Public-Private Partnership: Joint recommendations to improve downloads of large Earth observation data

Dr. Rahul Ramachandran
NASA/MSFC

Kevin Murphy
NASA/HQ

Dr. Chris Lynnes
NASA/GSFC

Katie Baynes
NASA/GSFC

Collaborative Effort between Amazon, Google, NASA and Microsoft

Contributors

- ***Amazon:*** *Jamie Baker, Jamie Kinney, Ariel Gold, Jed Sundwall, Mark Korver*
- ***Google:*** *Allison Lieber, William Vambenepe,* ***Matthew Hancher****, Rebecca Moore, Tyler Erickson*
- ***NASA:*** *Kevin Murphy, Rahul Ramachandran, Chris Lynnes, Katie Baynes*
- ***Microsoft:*** *Josh Henretig, Brant Zwiefel, Heather Patrick-Ahlstrom*

Overall Vision

- Enable efficient access and transfer of NASA's Earth science data to different cloud infrastructures to facilitate new data driven applications and foster new user communities

Specific Objectives

1. Shall assess NASA's Earth Observing System Data and Information Systems (EOSDIS) based on the recommendations from cloud providers
2. Shall develop an **internal** recommendations plan for ESDIS and the DAACs to improve NASA's Earth Observing System Data and Information systems (EOSDIS) based on assessment
3. Shall jointly develop a "***recommendations/best practices document***" that can be shared via USGEO with other organizations and agencies serving Earth science data

Methodology

- Amazon, Google and Microsoft provided NASA a set of “use cases/recommendations”
  - Covered not only how Earth observations are or will be used by these cloud providers but also documented their experiences interacting with data systems at different federal agencies and other organizations.
- Use cases/recommendations collated into a common set of recommendations
  - Recommendations have been used to assess the existing capabilities of NASA EOSDIS, identify existing gaps and plan for future improvements
- **Generalized these recommendations for the USGEO draft document**
  - Within the ***context of Common Framework for Earth-Observation Data***

Internal Assessment (Objectives 1&2)

- Select examples
- Presented as four components in subsequent slides
  - Recommendations/Suggestions from Cloud Providers
  - NASA (EOSDIS) Components Affected
  - Internal Assessment (by NASA team)
    - Reviewed existing system components on our own and/or reached out to leads
  - Internal Action items (by NASA team)

Catalog Mechanism

Recommendations from Cloud Providers

- Provide complete dataset metadata listing, or catalog file, for download as JSON
  - Should list all the data products, granules and access URLs
- Utilize an efficient hierarchy where needed to minimize paging

Components Affected

- Individual metadata catalogs at the DAACs
- Common Metadata Repository (CMR)

Internal Assessment

- CMR does not currently provide the full metadata in JSON format
  - Can access granules for data products via CMR API
  - Data access URL quality of the actual metadata records is poor
- CMR response does not provide nested Collection and Granule metadata records

Recommendations to ESDIS

- CMR provide complete output in JSON format
- RESTification of the CMR catalog
- Perform CMR quality assessment to clean up the data access URLs
- *Bulk download data scripts should always query the CMR catalog instead of pointing to individual metadata catalogs or file directories (Ensures uniform experience)*

Data Versioning

Recommendations from Cloud Providers

- Metadata field required to flag when dataset is COMPLETE
- Metadata field required to flag when the dataset has CHANGED
- Version number should be modified whenever the data has been updated

Components Affected

- Unified Metadata Model (UMM)

Internal Assessment

- Element <UMM-C:COLLECTION PROGRESS/> with 3 states - PLANNED, IN WORK, COMPLETE (not a required element) could be used to address this
- CMR API could be queried for Collection/Parameter Updates
- Could use <UMM-C: VERSION/> element but has consistency issues

Recommendations to ESDIS

- Collection Progress element should be a required element in UMM-C
- Create an additional CMR API functionality to find out which granules have been updated since a particular date
- Create a working group to suggest recommendations on consistent versioning at data collection and granule level

Generalized Recommendations

Common Framework for Earth Observation Data

- Section 2: Data Search and Discovery Services
- Section 3: Data Access Services
- Section 4: Data Documentation Services
- Section 5: Other Services

2.1 Search and Discovery Services

Recommendations

- Provide a single, authoritative catalog for federated data systems
- Ensure the catalog contains a complete and accurate representation of the data holdings
- Provide a complete dataset metadata listing for download in a simple machine-readable format
    - List all available data products, data granules with data access URLs
    - Utilize an efficient hierarchy where needed to minimize paging

Rationale

- Data that cannot be discovered through search via the catalog does not exist.
- For bulk downloads, machine-parseable catalog file is a simple, scalable, and robust mechanism that can reduce search and access times

2.2 Data Notification Services

Recommendations

- Develop a lightweight push notification feed to notify users of new datasets and granules, and version updates.
  - Be based on an HTTP post request to a URL that is registered to a user
  - Distribute a machine parseable notification message
  - Allow users to subscribe to a notification feed for individual datasets

Rationale

- Push mechanism places control of the curated archives directly in the hands of the data provider and eliminates polling latency
- Notification message with enough details will allow the user to immediately download and ingest the new or updated data if desired

3 Data Access Services

Recommendations

- Provide data access for file downloads without restrictions or with basic authentication headers (avoid stateful authentication mechanisms that require a login form)
- Allow parallel downloads.
  - Limit on parallel downloads - ensure that the limit is high enough to complete download in a reasonable amount of time.
  - Avoid requirements that all connections in a session come from the same IP address. Designate a technical point of contact to address questions concerning download limits

Rationale

- Cloud providers have a whole distributed system for managing the retrieval of data from partners at scale.
  - Does not interact well with stateful sessions, which generally operate under the assumption that there is a single machine at the other end of the line

4.1 Metadata: Data Version Information

Recommendations

- Provide metadata fields that flag when a dataset is complete and also flag when the dataset has changed.
- Ensure the version number in metadata fields is modified whenever the data has been updated

Rationale

- Important for data users to know when a particular dataset is complete and when it has changed.
  - Ensures users do not use stale data
  - Quickly determine if they have the most current version of the data

4.2 Metadata: Data Integrity

Recommendations

- Ensure metadata records include fields for both the file size and a checksum (preferably MD5) for each granule file.
  - Checksums can be used to track updates to granule files
- Provide a complete listing of files for datasets that include a variable number of tiles so that a user can determine which tiles are legitimately missing.

Rationale

- Metadata record should allow the user to verify whether they have obtained an accurate and complete copy of the data.
- File size, in conjunction with a strong checksum, can serve as an initial verification during the download stage and prevent unnecessary downloading and reprocessing of unchanged files after a version update.
- List of legitimately missing tiles clarifies for a user whether all available data has been downloaded.

4.3 Data: Formats

Recommendations

- Utilize standard file formats and minimize the use of esoteric file formats

Rationale

- Standard file format does not require special readers and as such can be quickly and easily used.
- Standard formats also allow data to be imported into major GIS packages without additional effort.

4.4 Data: Projections

Recommendations

- Utilize standard (common) map projections and coordinate systems where possible
- Provide and support software tools to handle specialized projections and coordinate systems (GDAL, Proj4, etc.)

Rationale

- Utilization of common coordinate systems and map projections minimizes the possibility of misuse of data and failure of tools during data use

5. Other: Technical Support Services

Recommendations

- Provide a clear point of contact for technical support and other questions related to data access mechanisms and networking issues.
  - Communicate directly with network engineers at cloud providers to support robust high-bandwidth connections.
  - Ensure TCP/IP configurations on servers are optimized to maximize transfer rates.
- Establish a communication channel such as a mailing list to provide updates on upcoming datasets, information about outages, and other relevant news related to data feeds

Rationale

- Clear point of contact enables cloud providers to distribute data with high reliability and ensures a response occurs in a timely manner.
  - Rapid response time is important due to the large number of end users that may be indirectly affected

Other Long Term Recommendations

- Provide mechanisms to enable cloud users to directly access data
- Enable semantic annotations on the metadata catalog to improve discoverability and use of data, especially for cross-domain users

Questions

Contact Information

Dr. Rahul Ramachandran

NASA/MSFC ZP11

rahul.ramachandran@nasa.gov

(256)961-7620