CS486C – Senior Capstone Design in Computer Science

Project Description

Project Title: Intelligent Search for AWS S3 Buckets	
Sponsor Information: 	Trent Hare, Cartographer Supporting NASA's Planetary Data System Astrogeology, United States Geological Survey thare@usgs.gov

Project Overview:

The Astrogeology Science Center, part of the United States Geologic Survey, provides support for several NASA satellite and robotic planetary missions (Mars, the Moon, Europa, Titan, etc.). Not only do we support the data processing software for the gathered data, but we also support the infrastructure required to archive and serve the data to scientists, students and the public (figure 1). To be able to handle the data volume we have transitioned all our data to Amazon's AWS S3 cloud storage.

Cloud storage buckets, for example S3, are foundational to modern cloud infrastructure. They power everything from static website hosting and big data pipelines to backup systems and media delivery. Their simplicity and scalability make them ubiquitous for large data sets and have been embraced by NASA. By focusing on S3, the most widely adopted object storage service and heavily used by NASA and the USGS, this project addresses a critical need for visibility, searchability, and security in cloud-native environments.

This project aims to build a platform that:

- Index publicly accessible AWS S3 buckets that we host for NASA.
- Enable intelligent search across bucket contents and metadata.
- Visualize bucket contents and allow users to tag a collection of files in an intuitive dashboard.
- Enhanced: audit security configurations to identify exposure risks.

Scope and Features

1. **Public Bucket Discovery**
   - Use known buckets.
   - Validate access using ListObjectsV2 and GetObject permissions.
2. **Metadata Indexing**
   - Extract object keys, sizes, last modified dates, and MIME types.
   - Optionally parse file contents (e.g., TXT, JSON, XML, PDF) for keyword indexing.
3. **Search Engine**
   - Implement full-text and metadata search using Elasticsearch or Meilisearch.
   - Support filters by file type, keyword, size range, and modification date.
4. **Visualization Dashboard**
   - Display bucket metadata and object distributions.
   - Search results with preview and metadata insights.
   - Allow users to locally tag files for sharing with others.
   - Optional: allow for visualization in Leaflet geospatial browser.
5. **Optional: Security Configuration Analysis**
   - Parse bucket policies, ACLs, and encryption settings.
   - Flag risky configurations (e.g., public writing, no encryption, overly permissive IAM roles).

Technical Stack

Layer	Technology Recommendations
Backend	Python (FastAPI)
AWS SDK	Boto3
Search Engine	Elasticsearch or Meilisearch
Frontend	Flask or React + D3.js or Svelte
Visualization	Optional: Geospatial, Leaflet (CartoCosmo fork)
Database	Database like SQLite or PostgreSQL or MongoDB
Deployment	Docker + AWS EC2 or Lambda

Knowledge, skills, and expertise required for this project:

- Basic knowledge of web applications and use of modern web application frameworks.
- Basic Python coding for example Flask and/or Boto3
- Basic database knowledge, indexing, and search
- Optional: knowledge of Leaflet, the technologies it relies on, and how to modify/extend Leaflet applications.

Equipment Requirements:

- There should be no equipment or software required other than a development platform and software/tools freely available online. For example, a local Docker implementation.
- The team can use their own machines/platforms during the development phase.
- We will list several publicly available S3 buckets for testing.

Deliverables

- Working prototype for public S3 bucket search and analysis.
- Dashboard with metadata indexing and visualization. The ability to refresh index when new files are added.
- Documentation and demo video.
- All code supported in GitHub (or like).
- Optional: security analytics report.

Learning Outcomes

- Mastery of AWS S3 APIs.
- Experience with search engine integration and metadata parsing.
- Skills in full-stack development and data visualization.
- Exposure to cloud security auditing.

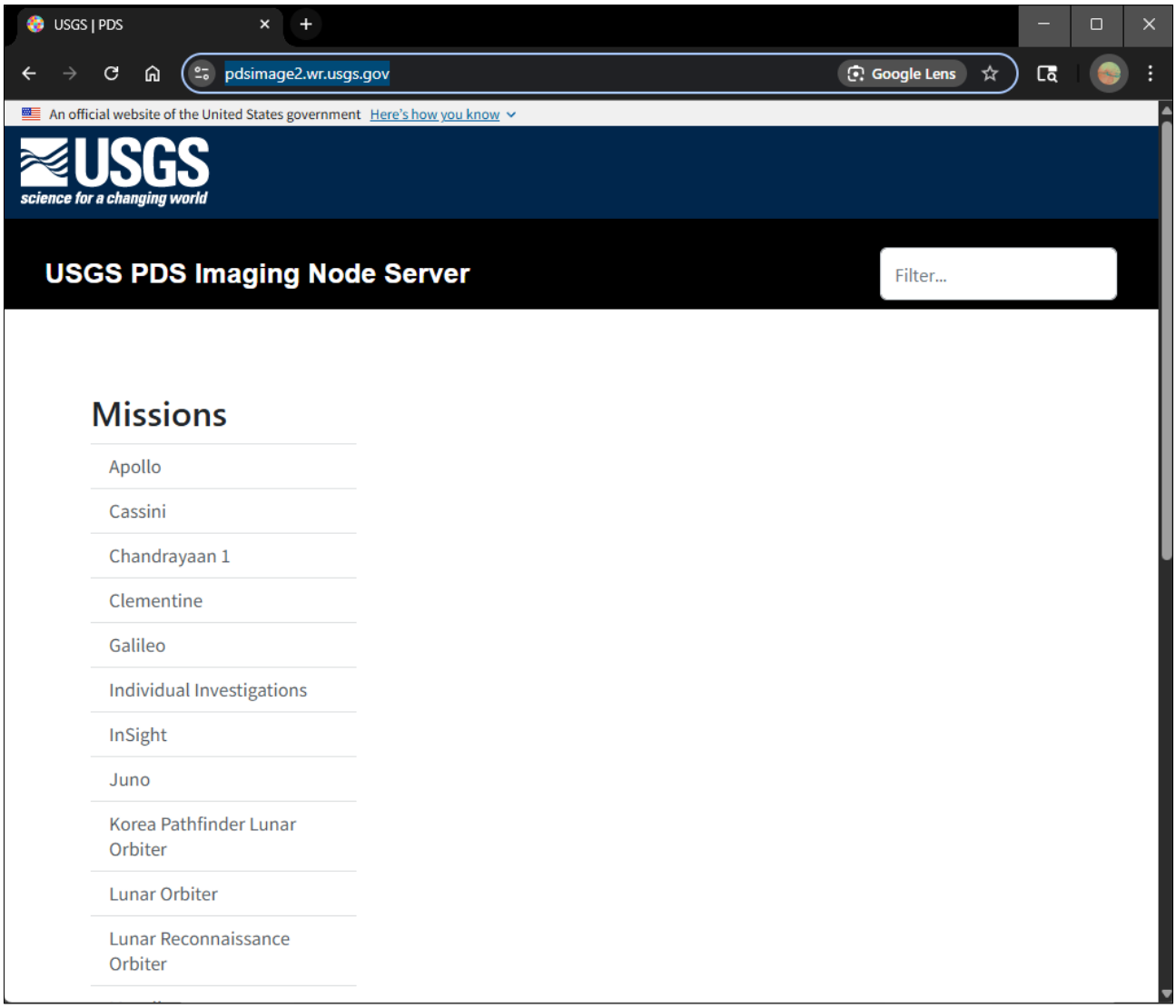

Figure 1. While we support a simple S3 Browser (https://pdsimage2.wr.usgs.gov/) there is no capability to search over the files. It would be great to have a very simple method to index the files and then support a simple search to find files.