Interagency Implementation and Advanced Concepts Team (IMPACT)
Earth Science Data and Information System (ESDIS)

NASA’s data traceability study

Rahul Ramachandran, **Manil Maskey**, Chris Lynnes
NASA

Arun John, T. Mukherjee
University of Alabama in Huntsville

CEOS WGISS-53
March 22-24, 2022

Summary: NASA Earth Observation data traceability study

Scope:

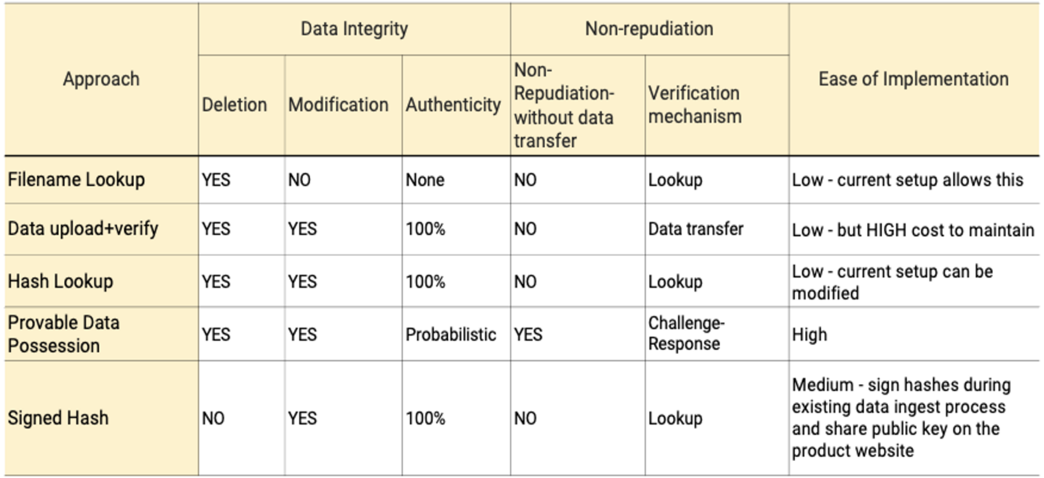
- Investigate problems that arise due to duplication and mirroring from a data user’s perspective
- Explore a range of technical solutions to address this data traceability problem
- Focus only on addressing a set of specified use cases

Study Facets:

- Data Integrity Checks
  - Authenticity
  - Data Deletion
  - Data Modification
- Non-repudiation
- Ease of Implementation
- Performance

Approaches:

- Filename lookup
- Data upload and Verify
- Hash lookup
- Provable data possession
- Signed hashes



Approach	Data Integrity			Non-repudiation		Ease of Implementation
	Deletion	Modification	Authenticity	Non-Repudiation-without data transfer	Verification mechanism	
Filename Lookup	YES	NO	None	NO	Lookup	Low - current setup allows this
Data upload+verify	YES	YES	100%	NO	Data transfer	Low - but HIGH cost to maintain
Hash Lookup	YES	YES	100%	NO	Lookup	Low - current setup can be modified
Provable Data Possession	YES	YES	Probabilistic	YES	Challenge-Response	High
Signed Hash	NO	YES	100%	NO	Lookup	Medium - sign hashes during existing data ingest process and share public key on the product website

Key takeaways:

- No approach addresses all the identified needs, both hash lookup and signed hashes are viable practical approaches to address some of the issues
- Hash lookup is suitable if the data user wants to process the full archive of a data set for any kind of large scale processing need

Basic Definitions

- Authoritative source: an entity that is authorized to develop or manage data for a specific purpose
    - Can also be an actively managed repository of valid or trusted data that is recognized by a community and supports data governance and stewardship practices.
- Trusted source can be a service provider or an organization that publishes data obtained from a number of authoritative sources
    - Are often compilations and subsets of the data from more than one authoritative sources
    - They are "trusted" because there is an "official process and/or agreement" for compiling the data from authoritative sources, data is actively managed, and documented
- Science users typically obtain data from authoritative or trusted sources.

Problem statement

- Cloud computing driving authoritative and trusted sources concept
- A data lake is a core component in cloud-based analysis paradigm providing a flexible data store to address data processing and analysis needs at scale.
  - Ideally there exists only one authoritative or trusted data lake on a cloud platform
- Each agency or data producing organization is using their own cloud platform based on their constraints or some are still using on premise infrastructure
  - Data products are being replicated and mirrored at different cloud platforms by third party actors
- Data duplication by third party actors, without proper data stewardship is threatening to create data swamps.
  - More critically it lacks data curation with little to no active data management throughout the data life cycle and little to no contextual metadata and data governance

Use Case Scenarios

- Ideal
  - all the authoritative data required is available on the same cloud platform
  - user only requires access to the cloud storage and notification mechanism
- Worst case
  - different data products are unavailable on the cloud platform
  - burden falls on the user to replicate the archive from different authoritative sources
  - user now requires a uniform scalable data access mechanism
- Common case
  - some data products may be available on the selected cloud platform, but their provenance is unknown.
  - user now has to not only move the unavailable data but also verify the authenticity of available data
  - in addition to a uniform approach for data access for the missing data, the user needs a uniform approach to verify the authenticity, completeness, correctness and consistency of the data available on the cloud

Actors

- Data Producers/Data Distribution Centers – source of the authoritative or trusted data
- Data Stewards – organizations within authoritative sources charged with the collection and maintenance of authoritative data
- Data Users – Consumers of the data
- Third Party Actors – Individuals or organizations replicating the data
- Verifier - Service or an actor that validates the integrity of the data

Study Facets

- Data Integrity Checks – three types:
  - Authenticity: A data user should be able to determine if the data file they accessed from a third party repository is from an authoritative source. For example, a user should be able to verify that a specific file is genuinely from NASA's data systems.
  - Data Deletion: A data user should be able to check if the file they accessed from a third party resource has been deleted at the authoritative source. For example, the original data product that has been mirrored to a cloud platform has undergone a major version update and the older version is no longer considered fit for use.
  - Data Modification: A data user should be able to check if the file they accessed from a third party resource has been modified or tampered with.
- Non-repudiation – the assurance that someone cannot deny the validity of something. For data users, it is the ability to prove their use of authoritative data despite claims to the contrary.
- Ease of Implementation – a qualitative assessment of the effort data producers and data stewards will need to expend if they adopt a specific technical approach.
- Performance - the overhead of extraneous data movement in or out of the cloud or of computing verification parameters must not outweigh the analysis performance gains.

Approaches

Filename Lookup

- filename is checked, not the contents

Data Upload and Verify

- users upload the files they accessed from a third party source to a verification service provided by the authoritative source
- involves high data movement, making it expensive and time consuming

Hash Lookup

- hash values for files are then stored in an authoritative source’s catalog or in a file that is accessible
- user can query the catalog or download the file for the hash and compare it with the hash generated for the file they have accessed from the third party source for verification

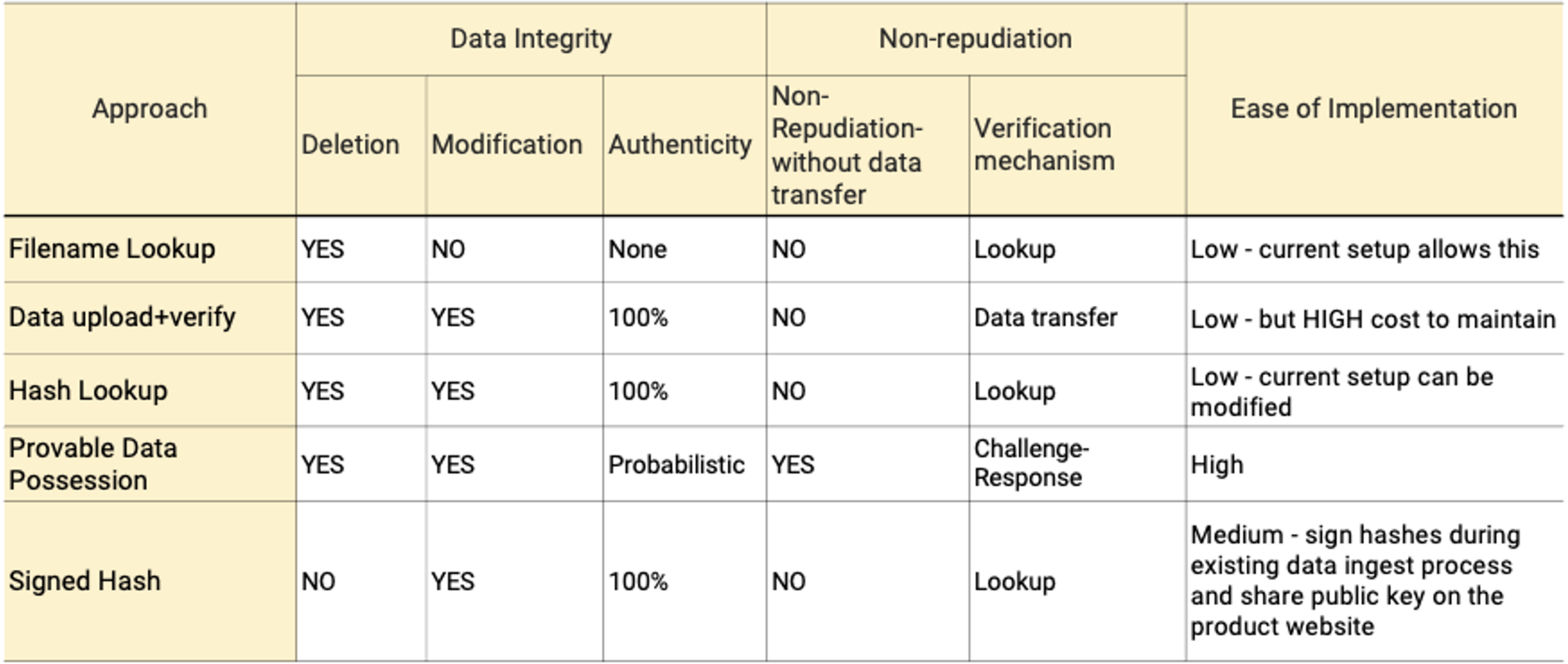
Provable Data Possession

- combines public key cryptography and cryptographic hashing. It has three steps: (1) producer key generation, (2) tag generation, and (3) verification

Signed Hashes

- instead of requiring the entire file to be signed, only the hash generated for the content is signed using the private key
- data users can authenticate the hash signature by using the public key made available by the authoritative source and verify the integrity of the file itself by comparing the signed hash against the hash they generated

Result Summary Table



Approach	Data Integrity			Non-repudiation		Ease of Implementation
	Deletion	Modification	Authenticity	Non-Repudiation-without data transfer	Verification mechanism	
Filename Lookup	YES	NO	None	NO	Lookup	Low - current setup allows this
Data upload+verify	YES	YES	100%	NO	Data transfer	Low - but HIGH cost to maintain
Hash Lookup	YES	YES	100%	NO	Lookup	Low - current setup can be modified
Provable Data Possession	YES	YES	Probabilistic	YES	Challenge-Response	High
Signed Hash	NO	YES	100%	NO	Lookup	Medium - sign hashes during existing data ingest process and share public key on the product website

Key Takeaways

- Both hash lookup and signed hashes are viable practical approaches to address some of the issues
- Hash lookup is suitable if the data user wants to process the full archive of a dataset for any kind of large scale processing need. Full listing of the files along with their hash values is required for completeness, correctness, and consistency checks
- Signed hashes are suitable when the data users only want subsets of the archive for analysis and completeness is not an issue. The data users can verify the data authenticity using the public key from the authoritative source and validate the integrity using the hash value
- It is the data user’s responsibility to verify the authenticity, completeness, correctness, and consistency of the data they are using