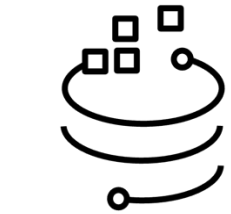
Nodeum
Data Management Software

IBM Storage Scale Days 2025 DE

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

Valery Guilleaume
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

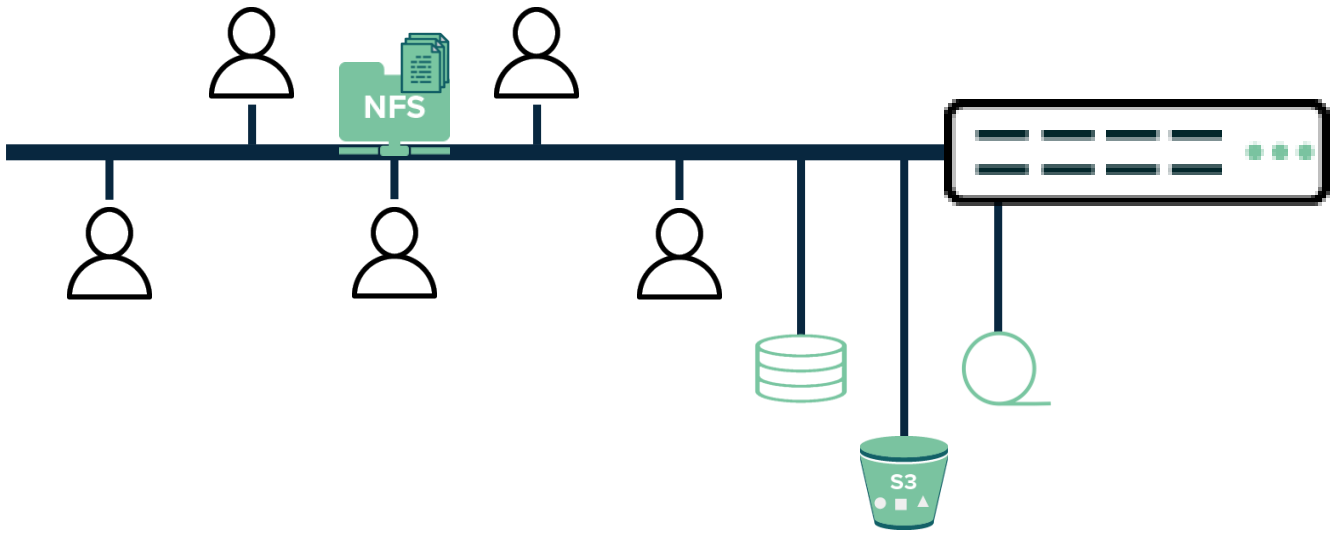
- Designed for **multi-directional** data movement across multiple storage media
- Supports on-premises and cloud storage
- Policy driven – providing accountability, traceability, and verification/validation
- Customers across all major industries

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?

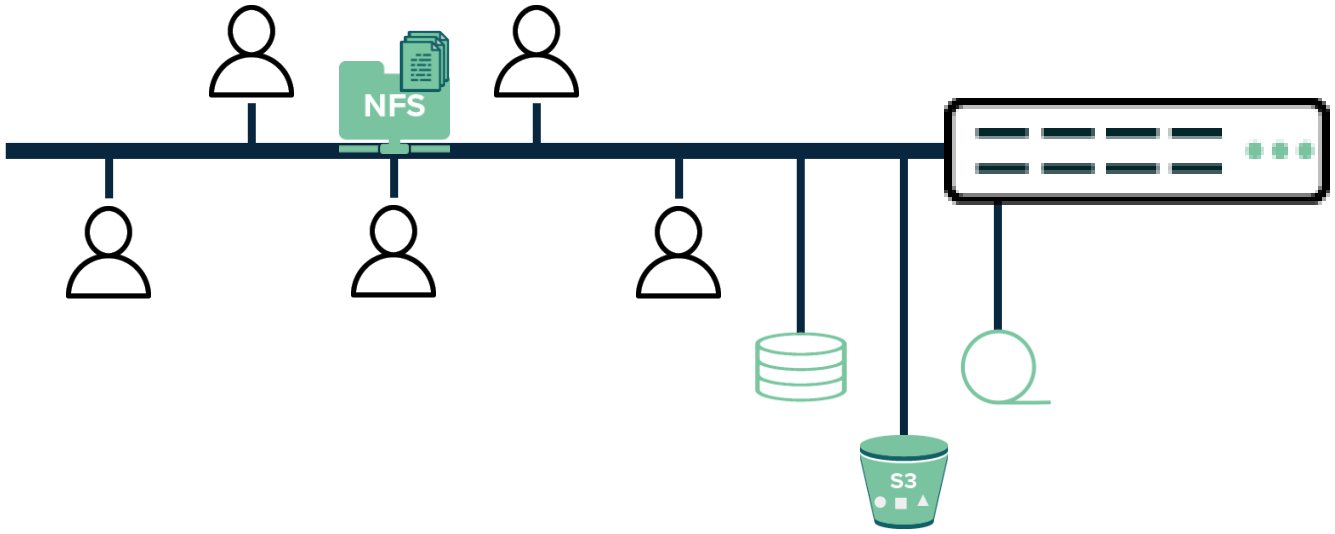
- Uni-directional data movers are perfect for:
  - Backup
  - Archiving
- They require an intermediary server like a backup server
- An administrator is required to manage this server
- Users have little or no autonomy as to how or where data is moved



WHAT IS NODEUM?

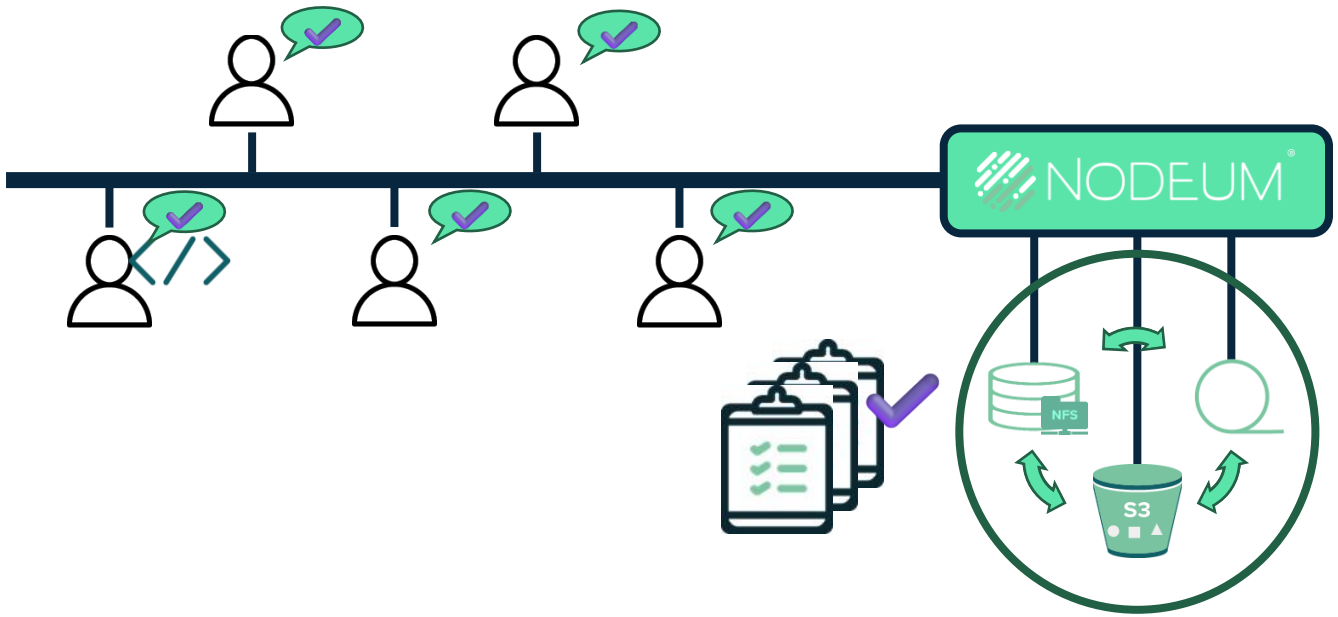
TECHNICAL USERS CAN MANUALLY MOVE DATA, BUT PERFORMANCE IS IMPACTED

- More technical users can move their own data with commands like:
  - Linux: *rsync, cp, mv*
  - Windows/Mac: drag and drop
- This solves the automony / 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	JÜLICH Forschungszentrum · JÜLICH SUPERCOMPUTING CENTRE · FENIX RI · VŠB TECHNICKÁ UNIVERZITA OSTRAVA · IT4INNOVATIONS NÁRODNÍ SUPERPOČÍTAČOVÉ CENTRUM · BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación · CINECA · cea
Research & University Life Science	University of Stuttgart Germany · UNIVERSITY OF HOHENHEIM · Universität Rostock Traditio et Innovatio · Medical University of Graz · KLINIKUM INGOLSTADT · IPG
Media Post Production	MPC FILM & EPISODIC · Flyover · MIKROS technicolor · WATCHNEXT · RTL BELGIUM
Earth Observation	envirosense the remote sensing professional · THEOS 2 · AIRBUS · Dewberry®

And also in additional sectors:

- **Financial Services**
- **Gov. / Administration**
- **Retail / Manufacturing**
- **Services**

USE CASE

FENIX-RI

www.humanbrainproject.eu – www.fenix-ri.eu

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.

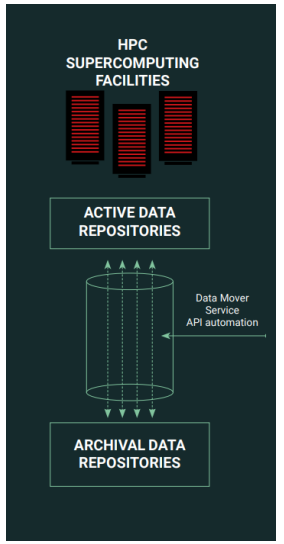
Sharing of data and FAIR data management is becoming mandatory. This adds to the requirements for modern data infrastructures.

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



JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

cea

CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre

CINECA

FENIXRI

PRODUCT OVERVIEW

MODERN ARCHITECTURE

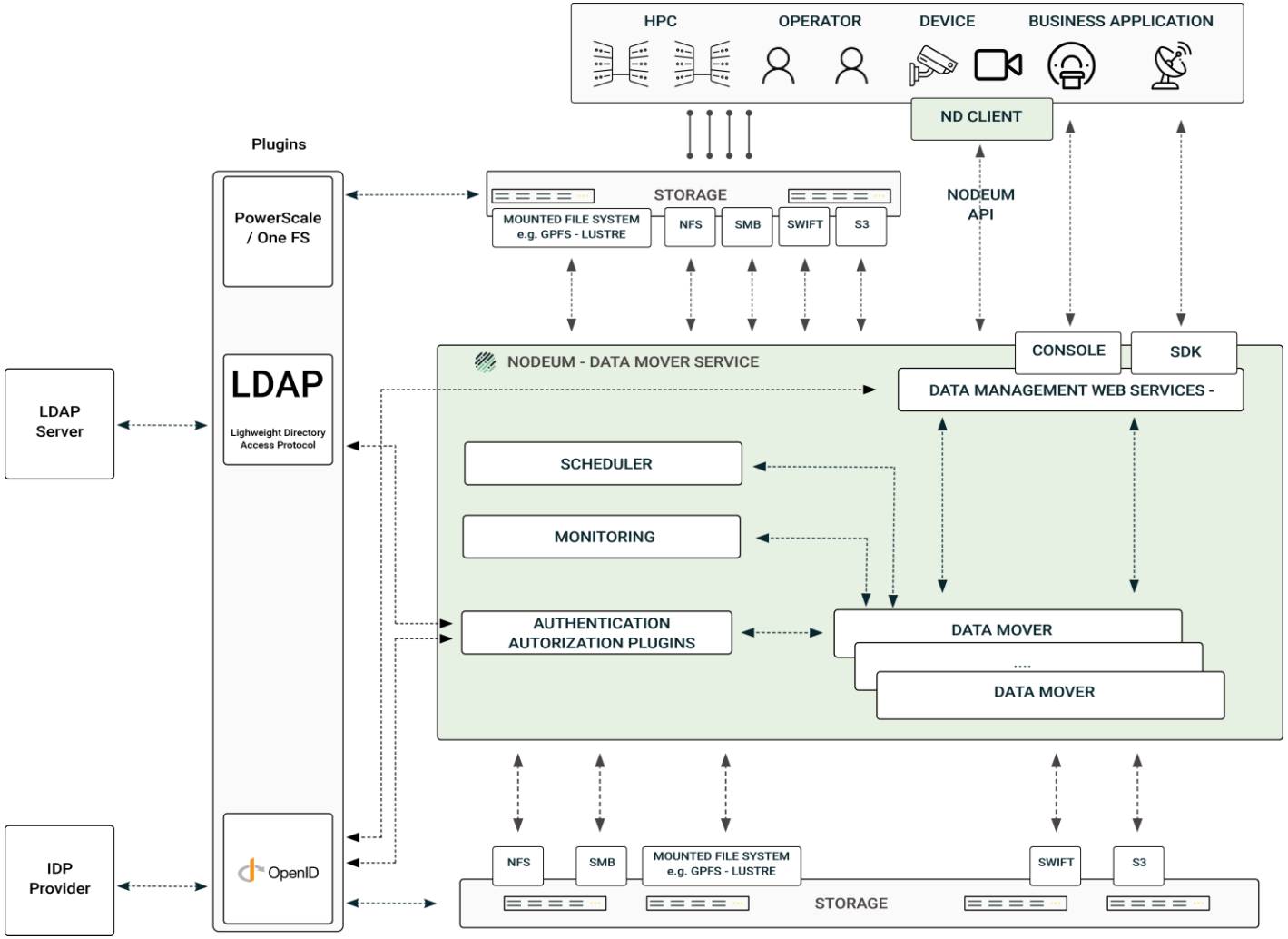
HIGHEST DATA MOVEMENT THROUGHPUT - READY FOR EXASCALE COMPUTING

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

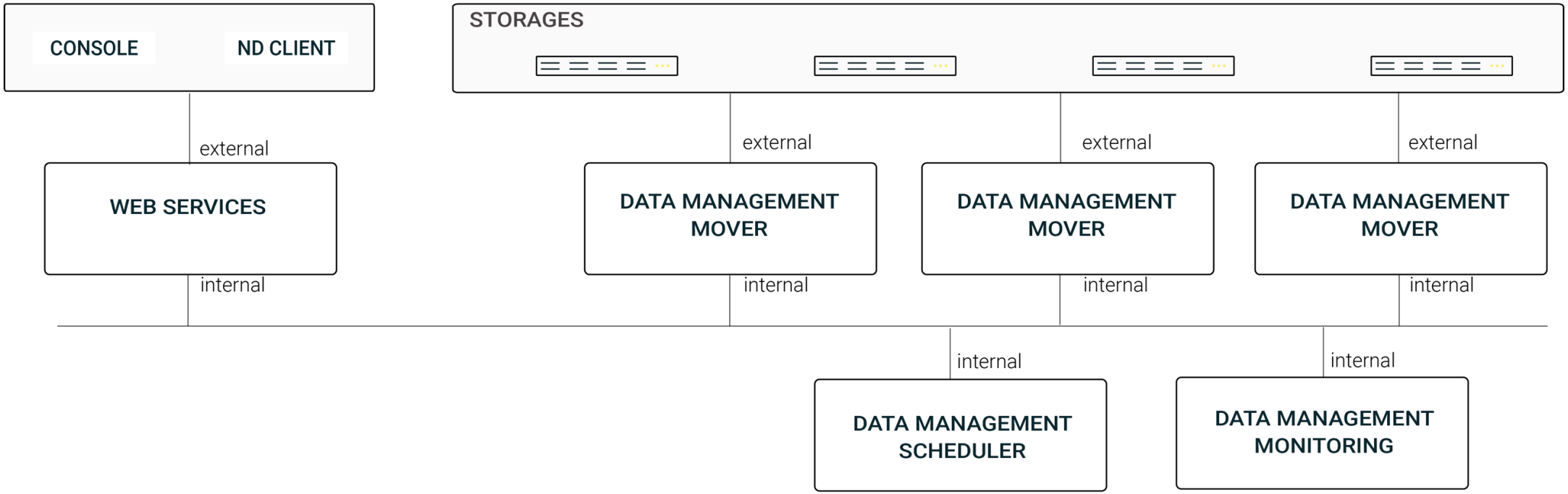
DOCKER AND APP MARKETPLACE READY



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.

IBM Spectrum Scale

LucidLink

And others …

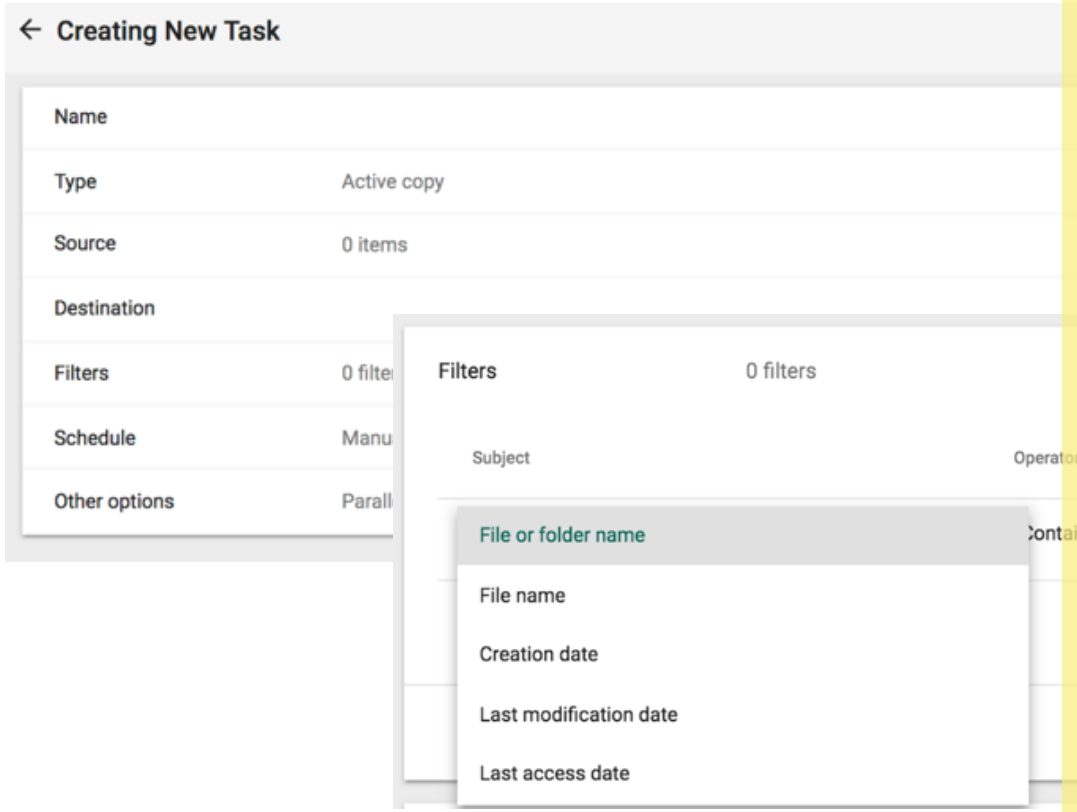
lustre™

BeeGFS®

StorNext®

POLICY-BASED WORKFLOW ORCHESTRATION

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



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

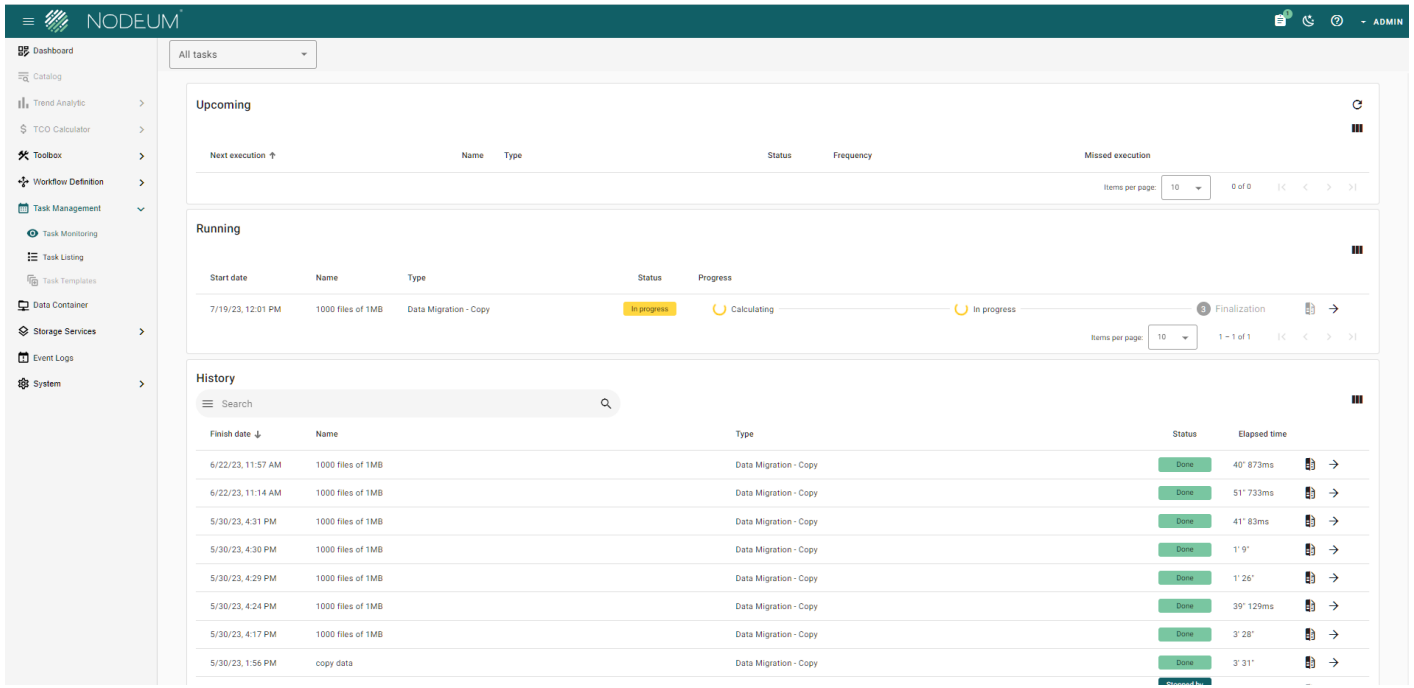
Schedule

- Automatic
- Manual
- Scheduled

CONSOLE

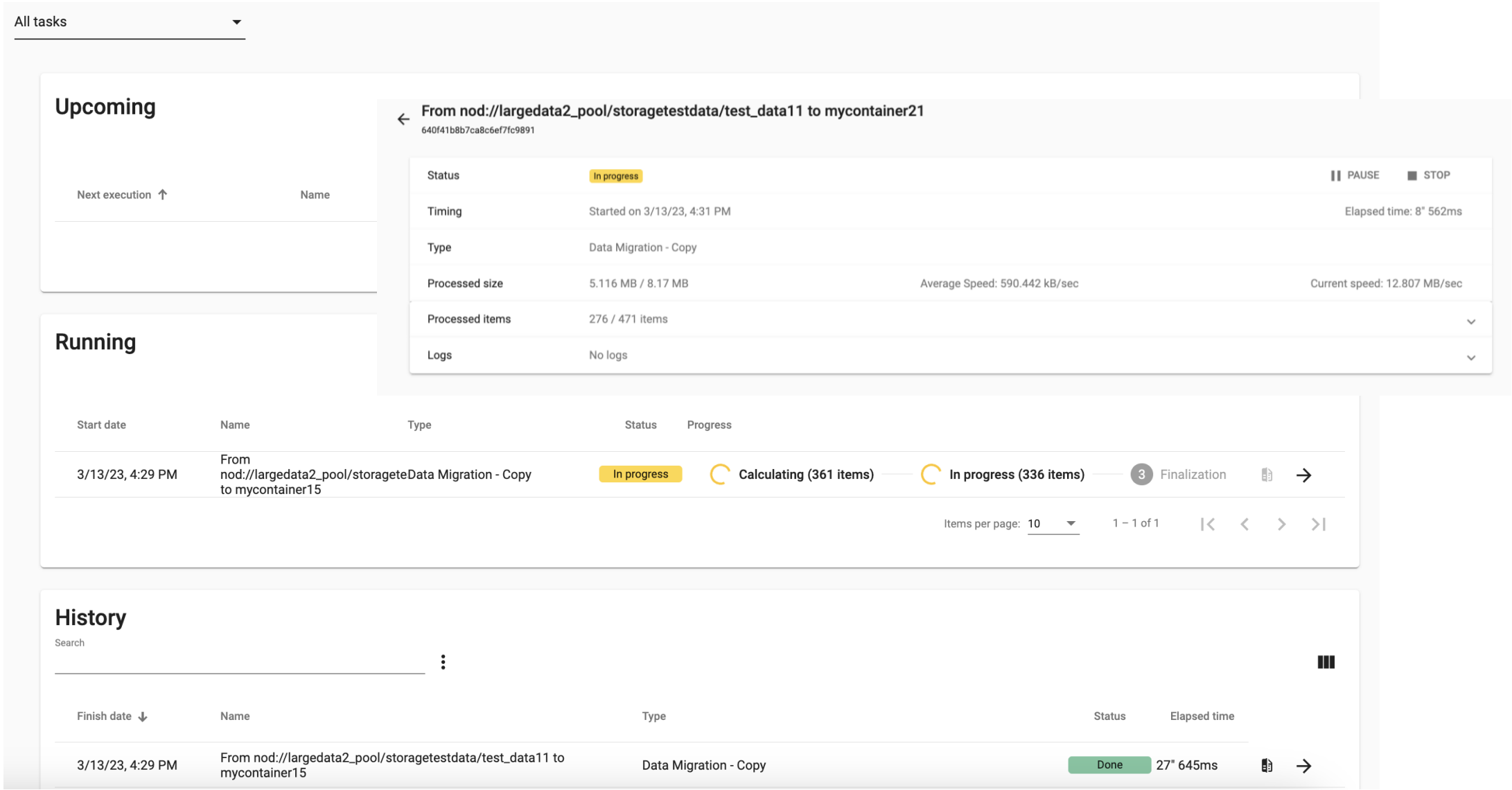
This is a secure HTML5 interface which provides a modern way to execute data management operations.

- 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.

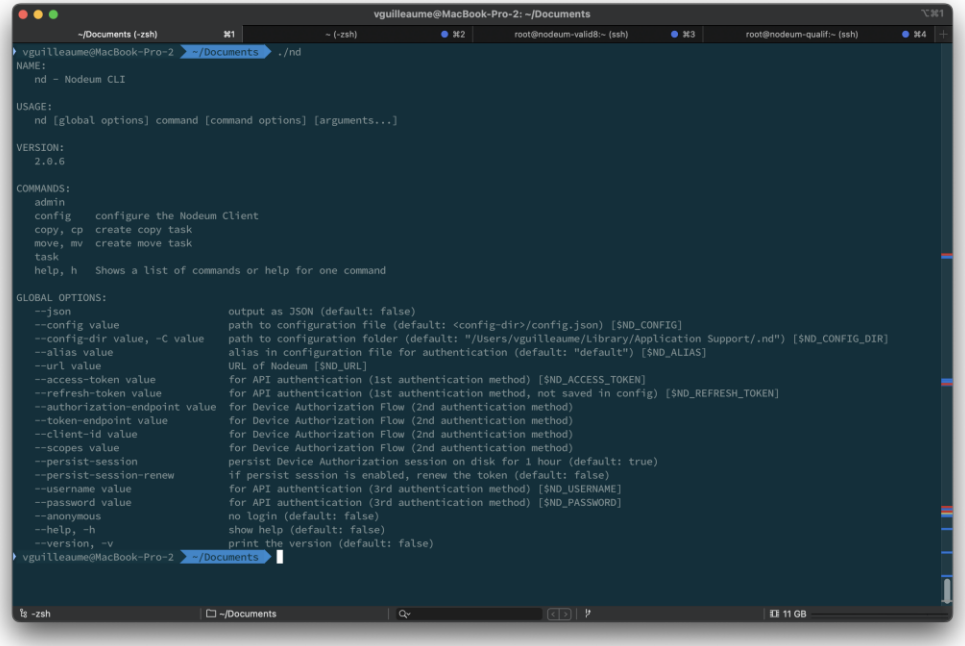
DATA MOVEMENT SUPERVISION



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.

INTEGRITY VERIFICATION

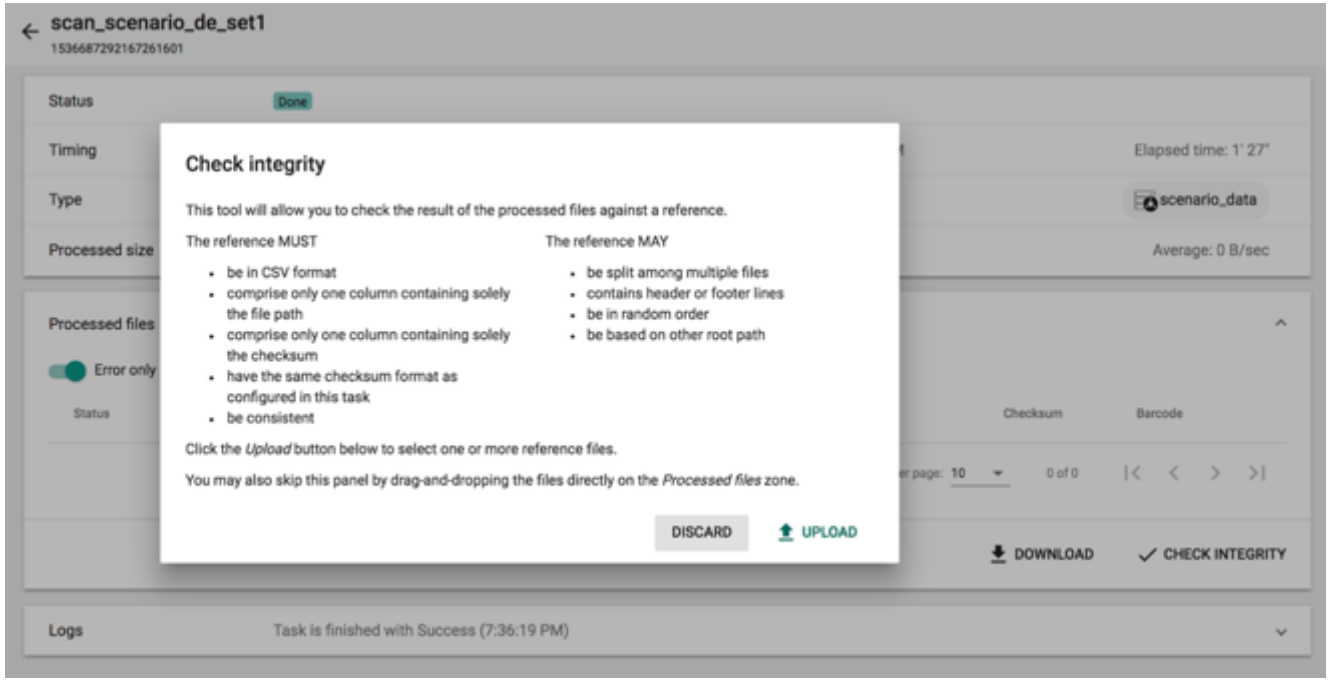
Data movement

Managed by Nodeum

Data is read from the source

Data is copied

Checksum is calculated



CALCULATION IN TRANSIT

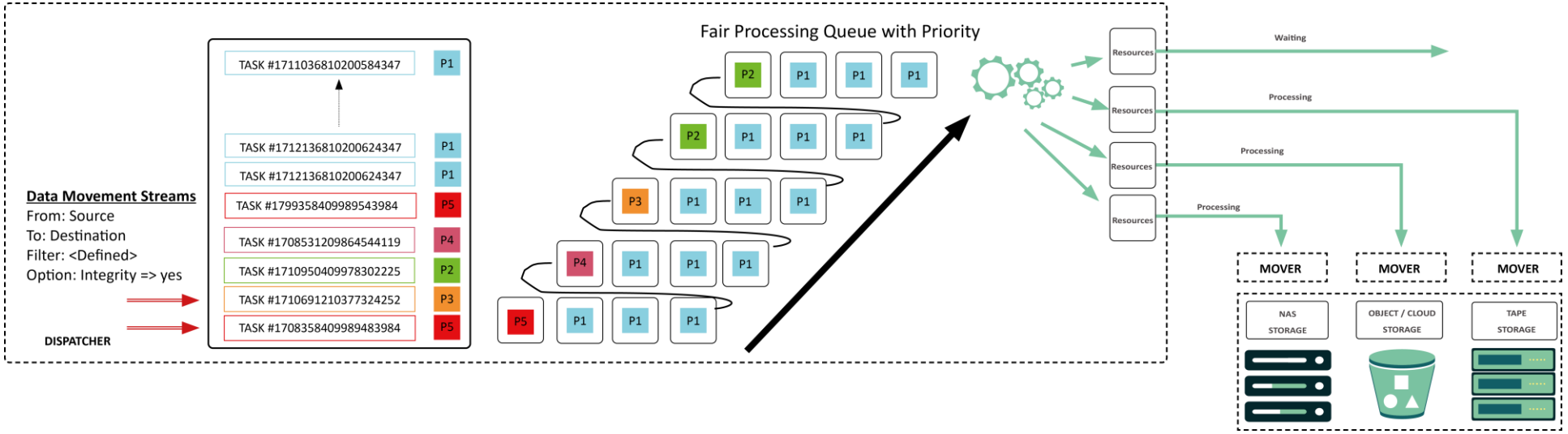
CRC32 – MD5 - XXHASH

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 mytestext -d
# file: mytestext
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”

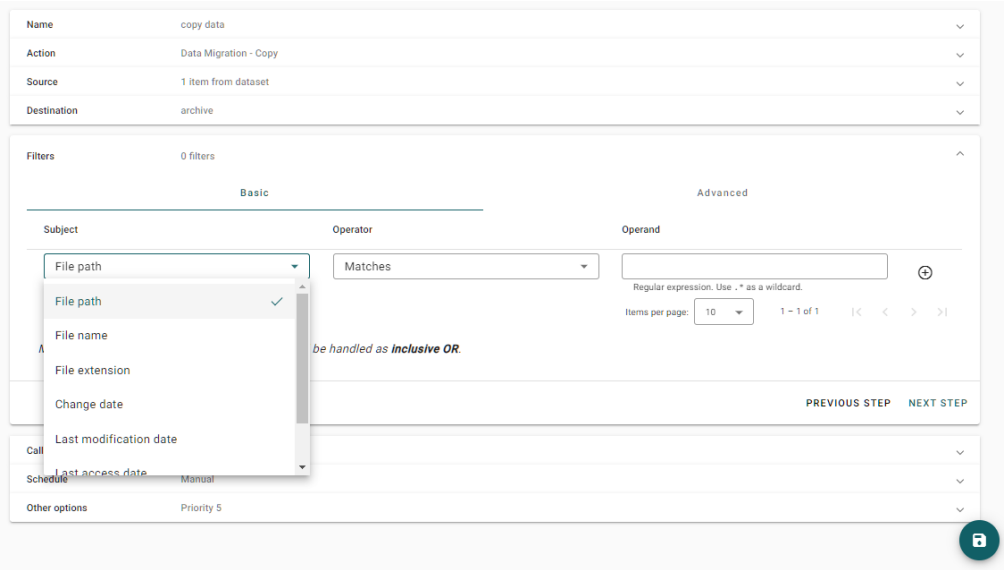
```
x-amz-meta-mode
x-amz-meta-uid
x-amz-meta-gid
x-amz-meta-atime
x-amz-meta-mtime
x-amz-meta-ctime
```

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.



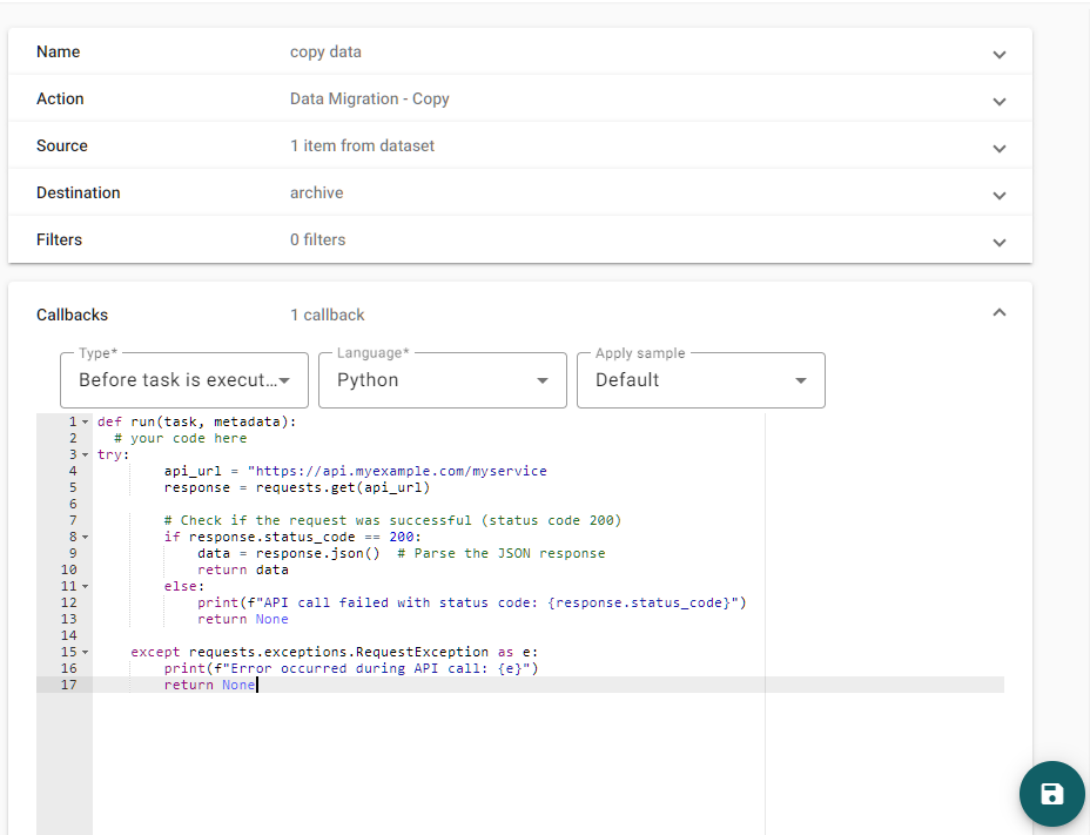
> Highly customizable, allowing users to create filters that are tailored to their specific needs. This can include setting up filters for specific file types, folders, or directories, as well as creating custom rules based on metadata or other criteria.

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.



CONTROL TREE STRUCTURE

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

Examples:

With --wd=.

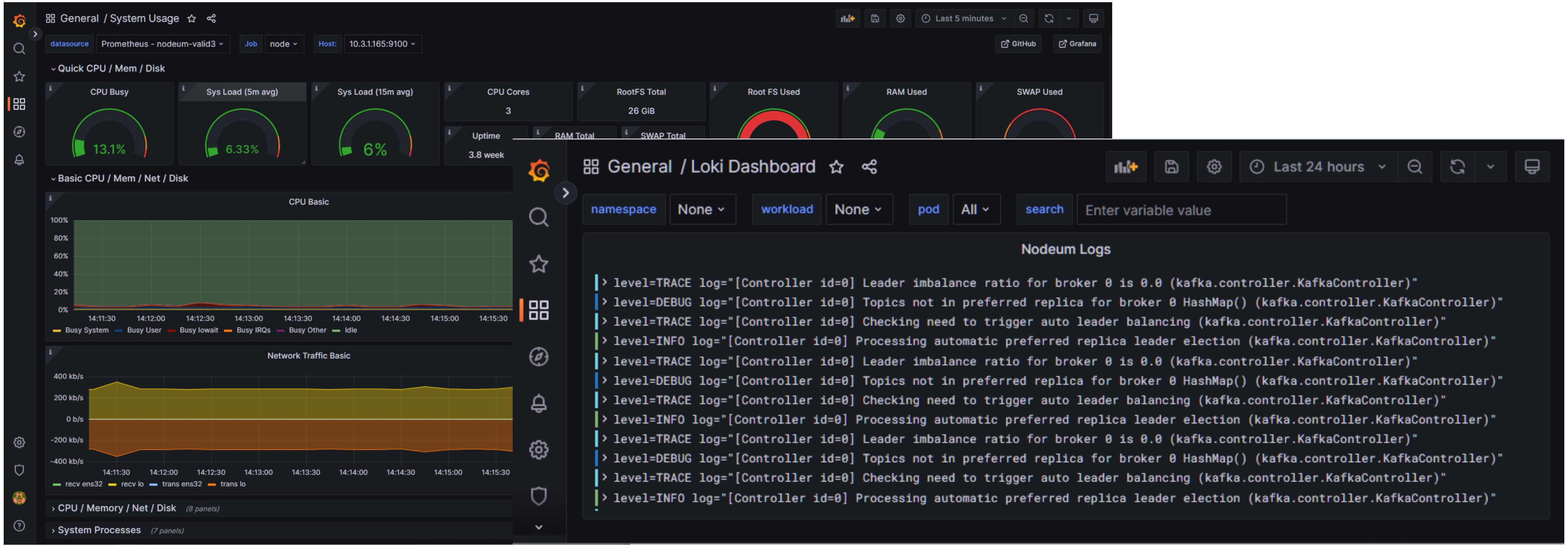
Source	Destination	Result
nod://source/folder/FILE.txt	nod://dest/directory/	nod://dest/directory/FILE.txt
nod://source/folder/FILE.txt	nod://dest/RENAMED.txt	nod://dest/RENAMED.txt
nod://source/folder/	nod://dest/directory/	nod://dest/directory/FILE.txt
nod://source/folder/	nod://dest/directory	nod://dest/directory/FILE.txt
nod://source/folder	nod://dest/directory/	nod://dest/directory/folder/FILE.txt
nod://source/folder	nod://dest/directory	nod://dest/directory/FILE.txt

With --wd=..

Source	Destination	Result
nod://source/folder/FILE.txt	nod://dest/directory/	nod://dest/directory/folder/FILE.txt
nod://source/folder/FILE.txt	nod://dest/RENAMED.txt	nod://dest/RENAMED.txt
nod://source/folder/	nod://dest/directory/	nod://dest/directory/folder/FILE.txt
nod://source/folder/	nod://dest/directory	nod://dest/directory/FILE.txt
nod://source/folder	nod://dest/directory/	nod://dest/directory/source/folder/FILE.txt
nod://source/folder	nod://dest/directory	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.

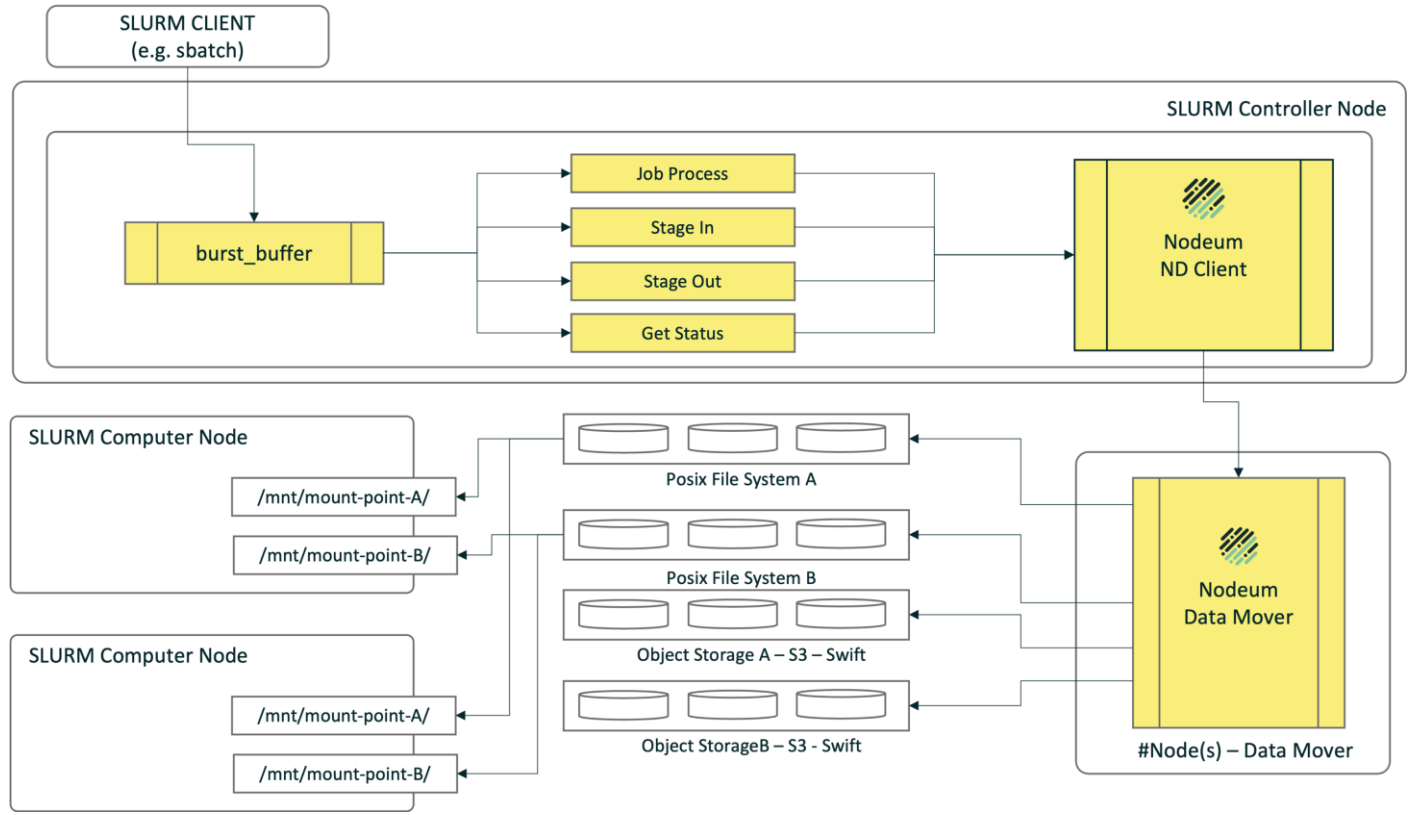


WORKLOAD MANAGER INTEGRATION

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:

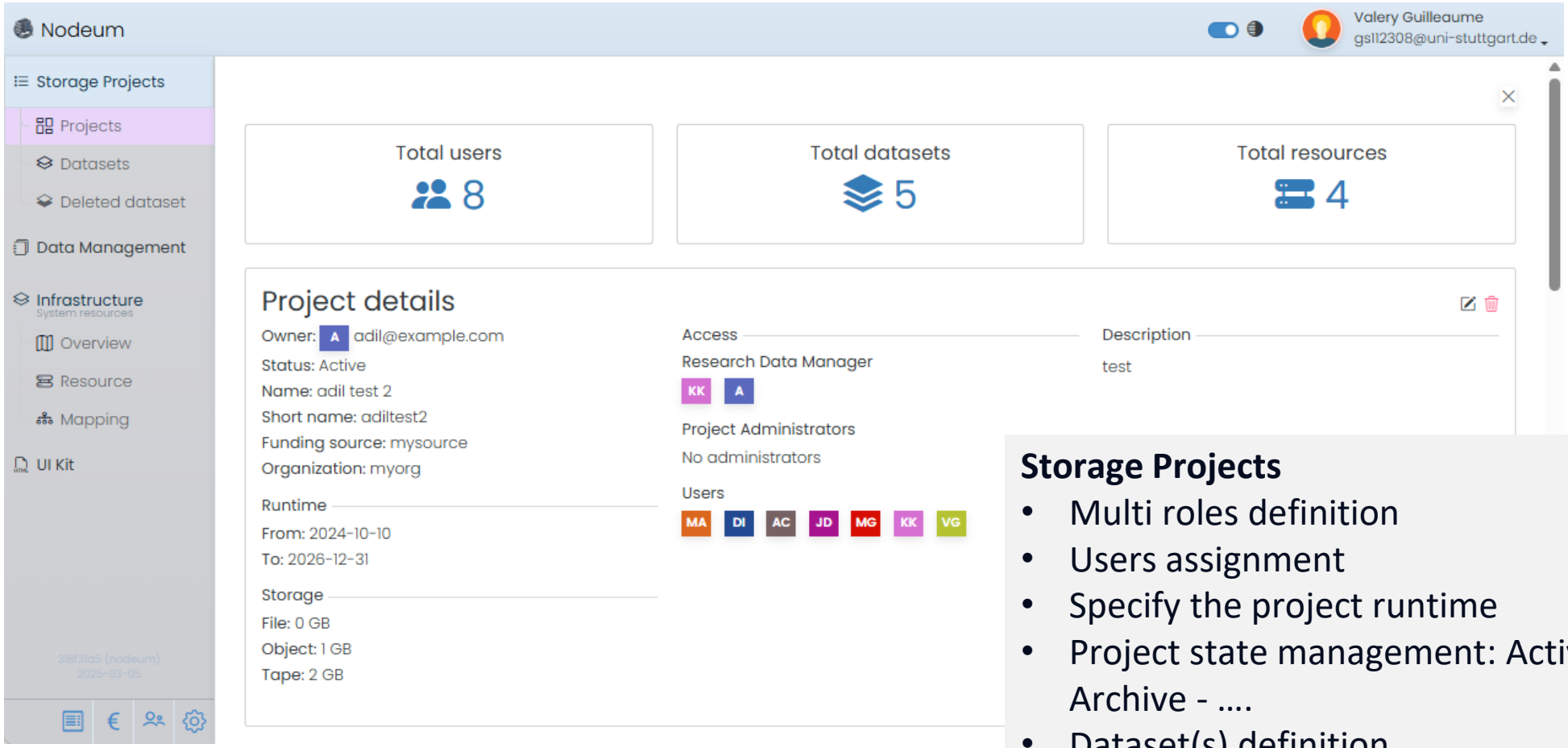
Help organizations to adopt F.A.I.R. data principles (Findable, Accessible, Interoperable, Reusable) and address evolving scientific requirements closely linked to storage technologies. Strengthen the connection between storage data movement, archiving, and project storage management, including metadata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

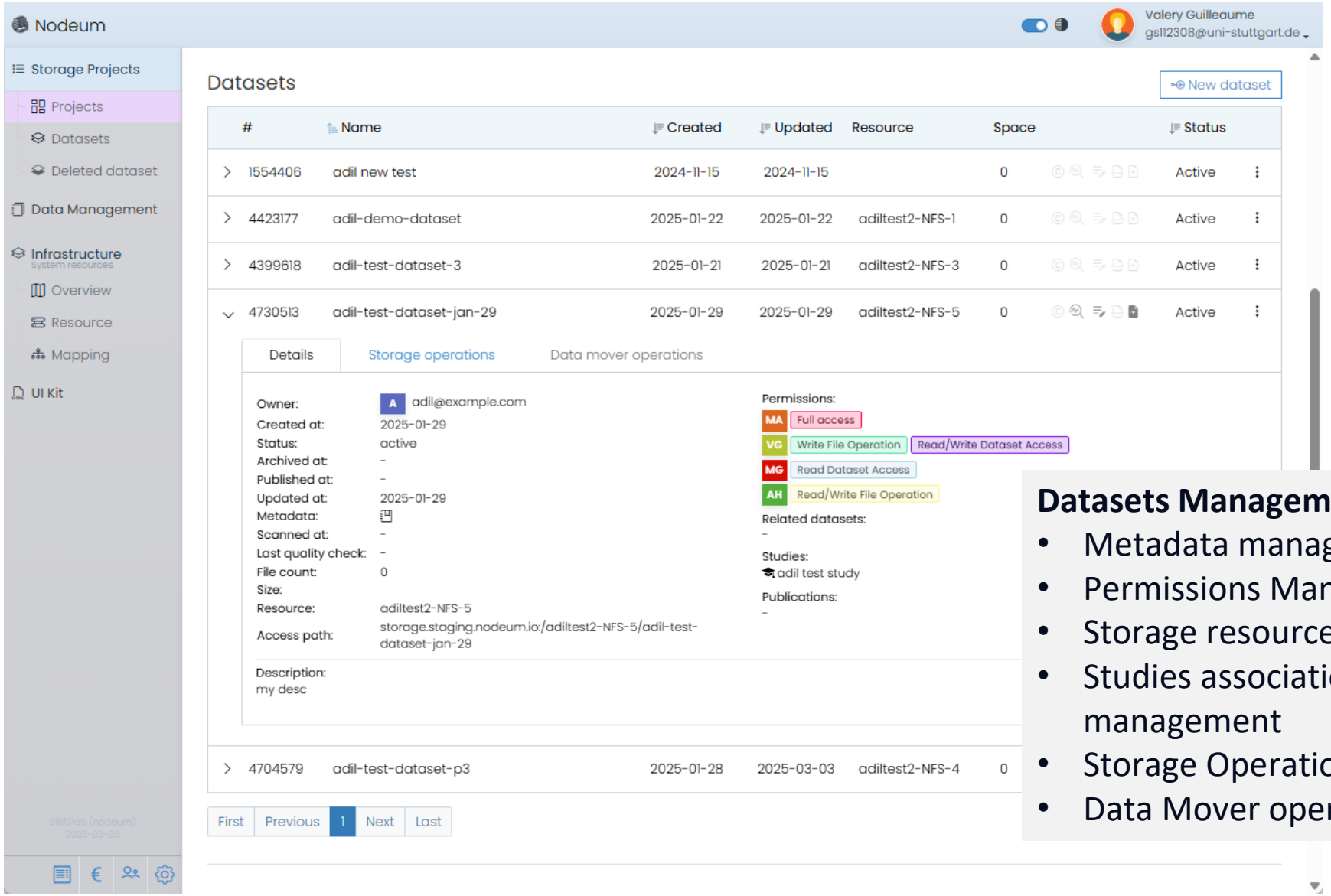


Beta version 0.4.0

Storage Projects

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

DATASETS MANAGEMENT

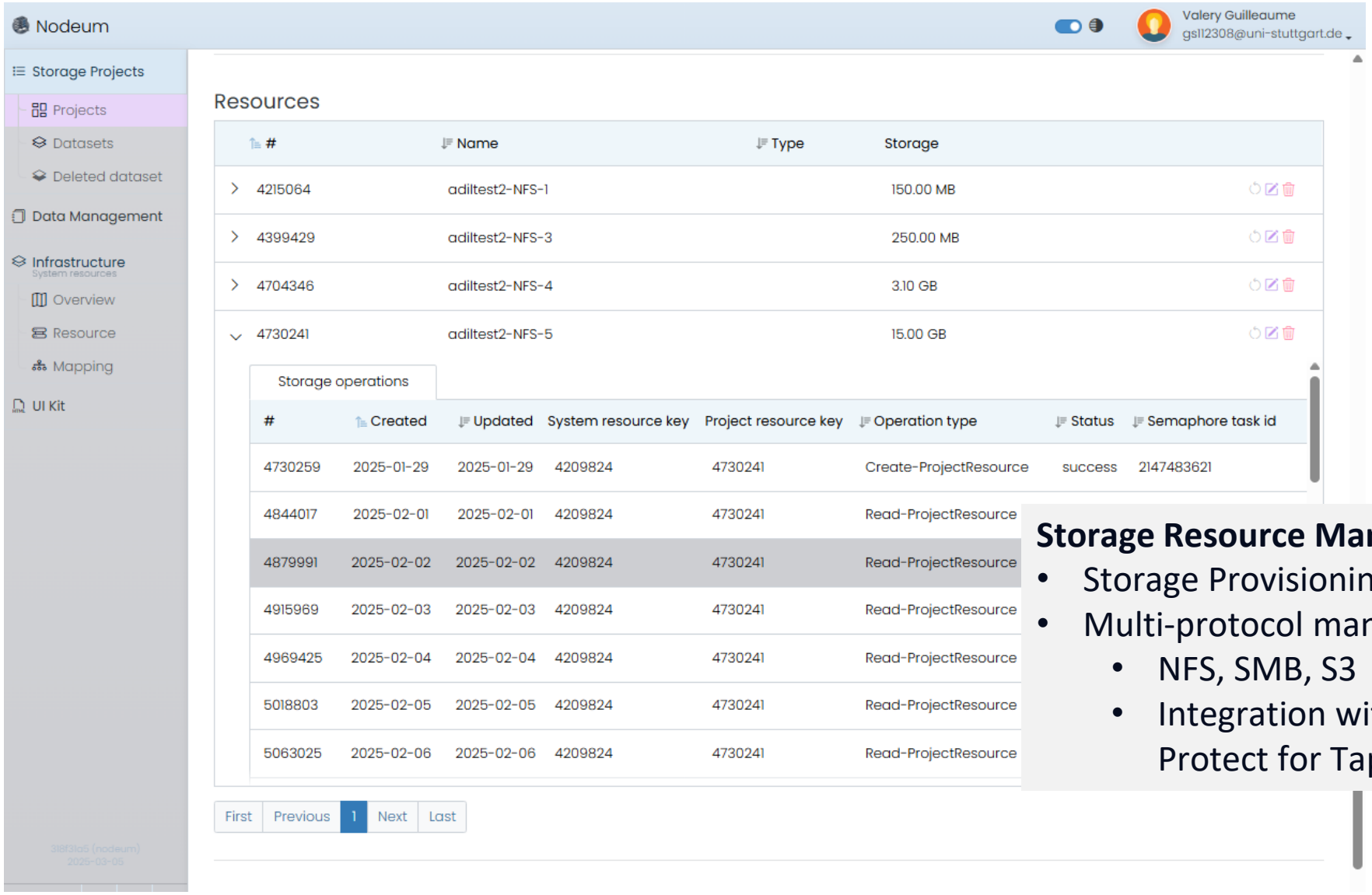


Datasets Management

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

Beta version 0.4.0

STORAGE SYSTEM MANAGEMENT

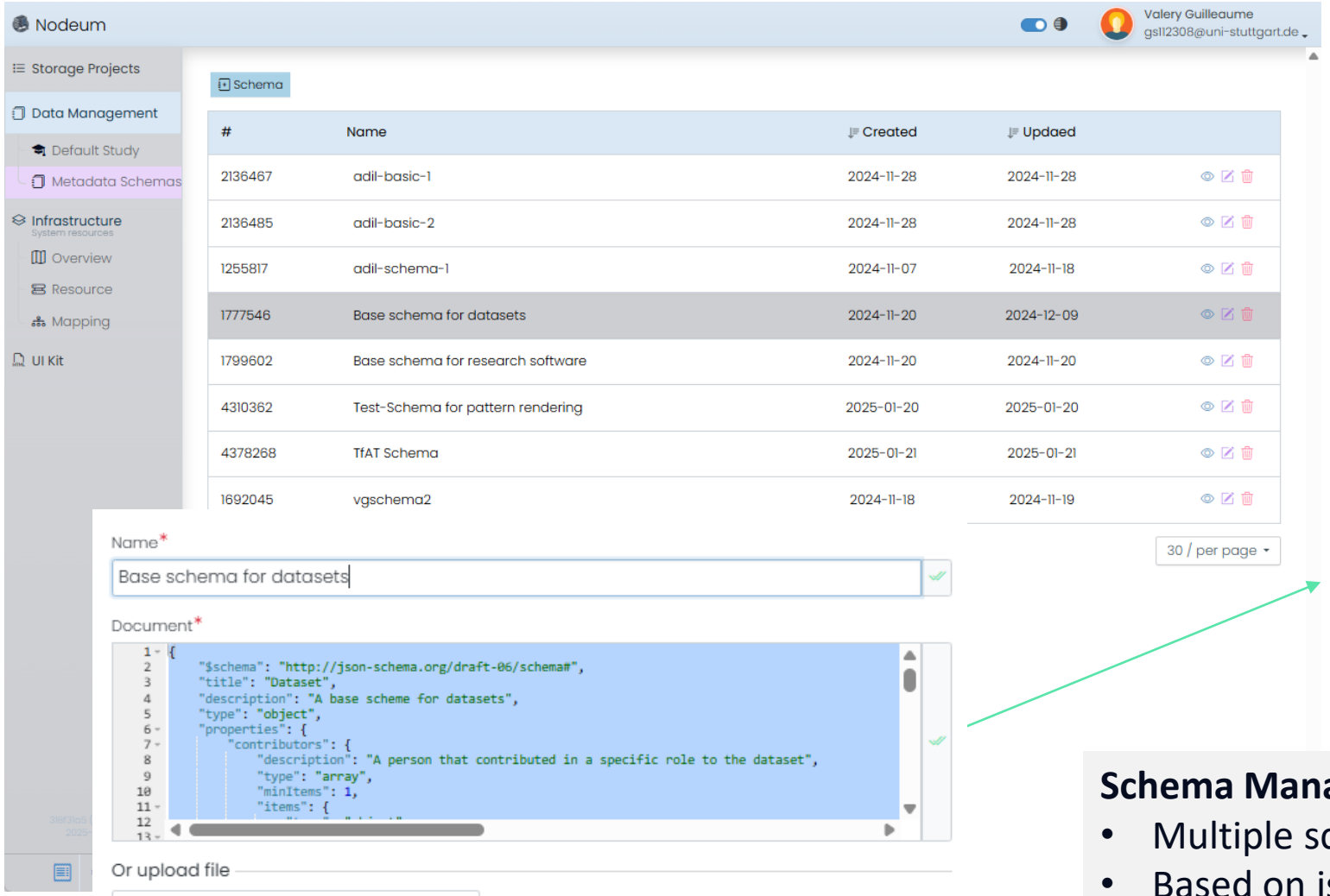


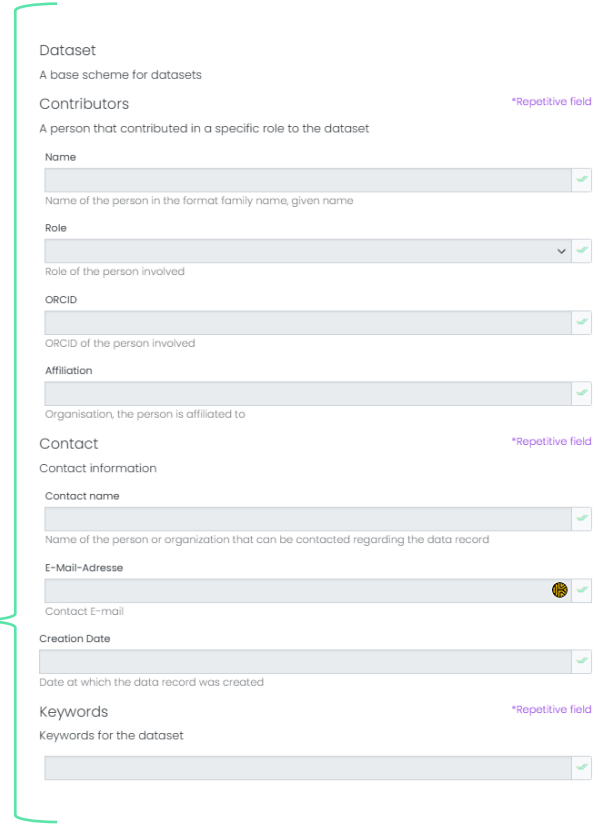
Beta version 0.4.0

Storage Resource Management

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

METADATA SCHEMA MANAGEMENT





Beta ver

Schema Management

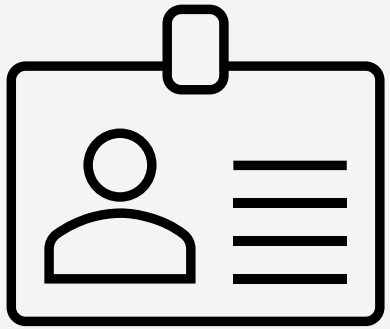
- Multiple schema capabilities
- Based on json-schema definition
- Export / Import schema

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 Guilleaume
Nodeum, CEO
valery@nodeum.io