Managing data reliability and integrity in federated cloud storage

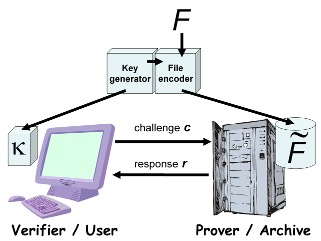
Krzysztof Styrc ACC CYFRONET
kstyrc@gmail.com

Promotor:
dr inż. Marian Bubak

Data validation is a well explored area:

- hash functions (MD5, SHA-1, etc),
- message authentication codes (MAC),
- error correcting codes (ECC).

However, emerging popularity of cloud storage services carry on new challenges. Storing vast amounts of data on external resources will probably result in network ineffciencies when calculating the whole file checksums. Proof of Retrievability (POR) algorithms try to solve this performance issues.



Long-term persistence of medical data sets requires reliability and integrity mechanisms to be built on top of Cloud storage. DRI is designed to fullfill these requirements by performing the following tasks:

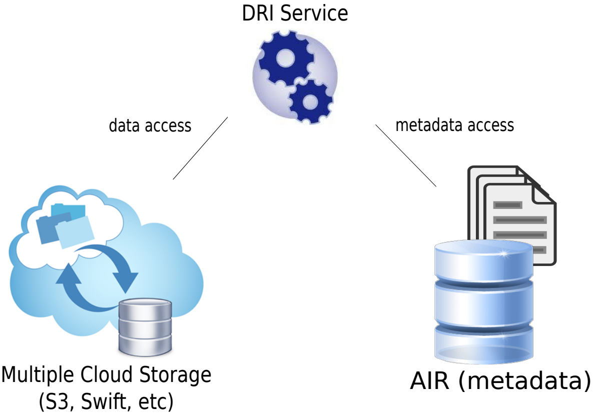
- periodic and request-driven integrity checks on data sets,
- facilitating storage of multiple copies of data on various Cloud platforms,
- tracking the history and origin of binary data sets.

Datasets validation:

- availability of each dataset at its (multiple) locations,
- the integrity of each dataset's file (checksum-based).

DRI Service:

- stateless REST WebService,
- built on top of AIR registry and multiple of cloud storage,
- self-running daemon (periodic checks) and API access,



Current functionalities:

- supported requests: dataset validation and adding dataset under management (computing its checksums),
- integrity checks based on SHA-256 of the first 1KB of a file – for fast testing on large datasets,
- asynchronous operations added to queue of a simple scheduler,
- report notifications (validation success/failure or finished adding dataset under management) as emails,

Current REST API:

- ***add_dataset_under_management/*{*datasetID*}** – asynchronously computing dataset's files checksums and notify by email,
- ***validate_dataset/*{*datasetID*}** – asynchronously validate dataset and notify by email.

On-going work:

- further integration with AIR registry,
- support for S3 storage besides Swift,
- efficient validation algorithm research,
- periodic validation based on scheduler,

Efficient validation – DRI challenge:

- the number and size of stored datasets can be vast,
- existing Cloud storage API's (Swift, S3) flexibility is limited,
- serving data – primary service , validation – secondary service,
- need for efficient and effective validation algorithm.