Storage

Storing big data requires highly scalable solutions that can handle data before and after processing. These solutions are accessible to a variety of processing and analytics services and can typically be tiered to help you reduce storage costs.

In AWS, big data storage is supported by the following services:

- S3 and Lake Formation for object storage
- S3 Glacier and Backup for backups and archives
- Glue and Lake Formation for data cataloging
- Data Exchange for third-party data

What is

In a global data ecosystem that continuously proliferates paradigms, platforms, services, and programming languages, it’s challenging to select optimal technology. Amazon Simple Storage Service (S3) is a cloud storage solution provided by Amazon Web Services (AWS), the e-commerce giant’s distributed computing and web infrastructure arm.

Using reliable and scalable infrastructure and a key-based object storage architecture, Amazon S3 is well suited to host massive amounts of structured and unstructured data in the form of data lakes.

Working with AWS S3 buckets

Amazon S3 objects are organized in buckets. Buckets are the main containers in S3, and every object must be stored in one. All of S3's main features, such as the interfaces and APIs, can act either on buckets or individual objects.

When users upload data, they create and name a bucket first, then move however many objects they need into it. Object names (keys), along with a version identifier, uniquely identify objects within buckets. They help users to organize data.

Organizations may use naming conventions to identify data owners, improve access control, and make the store more navigable for end users.
Amazon S3 users can work with the console or web-service interfaces to access raw objects or buckets.

Capacity and data structures

Amazon sets no cap on the total volume or number of items that can be stored in S3, but individual objects can’t be larger than 5 gigabytes, the limit on a single upload. S3 provides tools for uploading large objects in parts and migrating big data into storage.

AWS S3 is a key-value store, one of the major categories of NoSQL databases used for accumulating voluminous, mutating, unstructured, or semistructured data. Uploaded objects are referenced by a unique key, which can be any string. This high-level and generic storage structure affords users near-infinite flexibility.

S3 is capable of storing diverse and generally unstructured data, but it’s also suited for hierarchical data and all kinds of structured information. Features such as metadata support, prefixes, and object tags allow users to organize data according to their needs.

Manageability

Amazon S3’s simple underlying architecture and web service interface make initial deployment and configuration easy.

Management of AWS S3-hosted stores is straightforward yet flexible. From a graphical console customers can work directly with objects. The platform provides a REST interface that lets developers manage stored information at the account level, within buckets, or within individual objects.

S3 also supports batch operations across all levels, and a related service, AWS Lambda, can allow these operations to perform arbitrarily complex tasks.

Storage types and pricing

S3 offers several storage classes for different use cases and expected volumes. The standard storage class guarantees 99.99% availability, but several other options are available:

- Standard Infrequent Access is ideal for average data archival, backup, and recovery use cases.
- One Zone-Infrequent Access is best for data that is rarely requested but still needs to be quickly retrieved.
- Amazon Glacier is optimal for long-term data storage, and has the highest durability, with no latency requirements.

S3 pricing generally scales with usage. Price is reduced in regions where Amazon’s infrastructure is less costly to maintain.

Security and compliance

Amazon S3 keeps data secure with the aid of several tools. Built-in support for user permissions regulates who and what can download or upload data and prevents unauthorized access.
More sophisticated access policies, as well as several encryption options, are available:

- AWS Identity and Access Management (IAM) designates specific users and manages their data clearance.
- Access control lists enable object-level granularity when setting permissions.
- Security policies can be set at the bucket level or globally.
- Authentication options are available for queries, and both server-side and client-side encryption can be enabled for uploads.
- Managers can access audit logs to review data access and activity.
- Customers can use Amazon Macie to detect sensitive data in uploads, such as other people's intellectual property or personally identifiable information.

APIs and integrations

The REST API for managing Amazon S3 buckets allows developers to connect stored data to other web applications and services. Objects are available to HTTP clients (S3 can be used as a static website host), and URLs can point directly to stored resources.

File-sharing users can employ BitTorrent protocol for downloading data, leveraging peer-to-peer bandwidth savings instead of HTTP GET requests.

Using AWS S3 as a data lake

AWS S3’s features align well with the benefits of setting up and administering a data lake:

- The flexibility to store relational, hierarchical, semi-structured, or completely unstructured data saves resources. There’s no need to transform data to fit into a standardized schema, so stakeholders are free to focus their energies on using the data.
- With a centralized and common data store, issues stemming from rigid business boundaries and opaque silos can disappear. All members of an enterprise can bring their preferred tools to bear on the data relevant to them.
- A data lake creates clear separation between storage and processing. Because transformation takes place downstream, with data analysts, complex transformation processes are eliminated from the data pipeline.