Exploring the Feasibility of Indexing Campaign Store in Elasticsearch

Emily Mc Nett

Mentors: Nathan Hook, Eric Nienhouse, Jason Cunning

NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

SIParCS Summer Internships in Parallel Computational Science BOULDER • CO

University of Wisconsin-Stout
Wisconsin's Polytechnic University

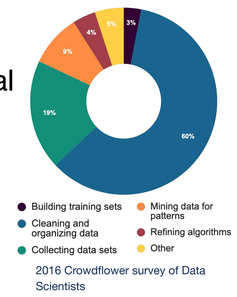
BACKGROUND

Campaign Store

NCAR Campaign Storage 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.

The Problem

The volume, variety, and occasional vagueness of Campaign Store's files has made it difficult for NCAR and external scientists to locate, clean, and organize the data they require to "do their science."



2016 Crowdflower survey of Data Scientists

The First Step (Towards a Solution)

With aspirations to reduce 'The Problem' comes the initial undertaking of indexing metadata found in the repository's files. This project is focused on determining the feasibility of using Elasticsearch to do just that.

QUESTIONS ANSWERED

Already

- Is Elasticsearch capable of holding campaign store within an index?
- Is Kibana a reliable and useful visualization and evaluation tool?

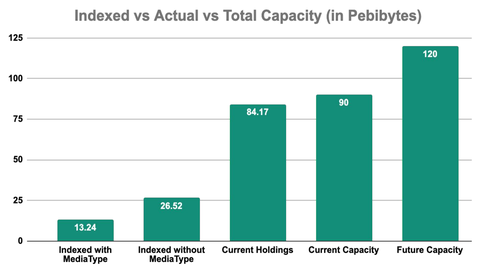
Through Analysis of Metadata

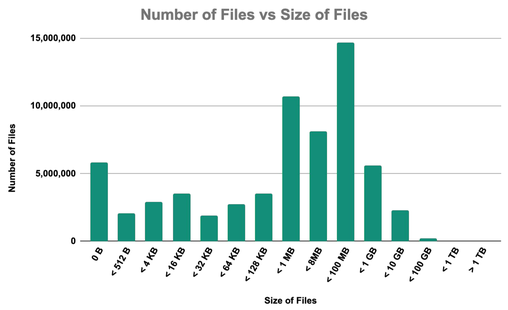
- What files do NCAR scientists majorly use?
- How much data does NCAR have?
- What is the distribution of file sizes at NCAR?

Through Continuation of the Project

- How can more metadata be more easily extracted from a zip file? a NetCDF file? etc. . .
- How can the metadata of files be used to understand the workflow of scientists?
- How often are files on Campaign Store added, updated, and deleted?

FINDINGS



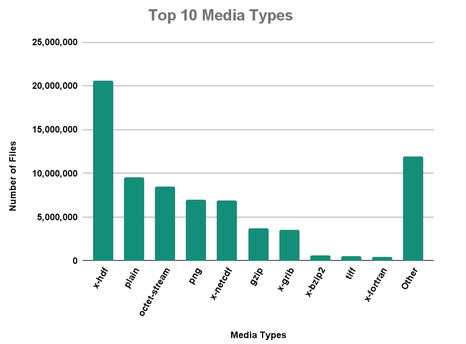


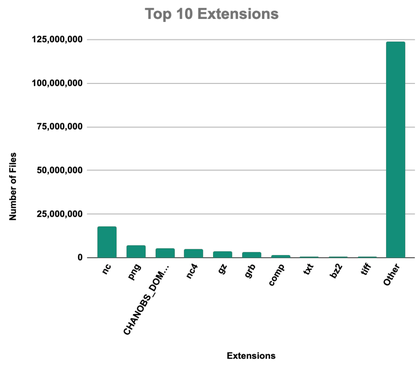
Labs	Count of Files	Sum of size
ACOM	13,671,527	1.10PB
CISL*	10,312,467	6.67PB
CGD	6,380,954	2.04PB
EOL	3,928,540	543.82TB
HAO	914,582	1.10PB
MMM	13,176,168	876.61TB
RAL	14,259,973	832.03TB

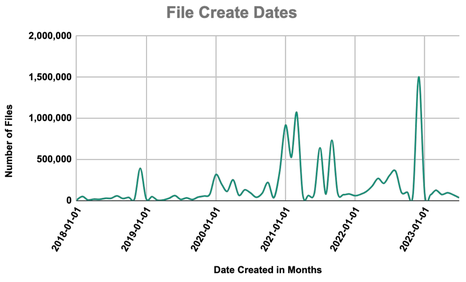
*CESM and Collections Directories Only

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	Unique Media Type Count
247,572	277







View the data on your own*

Username: siparcs
Password: siparcs2023

https://sagepoddev2.ucar.edu:5601/app/r/s/L3X1e

* Requires NCAR VPN access, best viewed in Firefox

IMPLEMENTATION

Technologies Used

- Java Spring Boot
- Podman
- Docker
- Elasticsearch
- Kibana

Software Principles

- Object Oriented
- Event Driven
- Layered Architecture
- Unit Testing
- Design Patterns

Methodology

- Agile Scrum
  - One week sprints

CONCLUSION

Conclusion

- Indexed and visualized metadata for over 26.5 pebibytes of files. (About 244 years of binge watching 4k movies)
- Provided answers and insight for initial inquiries on campaign store and elasticsearch.
- Confirmed feasibility of indexing Campaign Store file's metadata in Elasticsearch.

Future Work

- Incorporate Spatial, Temporal, Variable, Identification, and more extracted metadata.
- Allow for continuous traversals of the repository to monitor additions, changes, and deletions in files over time.
- Disseminate to web based data repositories in order to support external scientists.

Future Objectives

- Decrease time spent obtaining and cleaning data from 19% and 60% alike.
  - Thus, allocating more time for science.
- Provide continuously updated, greater insight on Campaign Store for its data managers.

ACKNOWLEDGMENTS

Thank you to my mentors Nathan Hook, Eric Nienhouse, and Jason Cunning, to Ken Cote, Nick Wehrheim, Bill Anderson, and Joseph Mendoza for assistance in getting my project to run, to the SIParCS coordinators Virginia Do, Julius Owusu Afriyie, and all other admin, to my fellow SIParCS interns, to Jerry Cyccone for PDWS, to Kristen Pierri for assistance at NCAR, to the NSF for this project, and to NCAR, CISL, and the Sage Team for their support during SIParCS.