CLUMIO

Building Resilient Amazon S3 Data Lakes

Designing Efficient Backup, Recovery, and Compliance Strategies for Multi-Petabyte Amazon S3 Data Lake Environments.

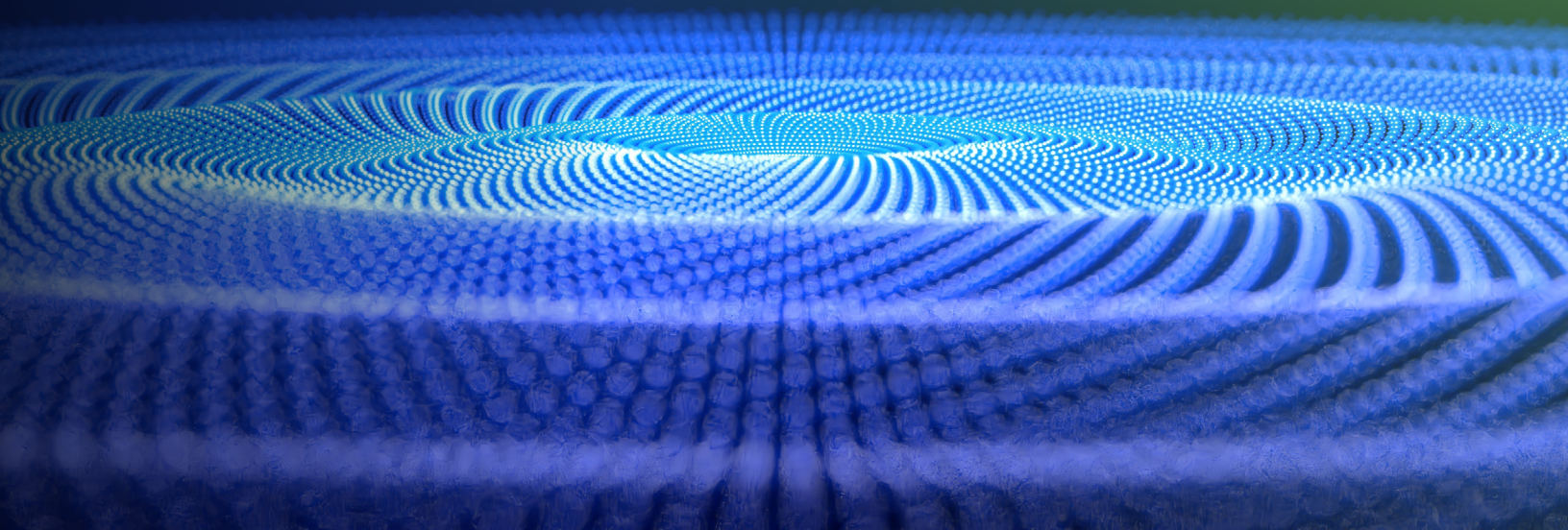


Table of Contents

Introduction

Data lakes on Amazon S3 are essential for building modern business intelligence and machine learning applications. Whether it's an infiltration detection engine powered by petabytes of security logs, therapeutics discovery leveraging miscellaneous genomic and medical data, or regulatory compliance across millions of financial transactions and records, Amazon S3 data lakes are now at the center of leading enterprises' innovation engine.

Data lakes have reached the point of maturity when resilience, recoverability, and compliance of its resident data are now fundamental to its architectural principles. No matter your use case or stage of adoption, the data powering your data lake needs to be able to withstand operational disruptions, accidental data loss, and malicious events. And when cloud operations and platform services teams think about data resilience, the first thing that comes to mind is backup and recovery.

At first glance, backing up data lakes doesn't sound very practical. In fact, backup for data lakes presents several architectural and cost challenges. For example:

- Do you need to back up everything in the data lake?
- How do you apply specific policies to individual constituent items?
- How does your backup scale in a cost-effective manner to protect billions of objects across millions of prefixes and buckets?
- How do you ensure compliance of these disparate data types on an ongoing basis?

In this whitepaper, we'll answer the above questions, and discuss how AWS customers can plan and design an efficient backup and recovery strategy for their Amazon S3 data lakes.

Amazon S3 data lakes are central to modern data architectures. Source

1. The 3 Key Challenges of Data Lake Backups: Data Velocity, Data Volume, Data Variety

Velocity: Depending on the use-case, Amazon S3 data lakes can process millions of new PUT and DELETE requests every hour. This high level of object churn can make inventorying itself a huge challenge, much less continuously backing it up with high fidelity. Simply creating a daily archive, point-in-time copies, or relying on object versions does not constitute resilience. In fact, none of these methods will likely be able to provide operational recovery. For true application resilience, the data in your data lake needs to be able to continuously track events, changes, and metadata and reflect them in an application-consistent manner to a secondary data environment.

Volume: Large data lakes are usually tens of petabytes to exabytes in scale, containing tens to hundreds of billions of objects. This presents two problems when it comes to backups—it's not cost-effective to backup everything, and it's an extremely difficult engineering problem to backup and restore petabytes of data. Data volume, therefore, presents another challenge to data lake resilience.

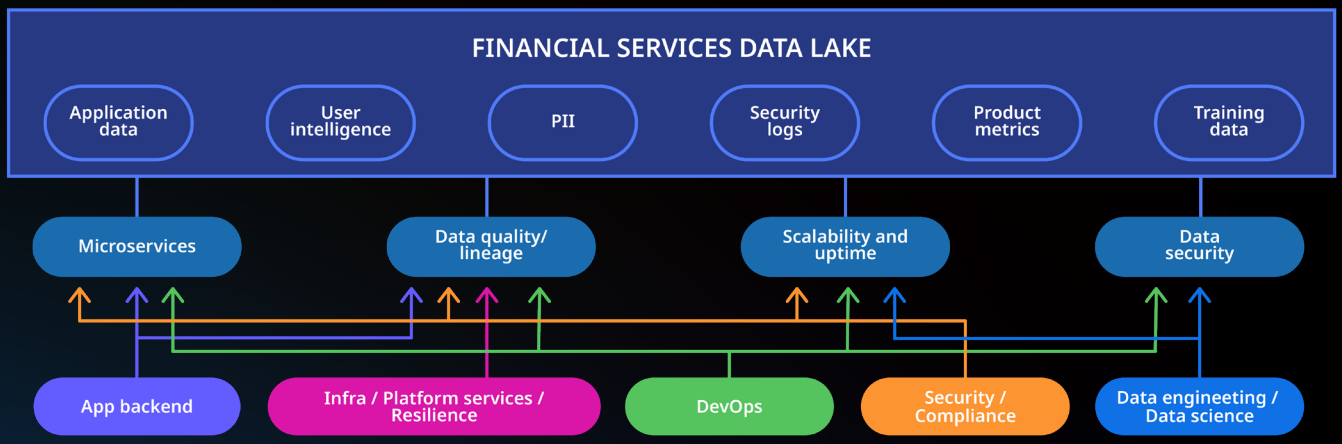
Variety: Disparate object types, varying lifecycle rules, and varying change rates make it difficult to set a standard backup procedure across the data lake. For example, a compliance backup can tolerate hours of retrieval time but need retention periods of several years, and operational backups need data restored in minutes but can be lifecycled after a month. This requires classifying data in your data lake across policies and storage tiers, which can be challenging at the scale of petabytes.

Through the course of this whitepaper, we will lay out how to address each of these challenges through architectural as well as personnel best practices.

Velocity	Volume	Variety
• Inventory • Data Integrity • Traceability	• Scalability • Cost • Restore Performance	• Governance • Policy Standardization • Classification

2. Finding Consensus Among Teams to Determine What Data is Critical

To create an effective data lake resilience plan, it's essential for the platform services or cloud infrastructure team to coordinate with other teams that are leveraging that data—these could be application owners that use this data for test and dev, data teams that use that feed insights to product and GTM, and the SecOps team responsible for safeguarding it. This could be implemented as a shared data catalog, queryable Amazon S3 access or CloudTrail logs, or simply periodic meetings to align key stakeholders. However you approach it, the objective is to create a 'living list' of data assets crucial to your organization's operations and prioritize datasets for backups and compliance.



A financial services data lake with the different types of data it houses, the use-cases those data types support, and the different teams that interact with the data

For instance, consider a fintech company leveraging a financial data lake to support services such as payment processing, credit risk assessment, and fraud detection. The platform engineering / CloudOps team would initiate a cross-functional effort to identify critical data by:

- Engaging with **application owners / backend engineers** to understand the data dependencies for each (micro) service's functionality. For example, the application owner for the fraud detection system would provide insights into what kind of historical transaction data, customer profiles, and behavior patterns are being used, and how long they require it for.
- Collaborating with **data teams** to identify essential data sources. These teams can provide input on the data quality, lineage, and processing needs, which will help determine the appropriate backup frequency and retention objectives.
- Consulting with **security teams** to identify sensitive data and what compliance guidelines this data needs to adhere to. This could include PII, material nonpublic information, or transaction records subject to retention as laid out by FINRA. The SecOps team can also provide guidance on encryption, access controls, air gaps, and other specific compliance requirements.
- Coordinating with **DevOps teams** to understand the operational requirements of the financial data lake. DevOps teams can offer insights into required scalability, uptime, and other operational SLAs, helping the platform services team architect its storage appropriately.

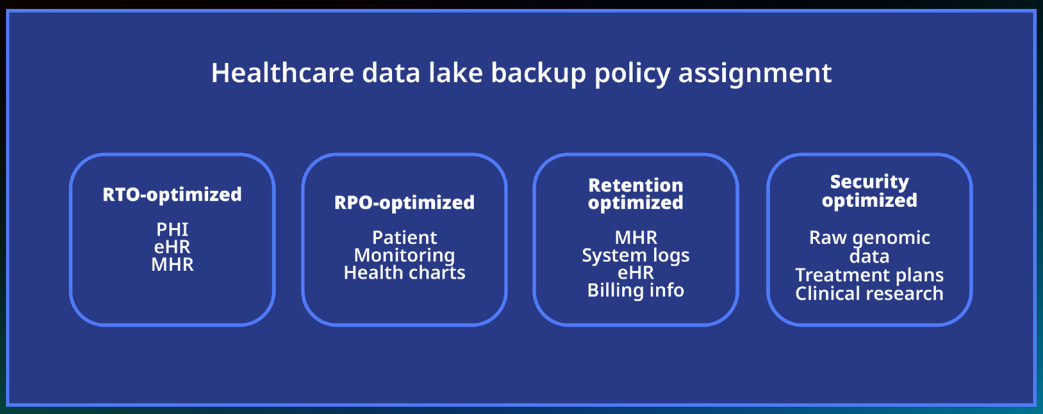
The resulting list will serve as the basis for designing and implementing data lake backup policies, ensuring that the most important data is protected and recoverable. By definition, data lake resilience is a cross-functional responsibility (data lakes → data engineering / data science, resilience → platform services, cloud infra). A collaborative approach ensures business continuity and minimizes the impact of potential data loss events.

3. Classifying Critical Data

Once you've identified what data is deemed critical, it's necessary to locate and classify it within the data lake.
Criticality is multifaceted—some data is critical because it can't be reproduced, other data can be deemed critical because your mission-critical applications depend on it. Therefore, it's important to create a framework of data classification before backups.

Resilience has 4 elements—recovery time (RTO-optimized), data recency (RPO-optimized), longevity (retention-optimized), and security (air gap-optimized).

1. Apply stringent **recovery time criteria** to files and objects that need to be restorable immediately. For example, in a healthcare data lake, this could mean patient information that needs to be always available to healthcare professionals. Any downtime could have catastrophic consequences. Identify the buckets and prefixes that contain this data, and set an instant recovery policy across this data.
2. For data that can withstand some downtime, but close to zero loss, the optimized approach would be to set a **minimum RPO policy**. In a healthcare data lake setting, this could apply to health monitoring metrics. While it's okay for dependent applications to have some downtime, this data needs to be captured and loss-minimized to the greatest extent possible.
3. Some cloud data is subject to stringent compliance requirements. This could include consumer information (governed by GDPR, CCPA, and other laws catering to privacy, easy retrievability, and erasure of information), manufacturing and energy (governed by bodies such as EPA, subject to long-term retention), and financial data (subject to laws and regulations such as the Dodd-Frank Act, Sox, and GLBA, enforced by the FINRA and SEC). Given their sensitive nature, most compliance policies mandate over 5 years of longevity for these records. In the setting of a healthcare data lake, HIPAA mandates 6 years of retention for patient medical records, audit logs and system access records, and electronic health records. This kind of data should be **retention optimized**.
4. While all data needs to be secure, there is some data that you simply cannot afford to lose. In a healthcare data lake, this could be raw genomic sequence data, patient medical records, and treatment plans. Make sure to backup these datasets outside of your enterprise security domain—a process called **air gapping**. That way, even in a situation where your entire environment gets compromised, the irreplaceable data is still alive. On top of air gapping, ensure that the storage platform for these backups is immutable.



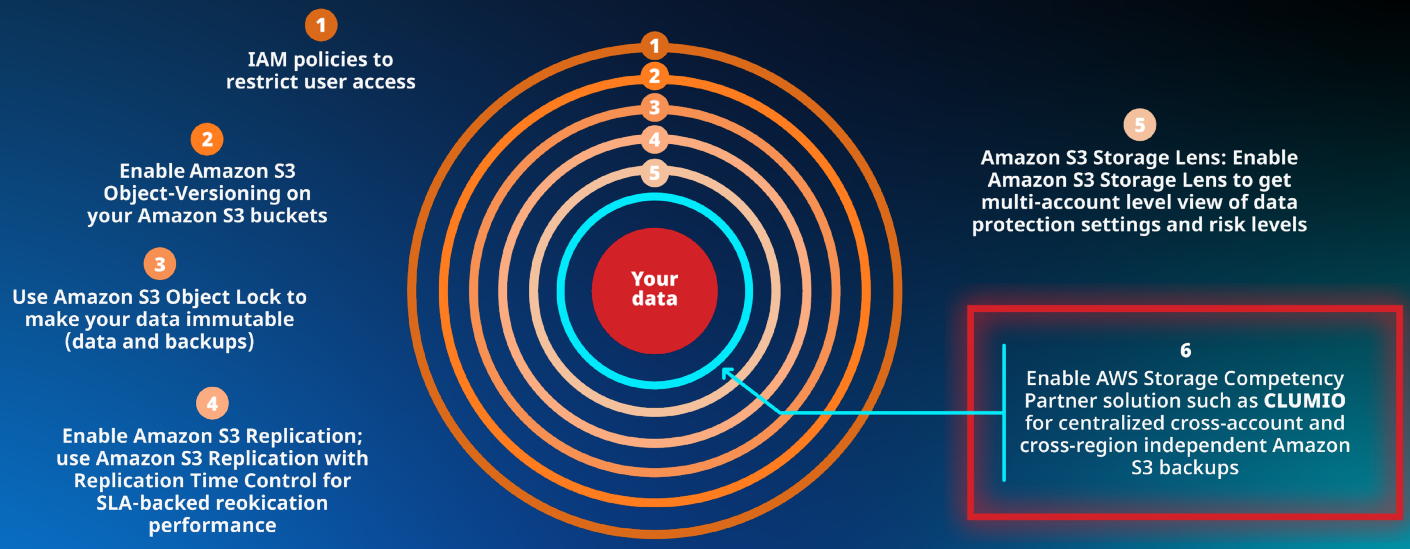
A healthcare data lake with the different types of data it houses, and the backup and retention framework that can be applied to each data type

Applying the above framework enables you to identify the right backup and retention strategies to align with the nature of the data. To streamline and automate this process, consider using a combination of LakeFormation to efficiently catalog and classify data, Macie to identify sensitive information, and Clumio to apply granular backup policies to your data and automate its protection and air gapping. Clumio is the only backup solution that's designed to handle PB-scale Amazon S3 data lake environments with up to 30 billion objects per bucket.

4. Enhancing Security with Amazon S3-native Capabilities

While cybersecurity tools are great for prevention of malicious events, it's important to be prepared for cases where prevention has failed, and your data lake storage is accessible to an adversarial actor. Fortunately, AWS and its partners offer multi-layered data protection options in Amazon S3.

- **Configuring visibility** of your Amazon S3 data lake environment is the first step. Unless necessary, you will want your buckets to be private to your organization by default.
- **Setting up IAM policies** that enable you to grant and restrict access to your users is the second step in configuring access control. If possible, enable multi-factor authentication to particularly sensitive buckets in your data lake. This builds an additional layer of resilience against social engineering.
- **Amazon S3 object versioning** can help you restore a corrupted, deleted, or unnecessarily modified object back to a suitable point in time. However, for data lakes with high object churn, use versioning selectively. Since versioning keeps older versions of modified or deleted objects, it can cause your storage volume (and costs) to rise unexpectedly.
- Regardless of whether your data lake is underpinned by **immutable storage**, your backups should be immutable by default. This ensures that even if the primary data is maliciously encrypted or deleted, the backups cannot be tampered with.
- To keep a live copy of your data outside of your primary region, use **cross-region replication**. Keep in mind that the mirror bucket will need to be versioned, and will immediately reflect each change that happens on your primary data. Replication is a great option for operational recovery from disasters, but if you need to roll back to a specific point in time, you are better off using a more compact and cost-effective out-of-region backup solution.
- **Bucket encryption** is another tool available to you in your arsenal, but you will have to weigh the need for encryption against performance requirements, particularly for near real-time access and processing. Regardless, your data lake backups should be always encrypted, both at rest and in-flight. You can choose to bring your own keys from your AWS account, or use third-party keys for an ironclad air gap setup.



Amazon S3 multi-layered data protection options

5. Long-term Retention for Data Compliance without Breaking the Bank

Enterprises in regulated industries often struggle with the costs and complexity of growing volumes of compliance-subject records that need to be retained for multiple years. The problem is exacerbated in high-velocity environments like data lakes. Some enterprises adopt the approach of taking periodic copies and storing them in Glacier Deep Archive. There are a few reasons why this is suboptimal, but we have suggestions for a better approach:

- **Compaction:** Simply taking daily / monthly copies of a subset of your data and archiving them to a cheaper tier can result in massive duplication and a pileup of large, raw, uncompressed files. Not only does this increase costs, it hampers retrievability during a recovery. Instead, use backup software that continuously and incrementally tracks event operations and backs up, de-duplicates, compresses, indexes critical data, and stores it intelligently so that during a recovery, only the metadata for those files can be queried and the exact data that's needed can be quickly found, un-frozen, un-bundled, and restored.
- **The small file problem:** Financial transactions, security logs, and sensor logs are a few examples of small, few-kB files that need to be retained for compliance adherence. However, most tiers of Amazon S3 require a minimum file size to store data and metadata (40kB+). When dealing with many such small files, the underutilized storage compounds quickly, and you may end up paying for up to 10X the storage you're actually using. Instead, explore solutions that chunk your files while retaining their metadata in a queryable format, so you're fully utilizing each byte that you pay for.
- **Recoverability:** During a DR / compliance audit or an actual recovery operation, it is far from ideal to have to manually scroll through URIs searching for the right data among reams of old archives and records. When an auditor is asking for specific information, your compliance solution needs to quickly sift through your data lake backups and provide you the exact version of the file that you're looking for through a simple search and calendar view.
- **Integrations with broader compliance tools:** Long-term copies are one piece of a larger compliance strategy, and this data needs to be easily retrievable by other compliance tools. With simple archived copies, organizations may struggle to maintain full visibility and control over their compliance data and processes. When investing in long term data retention for sensitive information in your Amazon S3 data lake, choose a solution that itself is compliant with high-benchmark compliance requirements (ISO 27701, HIPAA, etc.), has a well-documented API, and integrates with third-party compliance tools that your company may already be using.

Look for air-gapped, continuous backup solutions on the AWS Marketplace that satisfy the above guidelines to help you simplify the long-term data retention and archival of critical data in your data lake.

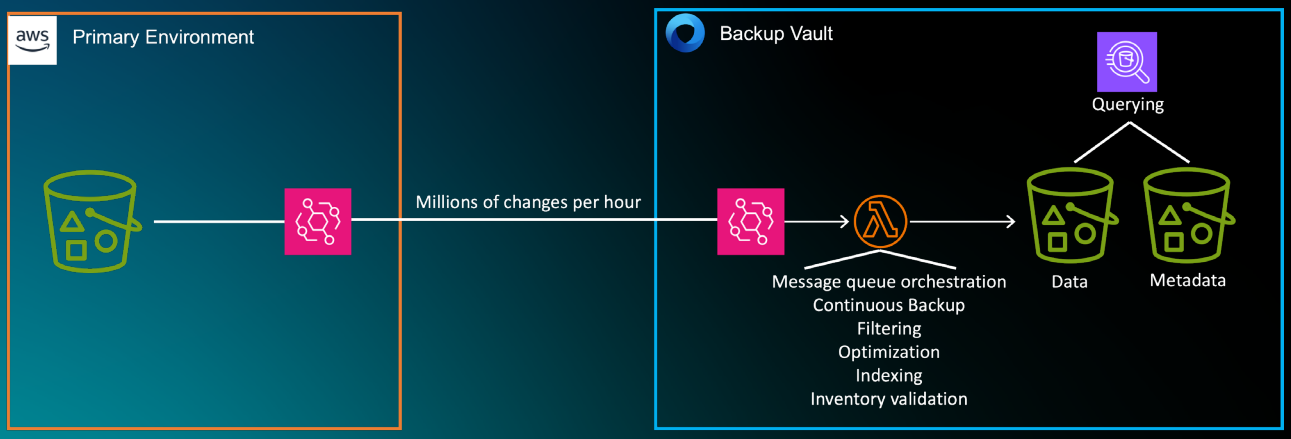
6. Automating Continuous Backups and Restore Testing for Your Data Lake

Once you have identified, classified, and applied the appropriate backup security and retention policies to your Amazon S3 data lake, it's time to set up automation. No matter your backup architecture, the following elements are good candidates to automate:

- Backup environment and policy automation for new data sources and accounts using Terraform or the backup service's REST API
- ID management, access control, and authentication using standard MFA tools
- Role based user access control integrating with your organizational units
- User provisioning using standard SCIM tools
- Logging and monitoring using CloudTrail
- Reporting using Storage Lens or the service's native dashboarding / reporting tools

Regularly testing your recovery process is also critical to ensure that your data lake is resilient. This includes:

1. Identifying objects, prefixes, and buckets of various sizes (both in GB and object count) in your primary data lake that you would like to test
2. Mapping out their criticality based on application dependencies and SLAs
3. Setting expectations around timing based on the size of the backup and the tier it's being thawed from
4. Timing the process of
   - Finding these datasets in the backed up copies
   - Finding the right point-in-time versions
   - Creating a restore environment to house the rehydrated data (this could be an existing account or new)
   - Kicking off restores at the levels of objects, collection of objects, prefixes, collection of prefixes, buckets, and collection of buckets
   - Completion of these restores
5. Monitoring success rates / retries of restore workflows
6. Monitoring any network or application performance issues in this period. This is crucial for large restores since you will likely hit API or bandwidth limits

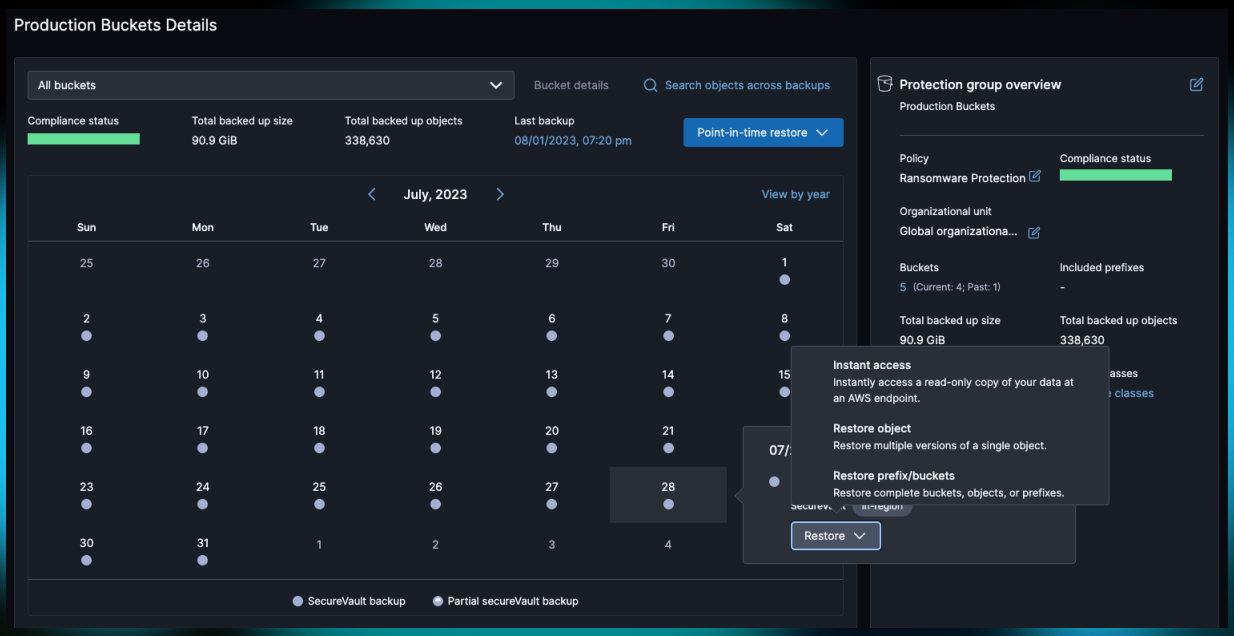


Simplified architecture of a Clumio backup

7. Clumio: an AWS Partner Building Resilient Data Lakes

Reading through the above, it can seem daunting to build data lakes with resilience in mind. But it doesn't have to be. AWS storage competency partner Clumio is a SaaS solution focused on helping AWS customers build more resilient data lakes (and more resilient applications, databases, and storage platforms too). Available on the AWS Marketplace, Clumio helps you get a fortified resilience posture for AWS data lakes set up in minutes.

Clumio's key innovation is its serverless workflow engine that orchestrates an army of Lambda functions that create a massively parallel data pipeline to inventory data sources, orchestrate, read, reduce, encrypt, and transfer data streams. Clumio's protection groups and rules engine help with data classification and automated policy assignment to existing and new data in your data lake, and its simple calendar views help you quickly identify exactly the data that needs to be recovered. Clumio itself is designed as an immutable, air gapped, and storage-optimized data lake that can scale to exabytes per customer and up to 30 billion objects per bucket.



Clumio simplifies recovery for data lakes, with a simple search and calendar view and the option to instantly access backed up data while full rehydration takes place

But even at the scale of large data lakes, Clumio is simple to use. It's an agentless SaaS solution that requires no management, upgrades, and installation tasks of traditional data protection. Clumio does not need versioning or replication to be enabled in your Amazon S3 buckets, and offers the ability for customers to protect only the data they need without having to back up entire buckets. During recoveries, Clumio offers customers a way to instantly access data in their backups without having to rehydrate their data lake back to the primary. Clumio even provides insights into hidden data protection costs and intelligently proposes ways to reduce spend. Pricing is publicly available, inclusive of support, and consumption-based down to the byte, with no minimum consumption requirements.

Assess your Amazon S3 recovery readiness and try Clumio today to experience a simple solution that ensures resilience for your data lakes.

clumio.com/demo

CLUMIO