

LimScaup

Open-Source Metadata Analytics for Lustre and Beyond

EOFS 2026

Janos Blum

Deutsches Klimarechenzentrum (DKRZ)

German Climate Computing Centre (DKRZ)

- Non-profit and non-commercial limited liability company since 1987
- The compute centre for German weather and climate research
- 100+ employees



HLRE-4 – Levante (2022-2028)



BullSequana, 3,000+ Nodes, 370,000+ Cores, AMD Milan, 14 PFLOPS

815 TB Main Memory, 120 PB Disk Storage,
30 PB Objectstore, 60 GPU Nodes

Storage at DKRZ

- Home
 - 300 TiB NVME (VAST)
- Project
 - 120 PiB lustre (DDN)
- Object Stores
 - 30 PiB MinIO (self maintained)
- HSM
 - 380 PiB LTO 8 (Stronglink / Versity)

The Problem

- Storage utilization grows
- Data becomes stale but is rarely removed by users
- Unix *du* can takes days to weeks for large projects

The Solution

- Store every single metadata per inode
- Compute the diff at regular intervals and store those for space efficient storage of historical metadata
- Offer efficient queries to the data
- Pre Aggregate common queries for web interactivity

```
[k202181@l40053 ~]$ limscaup-du /work/ks1522/limscaup/storage_metadata_lake/
```

path_like varchar	last_db_update timestamp	sum_size int128	sum_size_human varchar	count_paths int64
/work/ks1522/limscaup/storage_metadata_lake/	2026-09-11 00:00:00	6921485256802	6.2 TiB	719094

full_sub_dir varchar	last_db_update timestamp	sum_size int128	sum_size_human varchar	count_paths int64
/work/ks1522/limscaup/storage_metadata_lake/history	2026-09-11 00:00:00	6737877478087	6.1 TiB	691948
/work/ks1522/limscaup/storage_metadata_lake/current	2026-09-11 00:00:00	183460018997	170.8 GiB	27116
/work/ks1522/limscaup/storage_metadata_lake/daily	2026-09-11 00:00:00	147759718	140.9 MiB	30

Query took 0.41 seconds.



We have more than Lustre

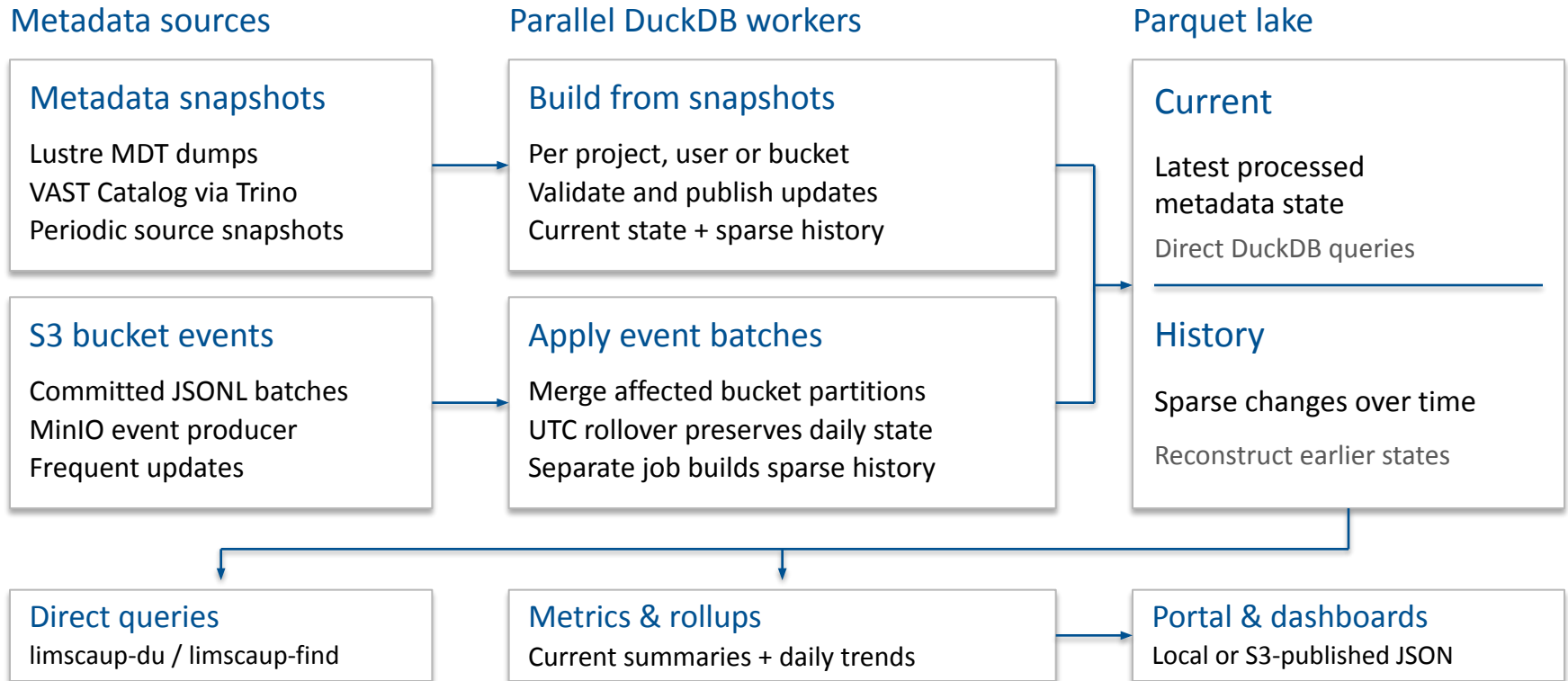
- VAST
 - /home
 - /shared
 - /s3
- MinIO S3
 - /minio-s3/<bucket>

Two new Adapters

- VAST
 - We gather a full metadata snapshot on a defined interval
 - As current as the specified interval (tested 3 hourly) but daily is sufficient
- MinIO
 - We send every single bucket event to a Galera Cluster
 - Write these logs in batches (flushed after 15 Minutes)
 - Advance the State with every processed log
 - As current as a batch is written to Disk

From Storage Metadata to Queries and Insights

Configuration + Slurm controller: schedule stages when inputs are ready



CLI: limscaup-du, limscaup-find

■ support of SQL predicate filtering

- `limscaup-du --filter "atime_ms > '2026-01-01'"`
- `limscaup-du --filter "uid = 20865"`
- `limscaup-du --filter "size = 0"`
- `limscaup-du --filter "path GLOB '*.png'"`

shows size of dirs accessed since 2026-01-01

shows size of dirs for the given user

shows dirs with 0 size files

shows dir size respecting only PNG files

■ show additional metadata

- `limscaup-du --columns </mount>`
- `limscaup-du --add`

shows which metadata is available for that mount

adds additional metadata like *atime_ms*, *uid*, *mode*

■ aggregate additional metadata

- `limscaup-du --add size:avg --add atime_ms:min`

last, first, any, max, min, avg, sum, count

■ store results

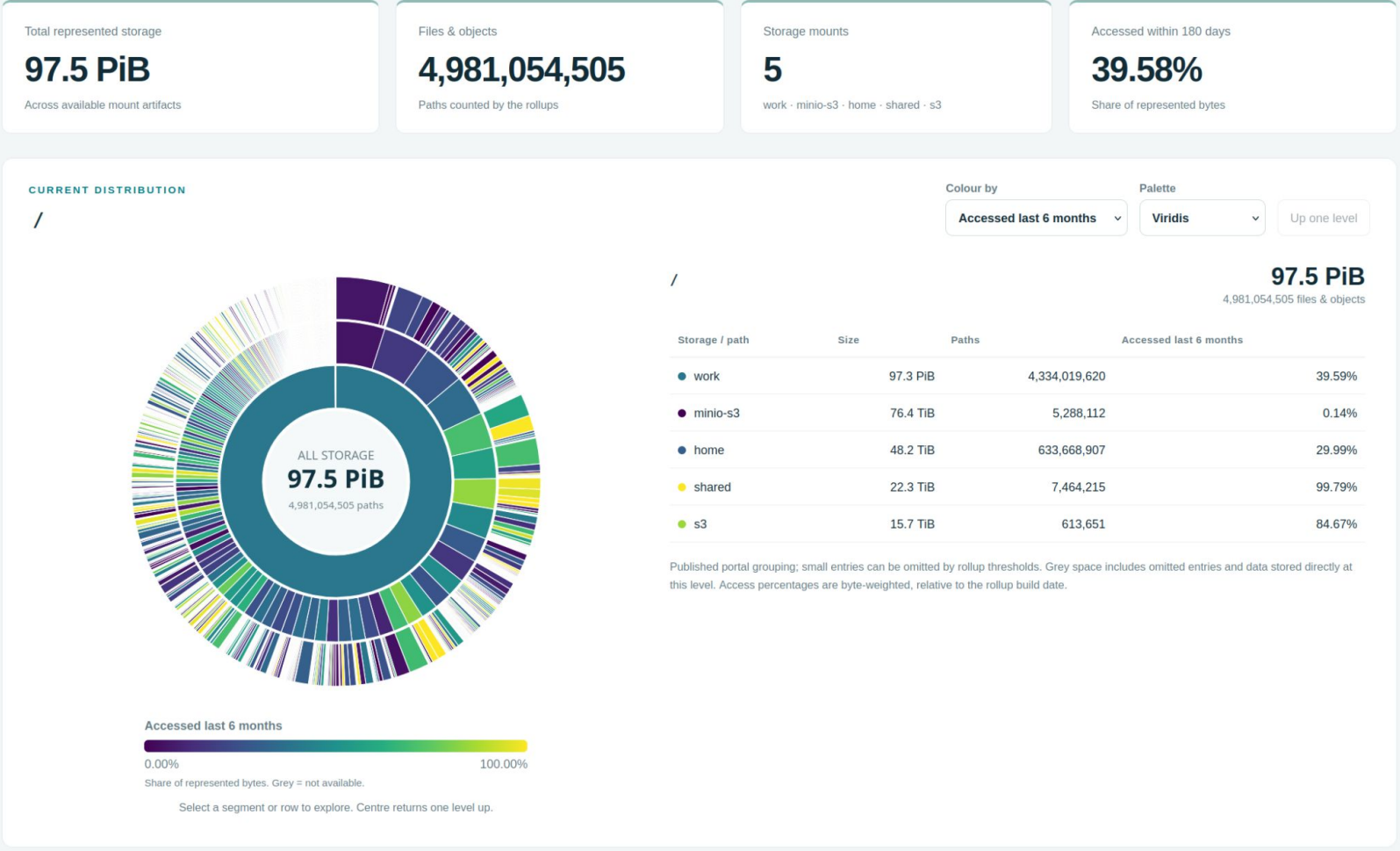
- `limscaup-du --save /path/file.csv`

store as csv, json or parquet

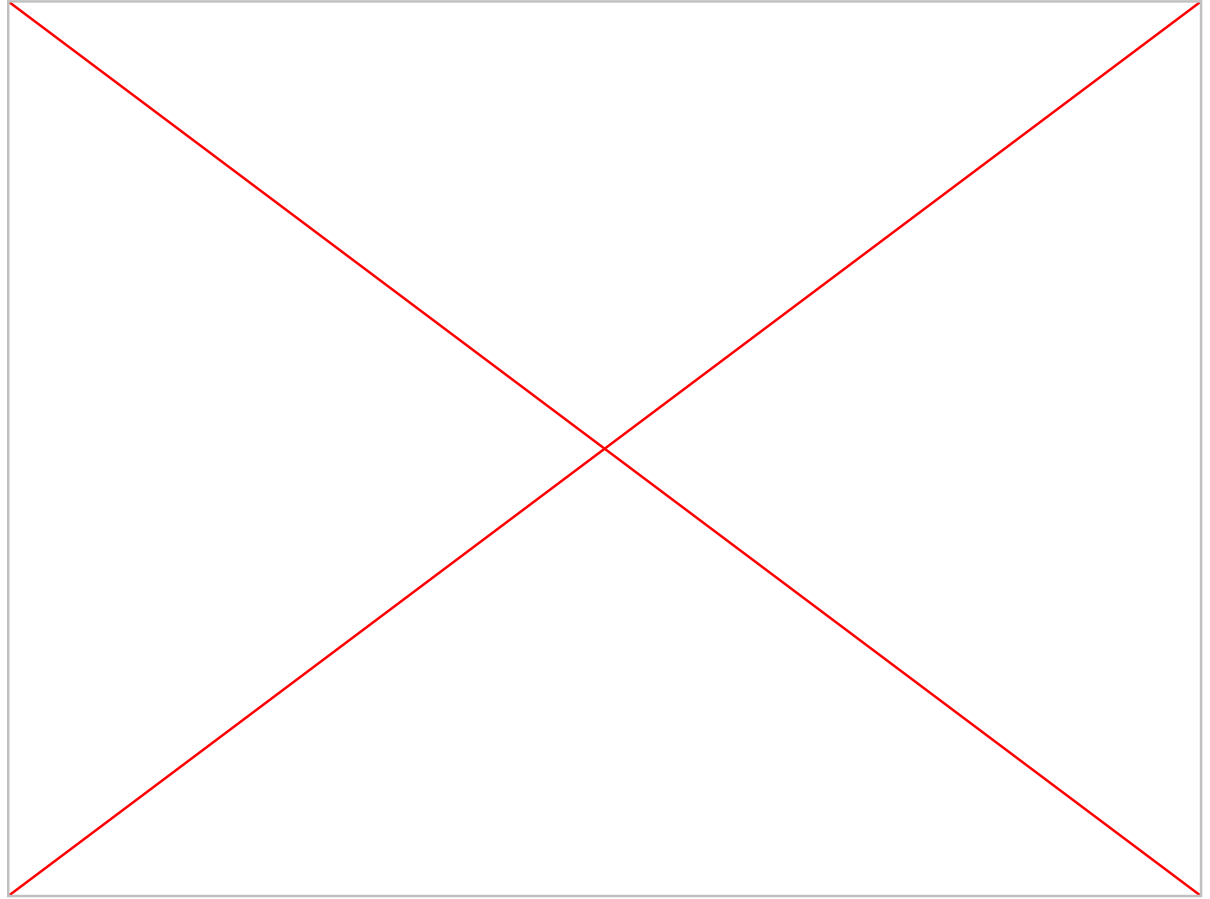
Data Privacy

- Maintaining access rights is critical on a shared file system, even if it is just metadata
- As the data lake consists of files, we apply group rights such that users can only query those parts of the datalake that match with their group or user memberships.
- History lake is for internal use only

Current Dashboard



History Dashboard



Performance

	Throughput	Memory Footprint	Parallelization
unix du	200 inodes/s	< 10 MiB	single threaded
limscaup-du (current)	10 - 30M inodes/s	100 - 200 MiB	utilizes a full node if permitted
limscaup-du (history)	100.000 to 1M inodes/s	0.5 - 1.5 GiB per 1M inodes	utilizes a full node if permitted

Compute Costs

work / Lustre · average snapshot · DKRZ Levante

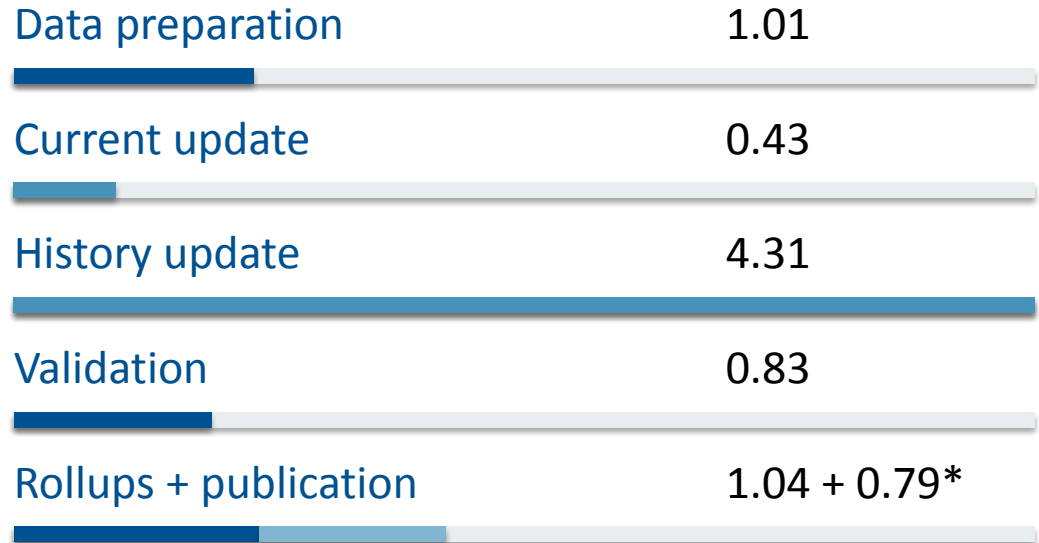
7.6

work node-hours
per snapshot

+ 0.8

shared node-hours
for rollups + publication

Node-hours by stage



*1.04 h work metrics + 0.79 h shared rollups/publication across mounts.

Storage Cost

Indexed file sizes · 15 Sep 2026 metadata snapshot

Mount	Current	History	Growth / 30 days
work	116.9 GiB	1.99 TiB	+75.2 GiB
home	53.8 GiB	101.0 GiB	+7.53 GiB
shared	746.4 MiB	3.42 GiB	+1.23 GiB
s3	72.1 MiB	186.2 MiB	+53.3 MiB
minio-s3	491.2 MiB	882.0 MiB	No monthly baseline yet

Active lakes only; backup generations, staging and rollups excluded.

Benefits and Limitations of the Metadata Lake Approach

Benefits

- Fast metadata queries
- Decoupled from file system traversal
- Enables historical analysis
- Scales to billions of inodes
- Minimal operational overhead
- Cross-filesystem
- High query flexibility
- No centralized database server
- Open data formats (Apache Hive Parquet)

Limitations

- Close to real-time
- Currently only running at DKRZ
- Requires additional disk space (though relatively small)

Roadmap

- HSM Adapter (Versity)
- Grafana Dashboards
- GPFS Adapter
- Add DuckLake as optional store

Questions

Questions?

Janos Blum (blum@dkrz.de)

LimScaup (<https://gitlab.dkrz.de/workflows/limscaup>)