SOLUTION BRIEF

GIT-LIKE DATA VERSION CONTROL ON STORAGEGRID USING LAKEFS

Enable Git-for-data operations on StorageGRID

Artificial intelligence and machine learning (AI/ML) is one of the fastest-growing use cases for object storage today. The days of deep and cheap object storage archives are behind us, and enterprises are looking toward a future driven by data. New AI/ML workloads come with a new set of challenges for storage administrators and engineers.

NetApp StorageGRID is enterprise-grade, on-premises object storage that is perfectly suited for modern AI/ML workloads. It is massively scalable, supporting low-touch, nondisruptive expansions, and can store billions of objects. Lighting fast performance on dedicated all-flash appliances provides fast response times for AI/ML workloads. The StorageGRID industry-leading information lifecycle management (ILM) policy engine enables tiering cold data from all-flash to spinning disk to maximize storage efficiency and performance. StorageGRID supports the industry-standard Amazon Simple Storage Service (S3) API.

lakeFS provides version control over the data lake, using Git-like semantics to create and access those versions. If you know Git, you'll be right at home with lakeFS. With lakeFS, you can use concepts on your data lake such as branch to create an isolated version of the data, commit to create a reproducible point in time, and merge to incorporate your changes in one atomic action.

lakeFS and NetApp have partnered to bring powerful Git-like version control to the StorageGRID data lake, solving several common AI/ML hurdles by enabling data reproducibility, rollbacks, data CI/CD, and more.

Isolated Dev/Test environments with copy-on-write

lakeFS makes creating isolated Dev/Test environments for extract, transform, load (ETL) testing instantaneous, and through its use of copy-on-write, cheap. These isolated environments enable you to test and validate code changes on production data stored on StorageGRID, without affecting it. And you can run analysis and experiments on production data in an isolated clone.

Reproducibility

What did my data look like at a point in time? Being able to look at data as it was at a given point is particularly useful in at least two scenarios:

Reproducibility of ML experiments
ML experimentation is usually an iterative process, and being able to reproduce a specific iteration is important. With lakeFS you can version all components of an ML experiment, including its data, stored on StorageGRID.

Troubleshooting production problems
Data engineers are often asked to validate the data. A user might report inconsistencies, question the accuracy, or simply report it to be incorrect. Because the data continuously changes, it's challenging to understand its state at the time of the error. With lakeFS you can create a branch from a historical commit to debug an issue in isolation.

Rollback of data changes and recovery from data errors

Human error, misconfiguration, or wide-ranging systematic effects are unavoidable. When they do happen, erroneous data might make it into production, or critical data assets might accidentally be deleted.

KEY BENEFITS

lakeFS helps data practitioners effectively deduplicate NetApp® StorageGRID® space, double engineering efficiency, and improve production outage recovery by up to 99%.

Data engineers benefit from:

- Isolated Dev/Test environments
- Continuous integration and continuous deployment (CI/CD) for data lakes
- Rollback operations to fix critical errors immediately

Data scientists benefit from:

- Robust data preprocessing
- Deduplicated experimentation
- Reproducible feature engineering and model training

By their nature, backups are the wrong tool for recovering from such events. Backups are periodic events that are usually not tied to performing erroneous operations. Therefore they may be out of date and require sifting through data at the object level. This process is inefficient and can take hours, days, or even weeks to complete. By quickly committing entire snapshots of data at well-defined times, recovering data in deletion or corruption events becomes an instant one-line operation with lakeFS. Just identify a good historical commit and then restore to it or copy from it.

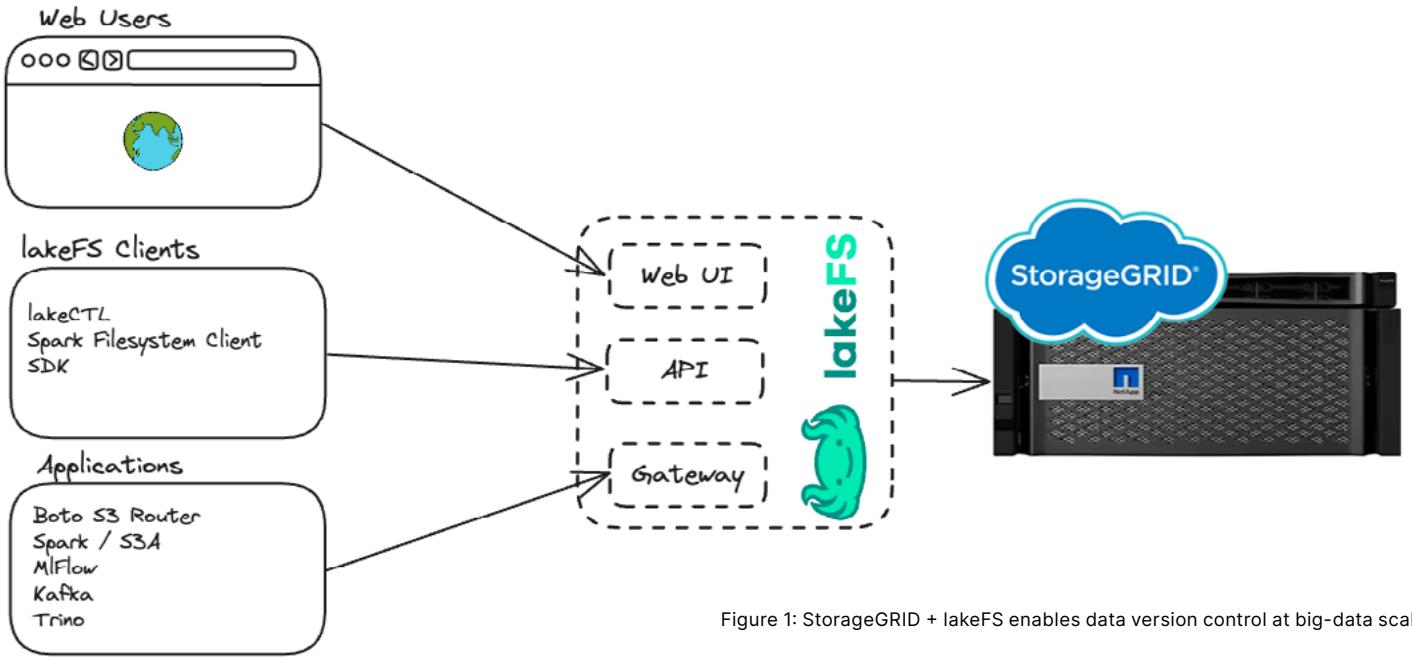


Figure 1: StorageGRID + lakeFS enables data version control at big-data scale.

Multi-table transaction guarantees

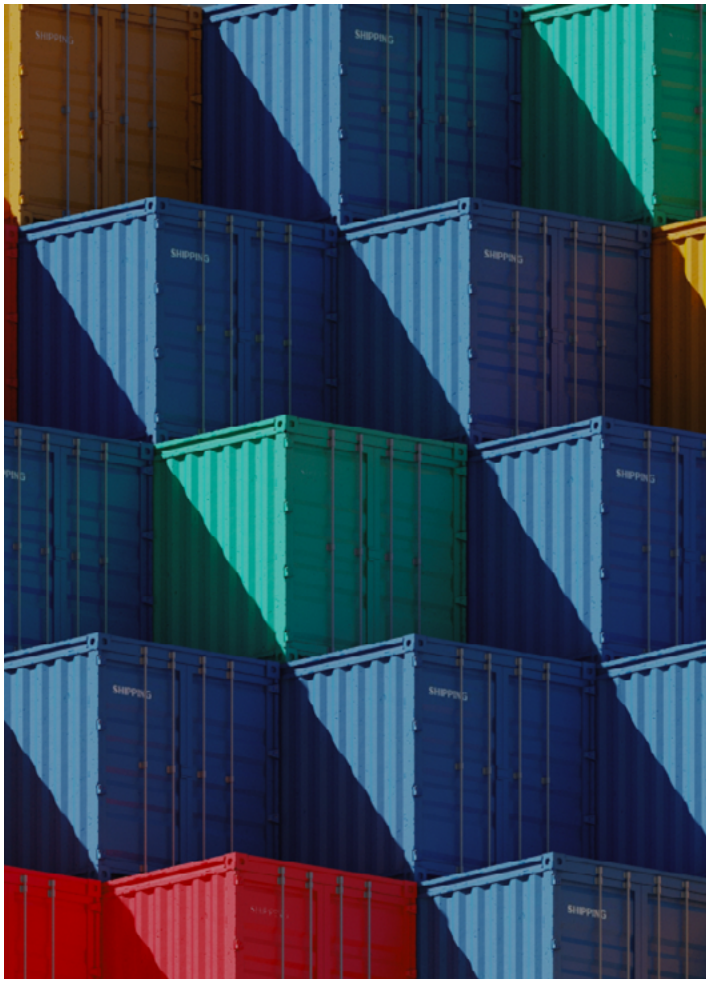
Data engineers typically need to implement custom logic in scripts to guarantee that two or more data assets are updated synchronously. This logic often requires extensive rewrites or periods during which data is not available. The lakeFS merge operation from one branch into another removes the need to implement this logic yourself. Instead, you can make updates to the desired data assets on a branch and then use a lakeFS merge to expose the data atomically on the “production” StorageGRID branch to downstream consumers.

Establishing data quality guarantees: CI/CD for data

The best way to deal with mistakes is to avoid them. A data source that is ingested into the lake introducing low-quality data should be blocked before exposure, if possible.

With lakeFS, you can achieve this block by tying data quality tests to commit and merge operations via lakeFS hooks. For example, you might trigger a file format validation, a schema check, or an exhaustive personally identifiable information (PII) data removal before data is promoted to production.

Learn more about lakeFS



About lakeFS

Founded in 2020 by Oz Katz and Dr. Einat Orr, lakeFS is the brainchild of Treeverse and has offices in Tel Aviv, New York City and San Francisco. Treeverse’s investors include Dell Technology Capital (DTC), Norwest Capital (NVP) and Zeev Ventures. www.lakefs.io

Contact Us

About NetApp

NetApp is the intelligent data infrastructure company combining unified data storage, integrated data services, and CloudOps solutions to turn a world of disruption into opportunity for every customer. NetApp creates silo-free infrastructure, then harnesses observability and AI, to enable the best data management. As the only enterprise-grade storage service natively embedded in the world’s biggest clouds, our data storage delivers seamless flexibility and our data services create a data advantage through superior cyber-resilience, governance, and applications agility. Our CloudOps solutions provide continuous optimization of performance and efficiency through observability and AI. No matter the data type, workload or environment, transform your data infrastructure to realize your business possibilities with NetApp. www.netapp.com