# Exploring NCAR's Campaign Store with Elasticsearch

## How much can we know about 120 PiB in a summer?


**Anh Nguyen[1,2], Nathan Hook[1], Eric Nienhouse[1], Jason Cunning[1]**

[1]National Center for Atmospheric Research, [2]Mount Holyoke College


## BACKGROUND

### NCAR Campaign Store:

- Resource for medium-term storage of project data, typically for three to five years, by NCAR labs and universities
- Large volume (120 petabyte capacity, 90% currently being used)

### Application:

- Java-based web application that utilizes Elasticsearch, Kibana, Spring Boot, Java FileVisitor interface to index files and directories

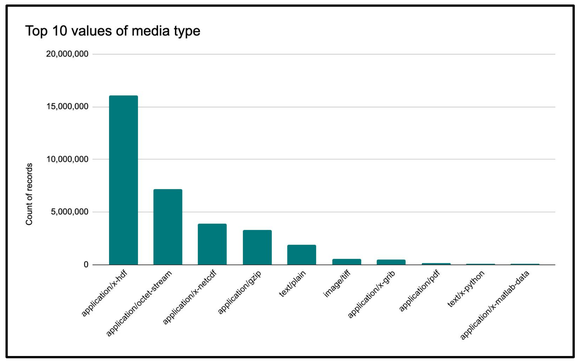
### Long-term goal:

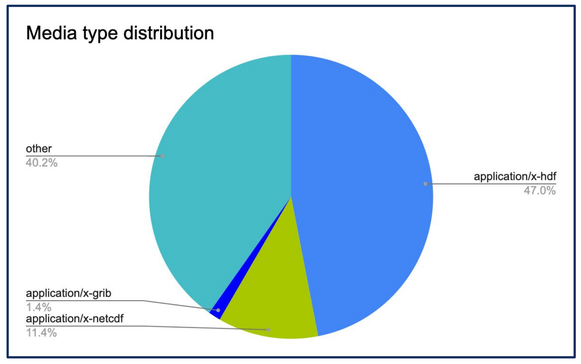
- Facilitating data searching within Campaign Store for NCAR scientists

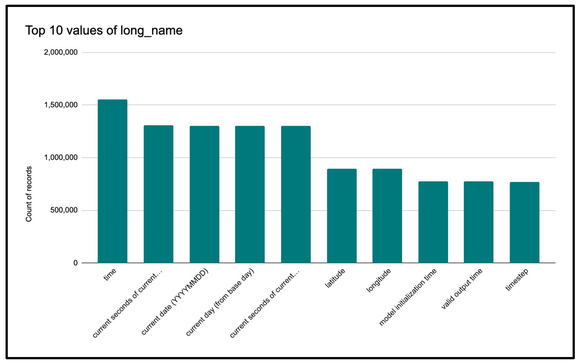
## OBJECTIVES

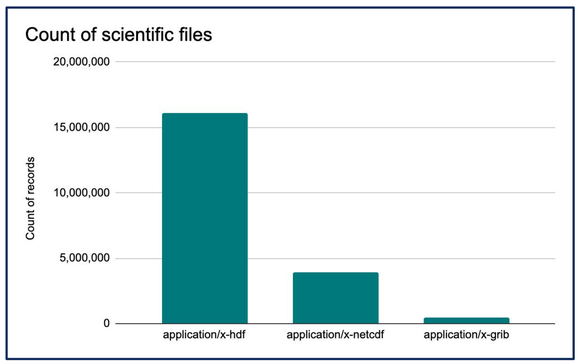
- **Index** : Index selected directories that contain a lot of scientific data
- **Restart** : Implement ability for file walkers to keep track of progress and restart
- **Update** : Improve efficiency of media type updates
- **Explore:** Explore scientific metadata extraction from NetCDF, HDF, Grib files

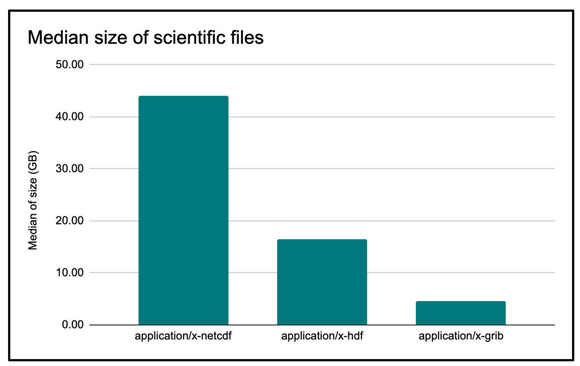
## FINDINGS

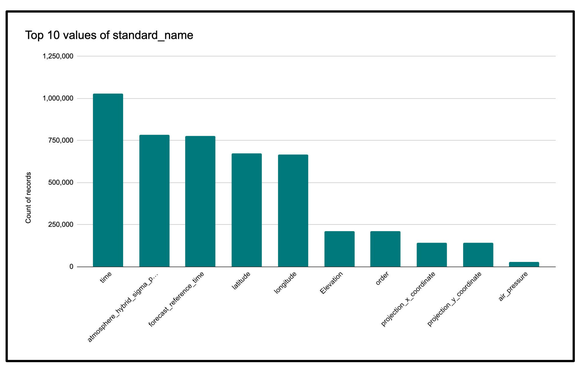


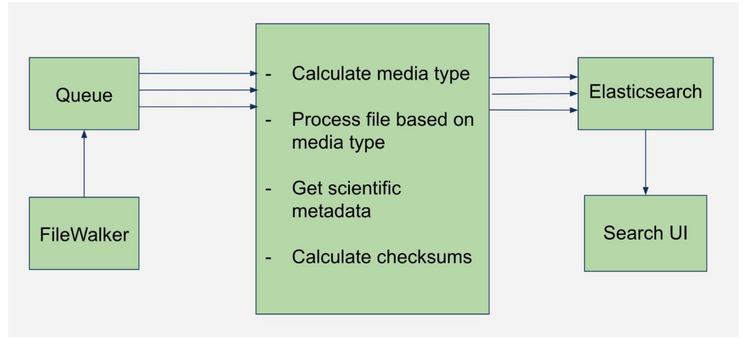




## CONCLUSION

- Implemented restart ability for the file walkers
- Indexed **486,899,650** files and **13,075,066** directories (**~37.1 PiB)**
- Calculated media type for **36,262,238** files, with max speed **120,000 updates/hr**
- **152** unique media types, **21,722,204** are hdf / netcdf / grib
- Checked **2,657,132** files for selected scientific metadata

## FUTURE WORK

- Adding functionalities and increasing efficiency





## ACKNOWLEDGMENTS

Thank you to my mentors Nathan Hook, Eric Nienhouse, and Jason Cunning for their constant support and guidance throughout this project. Thank you to Virginia Do, Jerry Cyccone, and the entire administrative team for all of their hard work to make SIParCS possible. Thank you to the NSF, NCAR, CISL, and the Sage Team for the opportunity and support during SIParCS.