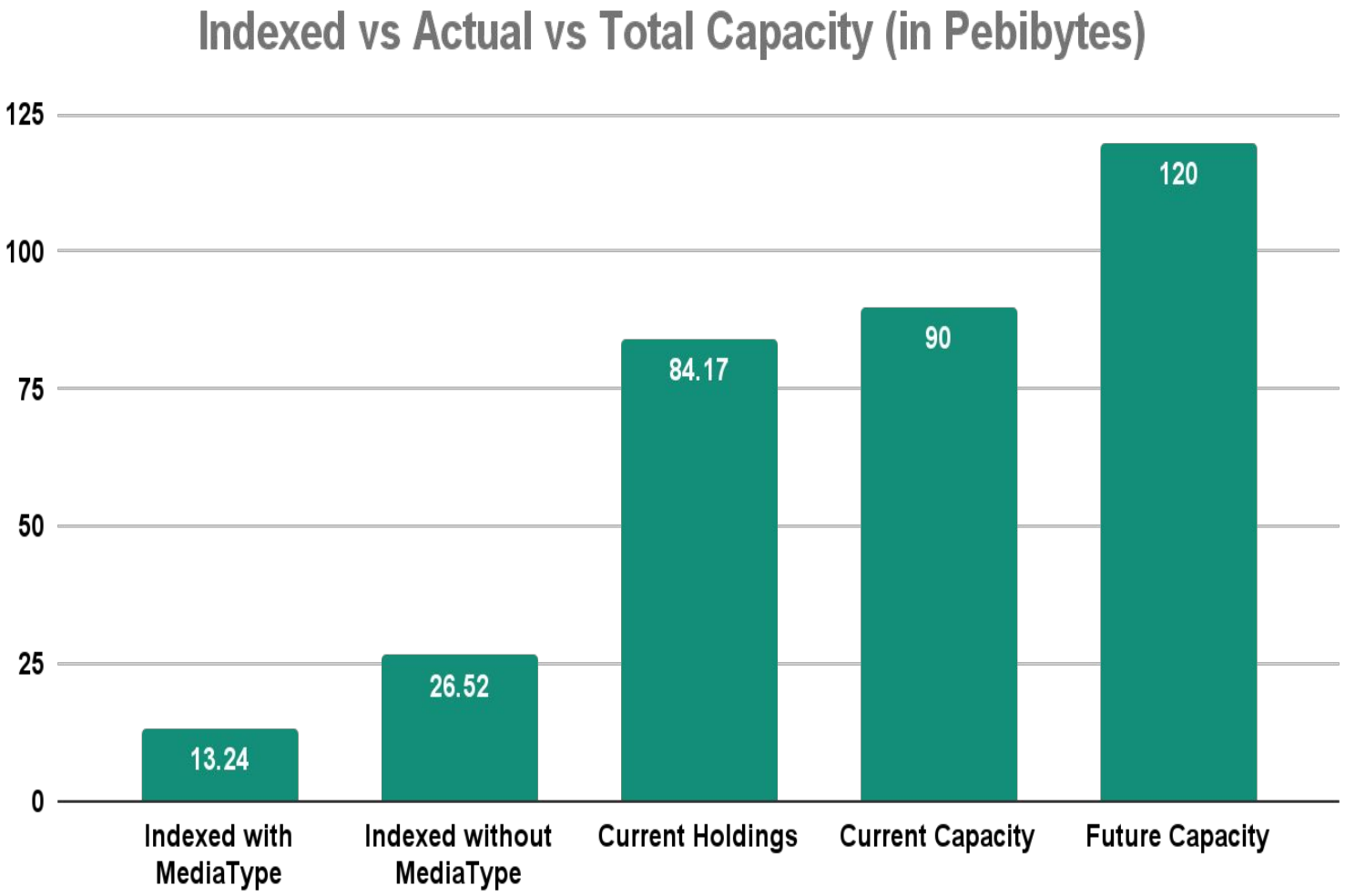
Challenges

Option One

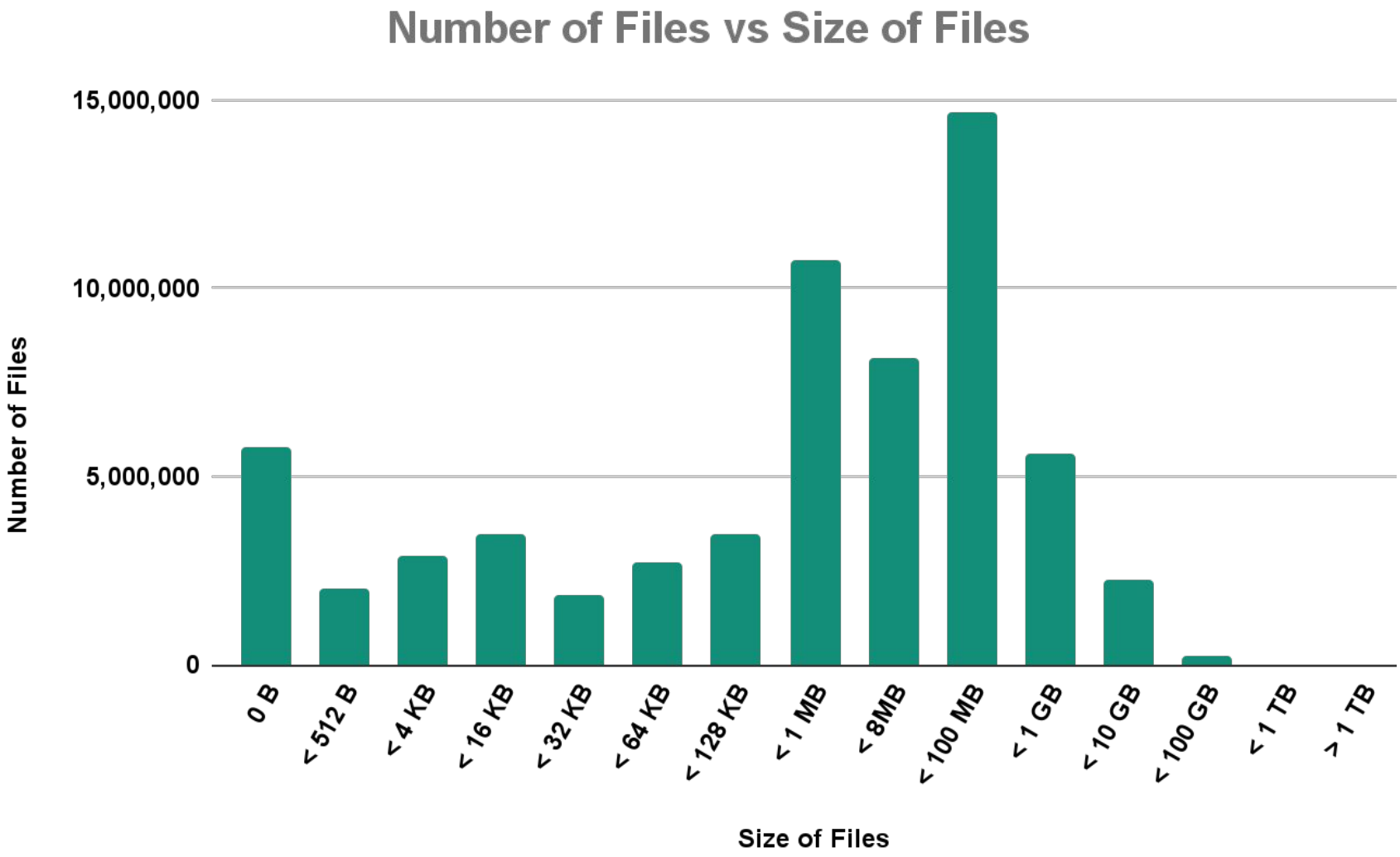
Option Two



Challenges


Indexed vs Actual vs Total Capacity (in Pebibytes)
125
100
75
50
25
0
13.24
26.52
84.17
90
120
Indexed with MediaType
Indexed without MediaType
Current Holdings
Current Capacity
Future Capacity

Previous Availability

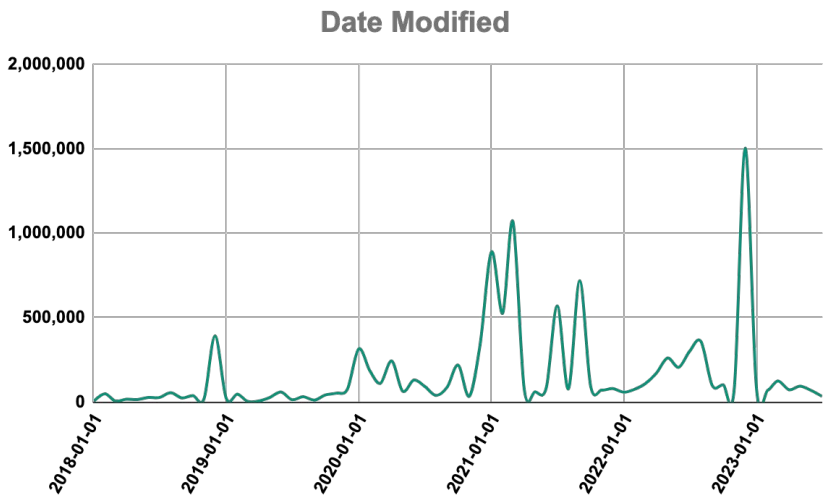


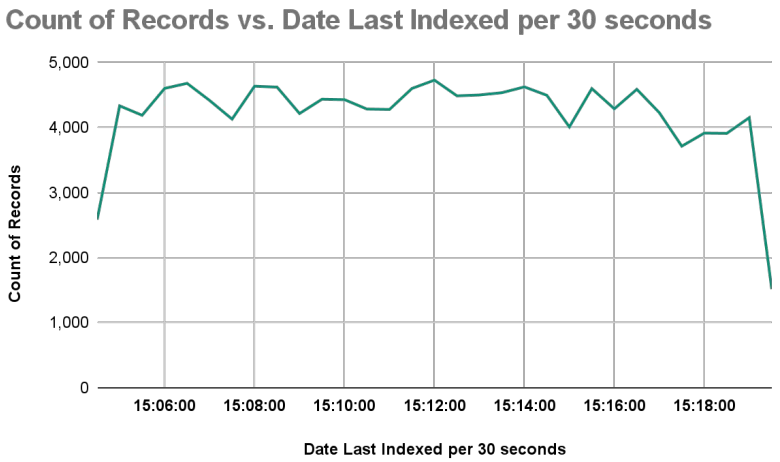
New Availability

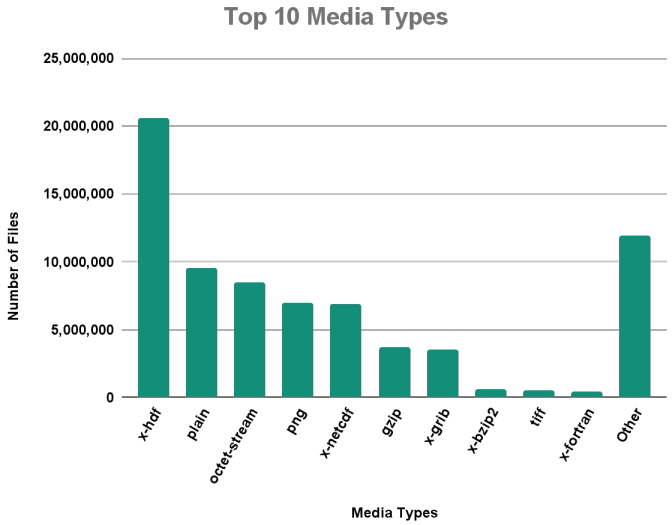
Labs	Count of Files	Sum of size
ACOM	13,671,527	1.10PB
CISL*	10,312,467	6.67PB
CGD	6,380,954	2.04PB
EOL	3,928,540	543.82TB
HAO	914,582	1.10PB
MMM	13,176,168	876.61TB
RAL	14,259,973	832.03TB

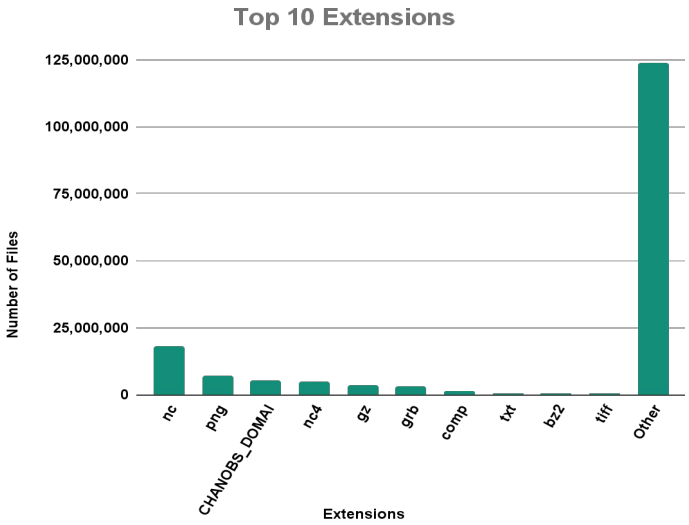
*CESM and Collections Directories Only

New Availability









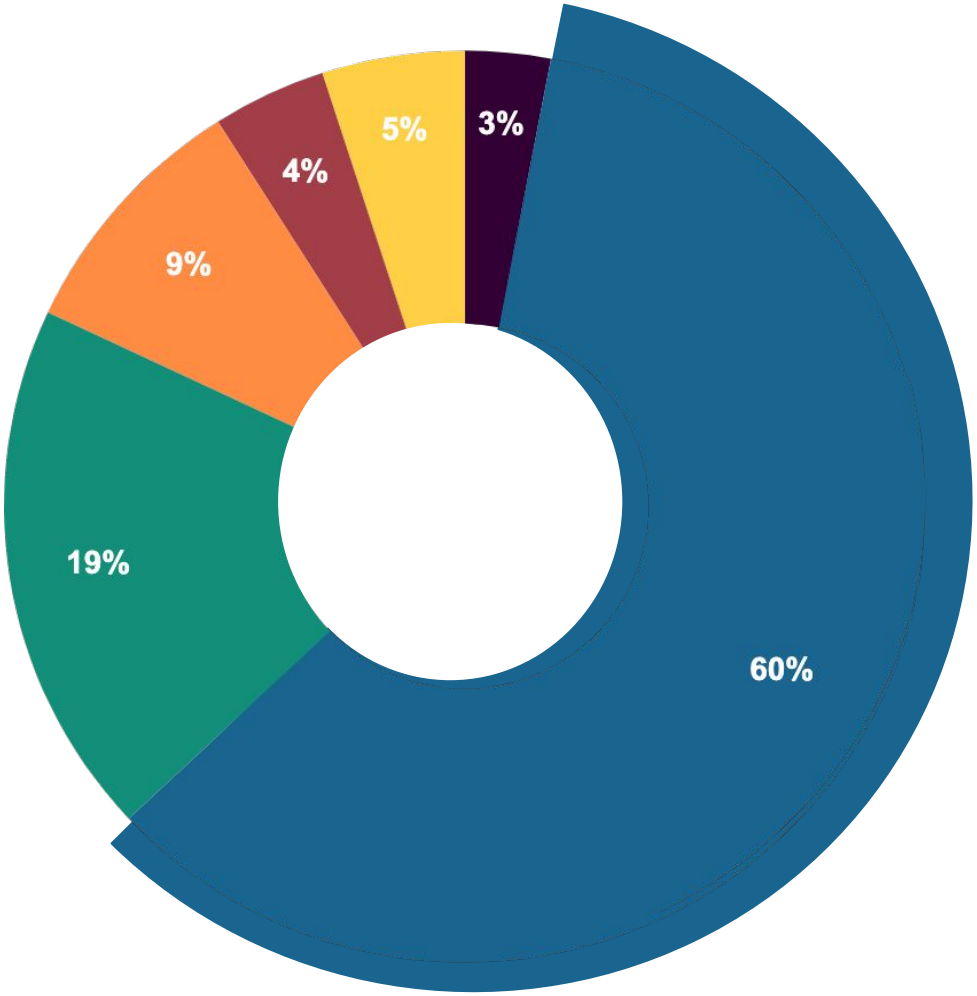
New Availability

Media Type	Count of Files	Extensions	Total Extensions
x-hdf & x-netcdf	27,391,357	nc, nc4, comp, LDASIN_DOMAIN1, h5, hdf, he5, ncf, CHRTOUT_GRID1, 0118, 0137, 0140, 0358, 0286, 1, 2016-10-01_00:00_DOMAI, 2016100100_DOMAIN1, 2, e001, e002. . .	1,476
x-grib	3,520,333	grb, grb2, grib2, tm00, subset, 01h, AAA, AAB, AAC, GFS, JMAGSM, NAM, NARR, f036, f042, ml, pl, raphrrr. . .	11,253

Unique Extension Count
247,572

Unique Media Type Count
277

New Availability



Unique Media Type Count
277

Unique Extension Count
247,572

Conclusion

- Indexed and visualized metadata for over 26.52 pebibytes of files. (About 244 years of binge watching 4k movies)
- Provided insight on Campaign Store.
- Discussed possible supports to reduce time searching for and organizing data.
- Confirmed feasibility of indexing Campaign Store file's metadata in Elasticsearch.
- Visualized findings with Kibana.

Future Work and Objectives

Future Work

- Incorporate Spatial, Temporal, and more extracted metadata.
- Allow for continuous traversals to monitor data churn.
- Disseminate to web based data repositories.

Future Objectives

- Decrease time spent obtaining and organizing data from 19% and 60%.
- Provide updated, greater insight on Campaign Store.

Acknowledgements

Mentors: Nathan Hook, Eric Nienhouse, Jason Cunning

Assistance from: Ken Cote, Nick Wehrheim, Bill Anderson, Joseph Mendoza

SIParCS Team: Virginia Do, Julius Owusus Afryie, all other admin, Fellow 2023 Interns.

Internship Assistance: Jerry Cyccone, Kristen Pierri

NCAR, CISL, & the Sage Team

Questions?