DISTRIBUTED MANAGEMENT AND PROCESSING OF ALICE MONITORING DATA WITH ONEDATA

Michał Orzechowski, Bartosz Baliś, Łukasz Dutka, Jacek Kitowski

ACC Cyfronet AGH
AGH University of Science and Technology Institute of Computer Science

On behalf of the ALICE Collaboration

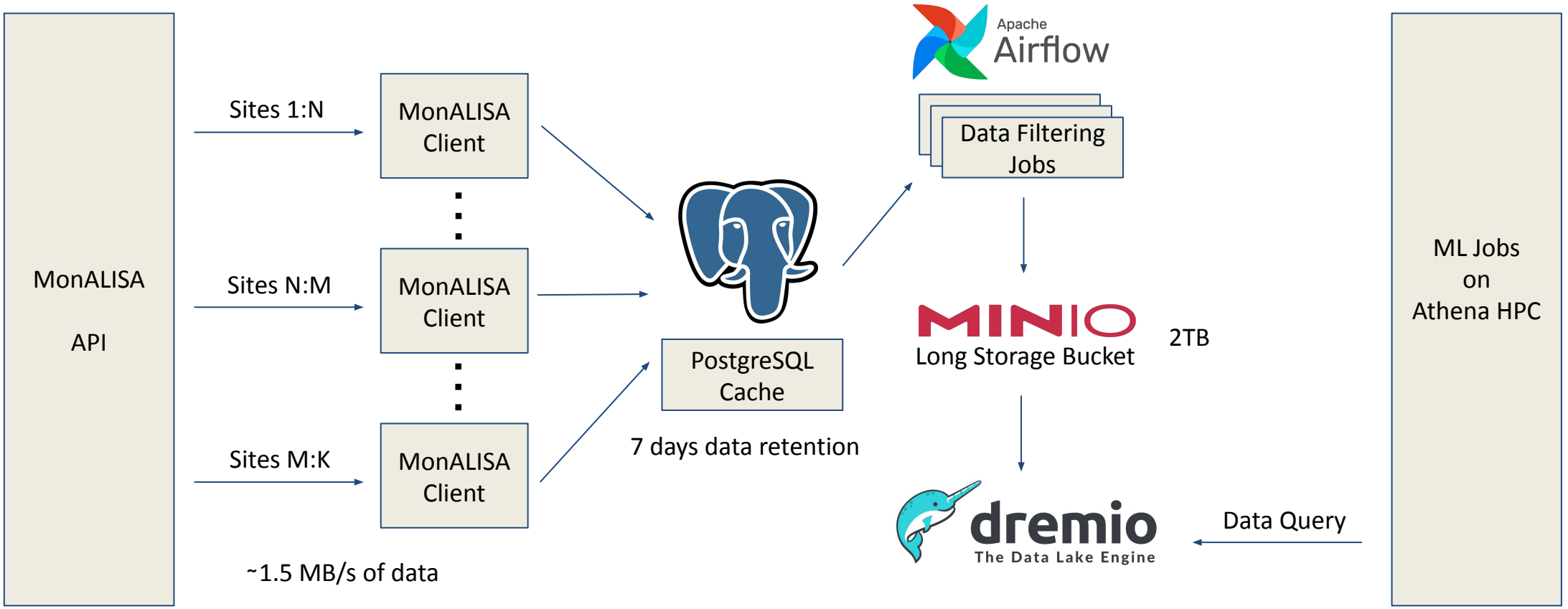
October 19 - 25, 2024
CHEP 2024

OBJECTIVES

- develop a datalake architecture capable of ingesting MonALISA data
- provide means to query data (features, time)
- make data directly available to the local HPC system for ML training[1]
- allow to share data with other data centers/clouds - at some point move ML training to CERN

1. Poster: *Towards more efficient job scheduling in ALICE: predicting job execution time using machine learning*

SYSTEM ARCHITECTURE



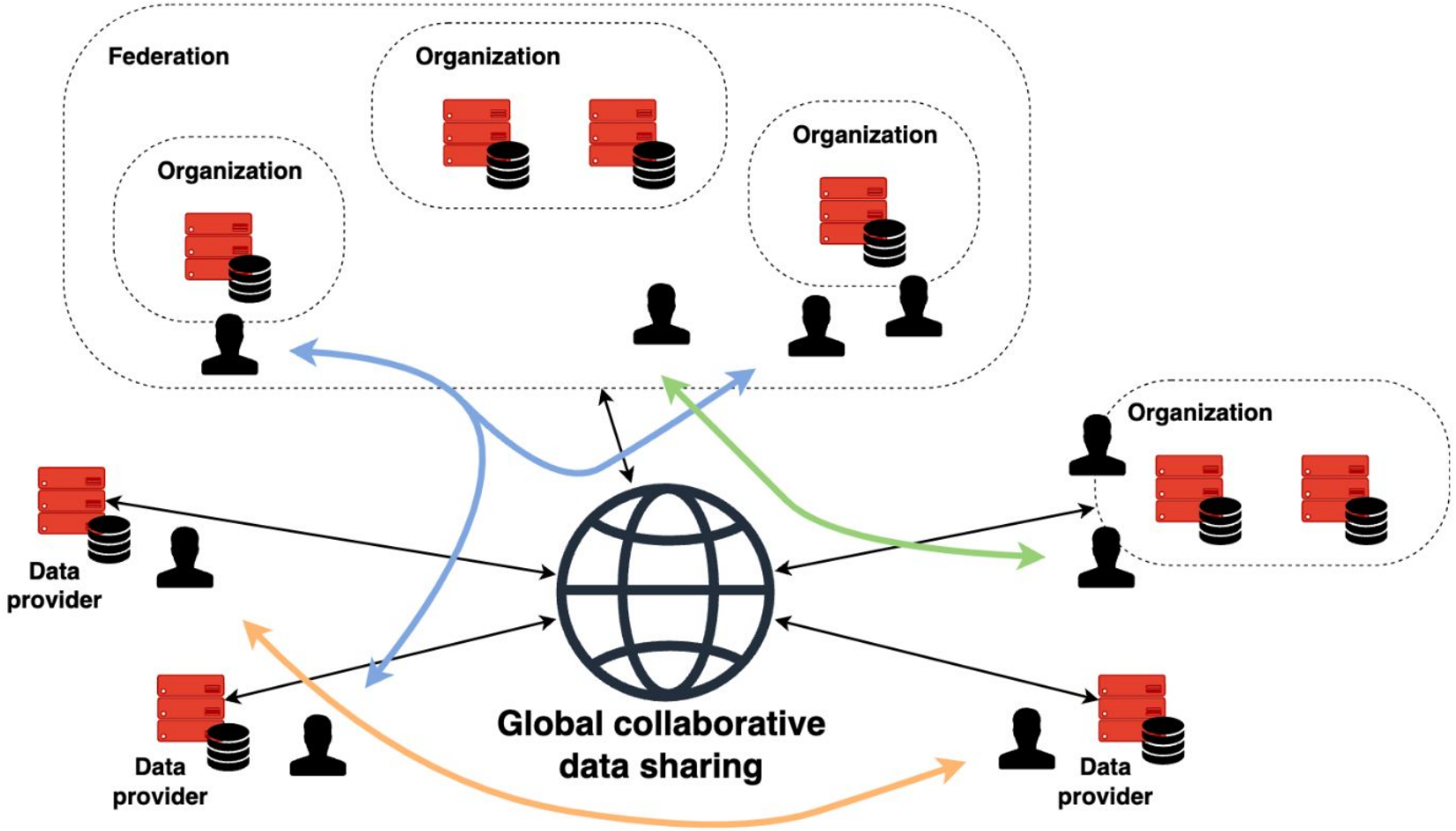
kubernetes

CYFRONET

ONEDATA - DISTRIBUTED, HIGH-PERFORMANCE DATA MANAGEMENT SYSTEM

Generic approach:

- unified access across autonomous data providers
- data sharing between organizational domains
- troublesome data transfers
- trust-driven approach privacy and security guarantees



WHO WE ARE?

ONEDATA

- 10+ years of devoted development - see github.com/onedata
- Open-source, developed at the **AGH University of Krakow** and **Cyfronet** data center
- We work tight with scientific communities on a case-by-case basic
- Our vision is to:
  - deliver a **data management** platform for large-scale and **distributed** problems,
  - address the challenges of global collaborative data sharing across **federated** organizational domains,
  - streamline data processing in **heterogeneous** data storage setups.
- Our funding comes from Polish and European grants and partnerships

SUPPORTED BY SCIENTIFIC COMMUNITIES

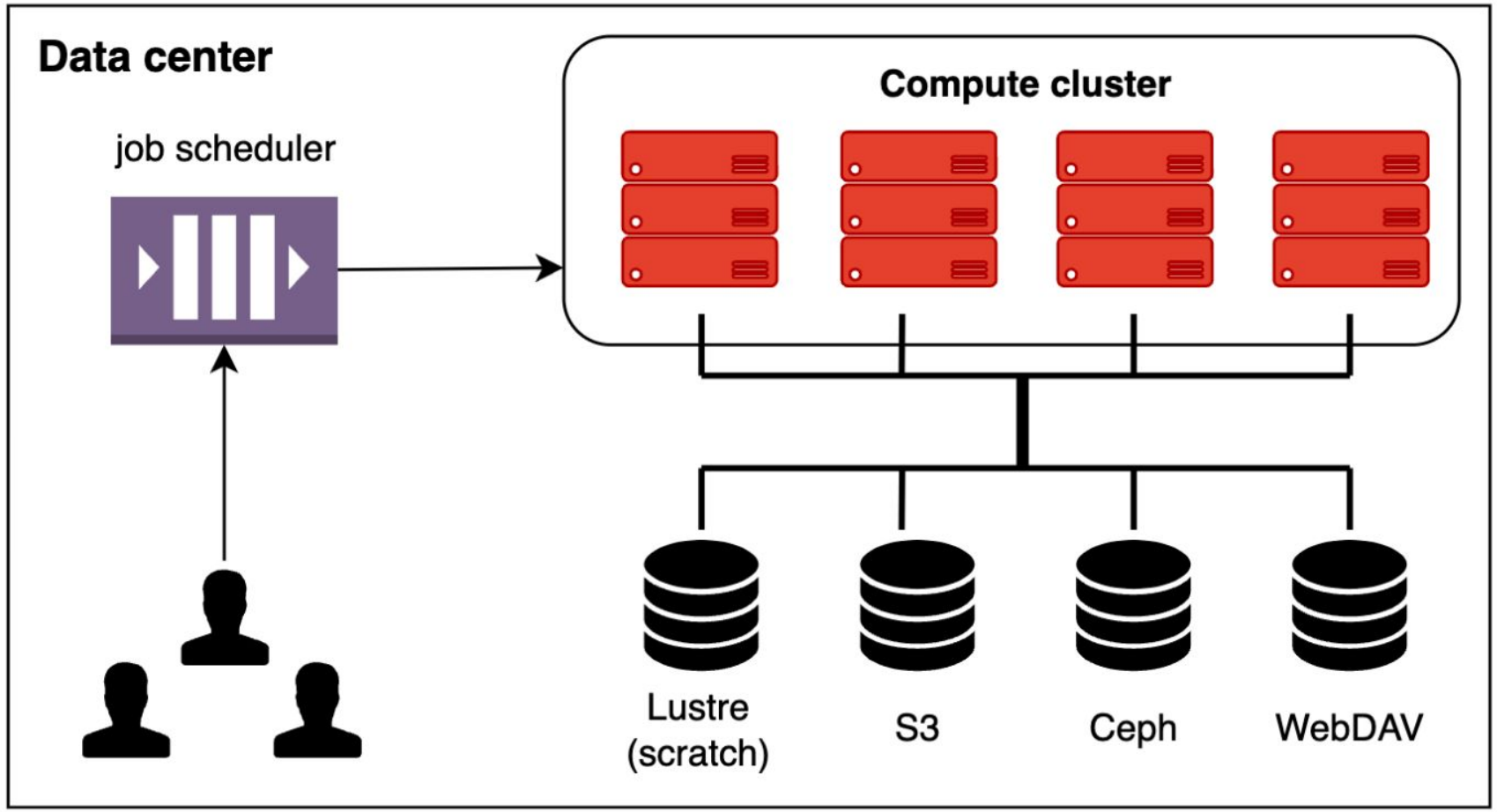
- We are always looking for new partnerships and projects in order to:
  - keep the project running (of course),
  - gain invaluable experience cooperating with experts, solving real usecases, and working on authentic large datasets (big files & large number of small files)
- Our supporters and partnerships:

ONEDATA NETWORK

- Between 5 and 10 active Zones in Poland and EU (depending on project lifecycles).
- Several instances not maintained by us.
- EGI DataHub (on the map), long haul project:
  - 20 sites (Oneproviders)
  - 2150 data spaces
  - ~1.77PB total storage size
  - 700+ users
- Archive for Polish National Museums:
  - 5PB of data — the current phase
  - 10PB of data — target scale
  - ~100M files

ONEDATA - DISTRIBUTED, HIGH-PERFORMANCE DATA MANAGEMENT SYSTEM

- heterogeneous storage systems
- manual data management
- **need for unified data access**



MULTIPLE STORAGE BACKENDS

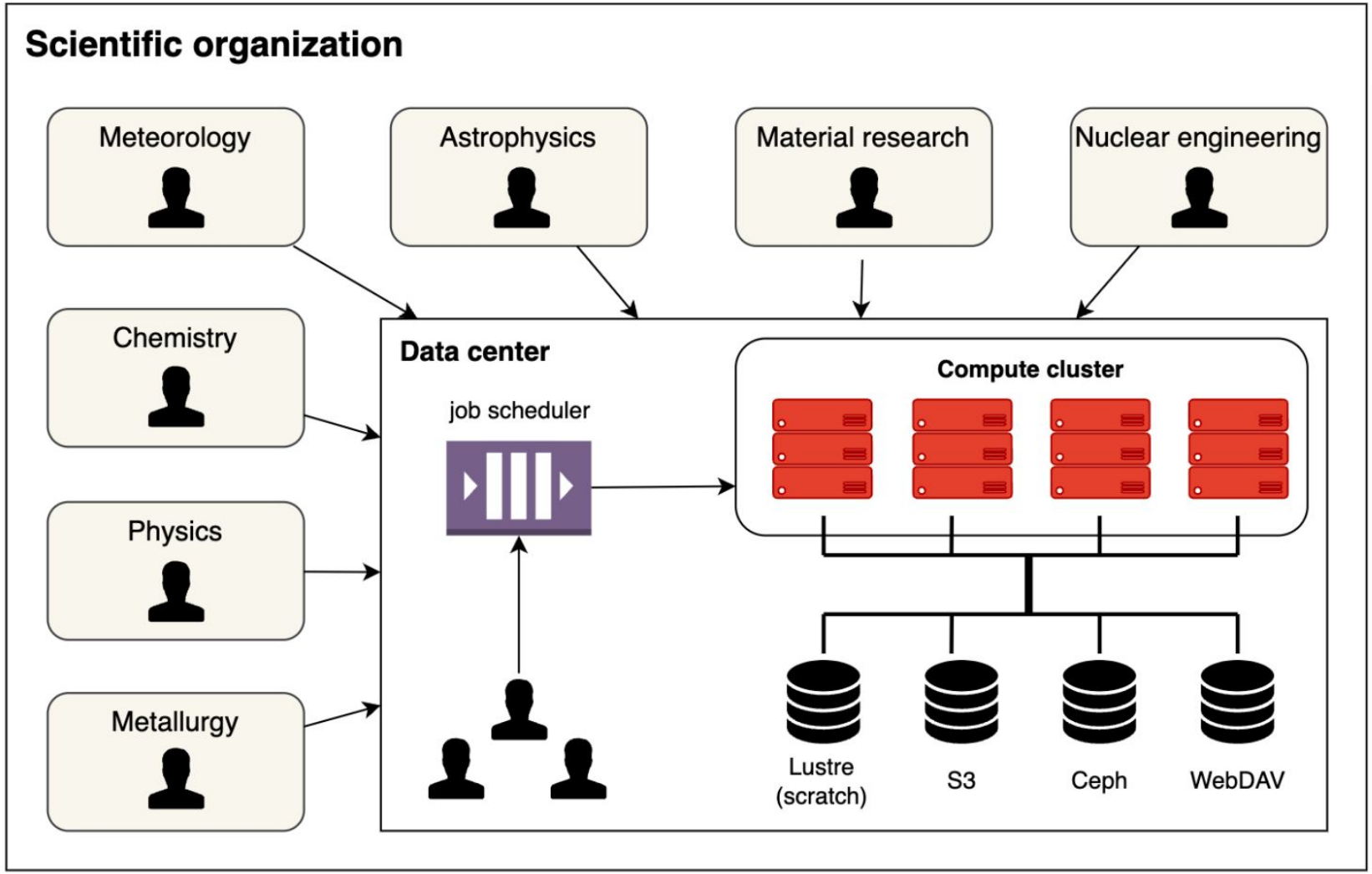
Storage backends are used to store the physical data. Oneprovider accesses the storage backends via "helpers" (drivers) implemented for each supported type of storage. Helpers serve as a POSIX-like abstraction, building a layer over different storage backend APIs and access methods.

Currently supported storage backends:

- **POSIX** — any POSIX compatible filesystem accessible by Oneprovider via a mount point
- NFS — filesystem exported via the NFS protocol — no need to mount it locally
- S3 — Amazon S3 compatible storage
- Ceph RADOS — versions 14, 15, 16
- **HTTP** — any server exposing data via HTTP or HTTPS
- XRootD — CERN's data management protocol for LHC data
- WebDAV — experimental
- dCache — experimental

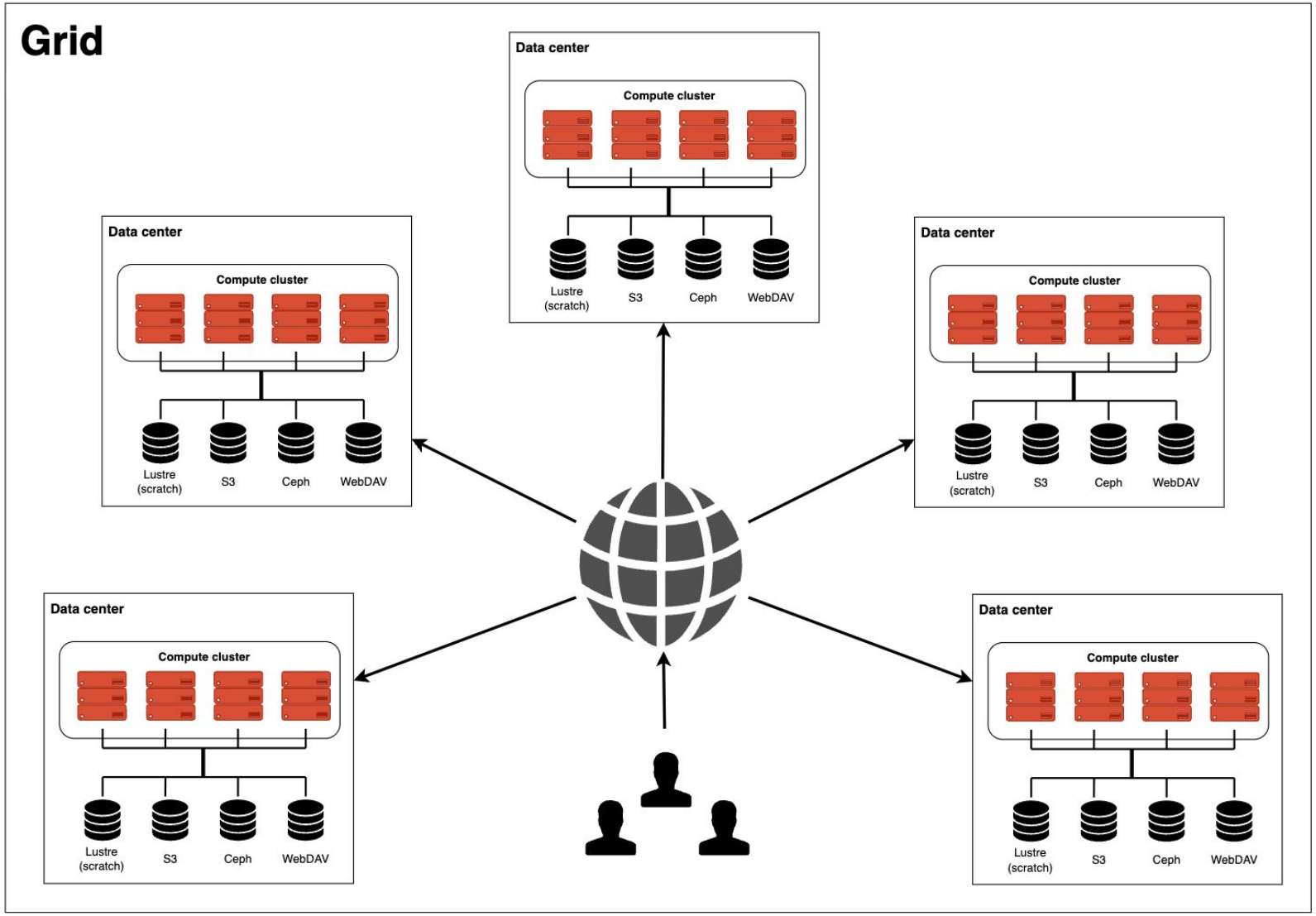
ONEDATA - DISTRIBUTED, HIGH-PERFORMANCE DATA MANAGEMENT SYSTEM

- heterogeneous storage systems
- manual data management
- required IT skills
- **need for user-friendly approach**



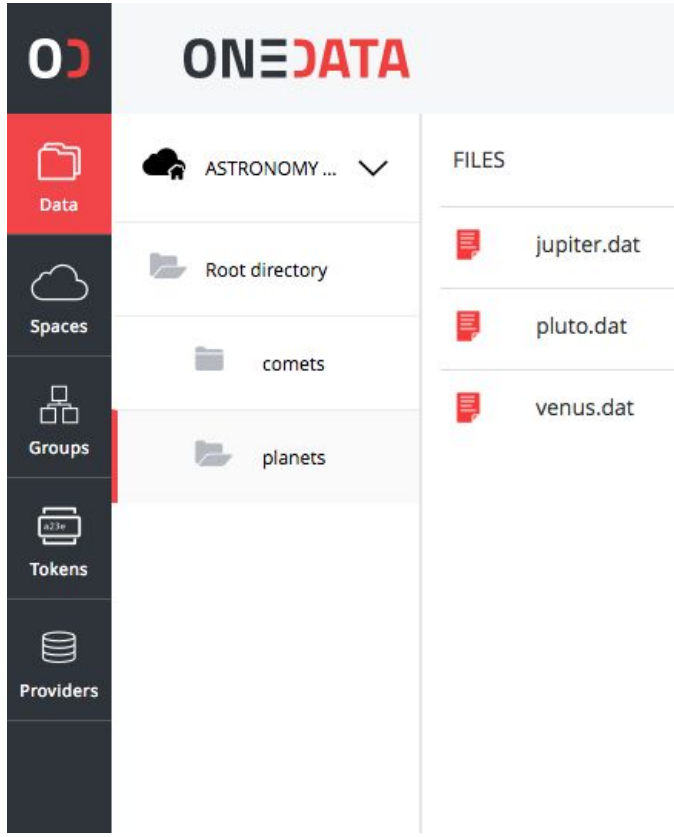
ONEDATA - DISTRIBUTED, HIGH-PERFORMANCE DATA MANAGEMENT SYSTEM

- heterogeneous storage systems
- manual data management
- required IT skills
- geographical data distribution
- troublesome data transfers
- different access control policies
- **need for suitable distributed data management tools**



POSIX ACCESS TO DATA WITH ONECLIENT

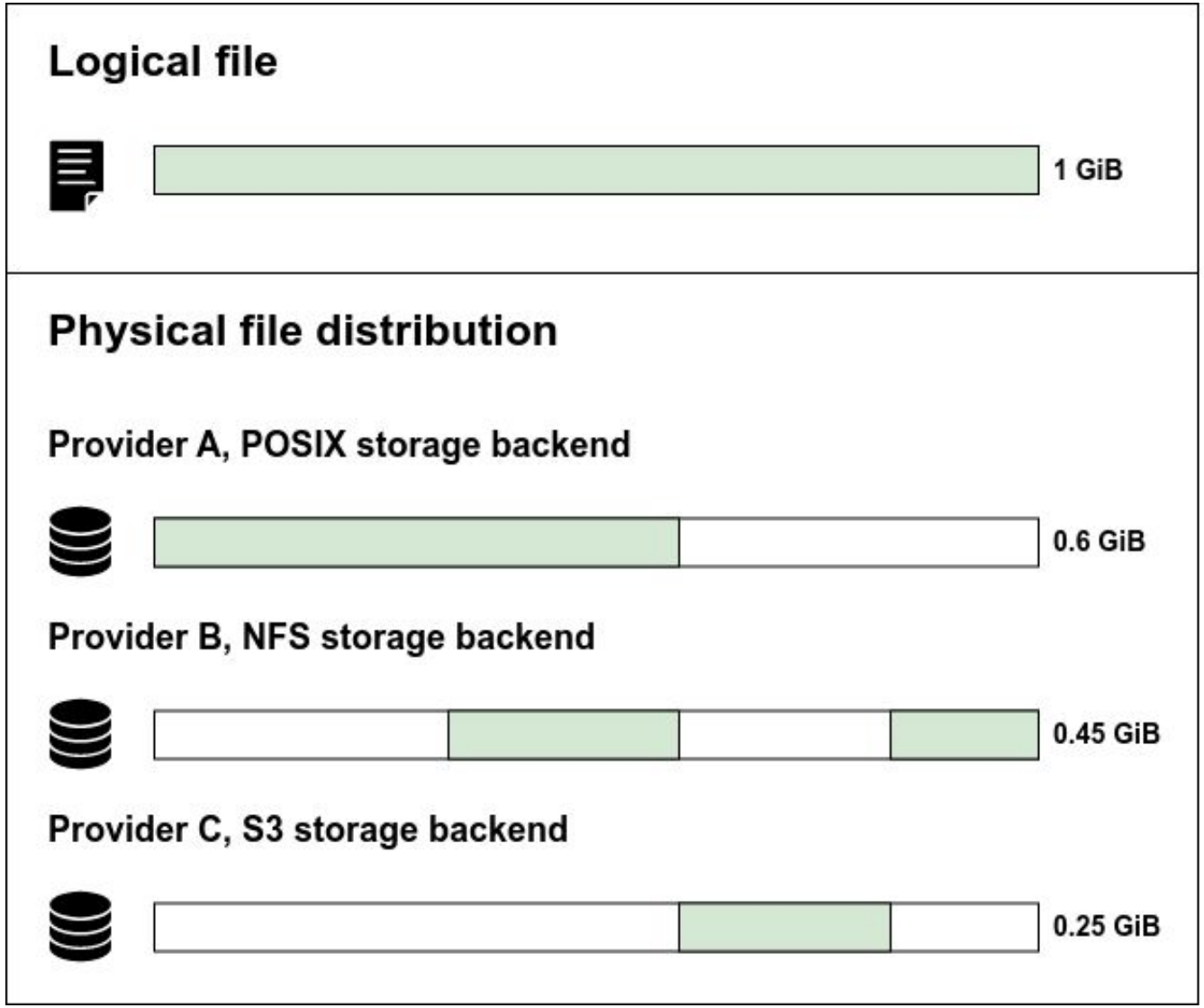
- presents Onedata virtual file system as POSIX
- support for most of the POSIX operations on globally distributed virtual file system
- all data accessible via a unified file system mountable on virtual machines, grid worker nodes and containers




```
[root@1f87c053280e oneclient]# ls
Astronomy Datasets  Big Data Experiment  Cancer Data
[root@1f87c053280e oneclient]# ls -lR
.:
total 0
drwxrwx--- 1 root 1733762 0 Sep 26 19:19 Astronomy Datasets
drwxrwx--- 1 root 1337123 0 Sep 26 19:14 Big Data Experiment
drwxrwx--- 1 root  608582 0 Sep 26 19:18 Cancer Data

./Astronomy Datasets:
total 0
drwxr-xr-x 1 1124656 1733762 0 Sep 26 19:20 comets
drwxr-xr-x 1 1124656 1733762 0 Sep 26 19:19 planets

./Astronomy Datasets/comets:
total 0
-rw-r--r-- 1 1124656 1733762 10000000 Sep 26 19:20 enck.dat
-rw-r--r-- 1 1124656 1733762 10000000 Sep 26 19:19 halley.dat

./Astronomy Datasets/planets:
total 0
-rw-r--r-- 1 1124656 1733762 10000000 Sep 26 19:07 jupiter.dat
-rw-r--r-- 1 1124656 1733762  5000000 Sep 26 19:08 pluto.dat
-rw-r--r-- 1 1124656 1733762  2000000 Sep 26 19:08 venus.dat

./Big Data Experiment:
total 0
-rw-r--r-- 1 1124656 1337123 10000000 Sep 26 19:08 cats_images.tgz
-rw-r--r-- 1 1124656 1337123  5000000 Sep 26 19:13 galaxies.img
-rw-r--r-- 1 1124656 1337123  5000000 Sep 26 19:14 spam_mails.tgz

./Cancer Data:
total 0
-rw-r--r-- 1 1124656 608582 5000000 Sep 26 19:15 brain_tumor.zip
-rw-r--r-- 1 1124656 608582 5000000 Sep 26 19:14 duct_cancer.zip
[root@1f87c053280e oneclient]#
```

DATA DISTRIBUTION

- the data in Onedata may be arbitrarily distributed among the storage backends of the supporting providers
- files are made up of parts of variable sizes — file blocks
- each provider holds a set of local file blocks, constituting a file replica
- when a file is read on a provider and the requested blocks are not present there, the missing ones are replicated on the fly from remote providers
- when a file is written on a provider, the overwritten blocks on other providers are invalidated. To read the file, the provider with invalidated blocks must once again replicate missing blocks from the provider with the newest version of the blocks



FILE DETAILS

download-files.json

Info | API | Metadata | Permissions | Shares | QoS | **Distribution**

Data distribution per storage

onedatify @ **Bari** /55eb75ab388522d56062f36fada12b40chb188/bw/download…	0 B — 45 KiB in 2 blocks — 60 KiB	75%
ceph @ **Krakow** /55eb75ab388522d56062f36fada12b40chb188/a/2/7/a271c…	0 B — 14.9 KiB in 2 blocks — 60 KiB	24%
onedatify @ **Lisbon** /55eb75ab388522d56062f36fada12b40chb188/bw/download…	0 B — 20.1 KiB in 2 blocks — 60 KiB	33%
posix @ **Paris**	0 B — 0 B in 0 block — 60 KiB	0%

This file was transferred manually 1 time – see history.

Block distribution

DATA TRANSFERS

Replicate files on demand and on the fly.

Migrate data between sites and storage backends on demand or with simple API interface.

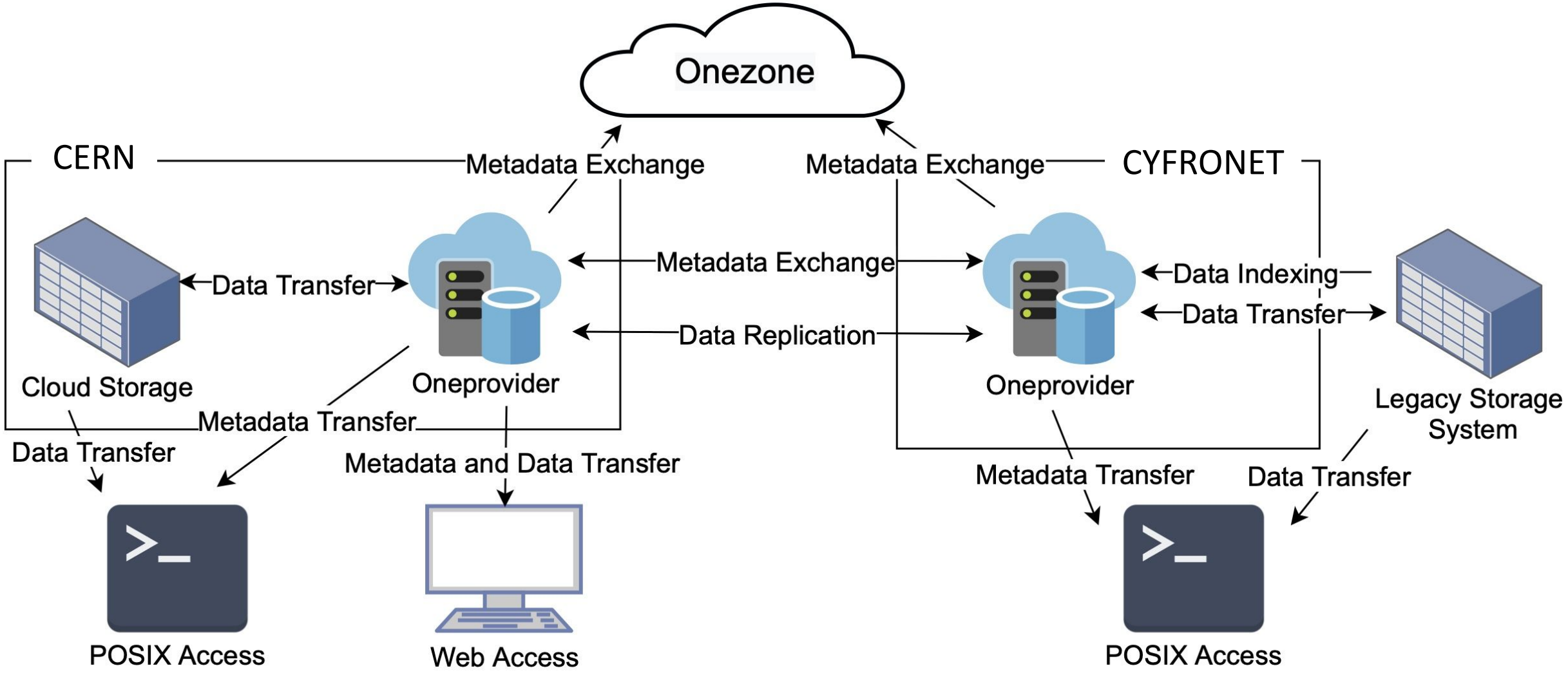
Easily check location of your data using GUI or API.

Types of data transfers:

- replication — copying (only the missing) data to achieve a complete replica at the destination. The data is copied from one or more providers holding the missing blocks.
- eviction — removing replica(s) from the specified provider. This operation is safe and will succeed only if there exists at least one replica of every block on other supporting providers.
- migration — replication followed by eviction. Replicates the data to the destination provider and then evicts the replica from the source provider.

DATA TRANSFERS

DATA INDEXING SUBSYSTEM AND DATA MIGRATION TO CERN



EXTRA RESOURCES

Improved **documentation** (in making) https://onedata.org/#/home/documentation

Dedicated **demo mode** for easy sandbox deployment:

- https://onedata.org/#/home/documentation/21.02/admin-guide/demo-mode.html

Extensive **training materials** (4 day workshop!) covering majority of Onedata:

- https://onedata.org/training

  user: training

  password: Oneworkshop58