

Use of CESNET/e-INFRA CZ data storage for laboratories and individual researchers

CESNET, as part of the Czech e-INFRA CZ infrastructure provides research data storage services, intended for scientific and academic users in the Czech Republic. It is useful for the storage of large volumes of scientific data, sharing between laboratories/institutions, supporting computational workflows and institutional data backup.

It is suitable for data volumes that are too large for standard local disk arrays (here I make a special pointing to microscopy data).

Individual users can request personal S3 object storage with a standard quota of 2TB. Laboratories/group/core facilities storage must be arranged through CESNET support, set up as a virtual organization.

They explicitly state that S3 object storage is not designed for long-term storage and publication, mainly because it's not a curated research record (no persistent identifier, rich metadata, license information, ...). For the purpose of long term and data publication, the National Repository Platform (NRP) is recommended (or other repositories)

CESNET storage definition

CESNET Data Storage department offers their services for storing large-scale **unstructured scientific data**. They support archiving and backup purposes.

It is primarily object-based and geographically distributed across Czech data centres. Data centres are in Jihlava, Plzeň, Ostrava, Brno, and Dolní Břežany, with raw capacities of approximately 25–34 PB each and certification under ISO/IEC 27001.

S3 buckets can be accessible by users using S3 clients including graphical tools, which makes the service more user friendly (3 Browser, CloudBerry Explorer, AWS CLI, Rclone, S3cmd, S5cmd, WinSCP, and CyberDuck)

More information on the specifics can be found here:

<https://docs.du.cesnet.cz/en/docs/introduction/description>

Prospective users

Institutions/organizations connected to the CESNET infrastructure (academics, students, staff of Czech research institutions) for **research and educational purposes**.

Storage capacity and specifics for user tiers

Individual users

Personal S3 storage with 2TB quota, which can be increased on request. Access is granted with a simple request (https://einfra.cesnet.cz/allfed/registrar/?vo=VO_s3&group=s3_cl4).

Sharing is not allowed between users with personal storage.

Laboratories, facilities and groups

Larger groups can make use of Group S3 storage, prepared individually through a Virtual Organization (VO), where a VO manager can invite and manage users (documentation for management here: <https://docs.du.cesnet.cz/en/docs/perun/user-management>).

The service is designed for group or cross-institution sharing and can be used for tasks such as lab/project storage, shared datasets, backup, processing workflows or medium data retention.

To set up this option, the responsible user should contact CESNET support for group storage.

For a team producing larger datasets, this means:

1. Define expected data volume, lifetime and access pattern. Users should consider needs
2. Contact CESNET support via support@cesnet.cz to request a group/VO S3 service with an appointed VO administrator.
3. VO administrator defines user groups, buckets, permissions and data handling rules.

Even though CESNET does not currently put limitations on storage length, such restrictions may apply in the future. Also, they state that storage resources cannot be reserved for future use unless agreed with CESNET.

CESNET provides an accounting system where users or VO administrators can monitor quota, usage and number of objects. Laboratories should define who is responsible for monitoring the quota and reacting before limits are reached.

(Full rules available here: https://du.cesnet.cz/en/provozni_pravidla/start)

Storage time

- Temporary data – maximum storage period of 365 calendar days per file/object. Exceeding this period may result in data removal by data storage administrator (it is not clear how do they decide what is temporary and what medium term)
- Medium-term data storage – quotas are set by the administrator (file/object volume and number of files quota).
- Long-term data storage (archival) primarily served by repositories within the NRP (metadata-rich, DOI, citable).

Another interesting point on storage is the membership lifecycle – CESNET e-INFRA accounts are renewed yearly. Data stored by a user is retained for the duration of the membership of the given user and for at least 6 months after the membership ends. In case of VO, is the administrator of the VO who decides whether the data should be transferred to another member, deleted or handled in a different way (https://du.cesnet.cz/en/provozni_pravidla/start).

What is the data stored suitable for? Use cases

CESNET S3 object storage can be used for active scientific workflows, so:

- Data storage
- Data sharing (for VO, not for individual storage)
- Data processing (e.g. in Metacentrum jobs - <https://docs.metacentrum.cz/en/docs/data/storage-department>): CESNET S3 can also be integrated into MetaCentrum computational workflows. Data can be staged from an S3 bucket into a job working directory and results can be uploaded back to S3 after processing.
- Automated backups (e.g. from our synology drives - <https://docs.du.cesnet.cz/en/docs/object-storage-s3/synology>)

Data protection

CESNET storage is designed with redundancy, but it cannot be taken as a backup. Data centers are distributed throughout Czech Republic, protecting the data against hardware failures, natural disasters and other errors. Backups are of course maintained (source information: <https://docs.du.cesnet.cz/en/docs/introduction/description>) .

CESNET operated in “best effort” mode – that means they make efforts to ensure data availability and security, but no additional guarantees. If additional requirements need to

be set, it should be negotiated with CESNET support (https://du.cesnet.cz/en/provozni_pravidla/start).

So, what is the level of protection for a given laboratory bucket should be defined with CESNET support.

S3 distinguishes two redundancy approaches (<https://docs.du.cesnet.cz/en/docs/object-storage-s3/s3-use-cases>):

- Replicated storage: data are handled in 3 copies within the data center. This may improve read performance because data can be retrieved from multiple locations, but write performance can be slowed down. This is suitable for smaller data volumes where speed is important.
- Erasure-coded storage: data are split into fragments and distributed across the storage system. This is suited for larger data volumes.

Advanced features of the CESNET S3 Data storage

There are several advanced features of the CESNET S3 data storage that could be interesting for the laboratories/facilities. Here are some listed. The full list available at:

Object versioning

By default, data stored in CESNET S3 is unversioned, so a deleted object is permanently removed and cannot be recovered. However, it can be set up upon agreement.

Bucket policies

VO administrator can set up who can access a bucket and which actions they are allowed to perform (read, write...)

Service accounts

They are non-personal accounts that can be set for technical support, such as backup servers, automated pipelines...

Presigned URLs

Temporary links can be generated, allowing for access to specific S3 objects without user account. This can be useful for short-term sharing (collaborators, reviewers...).

Other recommended tools

CESNET also provides collaborative tools such as FileSender and ownCloud/OnlyOffice. These may be useful for document collaboration or occasional file transfer, but they should

be treated as complementary services rather than the main storage solution for large microscopy datasets.

References

References

1. CESNET. Data Storage service page.
<https://www.cesnet.cz/en/services/data-storage-3/data-storage-13>
2. CESNET Data Storage documentation. Introduction.
<https://docs.du.cesnet.cz/en/docs/introduction/introduction>
3. CESNET Data Storage documentation. Infrastructure description.
<https://docs.du.cesnet.cz/en/docs/introduction/description>
4. CESNET Data Storage. Terms and operating rules.
https://du.cesnet.cz/en/provozni_pravidla/start
5. CESNET Data Storage documentation. S3 service overview.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/s3-service>
6. CESNET Data Storage documentation. Getting access to CESNET S3.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/s3-getting-access>
7. CESNET Data Storage documentation. Basic S3 use cases.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/s3-use-cases>
8. CESNET Data Storage documentation. Advanced S3 features.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/s3-features>
9. CESNET Data Storage documentation. S3 clients overview.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/s3-clients>
10. CESNET Data Storage documentation. AWS CLI for S3.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/aws-cli>
11. CESNET Data Storage documentation. Rclone for S3.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/rclone>
12. CESNET Data Storage documentation. S3cmd command-line tool.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/s3cmd>
13. CESNET Data Storage documentation. Boto3 Python library.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/boto3>

14. CESNET Data Storage documentation. Synology Backup to S3.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/synology>
15. CESNET Data Storage documentation. Veeam Backup Suite.
<https://docs.du.cesnet.cz/en/docs/object-storage-s3/veeam-backup>
16. CESNET Data Storage documentation. Veeam using WORM model.
https://du.cesnet.cz/cs/navody/object_storage/cesnet_s3/veeam_worm/start
17. CESNET Data Storage documentation. Perun user management.
<https://docs.du.cesnet.cz/en/docs/perun/user-management>
18. CESNET Data Storage documentation. Creating a group in Perun.
<https://docs.du.cesnet.cz/en/docs/perun/creating-group>
19. CESNET Data Storage documentation. RBD block storage service.
<https://docs.du.cesnet.cz/en/docs/object-storage-rbd/rbd-service>
20. MetaCentrum documentation. Storage Department services.
<https://docs.metacentrum.cz/en/docs/data/storage-department>
21. MetaCentrum documentation. Main documentation page.
<https://docs.metacentrum.cz/en/>
22. National Repository Platform documentation. Introduction to NRP.
<https://docs.nrp.eosc.cz/en/docs/intro-to-nrp/intro-to-nrp>
23. National Repository Platform documentation. NRP use cases.
<https://docs.nrp.eosc.cz/en/docs/intro-to-nrp/use-cases>
24. CESNET Data Storage documentation. ownCloud and OnlyOffice service guide.
<https://docs.du.cesnet.cz/en/docs/collaboration/owncloud>
25. CESNET Data Storage documentation. FileSender service guide.
<https://docs.du.cesnet.cz/en/docs/collaboration/filesender>