Google Cloud Storage (GCS) - Michaël Bettan

Definition: A highly-reliable, scalable, and fully-managed object storage service.	**Availability**: Global	**SLA**: >=99.95%

Use Cases:

- **Media storage and delivery**: Streaming video, image hosting, content delivery networks (CDNs).
- **Backup and disaster recovery**: Data backups (e.g., regulatory), system images, application backups.
- **Data lakes and analytics**: Storing large datasets for analysis using tools like BigQuery, Dataproc, and Dataflow.
    - Hadoop HDFS replacement

Billing:

- **Storage Costs:** storage class, region, and data volume.
- **Network Usage:** data reads and transfers between regions.
- **Operations Usage**: listing, deleting, and other object operations.
- **Retrieval Fees:** Nearline, Coldline, and Archive storage classes
- **Early Deletion Fees:** Applicable to data deleted before a minimum storage duration in certain classes.
- **Autoclass** management fee (+ one-time enablement)

Multiregion	***Min. Storage***	***Retrieval fees***	***Retrieval time***	***Availability***	***Durability***	***Early Deletion Fee***	***Use case***
Standard	*None*	*None*	*Milliseconds*	*99.99%*	*99.999999999%*	*No*	*Frequently accessed*
Nearline	*30 days*	*Yes*	*Seconds*	*99.9%*	*99.999999999%*	*Yes*	*Infrequently accessed data*
Coldline	*90 days*	*Yes*	*Minutes*	*99.9%*	*99.999999999%*	*Yes*	*Infrequently accessed data*
Archive	*365 days*	*Yes*	*Hours*	*99.9%*	*99.0%*	*Yes*	*Archiving, backup, and DR*

GCS Storage classes

- *Drivers → **Availability**, **Cost**, **Access**, **Performance***
- ***Geographic placement:***
    - ***Multi-Region:** Highest availability in largest area*
    - ***Dual-Region:** highly-available and low latency*
    - ***Region**: high local performance for single region*
- ***Autoclass**: optimizes storage costs and performance by automatically transitioning objects between storage classes based on access patterns. Frequent access promotes objects to Standard for faster retrieval, while infrequent access demotes them to Nearline or Coldline for cost savings, eliminating the manual copy-and-delete process,*
- ***Standard class** has **99.95% availability** (vs others 99.9%)*

Dual-Region: Turbo Replication

- ***Enhanced Data Durability and Availability**: Offers faster redundancy across two regions, minimizing data loss risk and ensuring uninterrupted service during regional outages.*
- ***Rapid Replication**: Replicates 100% of newly **written** objects to two regions within a **15-min Recovery Point Objective**, regardless of object size → **low write latency***

GCS Data Model

- ***Global Namespace**: Unique names for buckets across the entire platform for all clients.*
- ***Buckets**: Basic containers holding your data.*
- ***Objects**: Individual files within buckets, accessible through unique URLs.*

Data Protection

- ***Object Holds**: Prevent objects from being deleted or overwritten, useful for legal or compliance reasons.*
- ***Encryption type** → Google-managed, CMEK, CSEK*
- ***Retention policy**: minimum retention period*
- ***Object Versioning:** Maintain versions of objects to protect against accidental deletion or overwrites.*

Consideration

- ***Object Lifecycle rule :** apply actions to a bucket's objects when certain conditions are met. For example, switching objects to colder storage classes when they reach or pass a certain age. 2 type of actions: **change class** or **delete***
- ***Object conditions :** age, storage class, created before, ...*
- ***Data Processing Integration**: GCS integrates seamlessly with various data processing services like Dataproc, Dataflow, BigQuery, etc.*

Access Control & Security

- *2 options → **Uniform** or **Fine-grained** (legacy method)*
- ***Uniform**: uniform access to all objects in the bucket by using only bucket-level permissions (IAM).*
- ***Fine-grained**: Specify access to individual objects by using object-level permissions (ACLs) in addition to your bucket-level permissions (IAM)*
- ***Signed URLs**: time-limited via generated url without IAM*
- ***Signed Policy Documents**: upload policy to your bucket*
- ***Control access**: Project, Bucket, Object level*

Loading & Moving the data

- ***Online transfer:*** *gsutil, console, APIs, etc. (< 1TB)*
- ***Transfer service:*** *fully-managed service to move the data from clouds (GCS, S3, Azure Blob) and on-premises (docker)*
- ***STS for on-premises data*** *designed for large-scale transfers (up to petabytes of data, billions of files)*
- ***Transfer appliance:*** *high capacity storage server leased from Google to ship to your DC (>20 TiB or than a week to upload)*

Source	Scenario	Solution
S3, Azure Blob Storage		Storage Transfer Service
GCS bucket		Storage Transfer Service
Data Center	Enough bandwidth to meet your project deadline	gcloud storage command
Data Center	Enough bandwidth to meet your project deadline	Storage Transfer Service for on-premises data
Data Center	Not enough bandwidth to meet your project deadline	Transfer Appliance

Reference: Deciding among Google's transfer options

Data Integrity

- *CRC32C is a cyclic redundancy check (CRC) algorithm to detect errors introduced during data transfer or storage.*
- *CRC32C hash is calculated for each object and stored as an object attribute. Retrieve this hash value to verify the data's integrity after download.*
- *Calculate a CRC32C hash for the downloaded data using a tool like* ***gsutil hash*** *or libraries like* ***crcmod for Python.***
- *Supports* ***MD5 calculations*** *for compatibility with older systems or specific use cases with stricter security demands.*