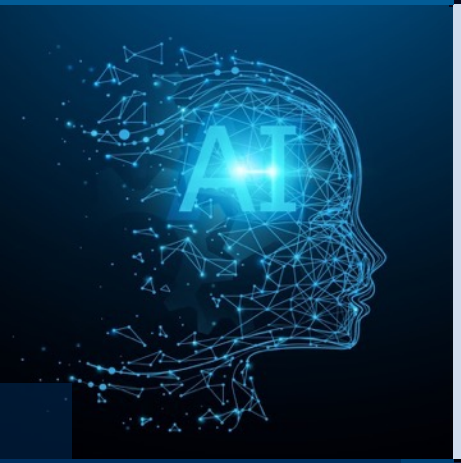
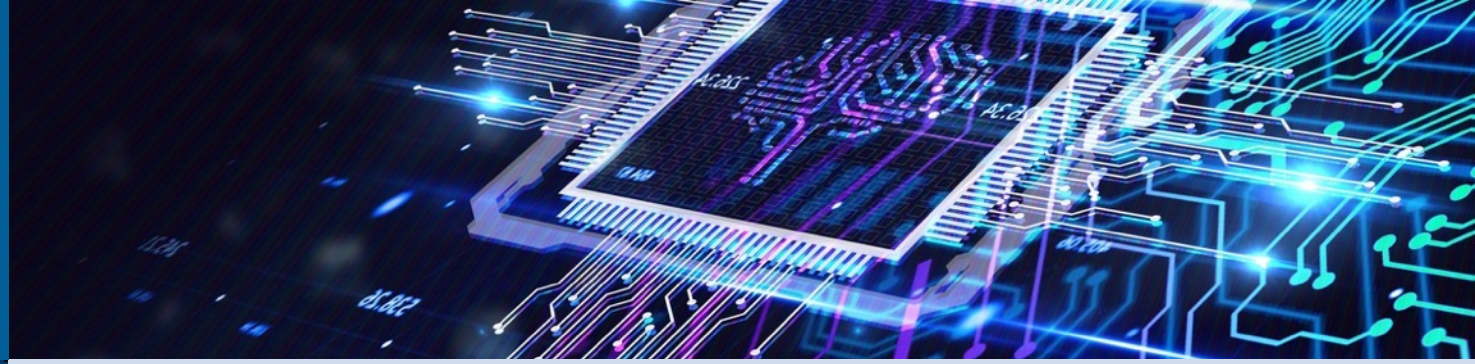
CSC

ICT Solutions for Brilliant Minds





Allas object storage service

Juha Törnroos, CSC – IT Center for Science Ltd.

20th SIG-NOC meeting, 07.05.2024

CSC cloud service offering for research

cPouta

Community Iaas cloud

General purpose

Accesible over internet

Powered by Openstack
ISO27001 Certified
In production, since 2013

WebUI, CLI & REST APIs supported

openstack.

ePouta

Community IaaS Cloud

Sensitive data

Accesible only from customer network
Powered by Openstack
ISO 27001 Certified
In production, since 2013

WebUI, CLI & REST APIs supported

openstack.

Rahti

Community PaaS Cloud leveraging containers
General purpose
HTTP(s) applications

Accesible over internet

Powered by Openshift OKD
In Open beta

Web UI, CLI & REST APIs supported

OPENSHIFT

Allas

Cloud storage service

General purpose

Accesible over internet

Powered by CEPH

In production, since 2018

CLI, S3 & Swift APIs

ceph

Data pool

or

Data warehouse

or

Data lake?

Data pool

”A data pool is a centralized repository of data where trading partners (retailers, distributors, or suppliers) can obtain, maintain, and exchange information about products in a standard format”
- IBM

Data warehouse

"A data warehouse is an enterprise system used for the analysis and reporting of structured and semi-structured data from multiple sources, such as point-of-sale transactions, marketing automation, customer relationship management, and more."

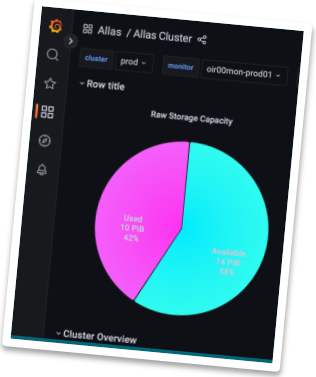
- Google

Data lake

"A data lake is a centralized repository that ingests, stores, and allows for processing of large volumes of data in its original form."

- Microsoft

Allas is a (object) storage service

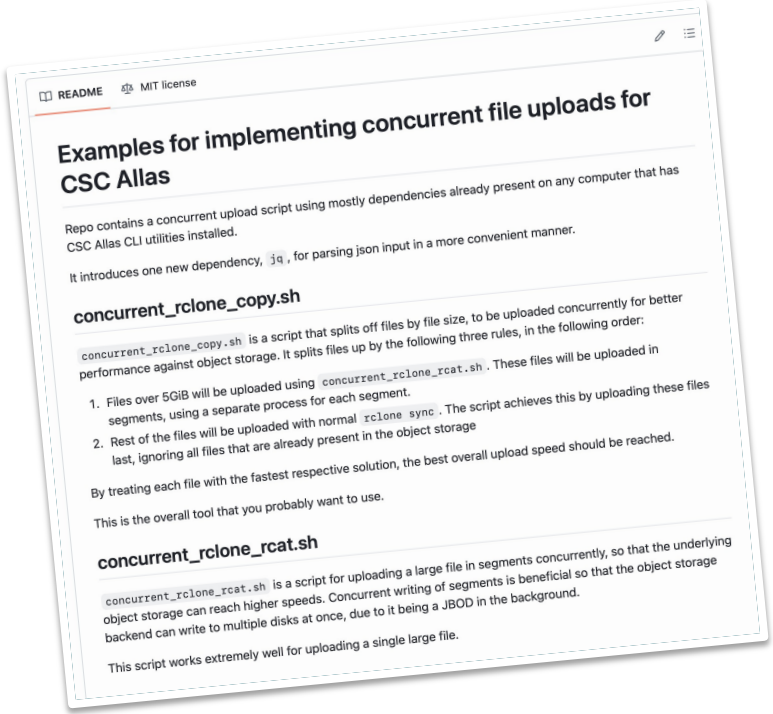


Some use cases

Main tasks for the new Allas GUI (use cases: https://wiki.eduuni.fi/x/Z6qqDQ)

- Browser Upload
- Browser Download
- Browser Create/Delete Bucket
- Browser Share/Unshare Bucket To Another Project
- Browser Share/Unshare Bucket With Open Read Access
- Browser Bucket Search
- Browser Object Search
- Browser Bucket Sorting
- Browser Bucket/Object Sorting

Some solutions



README MIT license

Examples for implementing concurrent file uploads for CSC Allas

Repo contains a concurrent upload script using mostly dependencies already present on any computer that has CSC Allas CLI utilities installed.

It introduces one new dependency, `jq`, for parsing json input in a more convenient manner.

concurrent_rclone_copy.sh

`concurrent_rclone_copy.sh` is a script that splits off files by file size, to be uploaded concurrently for better performance against object storage. It splits files up by the following three rules, in the following order:

1. Files over 5GiB will be uploaded using `concurrent_rclone_rcat.sh`. These files will be uploaded in segments, using a separate process for each segment.
2. Rest of the files will be uploaded with normal `rclone sync`. The script achieves this by uploading these files last, ignoring all files that are already present in the object storage

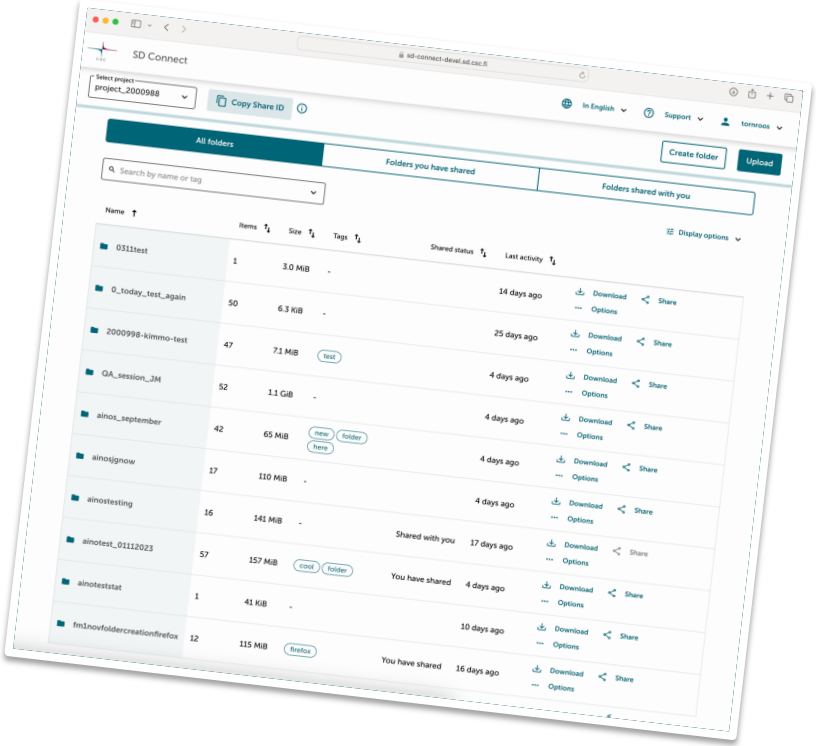
By treating each file with the fastest respective solution, the best overall upload speed should be reached.

This is the overall tool that you probably want to use.

concurrent_rclone_rcat.sh

`concurrent_rclone_rcat.sh` is a script for uploading a large file in segments concurrently, so that the underlying object storage can reach higher speeds. Concurrent writing of segments is beneficial so that the object storage backend can write to multiple disks at once, due to it being a JBOD in the background.

This script works extremely well for uploading a single large file.



And structural improvements

Interview structure

Created by Riina Salmivalli, last modified on Jan 18, 2022

Most relevant tasks to investigate:

- Browser file Upload
- Browser file Download
- Browser Create/Delete Bucket
- Browser Share/Unshare Bucket To Another Project
- Browser Share/Unshare Bucket With Open Read Access

Most relevant things to learn

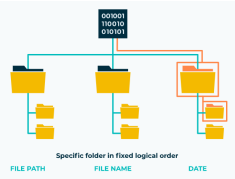
- what functions user assumes to have in general data mng tool
- how the user thinks GUI fits the user need
- what kind of factors/functions user finds important to support the usage
- what kind of factors the user feels insecure or unsure about
- how to user understands the terminology used

The user testing structure

1. Introduction for the testing session
2. Short semi-structured interview
3. Assignments and UX metrics
4. Additional questions
5. Wrap up

Assignment 3. Uploading a file (incl.encrypting)	SCS +1 / +0 / -1	Single ease question 0-6	Time on task	Other remarks
Assignment. Next upload the test file "testfile1" I sent you. In this case it includes sensitive data and should be encrypted while uploading. *Tell me when you know the upload has been completed.* *(In case user is unsure, tell him/her to refresh)* • *Does user understand the encrypt function* • *System progress visibility* • *Refresh function*				

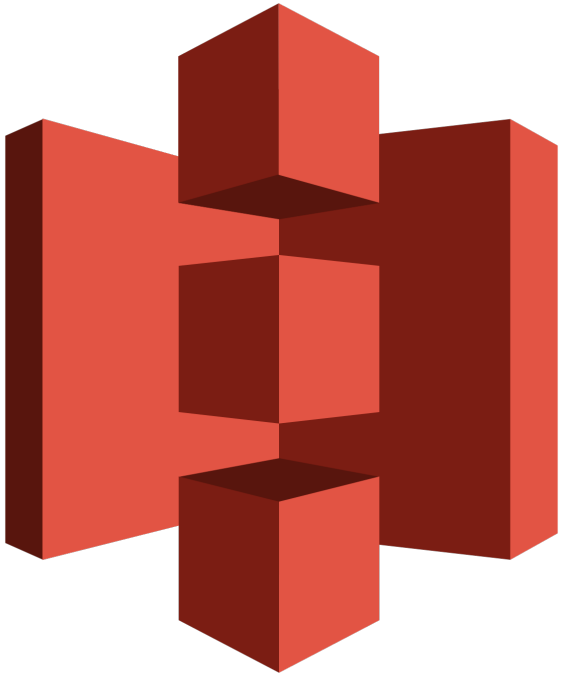
Different kind of storage services



- File storage
  - stores and organizes data as a single piece of information in a folder to help organize it among other data, similar to physical files stored in a paper filing system in an office.
  - is also called hierarchical storage.
  - when you need access to data, your computer system needs to know the path to find it.

- Object storage
  - saves files in a flat data environment, or storage pool, as a self-contained object.
  - does not actually use directories, folders, and other complex hierarchical organization, even if it can look like that.
  - works best for static storage, especially for unstructured data, where you write data once but may need to read it many times.

And different kind of object storages




WRAP
IT UP!