Cloud Storage as the data lake

Cloud Storage is well suited to serve as the central storage repository for many reasons.

Performance and durability: With Cloud Storage, you can start with a few small files and grow your data lake to exabytes in size. Cloud Storage supports high-volume ingestion of new data and high-volume consumption of stored data in combination with other services such as Pub/Sub. While performance is critical for a data lake, durability is even more important, and Cloud Storage is designed for 99.999999999% annual durability.

Strong consistency: One key characteristic that sets Cloud Storage apart from many other object stores is its support for strong consistency in scenarios such as read-after-write operations, listing buckets and objects, and granting access to resources. Without this consistency, you must implement complex, time-consuming workarounds to determine when data is available for processing.

Cost efficiency: Cloud Storage provides a number of storage classes at multiple prices to suit different access patterns and availability needs, and to offer the flexibility to balance cost and frequency of data access. Without sacrificing performance, you can access data from these various storage classes by using a consistent API. For instance, you can store infrequently used data to Cloud Storage Nearline, Cloud Storage Coldline, or Cloud Storage Archive using a lifecycle policy, and then access it later, maybe to gather training data for machine learning, with subsecond latency.

Flexible processing: Cloud Storage provides native integration with a number of powerful Google Cloud services, such as BigQuery, Dataproc (Hadoop ecosystem), Dataflow for serverless analytics, Video Intelligence API and Cloud Vision, and AI Platform, giving you the flexibility to choose the right tool to analyze your data.

Central repository: By offering a central location for storing and accessing data across teams and departments, Cloud Storage helps you avoid data silos that have to be kept in sync.

Security: Because data lakes are designed to store all types of data, enterprises expect strong access control capabilities to help ensure that their data doesn't fall into the wrong hands. Cloud Storage offers a number of mechanisms to implement fine-grained access control over your data assets.

Data ingestion

A data lake architecture must be able to ingest varying volumes of data from different sources such as Internet of Things (IoT) sensors, clickstream activity on websites, online transaction processing (OLTP) data, and on-premises data, to name just a few. In this section, you learn how Google Cloud can support a wide variety of ingestion use cases.

Pub/Sub and Dataflow: You can ingest and store real-time data directly into Cloud Storage, scaling both in and out in response to data volume.

Storage Transfer Service: Moving large amounts of data is seldom as straightforward as issuing a single command. You have to deal with issues such as scheduling periodic data transfers, synchronizing files between source and sink, or moving files selectively based on filters. Storage Transfer Service provides a robust mechanism to accomplish these tasks.

gsutil: For one-time or manually initiated transfers, you might consider using gsutil, which is an open source command-line tool that is available for Windows, Linux, and Mac. It supports multi-threaded transfers, processed transfers, parallel composite uploads, retries, and resumability.

Transfer Appliance: Depending on your network bandwidth, if you want to migrate large volumes of data to the cloud for analysis, you might find it less time consuming to perform the migration offline by using the Transfer Appliance.

See a more detailed overview of the ingest options and the key decision-making criteria that are involved in choosing an ingest option.

Architecture :