

Nodeum Data Management Software

IBM Storage Scale Days 2025 DE

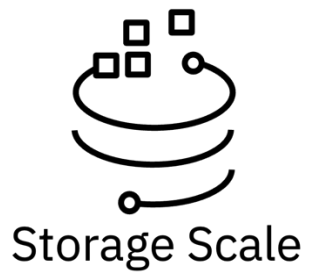
March 19th – 20th, 2025 | Heidelberg, Germany

Valery Guillaume

valery@nodeum.io



Disclaimer



- IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

AGENDA



- About Nodeum
- Use Case
- Product Overview
- Next Evolution
- Conclusion

ABOUT NODEUM



Nodeum empowers data owners with the ability to take self-ownership of how its data is organized, enriched, and shared with accountability, traceability, and verification.

WHAT IS NODEUM?



What Nodeum Does

- Designed for **multi-directional** data movement across multiple storage media
- Supports on-premises and cloud storage
- Policy driven – providing accountability, traceability, and verification/validation
- Customers across all major industries

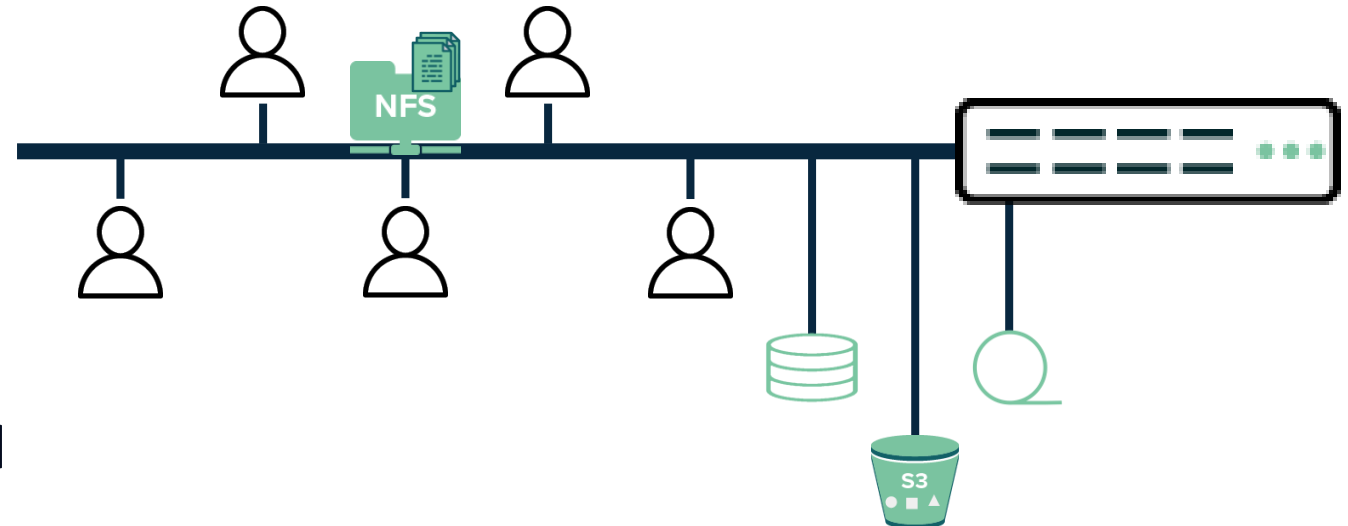
Who are Nodeum Customers?

- Users with massive datasets (>100 TB)
- Organizations with minimal IT support (often 1 to 2 IT admins)
- Users have to be self-reliant, take ownership, and individual responsibility
- Minimal impact on user productivity, performance

WHAT IS NODEUM?



- Uni-directional data movers are perfect for:
 - Backup
 - Archiving
- They require an intermediary server like a backup server
- An administrator is required to manage this server
- Users have little or no autonomy as to how or where data is moved

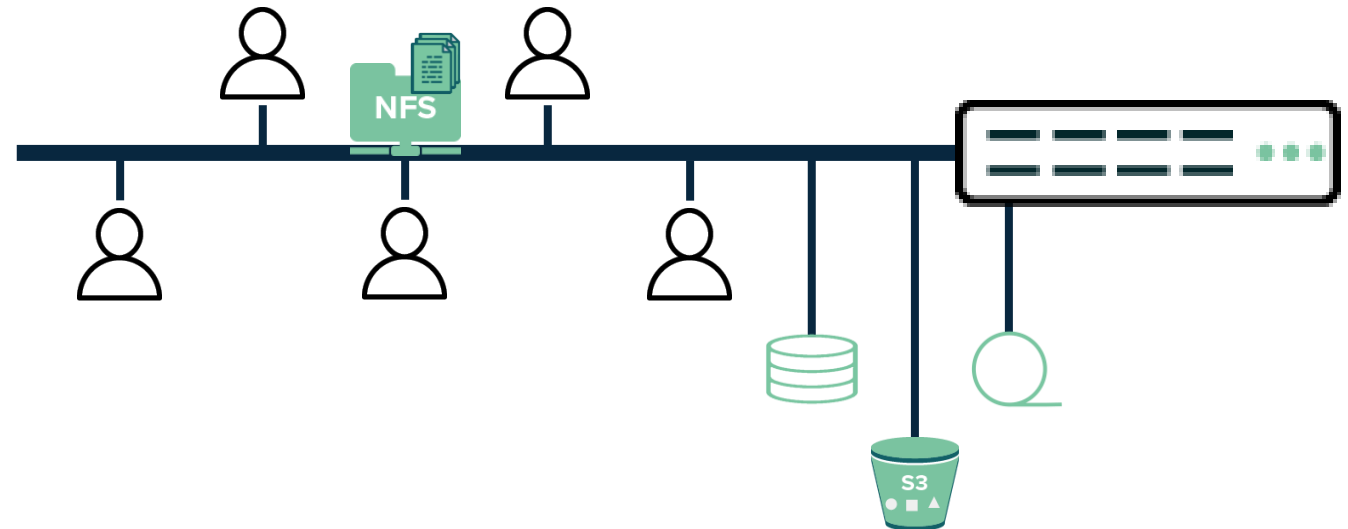


WHAT IS NODEUM?



TECHNICAL USERS CAN MANUALLY MOVE DATA, BUT PERFORMANCE IS IMPACTED

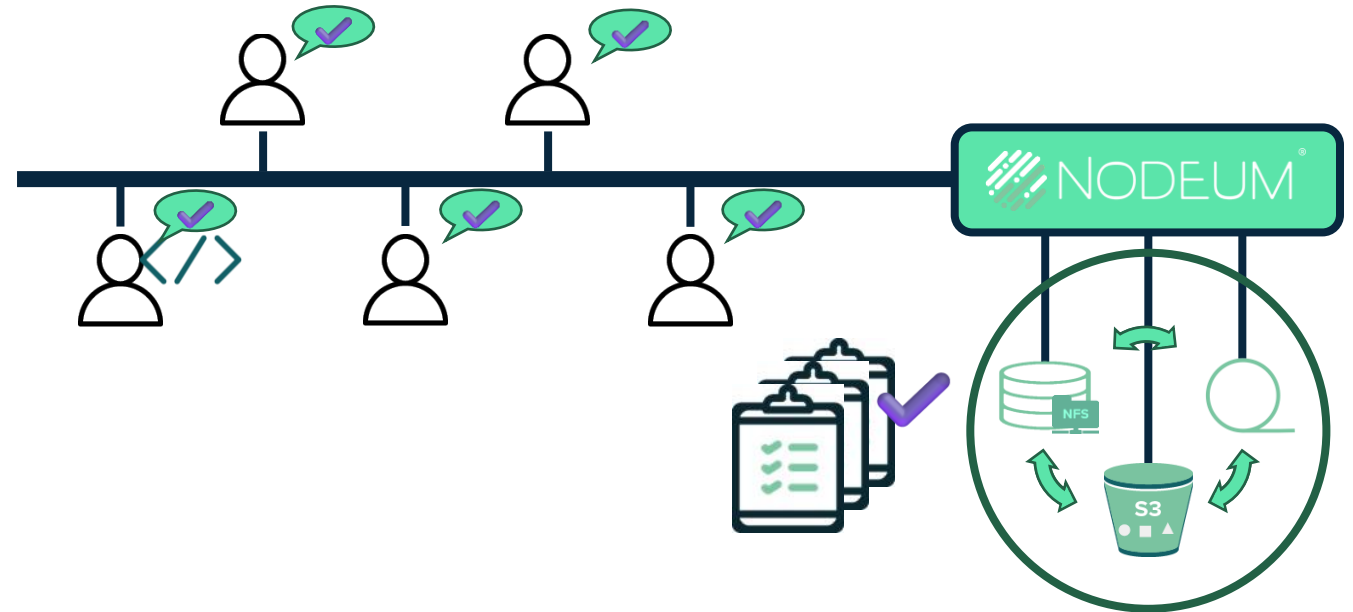
- More technical users can move their own data with commands like:
 - Linux: *rsync*, *cp*, *mv*
 - Windows/Mac: drag and drop
- This solves the autonomy / ownership problem
- But introduces new problems:
 - Performance
 - Accountability



NODEUM SOLVES ALL THESE PROBLEMS



- A user simply commands Nodeum to move the data
- Nodeum takes care of the movement
- Nodeum provides accountability
 - Catalog
 - Verification
- Everyone on the team can understand what has happened



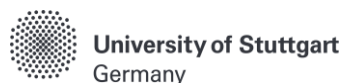
INDUSTRIES



HPC Super Computing



Research & University Life Science



Media Post Production



Earth Observation



And also in additional sectors:

Financial
Services

Gov. /
Administration

Retail /
Manufacturing

Services

USE CASE



NEED: Today, supercomputing systems are so performant and scalable that the speed of data generation has never been so fast. Research centers have to store the generated content in “data repositories” that are located close to each other and that are well integrated.

Two different categories of data repositories are used as storage tiers:

- Active Data Repositories which provide the performance when data is written by supercomputing systems
- Archival Data Repositories with interfaces used in Cloud systems, which are more suitable for data sharing

Research is becoming increasingly collaborative.

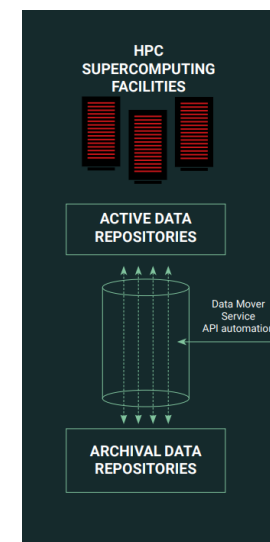
Sharing of data and FAIR data management is becoming mandatory. This adds to the requirements for modern data infrastructures.

ORGANIZE THE MOVEMENT OF THE DATA FROM THE ACTIVE TO THE ARCHIVAL DATA REPOSITORY

KEEP A DIRECT ACCESS BY THE USERS TO ACTIVE AND ARCHIVAL DATA REPOSITORIES

INTEGRATION WITH HPC WORKLOAD MANAGERS LIKE SLURM

PROVIDE A PUBLIC API AND SDK TO FACILITATE INTEGRATION WITH SPECIFIC RESEARCH APPLICATIONS



PRODUCT OVERVIEW

MODERN ARCHITECTURE



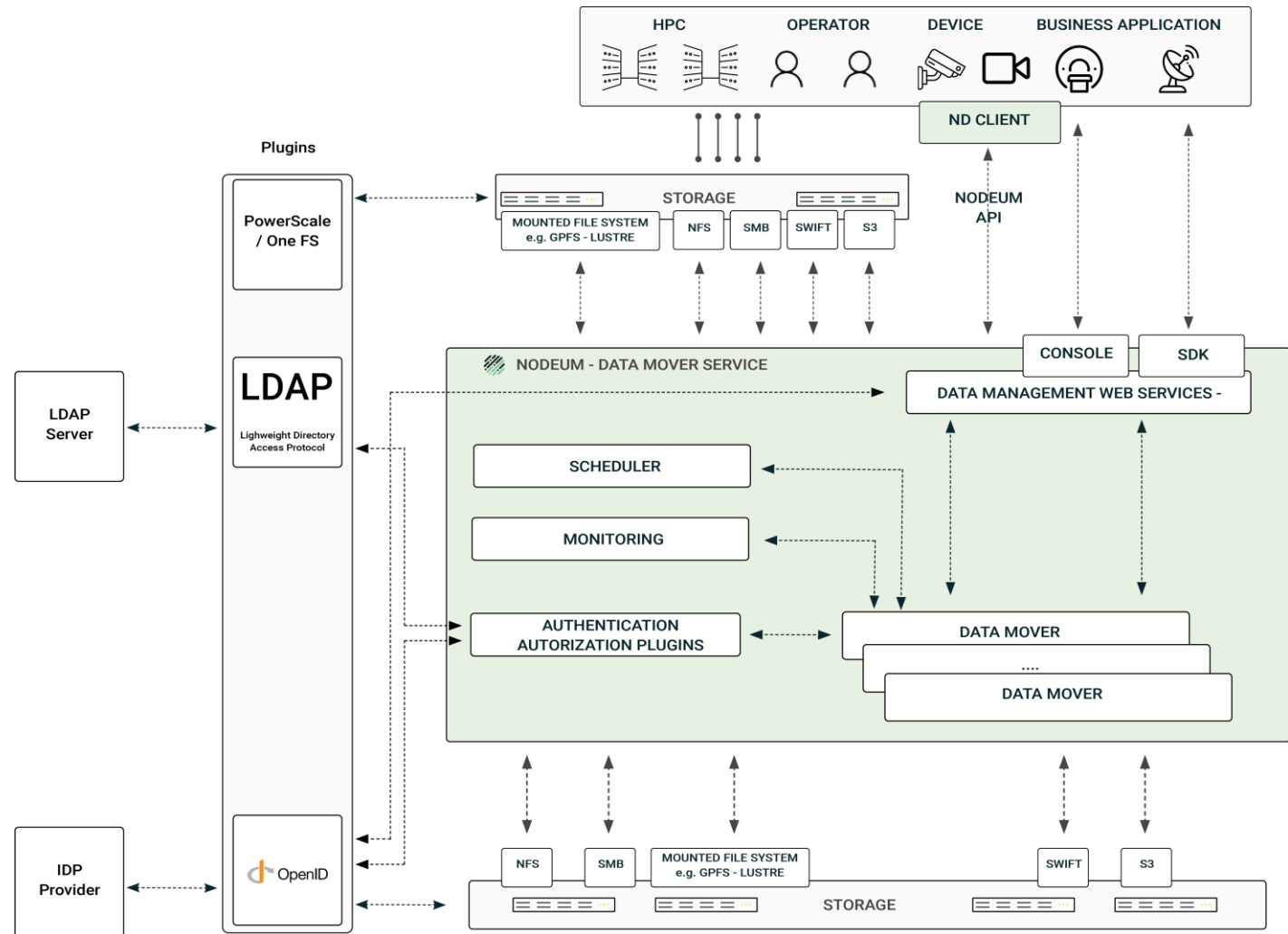
HIGHEST DATA MOVEMENT THROUGHPUT
- READY FOR EXASCALE COMPUTING

SUSTAIN 5,000 REQUESTS PER SECOND
AND HANDLE 10 MILLION SIMULTANEOUS
TRANSFER REQUESTS

COMPATIBLE WITH IDENTITY PROVIDER
(SERVICE AND STORAGE)

ND CLIENT TO ALLOW DATA MOVEMENT
DIRECTLY FROM COMPUTE NODE

DOCKER AND APP MARKETPLACE READY

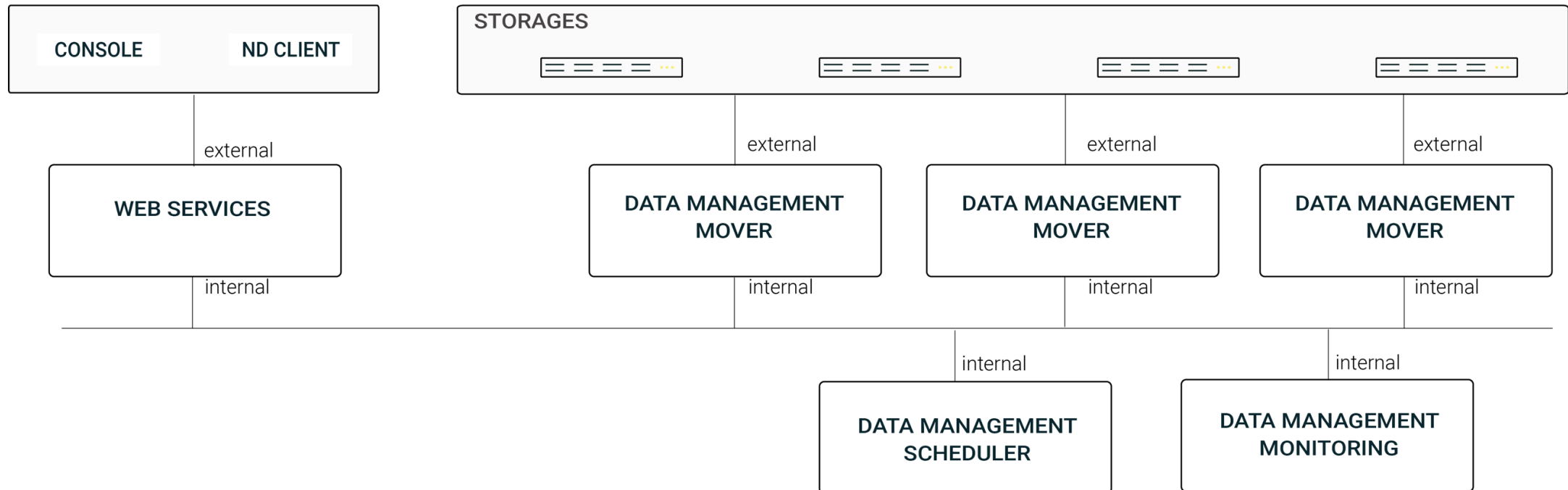


DATA MOVER - END-TO-END PARALLEL DESIGN



Nodeum is scalable horizontally and vertically. The solution can scale by adding nodes that perform the required roles.

- Add nodes that runs the mover nodes scale the movement throughput capability
- Run multiple mover services on a same nodes if you can increase the node resources.



MULTI PROTOCOLS & FILE SYSTEM



- File System based protocols such as SMB - NFS
- Object Storage protocols such as S3 – SWIFT
- But also POSIX Mounted File System capabilities

Nodeum extends its support for storage mounted by their proper client in addition to the storage types it already supports. This additional feature enables Nodeum to perform data mover operations on a wider range of storage types, providing greater flexibility and versatility in managing data.

It can now support storage devices that are directly connected to a client and mounted on a local directory. By supporting storage mounted by their proper client, Nodeum can provide more comprehensive data management capabilities and increase the efficiency of data transfer operations.



And others ...

POLICY-BASED WORKFLOW ORCHESTRATION



DESIGN THE MOST ADVANCED DATA FLOWS BASED ON THE USAGE OF THE CONTENTS

← Creating New Task

Name	
Type	Active copy
Source	0 items
Destination	
Filters	0 filters
Schedule	Manual
Other options	Parallel

Filters 0 filters

Subject Operator

File or folder name contains

File name

Creation date

Last modification date

Last access date

Flexibility in each defined workflow

Source : NAS – Cloud – Object Storage : All or Granular File

Selection including

- Filtering
- Enhanced Folder and Files selection

Destination : Tape, NAS, Cloud, Object Storage,

Basic (GUI) or Advanced (scriptable) filters :

- File Information :
 - File Name / Structure Folder
 - Regex capability
- Date : Creation date, Last modification date, Last access date

Options :

- Integrity Check : MD5 – CRC32 – XXHASH64
- Priority

Schedule

- Automatic
- Manual
- Scheduled

CONSOLE



This is a secure HTML5 interface which provides a modern way to execute data management operations.

- Cross-platform: The interface is compatible with most of modern browser such as Chrome, Firefox, Safari and Microsoft Edge. The interface is also mobile friendly.
- Multi Role: The role based management allow administration and user to access and operate with the interface.
- API based: The interface uses the Nodeum Web Services and its openly published REST API.

The screenshot displays the Nodeum console interface. On the left is a sidebar menu with options: Dashboard, Catalog, Trend Analysis, TCO Calculator, Toolbox, Workflow Definition, Task Management (expanded), Task Monitoring, Task Listing, Task Templates, Data Container, Storage Services, Event Logs, and System. The main content area is titled 'All tasks' and contains three sections:

- Upcoming:** A table with columns: Next execution, Name, Type, Status, Frequency, and Missed execution. It shows 0 of 0 items.
- Running:** A table with columns: Start date, Name, Type, Status, and Progress. It shows one task: '1000 files of 1MB' of type 'Data Migration - Copy' with status 'In progress' and a progress bar.
- History:** A table with columns: Finish date, Name, Type, Status, and Elapsed time. It lists several completed tasks, all of type 'Data Migration - Copy', with statuses marked as 'Done' and various elapsed times.

DATA MOVEMENT SUPERVISION



All tasks ▼

Upcoming

Next execution ↑

Name

Running

Start date

Name

Type

Status

Progress

3/13/23, 4:29 PM

From
nod://largetdata2_pool/storageteData Migration - Copy
to mycontainer15

In progress

Calculating (361 items)

In progress (336 items)

3 Finalization



Items per page: 10 ▼

1 – 1 of 1



History

Search



Finish date ↓

Name

Type

Status

Elapsed time

3/13/23, 4:29 PM

From nod://largetdata2_pool/storagetestdata/test_data11 to
mycontainer15

Data Migration - Copy

Done

27" 645ms



ND CLIENT



ND command line tool provides a modern set of commands to execute data movement operations with Nodeum.

- **Flexibility:** This flexibility enables users to perform various tasks in a customizable way.
- **Efficiency:** Users can perform multiple tasks simultaneously which can increase efficiency and productivity.
- **Control:** It allows users to manage and monitor the data movements.
- **Security:** It runs on a secure, encrypted channel
- **Compatibility:** The client is compatible with different O.S. including Linux, macOS, and Windows.

In summary, the Bash client provides a flexible, efficient, and secure way for users to manage and interact with Nodeum, allowing them to automate workflows, manage storage, and monitor the system with ease.

```
vguilleume@MacBook-Pro-2: ~/Documents
NAME:
nd - Nodeum CLI

USAGE:
nd [global options] command [command options] [arguments...]

VERSION:
2.0.6

COMMANDS:
admin
config  configure the Nodeum Client
copy, cp  create copy task
move, mv  create move task
task
help, h  Shows a list of commands or help for one command

GLOBAL OPTIONS:
--json                output as JSON (default: false)
--config value        path to configuration file (default: <config-dir>/config.json) [SND_CONFIG]
--config-dir value, -C value  path to configuration folder (default: "/Users/vguilleume/Library/Application Support/.nd") [SND_CONFIG_DIR]
--alias value         alias in configuration file for authentication (default: "default") [SND_ALIAS]
--url value           URL of Nodeum [SND_URL]
--access-token value  for API authentication (1st authentication method) [SND_ACCESS_TOKEN]
--refresh-token value for API authentication (1st authentication method, not saved in config) [SND_REFRESH_TOKEN]
--authorization-endpoint value for Device Authorization Flow (2nd authentication method)
--token-endpoint value for Device Authorization Flow (2nd authentication method)
--client-id value     for Device Authorization Flow (2nd authentication method)
--scopes value        for Device Authorization Flow (2nd authentication method)
--persist-session     persist Device Authorization session on disk for 1 hour (default: true)
--persist-session-renew  If persist session is enabled, renew the token (default: false)
--username value      for API authentication (3rd authentication method) [SND_USERNAME]
--password value      for API authentication (3rd authentication method) [SND_PASSWORD]
--anonymous           no login (default: false)
--help, -h            show help (default: false)
--version, -v         print the version (default: false)
```

INTEGRITY VERIFICATION

Data movement

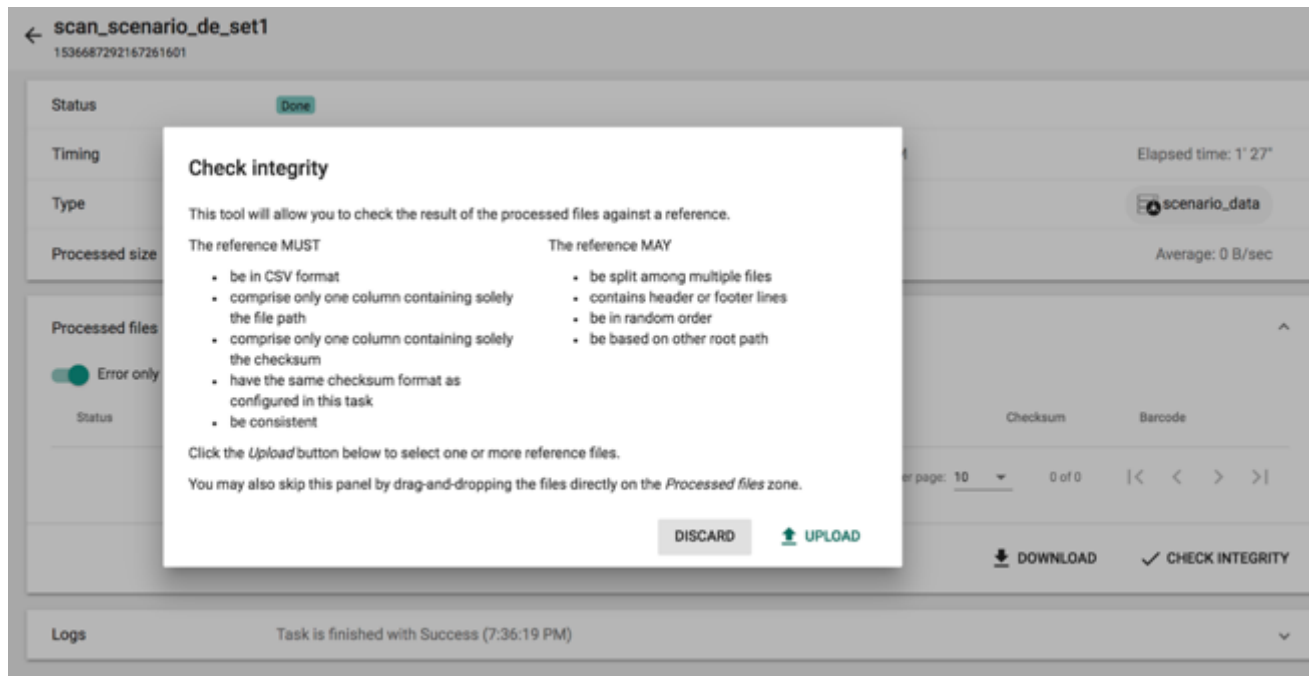


Managed by Nodeum

Data is read from the source

Data is copied

Checksum is calculated



CALCULATION IN TRANSIT

CRC32 – MD5 - XXHASH

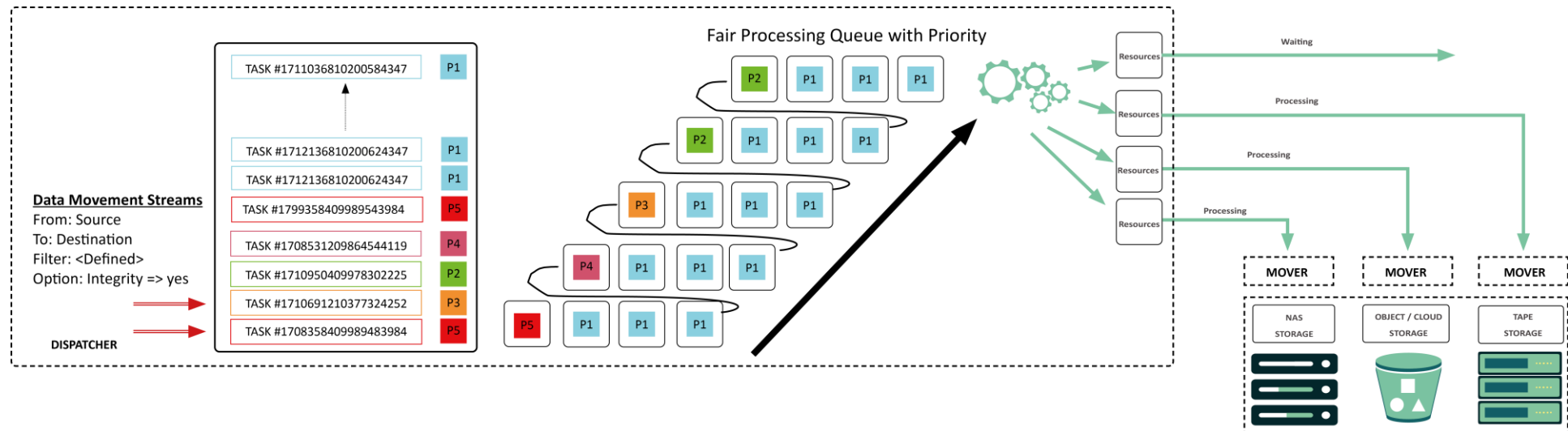
INTEGRITY COMPARISON

FAIRNESS PRIORITY MANAGEMENT



The implemented priority management allow prioritization of data movement workflows in preventing that highest priority which consume all available throughput in letting other workflows waiting indefinitely.

Fair Queuing involves allocating resources to different request, ensuring that each remaining request keeps an equal share of the remaining resources. This technique can help prevent one data movement workflows taking up all the available resources and ensure that all remaining workflows get fair access.



METADATA



Objective is to preserve the metadata in data movement processing. In addition, metadata can be used to filter the source content. This supports also the metadata which is included as file system extended attribute:

```
root@srv1:~# getfattr mytesttext -d
# file: mytesttext
user.ship="boat 1"
```

If the file is migrated to a S3 Object Storage based, then the following metadata will be defined: `x-amz-meta-custom-ship="boat 1"`.

In addition, the traditional Posix attribute are also preserved:

From Posix:

```
modeMetaKey   = "Mode"
uidMetaKey    = "Uid"
gidMetaKey    = "Gid"
atimeMetaKey  = "Atime"
mtimeMetaKey  = "Mtime"
ctimeMetaKey  = "Ctime"
```

To S3: "x-amz-meta-key"

```
x-amz-meta-mode
x-amz-meta-uid
x-amz-meta-gid
x-amz-meta-atime
x-amz-meta-mtime
x-amz-meta-ctime
```

FILTERING



The workflow Manager includes a powerful filtering module. This feature allows users to easily manage and organize their data movement workflows by filtering files based on specific criteria.

User can set up filters that automatically exclude or include files based on **file size**, **creation date**, **modification date**, **file type**, **metadata** and **more**. Complex filter rules combining multiple criteria are authorized.

Allow user to easily manage large volumes of data and automate their workflows. By setting up filters, users can ensure that only relevant files are included in their data movement tasks, which can help to speed up the transfer process and reduce the risk of errors or data loss.

The screenshot shows the 'copy data' workflow configuration. The 'Filters' section is currently empty (0 filters). Below this, there are two tabs: 'Basic' and 'Advanced'. The 'Basic' tab is active, showing a table with columns: Subject, Operator, and Operand. A dropdown menu is open under 'Subject', listing options: File path (selected), File name, File extension, Change date, Last modification date, and Last access date. The 'Operator' column has a dropdown set to 'Matches'. The 'Operand' column has a text input field with a placeholder 'Regular expression. Use . * as a wildcard.' and a '+' icon. Below the input field, it says 'Items per page: 10' and '1 - 1 of 1'. At the bottom of the table, it says 'be handled as **Inclusive OR**'. There are 'PREVIOUS STEP' and 'NEXT STEP' buttons at the bottom right of the filter section. The bottom of the page shows 'Call' and 'Schedule' tabs, with 'Schedule' set to 'Manual' and 'Other options' set to 'Priority 5'. A blue circular icon with a white 'B' is in the bottom right corner.

Highly customizable, allowing users to create filters that are tailored to their specific needs. This can include setting up filters for specific file types, folders, or directories, as well as creating custom rules based **on metadata or other criteria**.

HOOK SERVICE



Nodeum's hook is a feature that allows users to execute custom scripts or commands during specific events within each Data Movement, such as before or after a data movement task. These custom scripts or commands can be used to automate additional tasks, integrate with other systems, or perform specific actions based on the event that has occurred.

For example, a user can configure a hook to run a custom script before the start of a data movement task. This script could then perform additional actions such as sending an email notification, updating a metadata database, or triggering an event.

Using Nodeum hooks can greatly enhance the automation and integration capabilities of the solution, allowing users to customize workflows and extend the functionality of the platform to meet their specific needs.

Name	copy data	▼
Action	Data Migration - Copy	▼
Source	1 item from dataset	▼
Destination	archive	▼
Filters	0 filters	▼

Callbacks

1 callback

^

Type*

Before task is execut...

Language*

Python

Apply sample

Default

```
1 def run(task, metadata):
2     # your code here
3     try:
4         api_url = "https://api.myexample.com/myservice"
5         response = requests.get(api_url)
6
7         # Check if the request was successful (status code 200)
8         if response.status_code == 200:
9             data = response.json() # Parse the JSON response
10            return data
11        else:
12            print(f"API call failed with status code: {response.status_code}")
13            return None
14
15    except requests.exceptions.RequestException as e:
16        print(f"Error occurred during API call: {e}")
17        return None
```


CONTROL TREE STRUCTURE



Working directory configuration is available to control the structure of the directory you want at destination.

Examples:

With `--wd=.`

Source

```
nod://source/folder/FILE.txt
nod://source/folder/FILE.txt
nod://source/folder/
nod://source/folder/
nod://source/folder
nod://source/folder
```

Destination

```
nod://dest/directory/
nod://dest/RENAMED.txt
nod://dest/directory/
nod://dest/directory
nod://dest/directory/
nod://dest/directory
```

Result

```
nod://dest/directory/FILE.txt
nod://dest/RENAMED.txt
nod://dest/directory/FILE.txt
nod://dest/directory/FILE.txt
nod://dest/directory/folder/FILE.txt
nod://dest/directory/FILE.txt
```

With `--wd=..`

Source

```
nod://source/folder/FILE.txt
nod://source/folder/FILE.txt
nod://source/folder/
nod://source/folder/
nod://source/folder
nod://source/folder
```

Destination

```
nod://dest/directory/
nod://dest/RENAMED.txt
nod://dest/directory/
nod://dest/directory
nod://dest/directory/
nod://dest/directory
```

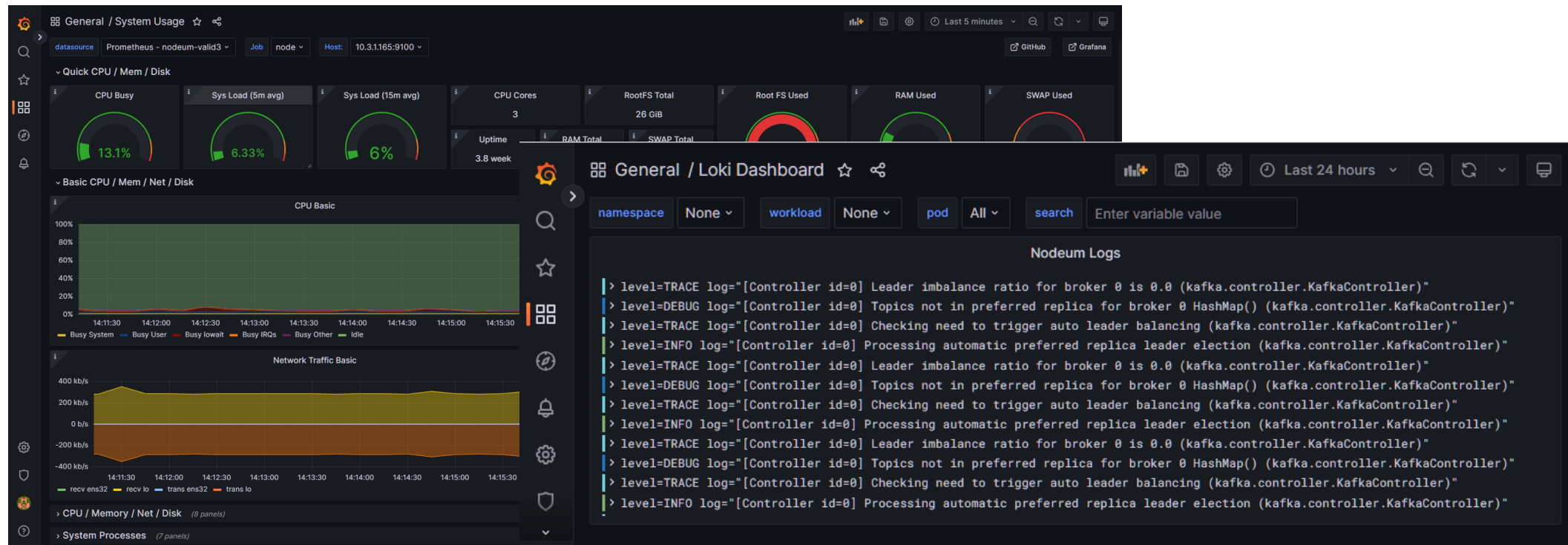
Result

```
nod://dest/directory/folder/FILE.txt
nod://dest/RENAMED.txt
nod://dest/directory/folder/FILE.txt
nod://dest/directory/FILE.txt
nod://dest/directory/source/folder/FILE.txt
nod://dest/directory/FILE.txt
```

MONITORING & ALERTS



Nodeum measures the status of all cluster nodes, including a set of metrics for system resource utilization. This data is stored in a local Prometheus database, guaranteeing long-term retention. These metrics can then be exported to Grafana visualization tools. It also allow the export of Nodeum logs to Grafana/Loki for log management.



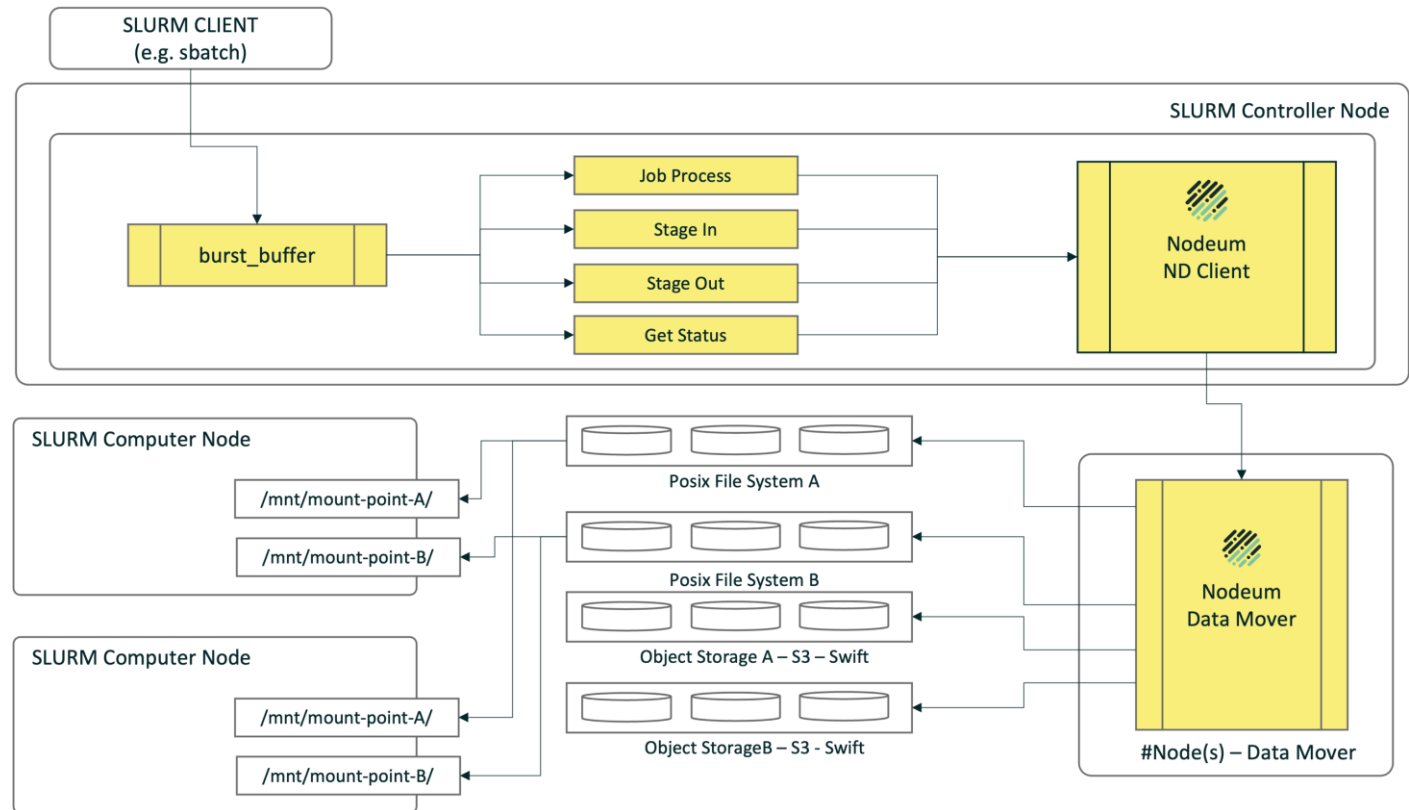
WORKLOAD MANAGER INTEGRATION



Workload Managers are used by many of the world's supercomputers and computer clusters. It provides key functions to allocate resources (computer nodes) to users for some duration of time so they can perform work.

Nodeum is integrated within SLURM which the most used Workload Manager.

SLURM user can schedule their job calculation with the definition of each data movement which has to be performed to get data at the edge.



NEXT EVOLUTION

Revolutionizing Data Lifecycle
Management with Nodeum &
FAIR Principles

INTRODUCTION

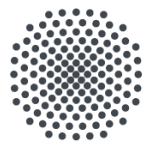


Extend Nodeum with effective data lifecycle management, featuring advanced capabilities for data discovery, inventory, and storage across short-, mid-, and long-term needs, along with comprehensive metadata management and publication.

Objective:

Help organizations to adopt F.A.I.R. data principles (Findable, Accessible, Interoperable, Reusable) and address evolving scientific requirements closely linked to storage technologies. Strengthen the connection between storage data movement, archiving, and project storage management, including metadata management and publication (e.g. Dataverse).

At Nodeum, we collaborate on the bwSFS-2 project, led by University of Stuttgart and University of Hohenheim.



University of Stuttgart
Germany



**UNIVERSITÄT
HOHENHEIM**

STORAGE PROJECT DEFINITION



Nodeum

Storage Projects

Projects

Datasets

Deleted dataset

Data Management

Infrastructure

System resources

Overview

Resource

Mapping

UI Kit

318f31a5 (nodeum)
2025-03-05

€

Total users

8

Total datasets

5

Total resources

4

Project details

Owner: adil@example.com

Status: Active

Name: adil test 2

Short name: adiltest2

Funding source: mysource

Organization: myorg

Runtime

From: 2024-10-10

To: 2026-12-31

Storage

File: 0 GB

Object: 1 GB

Tape: 2 GB

Access

Research Data Manager

Project Administrators

No administrators

Users

Description

test

Beta version 0.4.0

Storage Projects

- Multi roles definition
- Users assignment
- Specify the project runtime
- Project state management: Active – Archive -
- Dataset(s) definition
- Resource(s) definition

DATASETS MANAGEMENT



Nodeum

Storage Projects

Projects

Datasets

Deleted dataset

Data Management

Infrastructure

Overview

Resource

Mapping

UI Kit

Valery Guillaume

gs112308@uni-stuttgart.de

Datasets

New dataset

#	Name	Created	Updated	Resource	Space	Status
> 1554406	adil new test	2024-11-15	2024-11-15		0	Active
> 4423177	adil-demo-dataset	2025-01-22	2025-01-22	adiltest2-NFS-1	0	Active
> 4399618	adil-test-dataset-3	2025-01-21	2025-01-21	adiltest2-NFS-3	0	Active
✓ 4730513	adil-test-dataset-jan-29	2025-01-29	2025-01-29	adiltest2-NFS-5	0	Active

Details

Storage operations

Data mover operations

Owner: A adil@example.com

Created at: 2025-01-29

Status: active

Archived at: -

Published at: -

Updated at: 2025-01-29

Metadata:

Scanned at: -

Last quality check: -

File count: 0

Size: -

Resource: adiltest2-NFS-5

Access path: storage.staging.nodeum.io/adiltest2-NFS-5/adil-test-dataset-jan-29

Description: my desc

Permissions:

MA Full access

VG Write File Operation Read/Write Dataset Access

MG Read Dataset Access

AH Read/Write File Operation

Related datasets: -

Studies: adil test study

Publications: -

> 4704579

adil-test-dataset-p3

2025-01-28

2025-03-03

adiltest2-NFS-4

0

First

Previous

1

Next

Last

Datasets Management

- Metadata management
- Permissions Management
- Storage resources association
- Studies association for metadata management
- Storage Operation Overview
- Data Mover operations readiness

STORAGE SYSTEM MANAGEMENT



Nodeum

Storage Projects

Projects

Datasets

Deleted dataset

Data Management

Infrastructure

Overview

Resource

Mapping

UI Kit

Resources

#	Name	Type	Storage	
> 4215064	adiltest2-NFS-1		150.00 MB	
> 4399429	adiltest2-NFS-3		250.00 MB	
> 4704346	adiltest2-NFS-4		3.10 GB	
▼ 4730241	adiltest2-NFS-5		15.00 GB	

Storage operations

#	Created	Updated	System resource key	Project resource key	Operation type	Status	Semaphore task id
4730259	2025-01-29	2025-01-29	4209824	4730241	Create-ProjectResource	success	2147483621
4844017	2025-02-01	2025-02-01	4209824	4730241	Read-ProjectResource		
4879991	2025-02-02	2025-02-02	4209824	4730241	Read-ProjectResource		
4915969	2025-02-03	2025-02-03	4209824	4730241	Read-ProjectResource		
4969425	2025-02-04	2025-02-04	4209824	4730241	Read-ProjectResource		
5018803	2025-02-05	2025-02-05	4209824	4730241	Read-ProjectResource		
5063025	2025-02-06	2025-02-06	4209824	4730241	Read-ProjectResource		

First

Previous

1

Next

Last

Valery Guillaume

gs112308@uni-stuttgart.de

Beta version 0.4.0

Storage Resource Management

- Storage Provisioning
- Multi-protocol management
 - NFS, SMB, S3
- Integration with IBM Spectrum Protect for Tape Archiving.

METADATA SCHEMA MANAGEMENT



Nodeum

Valéry Guillaume
gsil2308@uni-stuttgart.de

Storage Projects

Data Management

Default Study

Metadata Schemas

Infrastructure

System resources

Overview

Resource

Mapping

UI Kit

Schema

#	Name	Created	Updated	
2136467	adil-basic-1	2024-11-28	2024-11-28	
2136485	adil-basic-2	2024-11-28	2024-11-28	
1255817	adil-schema-1	2024-11-07	2024-11-18	
1777546	Base schema for datasets	2024-11-20	2024-12-09	
1799602	Base schema for research software	2024-11-20	2024-11-20	
4310362	Test-Schema for pattern rendering	2025-01-20	2025-01-20	
4378268	TfAT Schema	2025-01-21	2025-01-21	
1692045	vgschema2	2024-11-18	2024-11-19	

30 / per page

Name*

Base schema for datasets

Document*

```
1- {
2-   "$schema": "http://json-schema.org/draft-06/schema#",
3-   "title": "Dataset",
4-   "description": "A base schema for datasets",
5-   "type": "object",
6-   "properties": {
7-     "contributors": {
8-       "description": "A person that contributed in a specific role to the dataset",
9-       "type": "array",
10-      "minItems": 1,
11-      "items": {
12-        "type": "object",
13-        "properties": {
14-          "name": {
15-            "type": "string",
16-            "description": "Name of the person in the format family name, given name"
17-          },
18-          "role": {
19-            "type": "string",
20-            "description": "Role of the person involved"
21-          },
22-          "orcid": {
23-            "type": "string",
24-            "description": "ORCID of the person involved"
25-          },
26-          "affiliation": {
27-            "type": "string",
28-            "description": "Organisation, the person is affiliated to"
29-          }
30-        }
31-      }
32-    }
33-  }
34-}
```

Or upload file

Click here to select or drop schema file here

Beta ver.

Close Update

Dataset

A base scheme for datasets

Contributors

A person that contributed in a specific role to the dataset

Name

Name of the person in the format family name, given name

Role

Role of the person involved

ORCID

ORCID of the person involved

Affiliation

Organisation, the person is affiliated to

Contact

Contact information

Contact name

Name of the person or organization that can be contacted regarding the data record

E-Mail-Adresse

Contact E-mail

Creation Date

Date at which the data record was created

Keywords

Keywords for the dataset

Schema Management

- Multiple schema capabilities
- Based on json-schema definition
- Export / Import schema

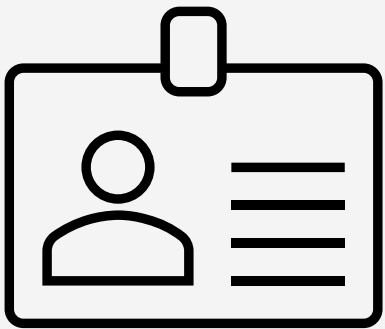
CONCLUSION

CONCLUSION



Nodeum address the challenges of **data discovery** and **data movement** with a **controlled** approach. This in a multi-tier storage architecture, when end-user focus on their business operation.

Thank you



Valery Guillaume
Nodeum, CEO
valery@nodeum.io

