A Cloud-Optimized GNSS Data Lakehouse for Scalable Ingestion, Processing, and Analysis

Markus Bradke, Nils Brinckmann

GFZ Helmholtz Centre for Geosciences

IGS Workshop 2026, Santiago de Chile, 1 June 2026

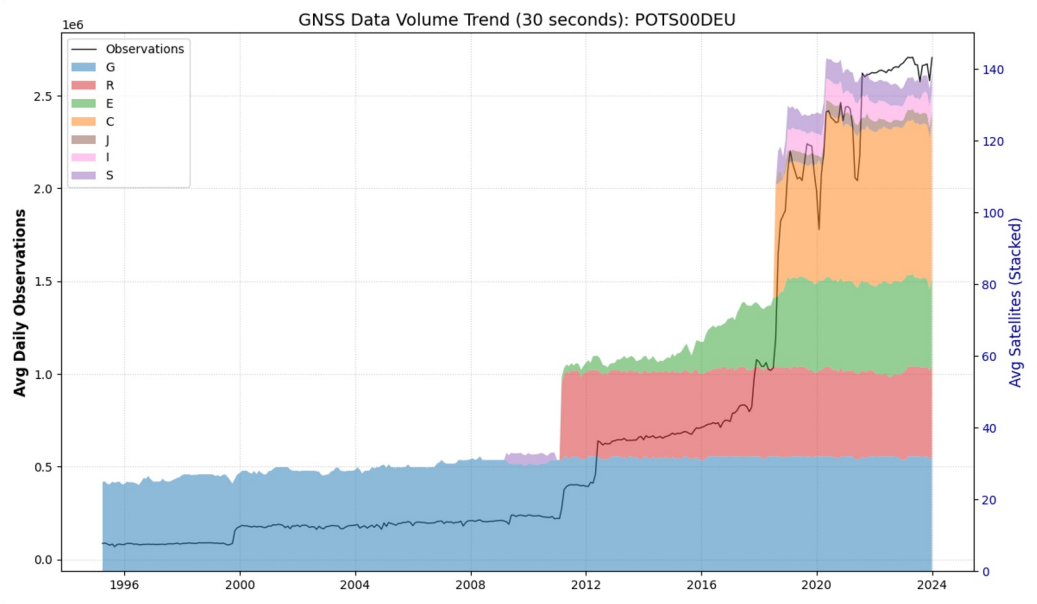
GFZ Helmholtz Centre for Geosciences

HELMHOLTZ

Motivation

Motivation

Data Volume



- Dense networks
- Multiple constellations
- Multiple frequencies and tracking modes
- Higher sample rates (1, 5, and up to 50Hz)

Motivation

RINEX vs. Object Store

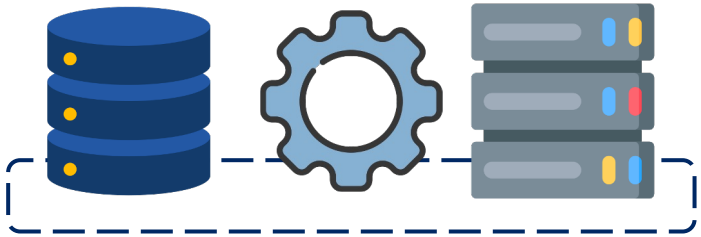
RINEX is optimized for file exchange, not cloud-scale analytics.

- RINEX is file-oriented (POSIX), S3 is an object store (HTTP-based)
- Small-files problem = High Cost & Latency
- Inefficient access patterns for analytics (full-file scans)
- Poor scalability for many files (slow file listing & metadata overhead)
- No flexible data access (querying subset not possible)

Motivation

Limitations of Traditional Data Centers

- Monolithic pipelines
- Tight coupling for
  - ingestion
  - processing
  - storage
- Hard to scale and hard to evolve
- Operational complexity (sampling, merging)
- Inflexible for analytical demands (e.g., adding 24:00 hour second)



Modern GNSS Platform

GFZ Helmholtz Centre for Geosciences

HELMHOLTZ

Modern GNSS Platform

Requirements

Functional

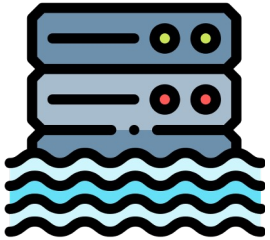
- Scalable ingestion
- Multi-engine access
- Efficient querying

Non-functional

- Cloud-optimized and cost-efficient
- Open formats
- Vendor-neutral

Data Lakehouse

Definition



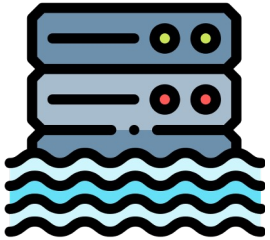
A **Data Lakehouse** is an architecture that combines:

- **Data lake** → scalable, low-cost object storage (e.g., S3, GCS, Azure Blob)
- **Data warehouse** → structured tables, ACID transactions, schema enforcement, fast SQL analytics

It allows you to store raw and processed data in object storage *while still having warehouse-level reliability and performance*.

Data Lakehouse

Features

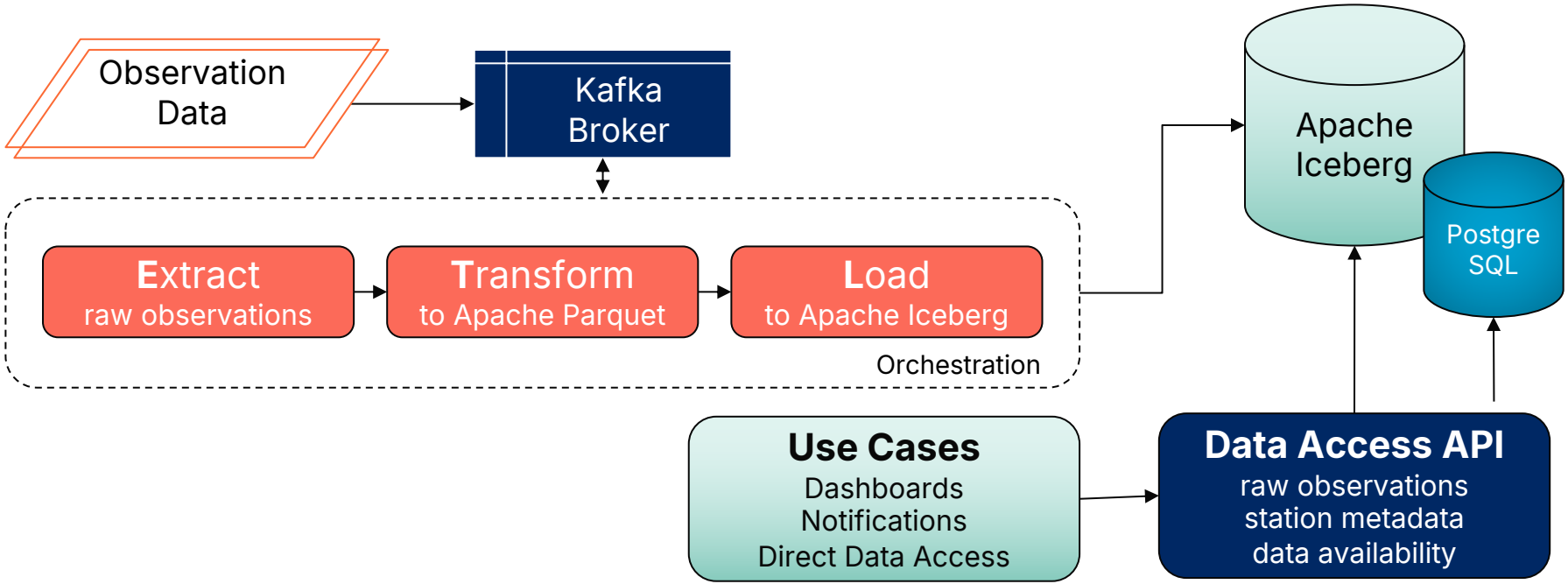


- **ACID transactions**: Safe concurrent writes
- **Schema evolution**: Update table columns without breaking history
- **Partition evolution**: Update table partitioning schema without rewriting table
- **Hidden partitioning**: Improves performance without users managing partition logic manually
- **Time travel**: Query historical versions of datasets
- **Snapshot isolation**: Reliable reads during concurrent writes

System Architecture

System Architecture

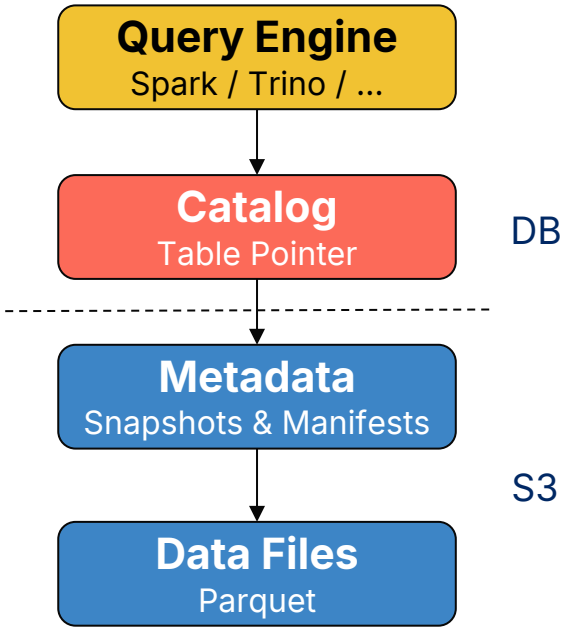
Simplified Design Overview



System Architecture

Storage Layer

- Apache **Iceberg** with **Nessie** as catalog
- **CephFS** as S3 storage
- Separation of data and metadata
- Data written in **Parquet** files with table specific partitioning (station, day)
- Data files are linked in manifest files
- Maintenance is performance-critical, especially if you write a lot of smaller files (e.g., 15M)



System Architecture

Observation Model

- Support for rapid spatio-temporal queries
- Orthogonal slicing (station × epoch × satellite × signal)

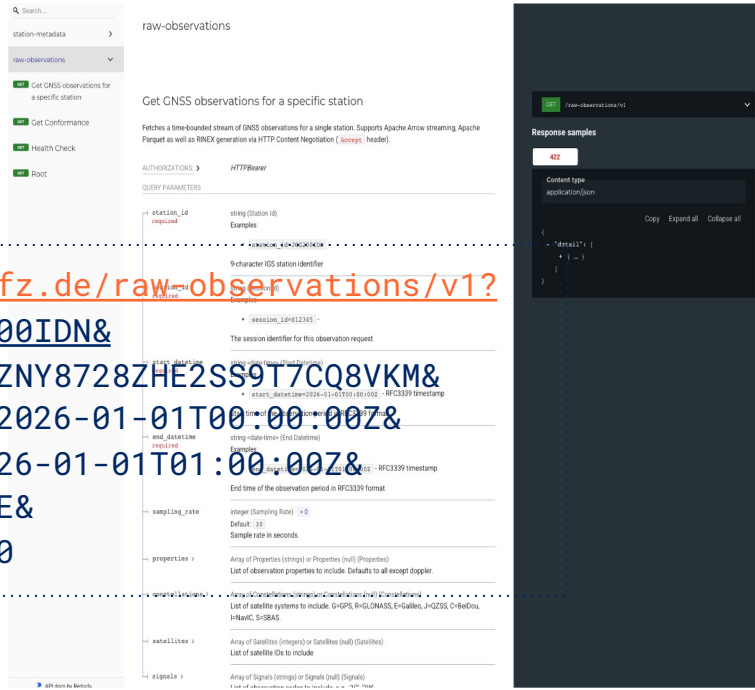
```
SELECT station_id, epoch, constellation, satellite, signal, phase, code, snr, loss_of_lock FROM data WHERE constellation = 'G'
```

	station_id	epoch	constellation	satellite	signal	phase	code	snr	loss_of_lock
1	BIK000KGZ	2023-01-01T00:00:00.000Z	G	8	1C	122388837.184	23289821.853	38.812	false
2	BIK000KGZ	2023-01-01T00:00:00.000Z	G	8	1W		23289820.879	34.25	
3	BIK000KGZ	2023-01-01T00:00:00.000Z	G	8	2L	95367927.212	23289822.689	34.75	false
4	BIK000KGZ	2023-01-01T00:00:00.000Z	G	8	2W	95367922.19	23289818.436	34.25	false
5	BIK000KGZ	2023-01-01T00:00:00.000Z	G	8	5Q	91394247.624	23289819.717	44.375	false

System Architecture

Data Access

- REST API connecting to Apache Iceberg and PostgreSQL metadata catalog
- Response types:
  - Apache Arrow Stream
  - Apache Parquet
  - RINEX (v3/4)
- Query parameters:
  - Station/Session
  - Time range
  - Sampling rate
  - Constellations/Satellites
  - Signals
  - Properties (phase, code, snr, ...)



https://igsws.gfz.de/raw-observations/v1?
station_id=JOG200IDN&
session_id=01KRZNY8728ZHE2SS9T7CQ8VKM&
start_datetime=2026-01-01T00:00:00Z&
end_datetime=2026-01-01T01:00:00Z&
constellations=E&
sampling_rate=30

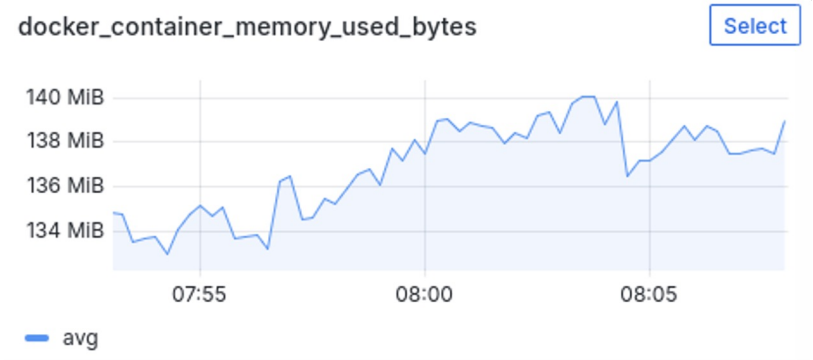
Service not publicly available yet.

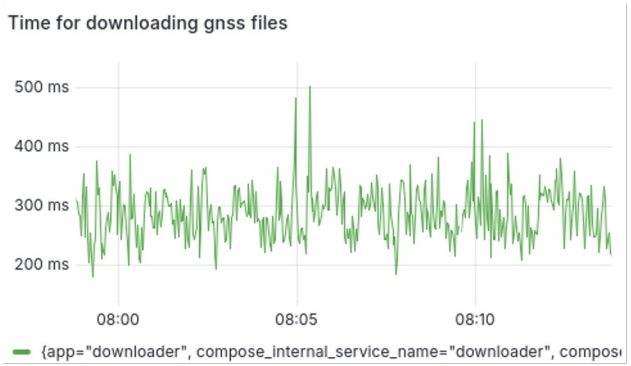
GFZ Helmholtz Centre for Geosciences

System Architecture

Observability Layer

- Prometheus for continuous logging of services metrics
- Alloy and Loki for the service logs
- Grafana for the visualisation and alerts





GFZ Helmholtz Centre for Geosciences

Benchmarks

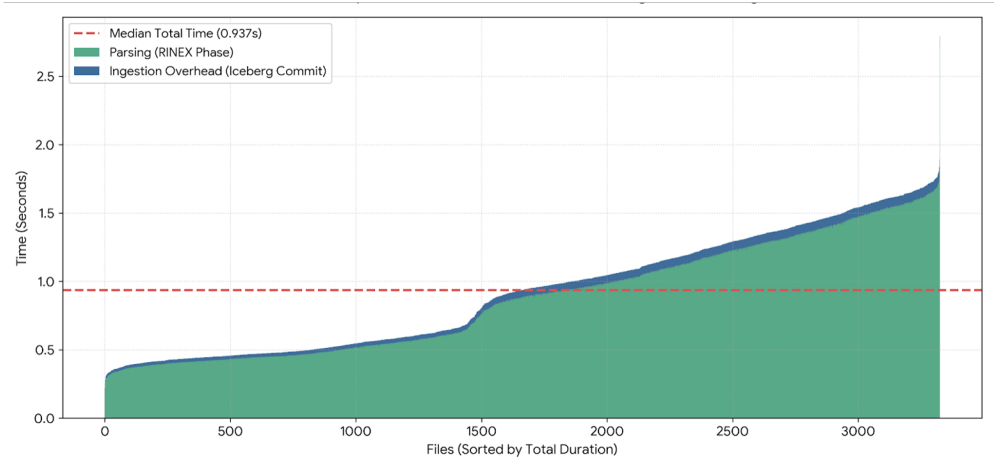
Apple M4 Pro
12 CPU
32G RAM

Benchmarks

Ingestion using Pylceberg

Ingestion scales horizontally with load.

- Using GFZ archive* to test the processing
  - 50 stations, 1 Hz, RINEX v3
  - 1 year of Multi-GNSS data
- 10 parallel worker
- Ingestion of 1.75M files in ~3 days

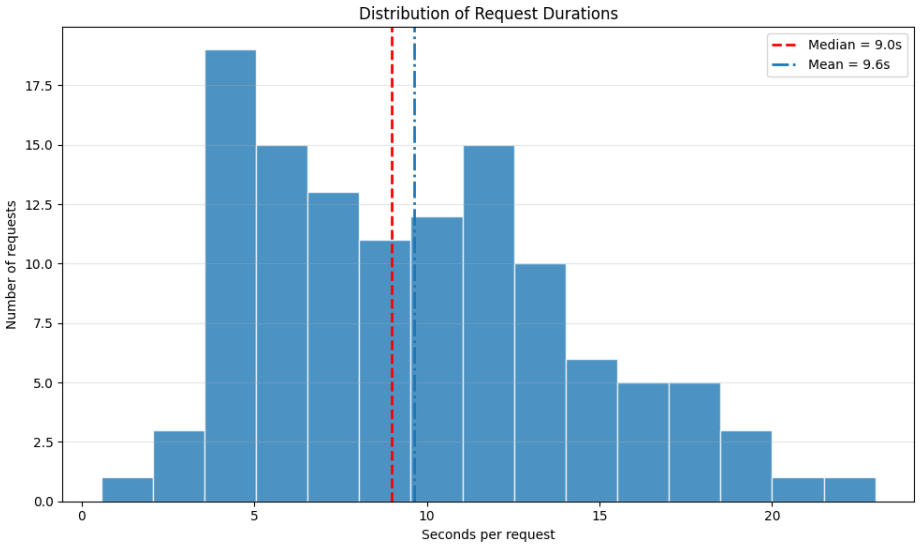


Benchmarks

Data Access using FastAPI and Trino

Access scales efficiently, handling high concurrent request loads.

- 120 parallel requests requesting daily 30S observations
- 4 parallel worker
- Response type: Apache Arrow
- Median request duration: 9-11s
- Total wall time: 5-6m
- Success rate: 100%



GFZ Helmholtz Centre for Geosciences

Conclusions

Conclusions

GNSS archives are becoming analytical datasets.

- Lakehouse architecture enables reproducible cloud-optimized GNSS analytics
- Object storage simplifies operations at scale
- Portable deployment procedure with individual evolvable micro-services
- Scalable data ingest and access
- Real-time monitoring with commonly used components

Lessons Learned

Building a GNSS lakehouse is about managing metadata, partitions, and analytics at scale – not storing files.

- Small files become the dominant scalability problem
- Metadata management matters more than raw storage capacity
- Partitioning strategy is critical
- Object storage changes the architecture model
- Maintenance is essential to keep the datalake healthy

GFZ Helmholtz Centre for Geosciences

Contact

markus.bradke@gfz.de
https://gnss.gfz.de

Follow us on

Bluesky https://bsky.app/profile/gfz.bsky.social

Instagram https://www.instagram.com/gfz_potsdam/

YouTube https://www.youtube.com/user/GFZvideos

LinkedIn https://www.linkedin.com/company/gfz-helmholtz-zentrum-fuer-geoforschung

Icons provided by:
https://www.flaticon.com