Rucio - S3-compatible access interface

IRIS-HEP Fellow Final Presentation

Kyrylo Meliushko

Mentors: Lukas Heinrich (TUM), Matthew Feickert (UWM), Mario Lassnig (CERN), Martin Barisits (CERN)
Agenda

1) Understanding Rucio
2) Project overview and goals
3) Required knowledge and my work
4) Personal takeaways
Figure 1: The cumulative ATLAS data volume approaches 500 Petabytes in early 2020. Growth has been linear with respect to the scale of the experiment, with considerable data deletion before longer observation periods.
Understanding Rucio

Scalable Data management

Data Replication

Data Distribution

Data Provenance

rucio.cern.ch
Understanding Rucio

Project overview and goals (S3)

S3 - Simple Storage Service

Lightweight & easy
Scalable
Widely used

Portable way to share less-important experiment data/notes between analysts

S3
Project overview and goals

Rewind: Interface to interact with S3 directly from Rucio.

Weeks 1-3: Explore Rucio, MinIO's "mc" tool, S3 API. Prepare development environment.

Weeks 4-5: Implement basic data transfer capabilities between Rucio and S3.

Weeks 5-7: Integration testing, credentials extraction.

Weeks 7-10: Additional features: access sharing, quotas, …

Weeks 10-12: close GitHub Issues, refine docs, prepare final presentation.

July 3 - Sept 22
Required knowledge & my work

- S3 -> python boto3 library
- General Rucio architecture

- Working with Python script as CLI
- Pytest framework -> remote testing with Docker
- General S3 usage, exceptions when calling API’s
- Triggering other Rucio clients (DIDClient, DownloadClient), as well as exploring LFN2PFN for reproducible scope:path naming
- Loading and caching of S3 Credentials
- Understanding of communication between code of Rucio components
Required knowledge & my work (Changes)

```
rucio s3 make-bucket user.lheinric:someuniquename

  →  rucio bucket create user.lheinric:/someuniquename/

rucio download user.lheinric:someuniquename/file.root

  →  rucio download user.lheinric:/someuniquename/file.root

rucio s3 credentials > ~/.mc/config.json

  →  removed since raises a lot of security issues
```
Required knowledge & my work (Structure)

General S3 implementation ->

DIDClient
UploadClient
S3Client
CredentialClient
DownloadClient

...
Required knowledge & my work

(Structure)

S3Client tests

Configuration loading
Personal takeaways

- Pytest: integration and unit tests
- Test driven development (TDD)
- Optimal architecture for big projects
- S3 and it’s identity and access management
- Improved python knowledge, remote debugging skills
- General Rucio understanding

...
Thank you for your attention!