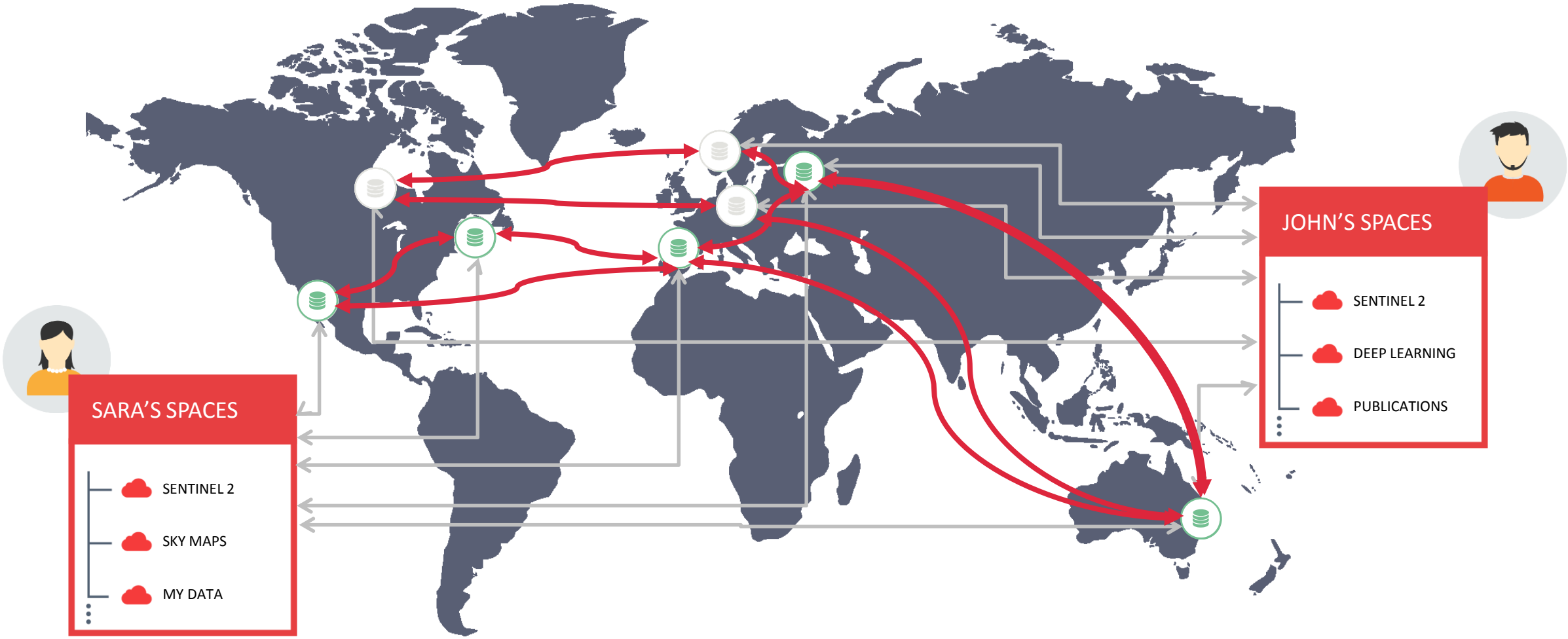
ONEDATA

# Onedata Platform for Transparent and Unified Data Access in Hybrid Clouds

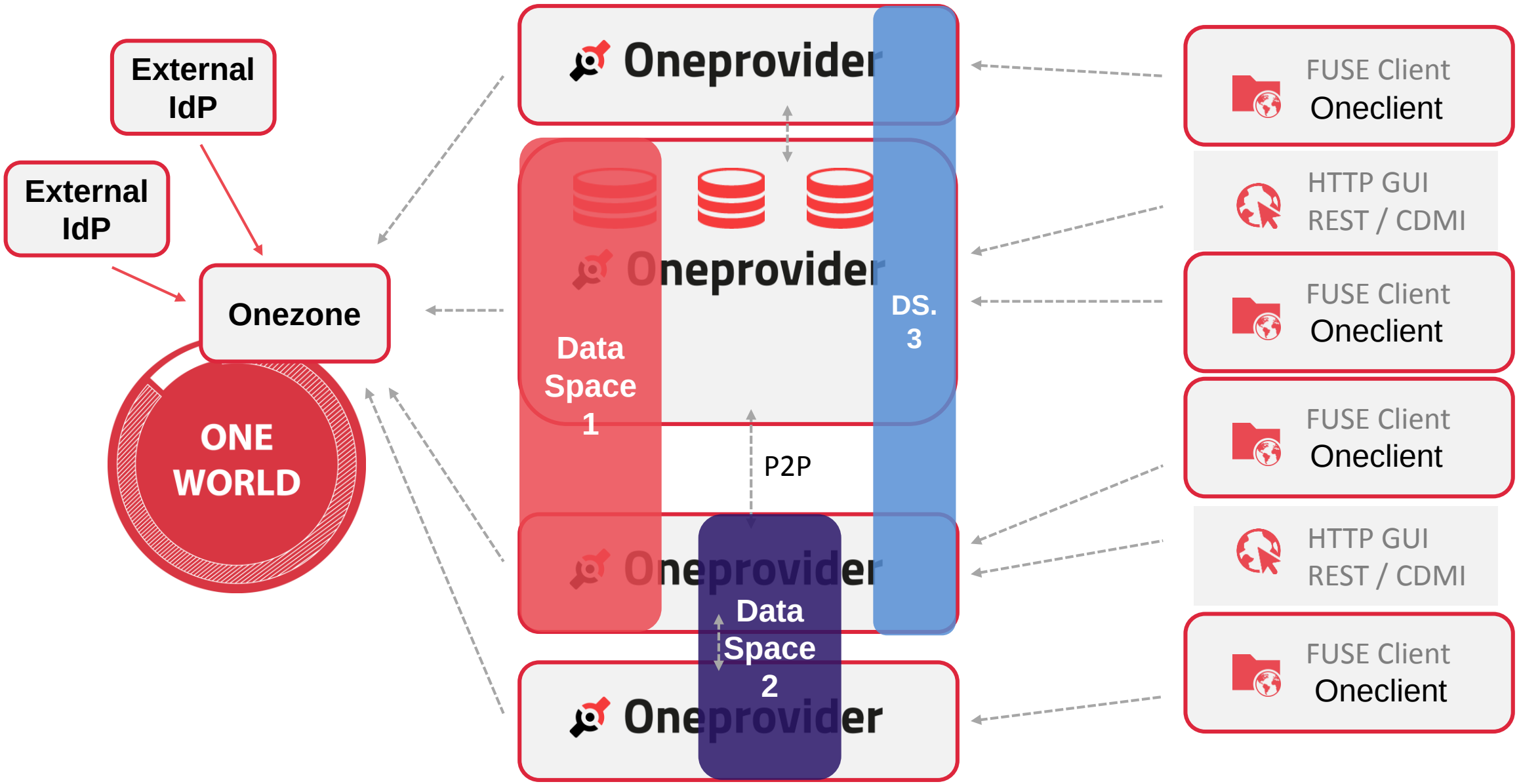
Presented by: Bartosz Kryza

# DATA IN HYBRID CLOUD ENVIRONMENTS



# ARCHITECURE OVERVIEW

# ONEDATA ARCHITECTURE



# MAIN FEATURES

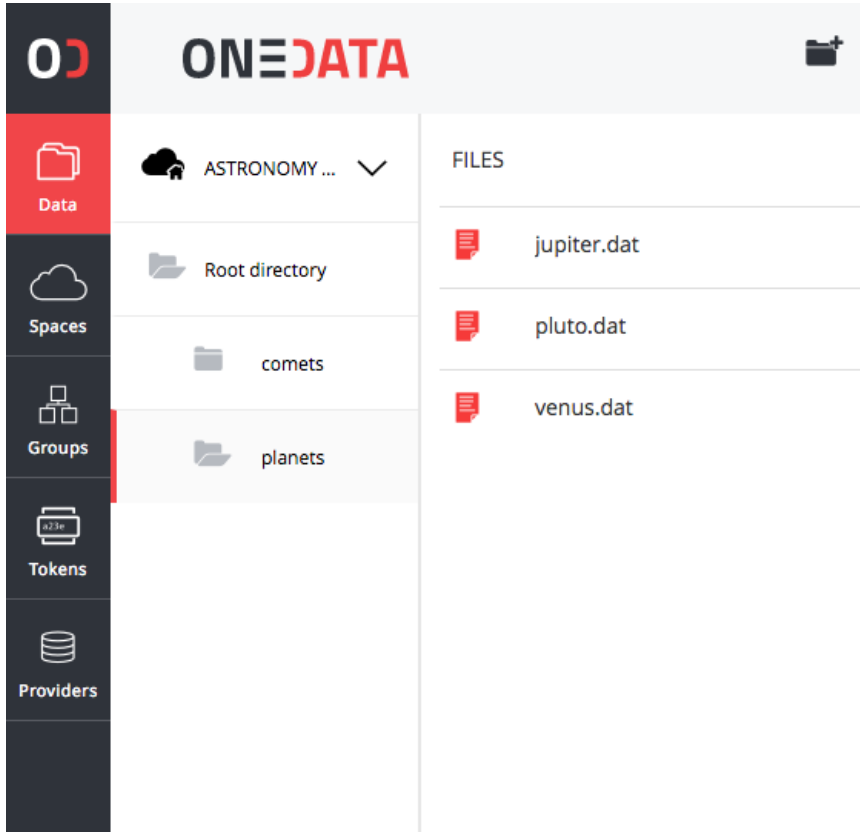
# PROBLEM: GLOBAL NAMESPACE AND TRANSPARENT ACCESS TO DATA IN MULTI-CLOUD ENVIRONMENTS

**Thanks to Onedata now you can:**

- Transparently access existing and create new data in multi-cloud environments
- Access data from anywhere
- Use many protocols to access the same data
- Use global namespace for globally distributed data collections

# […] BUT WE WANT POSIX

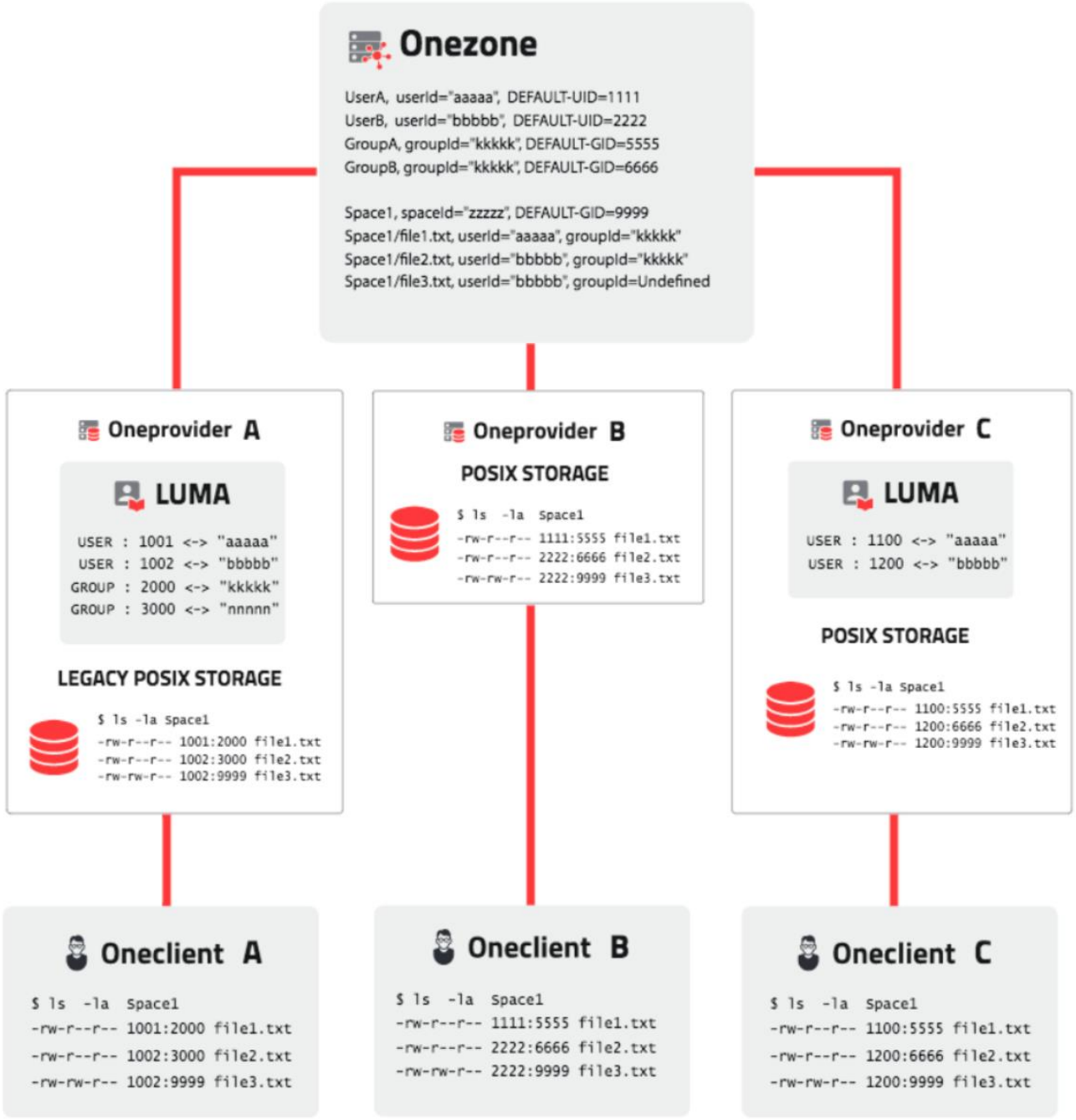
- Support for most of the POSIX operations on virtual file system
- Global namespace automatically mapped to a virtual file system mountable on VM, Grid nodes, Docker containers, desktops




```
[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 Experimen
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.d

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

./Big Data Experiment:
total 0
-rw-r--r-- 1 1124656 1337123 10000000 Sep 26 19:08 cats_ima
-rw-r--r-- 1 1124656 1337123  5000000 Sep 26 19:13 galaxies
-rw-r--r-- 1 1124656 1337123  5000000 Sep 26 19:14 spam_mai

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

# […] WITH THE SAME PERMISSIONS

- LUMA - flexible mapping of POSIX access rights among different storages
- Extensible – easy to develop integration with custom AAI systems (e.g. LDAP)

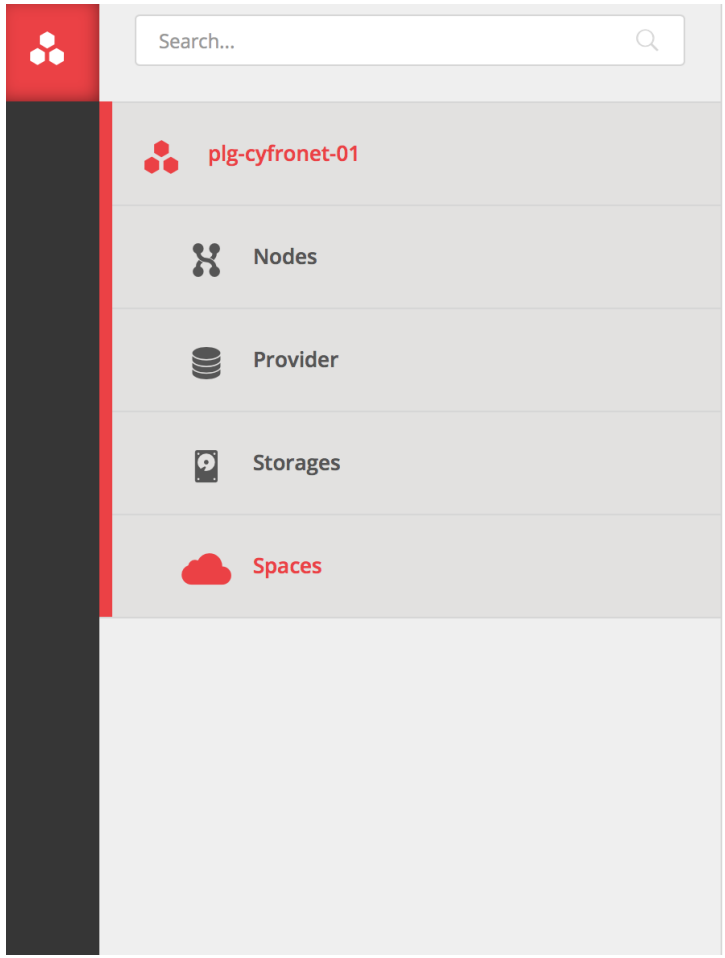


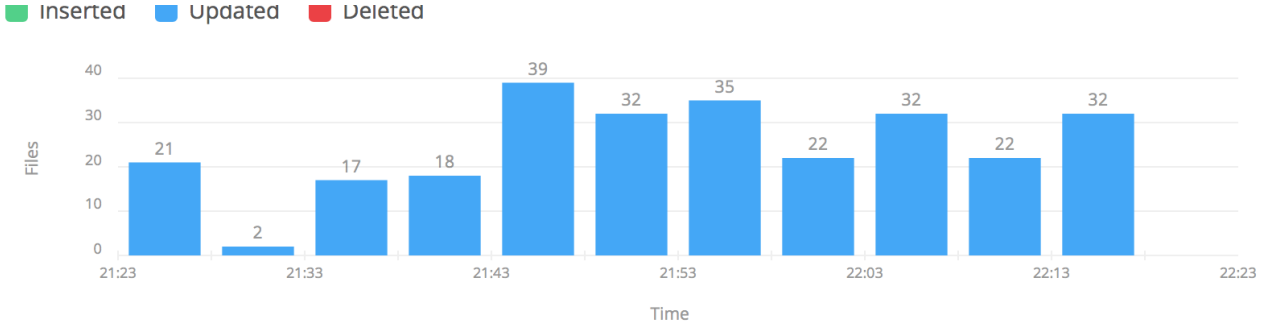
# PROBLEM: HOW TO ACCESS LOCKED-IN DATA

**Thanks to Onedata now you can:**

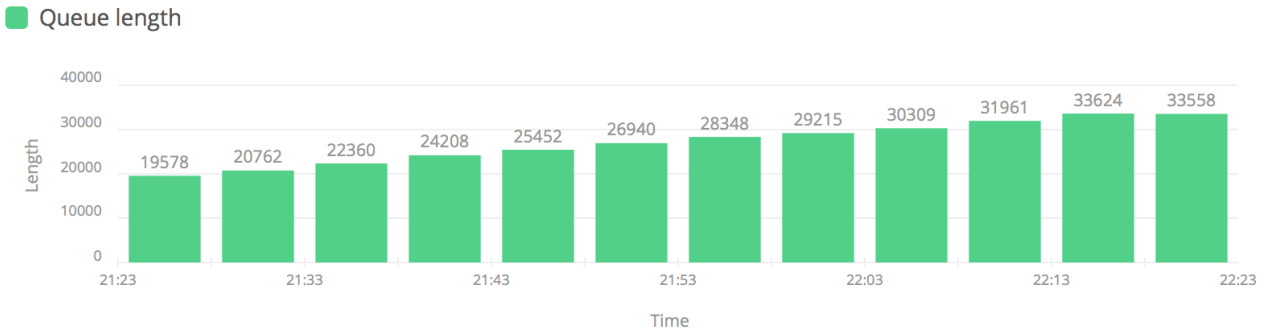
- Current data collections stored on locally available POSIX file system can be easily exposed
- Access data in public Cloud on-the-fly, without pre-staging, with smart caching
- Remotely created data can be replicated back to the private Cloud local storage automatically

# SYNCHRONIZATION OF PRIVATE STORAGE




**FILE OPERATIONS QUEUE STATISTICS**




**SYNCHRONIZATION THROUGHPUT**

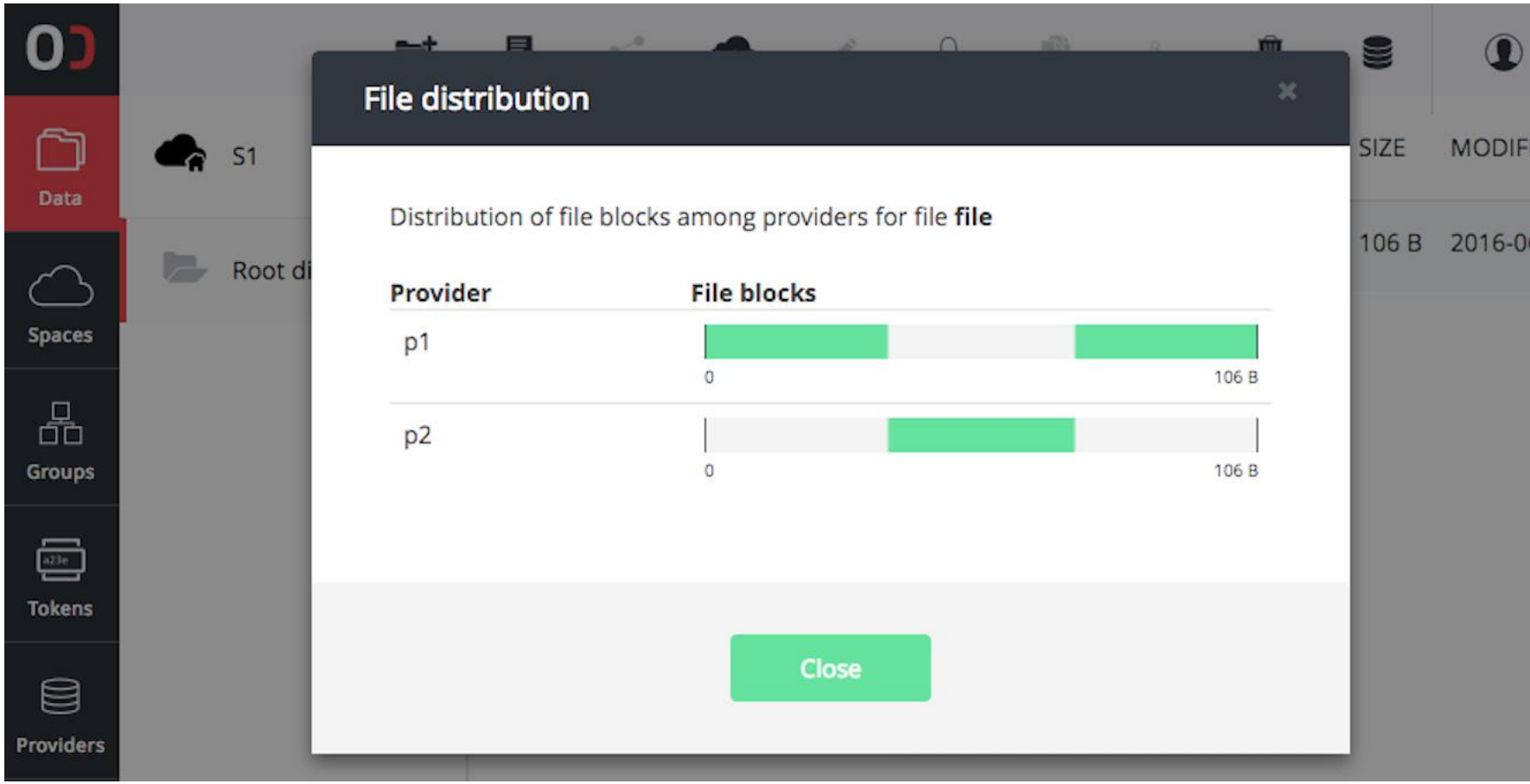
# PROBLEM: HIGH TROUGHPUT DATA TRANSFERS

**Thanks to Onedata now you can:**

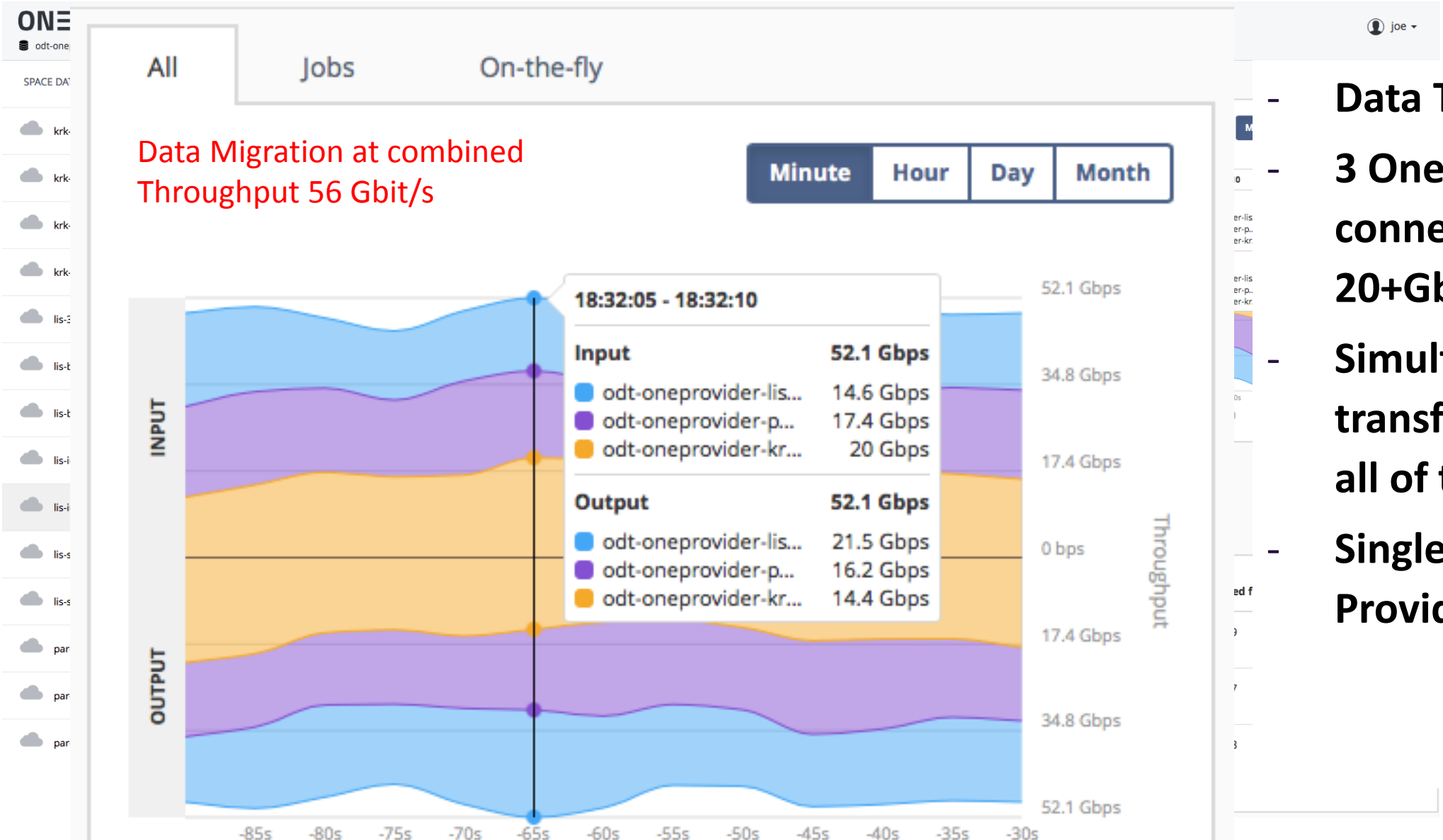
- Replicate files on demand and on the fly without any additional effort
- Migrate data between sites on demand with simple API interface
- Easily monitor transfers and current data distribution through Web GUI or REST API

# REPLICA MANAGEMENT SIMPLIFIED

- Manage files not replicas
- Files distribution between locations is hidden below the virtual filesystem structure
- Replica management on a block basis
- Missing blocks delivered on the fly
- API for replica management for pre-staging  and implementing external data policy management



# HIGH TROUGHPUT DATA MIGRATION IN P2P MESH



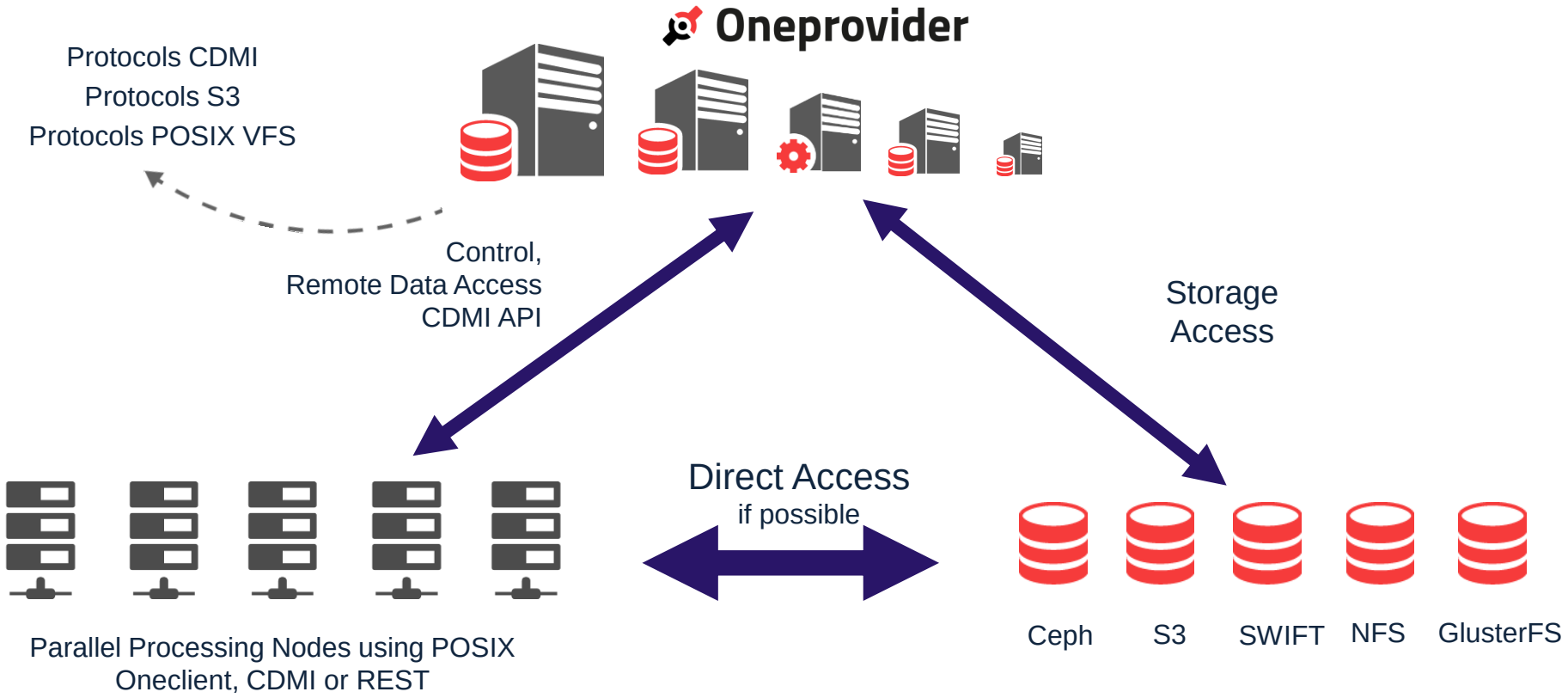

- Data Transfer Mesh
- 3 Oneproviders connected by 20+Gbit/s links
- Simultaneous data transfers between all of them
- Single VM Node per Provider

# PROBLEM: HIGH TROUGHPUT LOCAL DATA ON HETEROGENOUS STORAGE

**Thanks to Onedata now you can:**

- You can work with different type of storages: Ceph, S3, POSIX, SWIFT, GlusterFS, WebDAV (new)
- DirectIO architecture allows for high throughput data processing from all machines running Oneclient with only single Oneprovider deployment

# MULTIPLE STORAGE BACKENDS



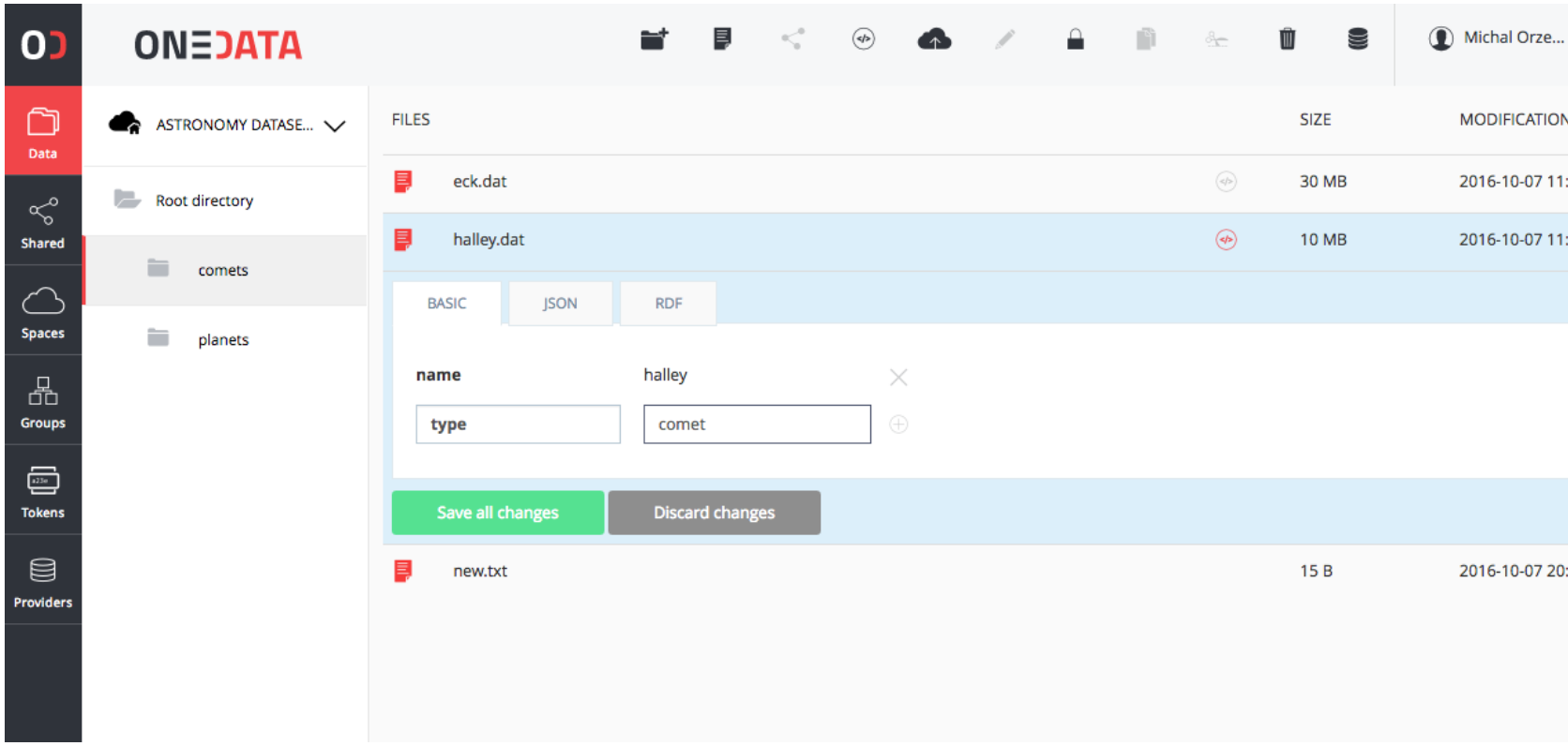
# PROBLEM: DATA DISCOVERY AND REPLICATION BASED ON DISCOVERY RESULTS

**Thanks to Onedata now you can:**

- Work with data and metadata in one system – avoiding problems of consistency
- Monitor metadata data changes through API in order to integrate with other services
- You can discover data within data collections using highly efficient indices built on top of Map-Reduce
- New interface for originating replicas based on data discovery results

# INTEGRATED METADATA MANAGEMENT

- All files and directories can have custom metadata
- API for metadata management
- API for data discovery based on metadata
- Virtual Folders based on metadata tags

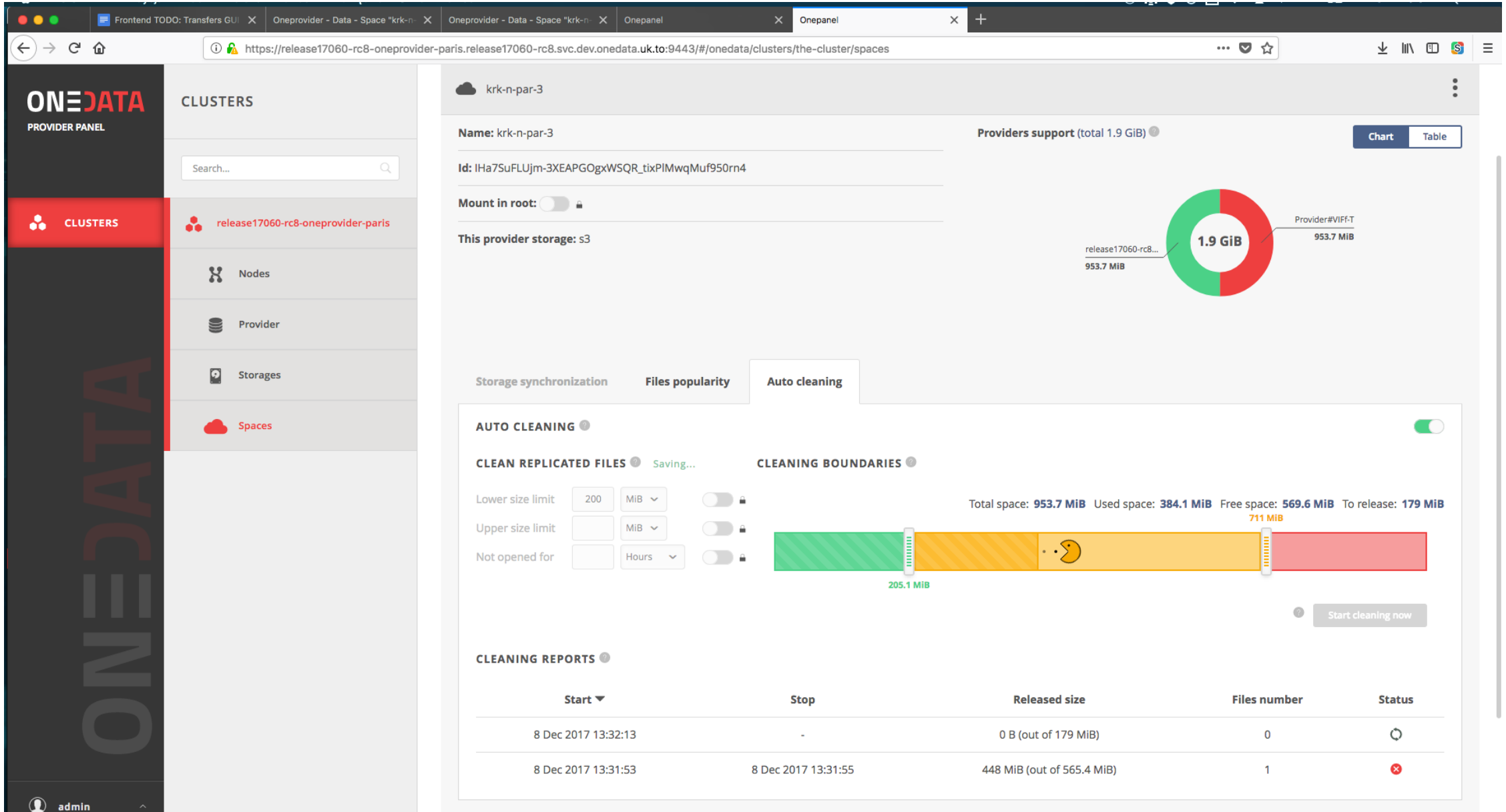


# PROBLEM: CACHE MANAGEMENT AND FILE POPULARITY

**Thanks to Onedata now you can:**

- Control and Monitor local storage usage
- Monitor local file popularity
- Automatically evict locally least popular replicated data
- Expose interface for external logic to cache eviction

# ONEDATA FILE POPULARITY AND SMART CACHING



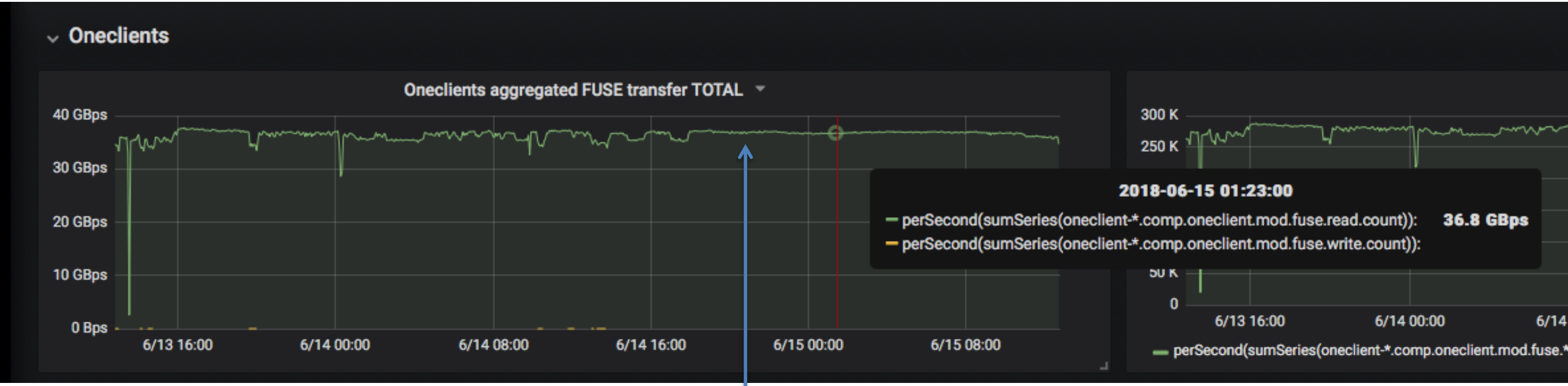
# ONEDATA IN HNSC

# RESULTS

**Onedata has been tested thoroughly using applications from:**

- **EMBL**
  - Large genome files processed in public Cloud on the fly from storage on private Cloud
- **INFN**
  - Large number of files processed in a sequence of random read operations with very small block size (~1KB) in public Cloud on the fly from storage on private Cloud with automatic prefetch
- **DESY**
  - Complex deployment including LUMA for UID/GID mapping from private Cloud storage to public Cloud and automatic replication of results back to the private Cloud storage with correct access rights

# ONEDATA HIGH TROUGHPUT DATA PROCESSING ON HNSC




Onedata Transparent POSIX File System
Processing transparently cached data - 37GBytes/sec

# LESSONS LEARNED

- Implementing transparent hybrid cloud virtual filesystem with pure POSIX interface for real life applications is challenging – but possible!
- Seemingly simple issues such as authentication or uid/gid mapping can be in fact very challenging and time consuming to get right
- High-performance backend storage and low latency network are critical to achieve high IOPS
- Performance issues are hard to track instantly in hybrid Cloud deployments, as they can depend on multiple factors (network issues, storage issues, high system loads from other users, etc.)

# COMMERCIALIZATION PLAN

- Spin-off company is being established, which will be in charge of evolution of commercial features of the project
- Several support plans will be provided with different reaction times and support levels
- Pricing will be based on the number of Oneprovider instances deployed and total storage size managed by them
- Tentative pricing, flat rate per 1PB/Oneprovider/month:
  - Silver – 48h reaction time
  - Gold – 24h reaction time
  - Platinum – 7h reaction time

# ACKNOWLEDGEMENTS

**We would like to thank everyone in the HNSC for this opportunity in particular:**

- All Buyers Group members who have participated in Onedata evaluation for their patience and vigilance in pointing out missing features and bugs

- Both T-Systems and RHEA consortia for trusting us and helping us with evaluation on production Cloud infrastructures

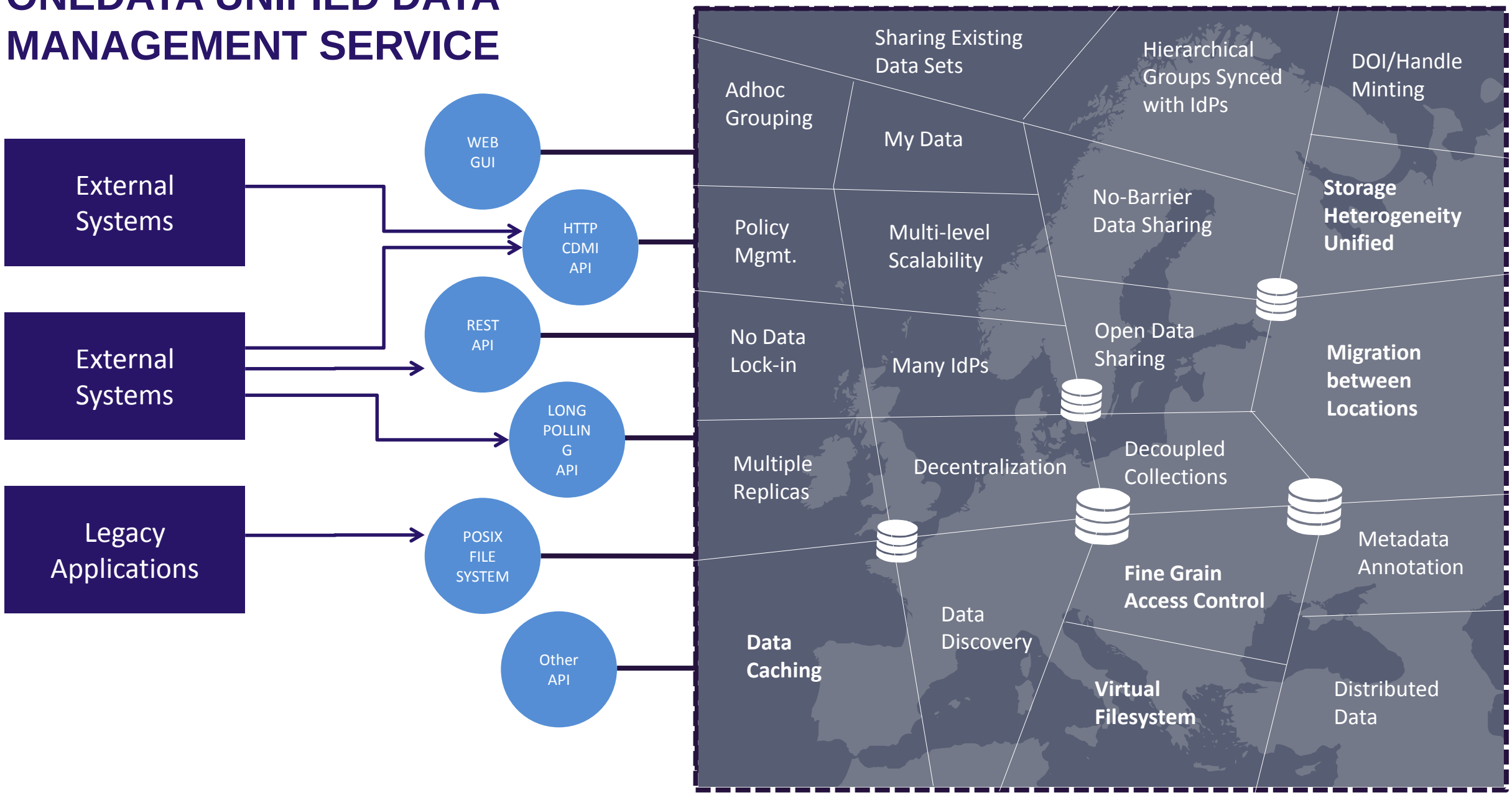
# QUESTIONS?

**Please visit:**

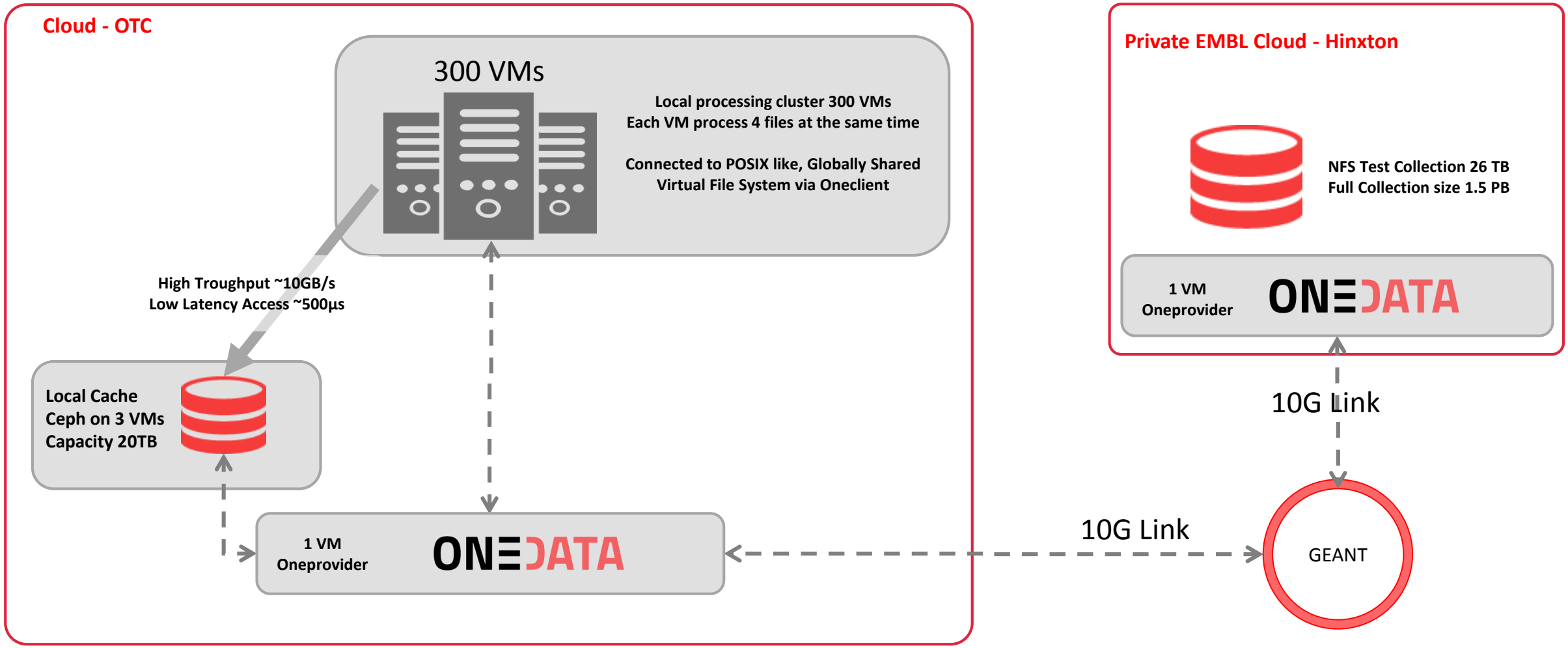
**www.onedata.org**

# BACKUP

# ONEDATA UNIFIED DATA MANAGEMENT SERVICE



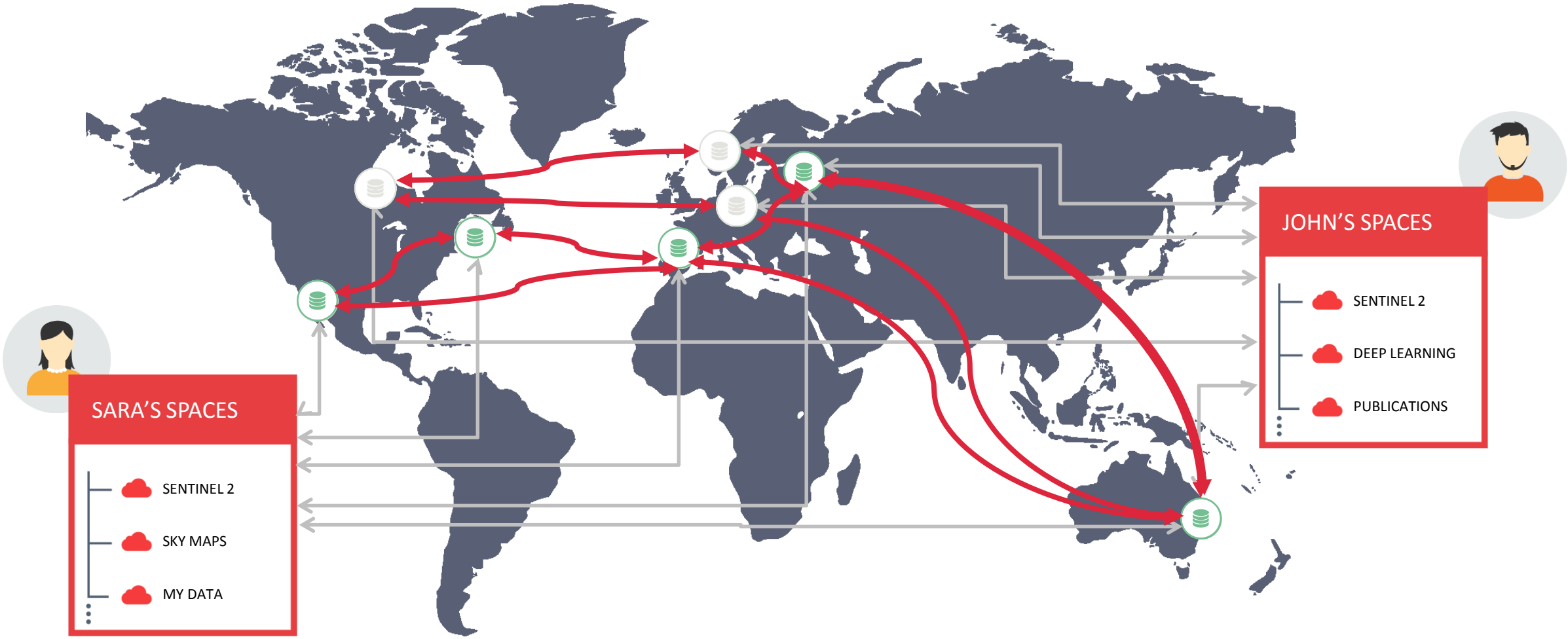
# EXAMPLE - EMBL-EBI APPLICATION DEPLOYMENT



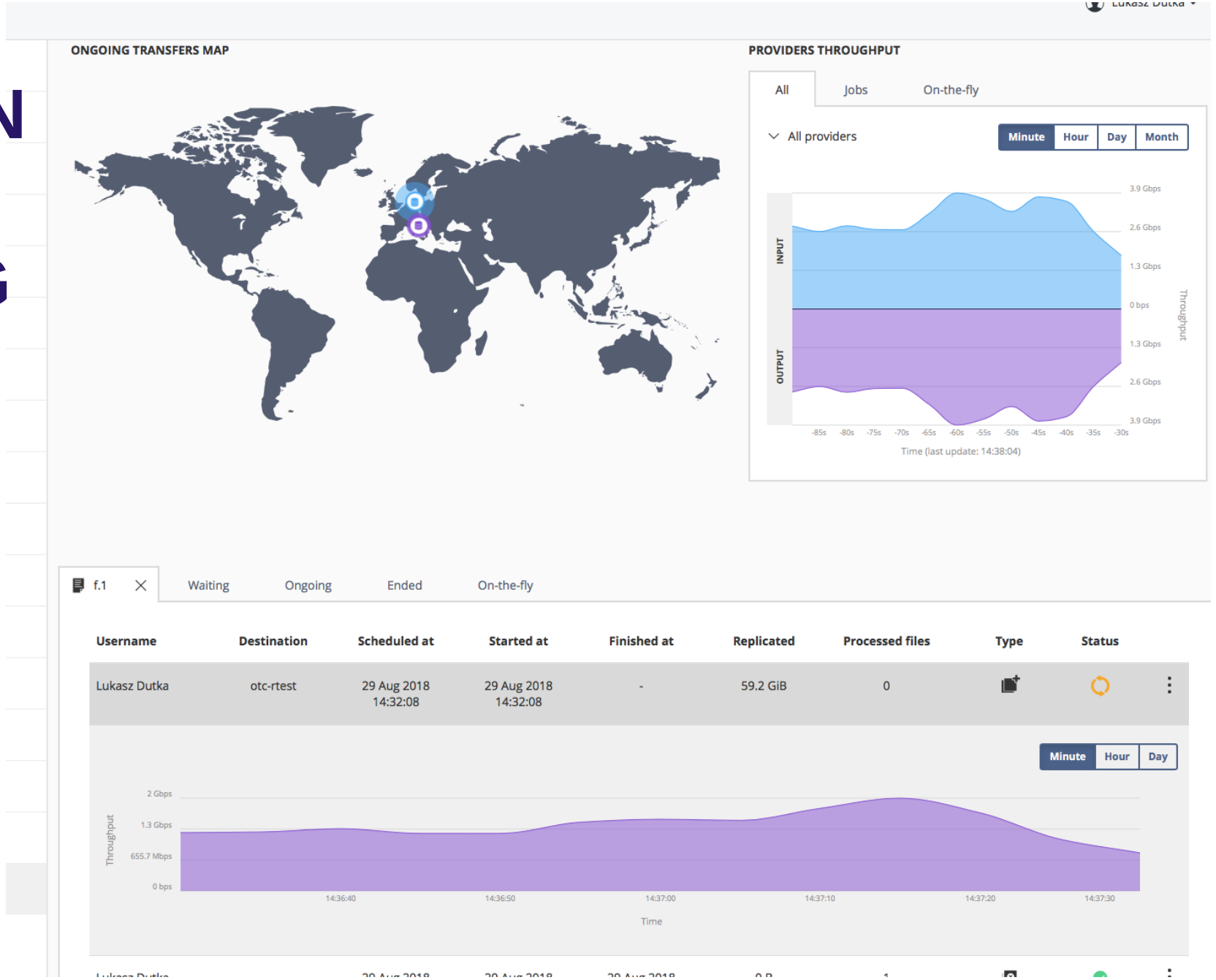
# CDMI HTTP ACCESS

| Operations | Capabilities |
|---|---|
| Basic object GET PUT DELETE | cdmi_dataobjects, cdmi_read_value, cdmi_modify_value, cdmi_delete_dataobject |
| Basic container GET PUT DELETE | cdmi_list_children, cdmi_create_container, cdmi_delete_container |
| Metadata (container&dataobject) | cdmi_read_metadata, cdmi_modify_metadata, cdmi_size, cdmi_(atime\|mtime\|ctime) |
| Access control lists (rwx) | cdmi_acl |
| Big folders | cdmi_list_children_range |
| File System Export (FUSE client) | - |
| Move and copy | cdmi_(move\|copy)_(container\|dataobject) |
| Big files | cdmi_read_value_range, cdmi_modify_value_range |
| Access by ObjectID | cdmi_object_access_by_ID |

# DATA IN HYBRID CLOUD ENVIRONMENTS



# REPLICATION JOBS AND MONITORING



# PROBLEM: AUTOMATIC DECISIONS ABOUT DATA PLACEMENT BASED ON EVENTS

**Thanks to Onedata now you can:**

- You can hook your customized decision logic into the system
- Onedata offers a long polling HTTP API for monitoring changes in the system
- The API is eventually consistent
- The customized logic can be used for instance for:
  - automatic replication to certain locations
  - metadata ingestion