



# Google Cloud

---

Construcción  
de un data lake

# Temario

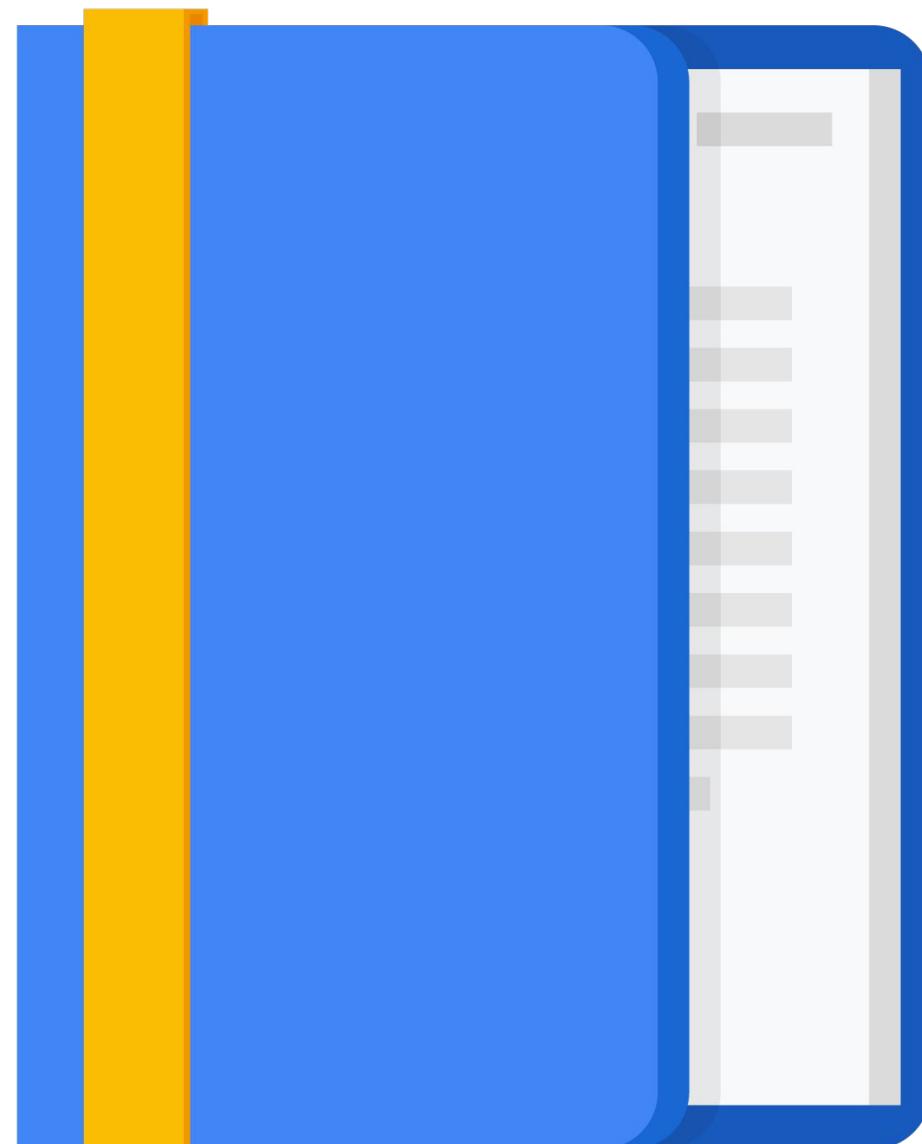
---

Qué es un datalake?

Opciones de almacenamiento

Opciones de Ingesta

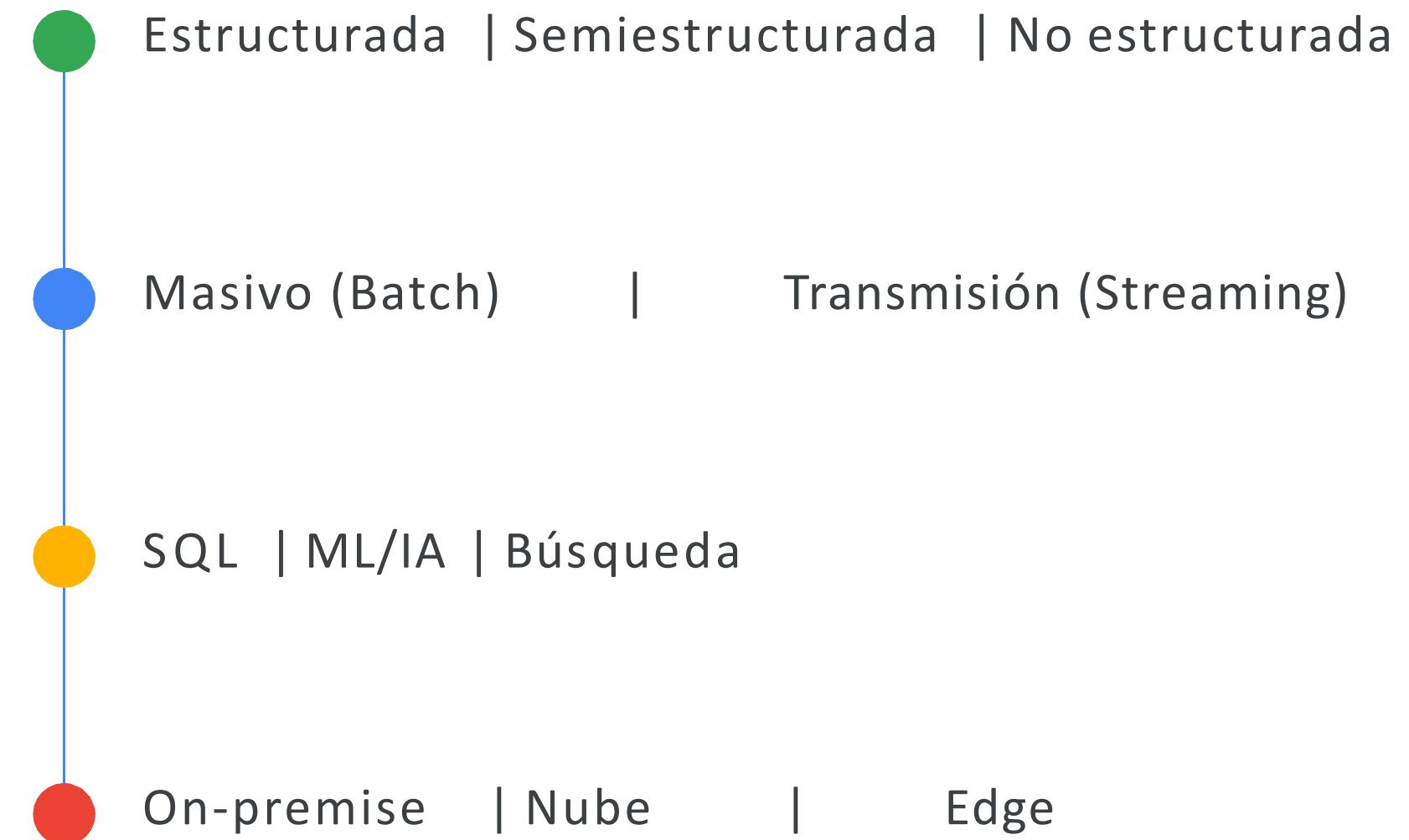
Construcción de un data lake con Cloud Storage



# ¿Qué es un data lake?

---

Es una plataforma de datos escalable y segura que permite a las empresas **transferir**, **almacenar**, **procesar** y **analizar** cualquier tipo o volumen de información.



# Data lake vs Data Warehouse

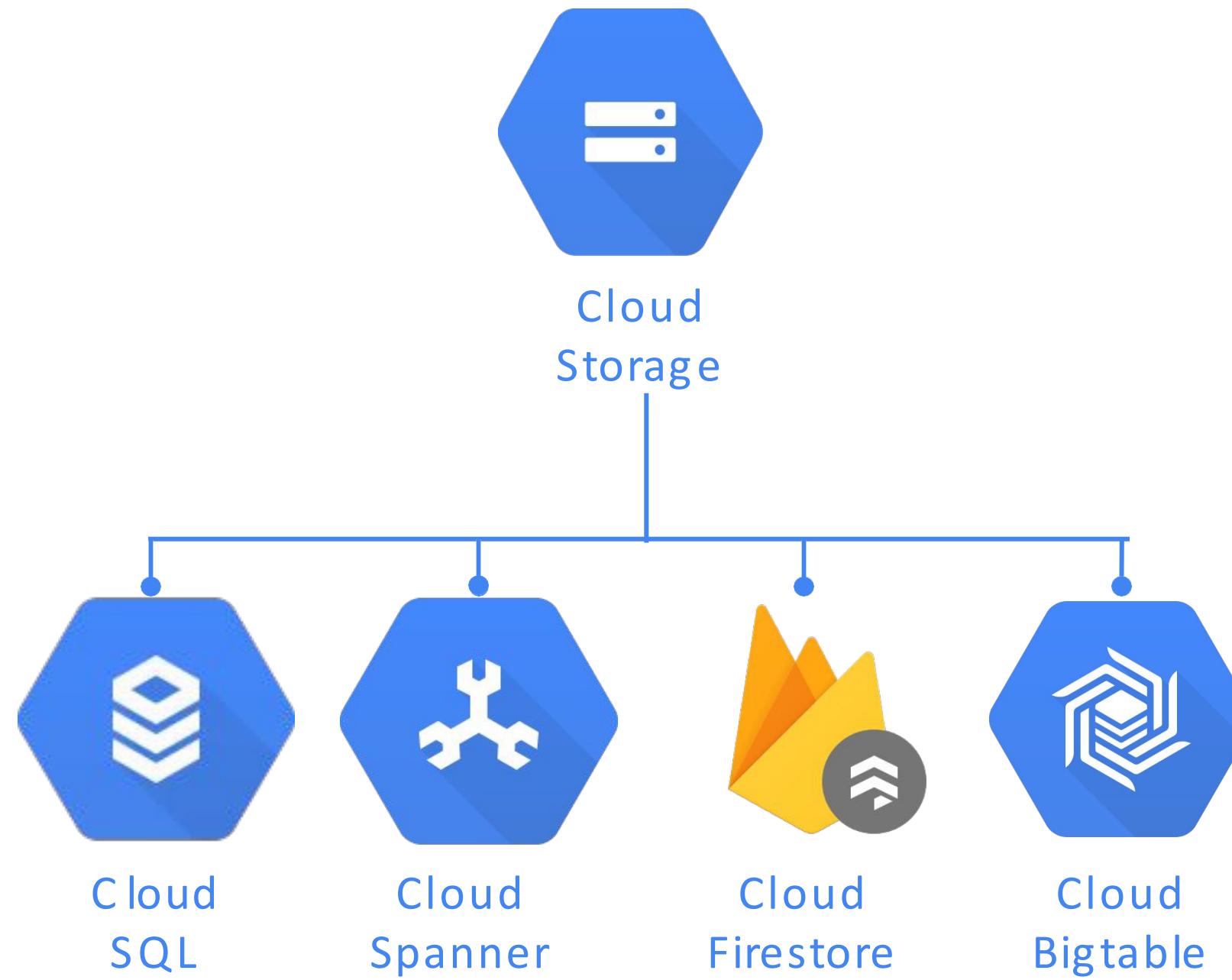
## Data Lake

- Retiene todos los datos en su formato nativo.
- Es compatible con todos los usuarios y tipos de datos.
- Se adapta a los cambios fácilmente.
- Suele ser específico de una aplicación.

## Data Warehouse

- Se carga solo cuando se define su uso.
- Está procesado, organizado y transformado.
- Proporciona estadísticas más rápidas.
- Aporta datos históricos y actuales para los informes.
- Suele compartir un esquema coherente entre las aplicaciones.

# Opciones de almacenamiento para sus datos en GCP



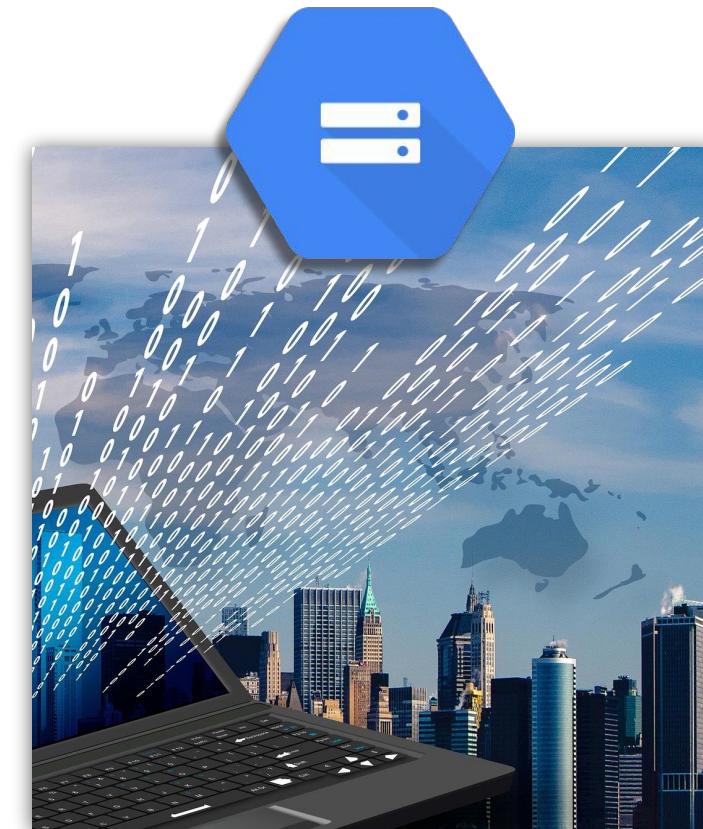
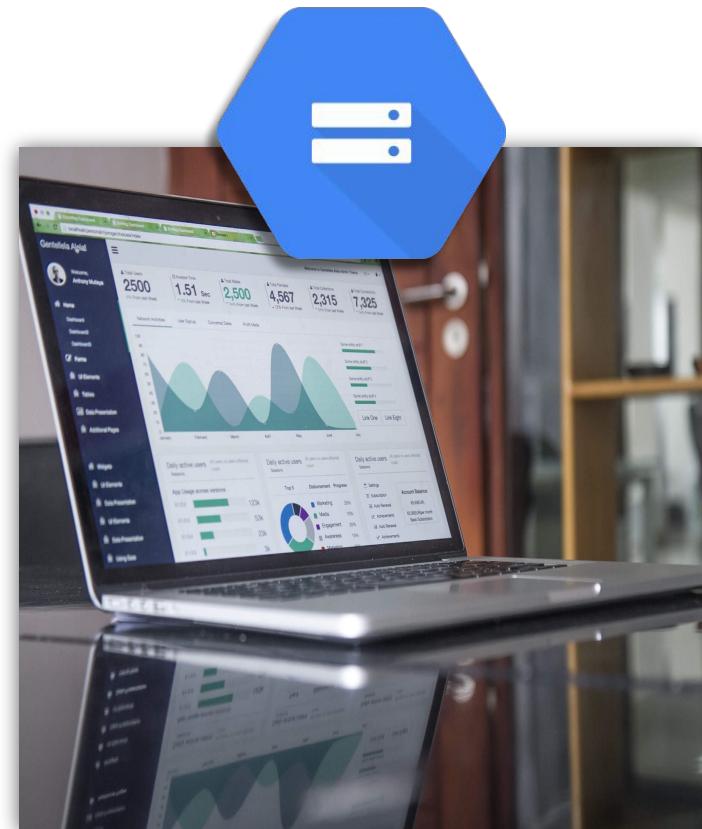
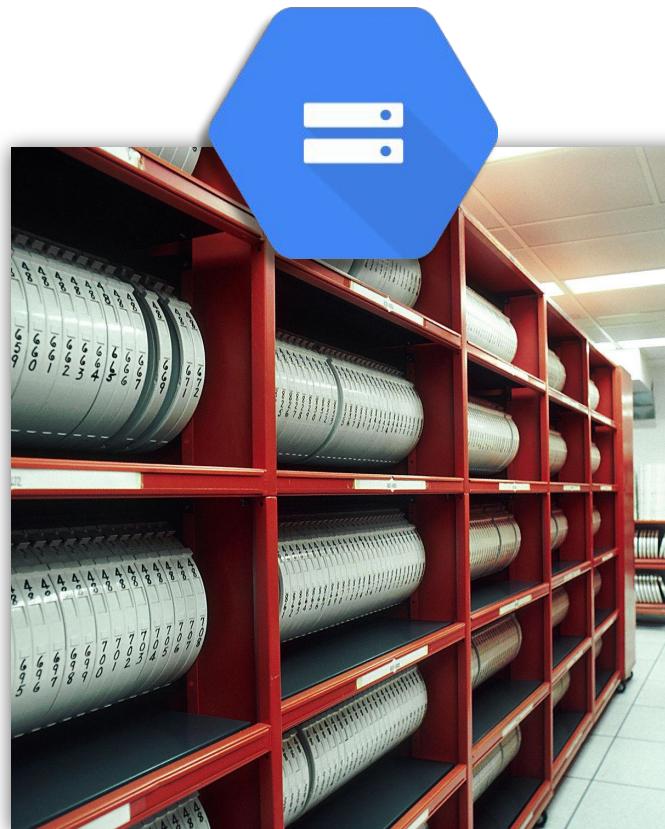
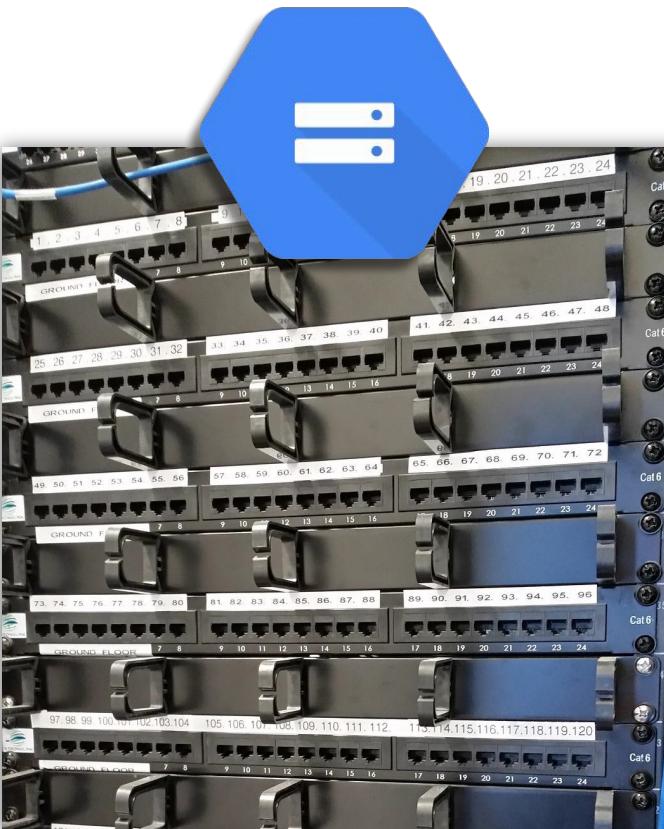
# Cloud Storage es un sistema de almacenamiento de archivos en la nube

- Alto rendimiento, a escala de Internet
  - Administración simple
- Sin administración de la capacidad
- Encriptación de datos en reposo y en tránsito (esta es por defecto)
- Servicios de importación en línea y sin conexión disponibles



\*Proveedores de almacenamiento en la nube orientado a usuarios finales.

# Cloud Storage se diseñó para tener una durabilidad anual de 99.9999999%



Copia de seguridad

Infraestructura  
de reemplazo o retiro  
de servicio

Analítica y  
Machine Learning

Almacenamiento  
y entrega de contenido

# ¿Cómo funciona Cloud Storage?

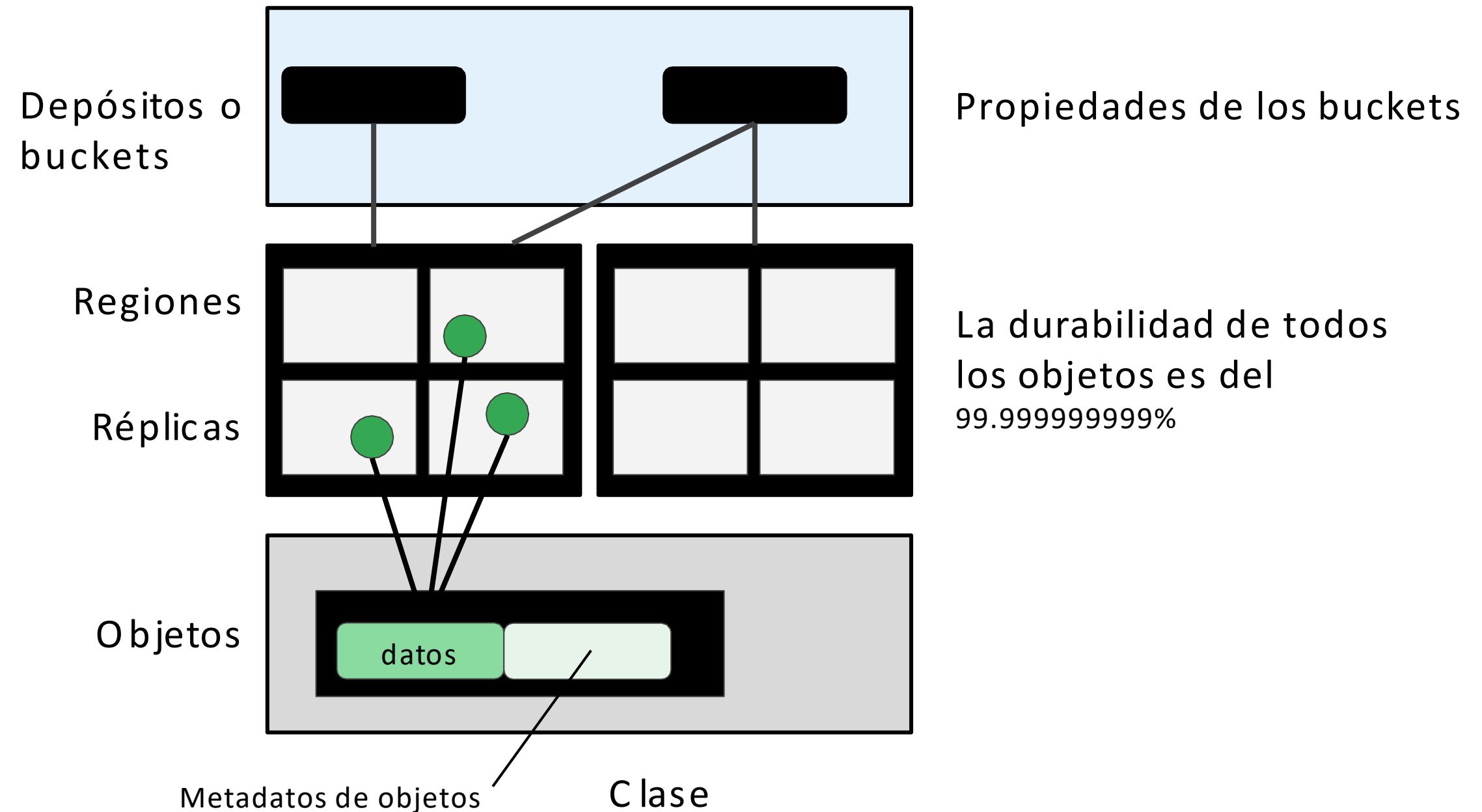


Un espacio de nombres global único simplifica la localización de depósitos y objetos

Ubicación para controlar la latencia

Durabilidad y disponibilidad

Los nombres de objetos largos simulan la estructura



# Las propiedades de los buckets dependen de sus requisitos

europe-north1  
asia-south1  
eu  
asia  
nam4 (us-central1,  
us-east1)A

Regional o Multi-Regional  
Nearline  
Coldline

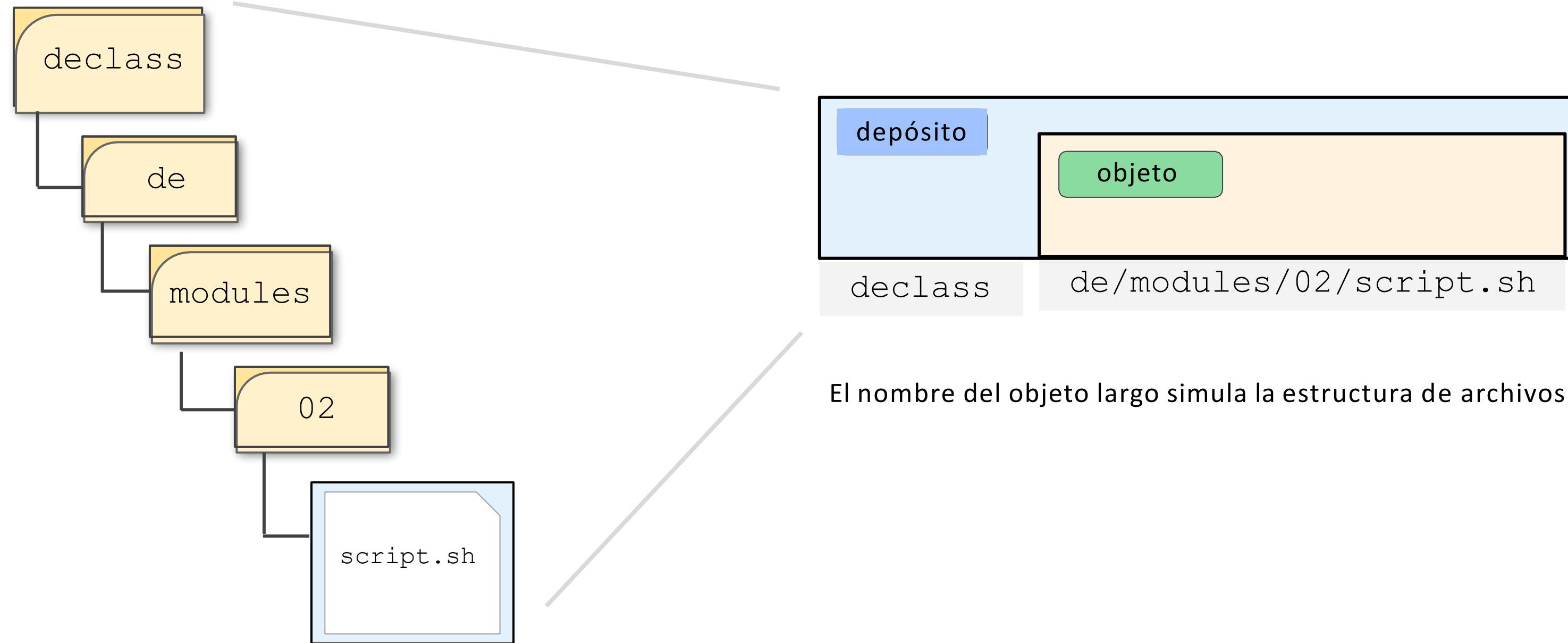
<https://cloud.google.com/storage/docs/locations#location-dr>



región única  
varias regiones  
dos regiones

<https://cloud.google.com/storage/docs/storage-classes/?hl=es>

# Cloud Storage simula un sistema de archivos



Acceso al archivo `gs://declass/de/modules/02/script.sh`

Acceso web `https://storage.cloud.google.com/declass/de/modules/02/script.sh`

# El problema: Las leyes de la física y el ancho de banda del enlace de la red.

Data Size	Bandwidth (assuming 100% utilization)						
	1 Mbps	10 Mbps	100 Mbps	1 Gbps	10 Gbps	100 Gbps	
1 GB	3 hrs	18 mins	2 mins	11 secs	1 sec	0.1 secs	
10 GB	30 hrs	3 hrs	18 mins	2 mins	11 secs	1 sec	
100 GB	12 days	30 hrs	3 hrs	18 mins	2 mins	11 secs	
1 TB	124 days	12 days	30 hrs	3 hrs	18 mins	2 mins	
10 TB	3 years	124 days	12 days	30 hrs	3 hrs	18 mins	
100 TB	34 years	3 years	124 days	12 days	30 hrs	3 hrs	
1 PB	340 years	34 years	3 years	124 days	12 days	30 hrs	
10 PB	3404 years	340 years	34 years	3 years	124 days	12 days	
100 PB	34048 years	3404 years	340 years	34 years	3 years	124 days	

# Cloud Console

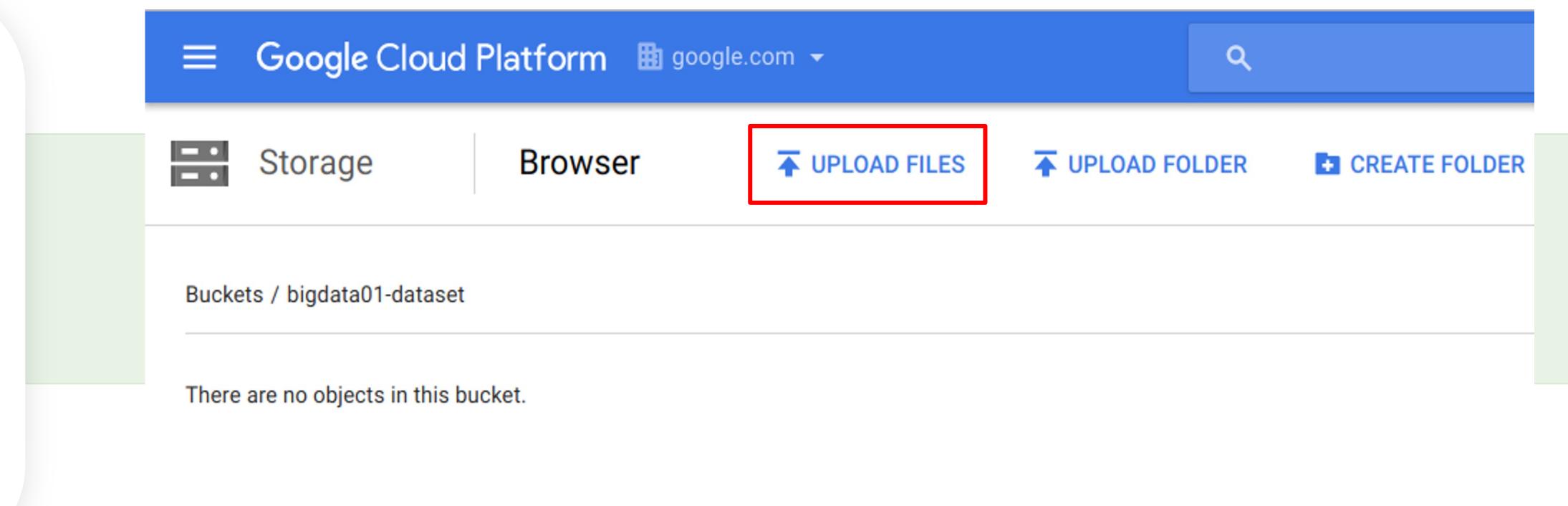
**El rendimiento está limitado por el ancho de banda de su red**

## Funcionalidad

- Importación de archivos a través de la interfaz de usuario web (via Web UI)

## Recomendado para

- Mover hasta 10TB
- Movimiento de datos On-premise
- Exploración de datos



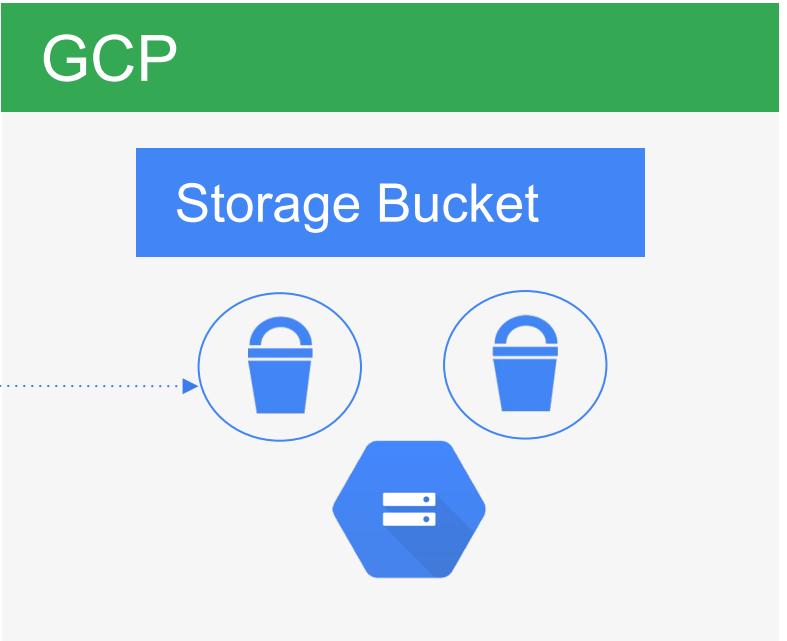
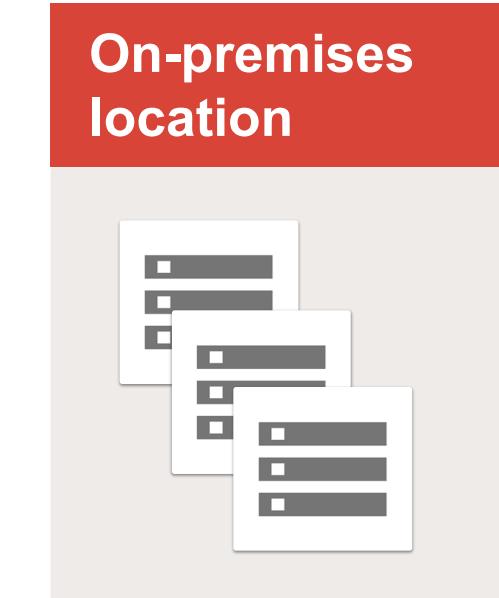
# gsutil

Cloud Storage proporciona **gsutil**, una utilidad de línea de comandos que puedes usar para mover datos basados en archivos desde cualquier sistema de archivos existente a **Cloud Storage**. Escrito en Python, **gsutil** se ejecuta en sistemas Linux, macOS y Windows. Además de mover datos a Cloud Storage, puede usar gsutil para crear y administrar depósitos de Cloud Storage, editar los derechos de acceso de los **objetos** y copiar objetos de Cloud Storage.

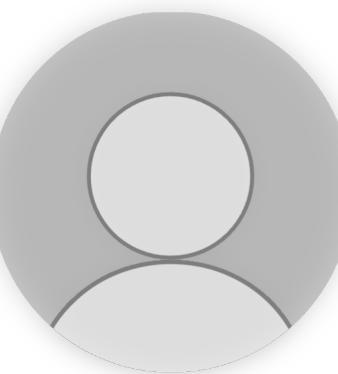
Viene con el SDK de Google Cloud

## Recomendado por

- **Recomendado para mover hasta 10 TB**
- También recomendado para el movimiento de datos on-premise



```
↳ gsutil -m cp *.csv gs://data-analytics-pocs/files
Copying file://data_file1.csv [Content-Type=text/csv]...
Copying file://data_file2.csv [Content-Type=text/csv]...
Copying file://data_file3.csv [Content-Type=text/csv]...
Copying file://data_file5.csv [Content-Type=text/csv]...
Copying file://data_file4.csv [Content-Type=text/csv]...
/ [5/5 files][ 2.2 MiB/ 2.2 MiB] 100% Done
Operation completed over 5 objects/2.2 MiB.
```



# Cloud Storage Transfer Service

## Funcionalidad

- Importa rápidamente online data a GCP
- Transfiere datos de ***data source*** a un ***data sink***
- ***Data sink*** (el destino) es siempre un bucket en Google Cloud Storage
- ***Data source*** puede ser un depósito de Amazon Simple Storage Service (Amazon S3), un blob de Microsoft Azure Storage, una ubicación HTTP / HTTPS o un depósito de Google Cloud Storage.

## Recomendado por

- Mover de 1 a 10 TB de datos: también se puede usar gsutil
- **Mover 10 TB o más datos**
- Cloud Storage a Cloud Storage (de depósito de almacenamiento multirregional a depósito de almacenamiento regional)

### Create a transfer job

You can transfer data to your Cloud Storage bucket from a source you specify here. Required permissions: You must be a project owner and destination bucket owner, and you need read access to the source. [Learn more](#)

#### 1 Select source

- Google Cloud Storage bucket  
 Amazon S3 bucket  
 List of object URLs

You must have read access to the source bucket.

Cloud Storage bucket

 bucket

Browse

#### Specify file filters

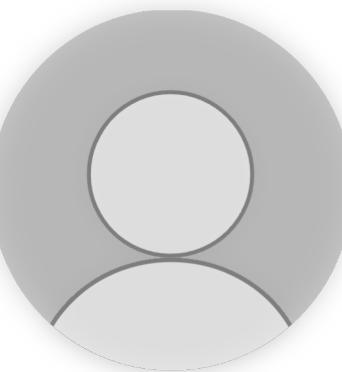
**Continue**

#### 2 Select destination

#### 3 Configure transfer

**Cancel**

Creating a transfer job grants a Cloud Storage Transfer Service account the necessary source, destination, and project permissions to complete the transfer. Your permissions will update to reflect this change.



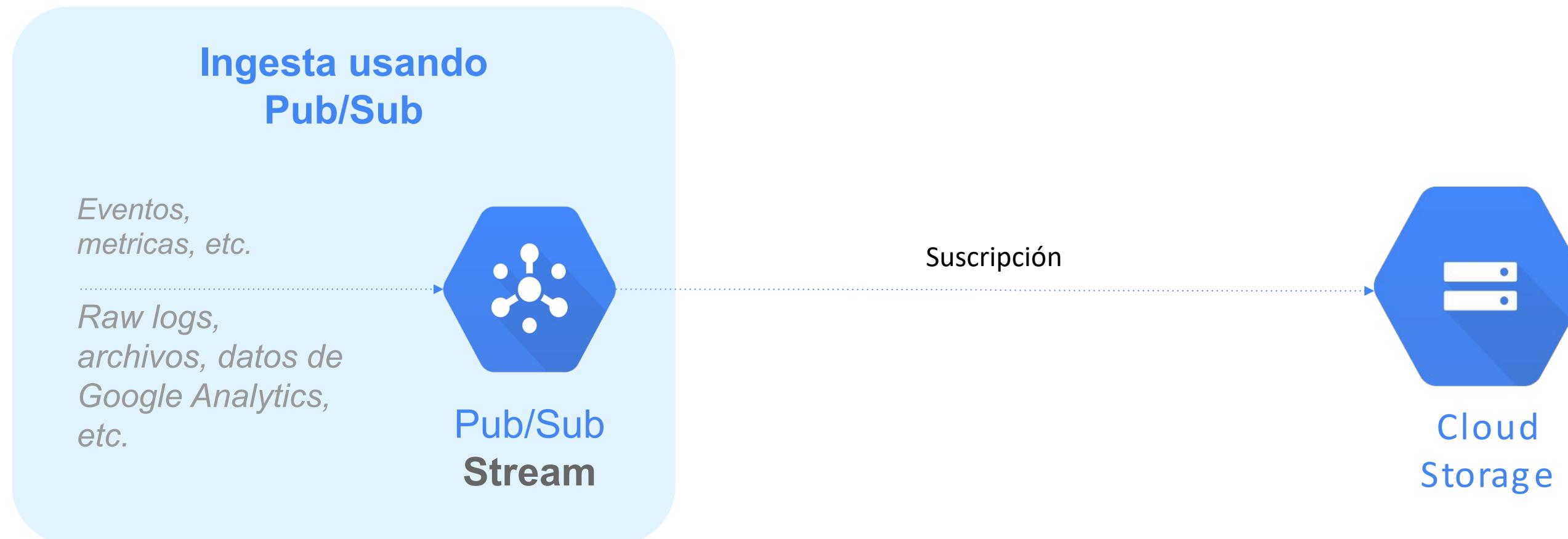
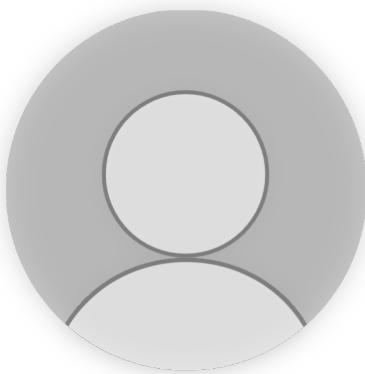
# Transfer Appliance

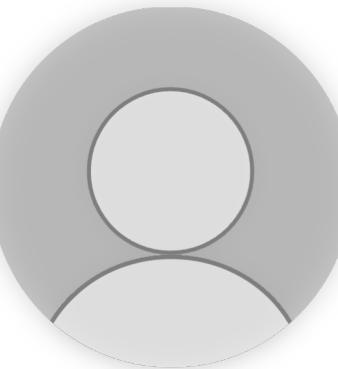
**El rendimiento está limitado por el ancho de banda de su red**

**Transfer Appliance** es un servidor de almacenamiento de gran capacidad que se alquila de Google. Lo conectas a tu red, cargas los datos y lo envías a un centro de carga donde los datos se suben a Cloud Storage. Transfer Appliance viene en varios tamaños



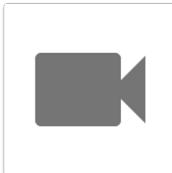
# Ingesta Streaming usando Pub/Sub



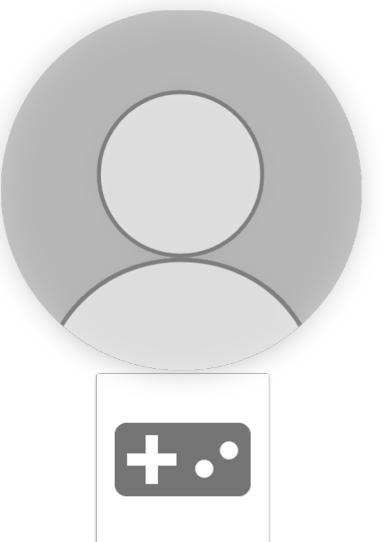
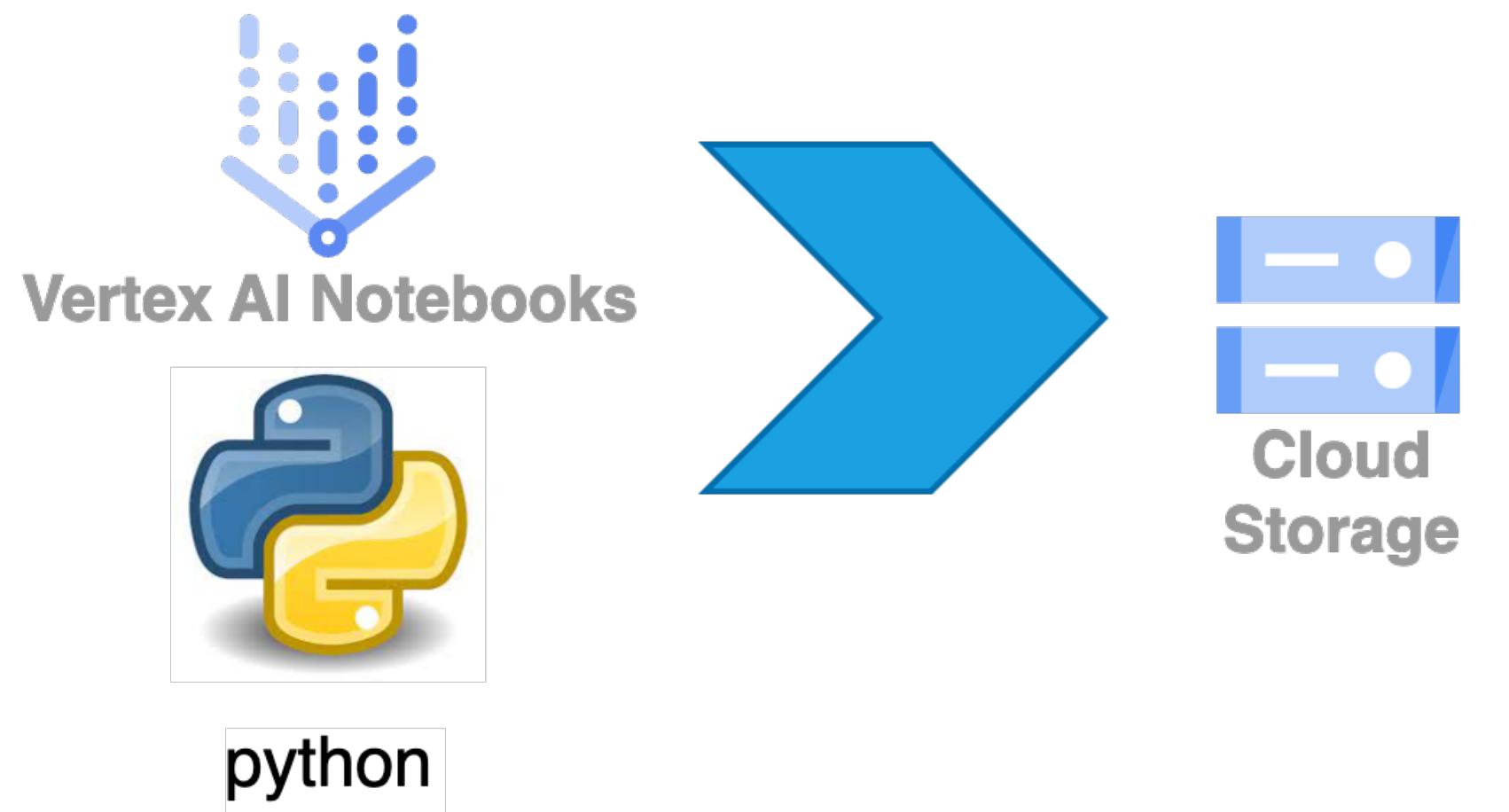


# Opciones de Ingesta

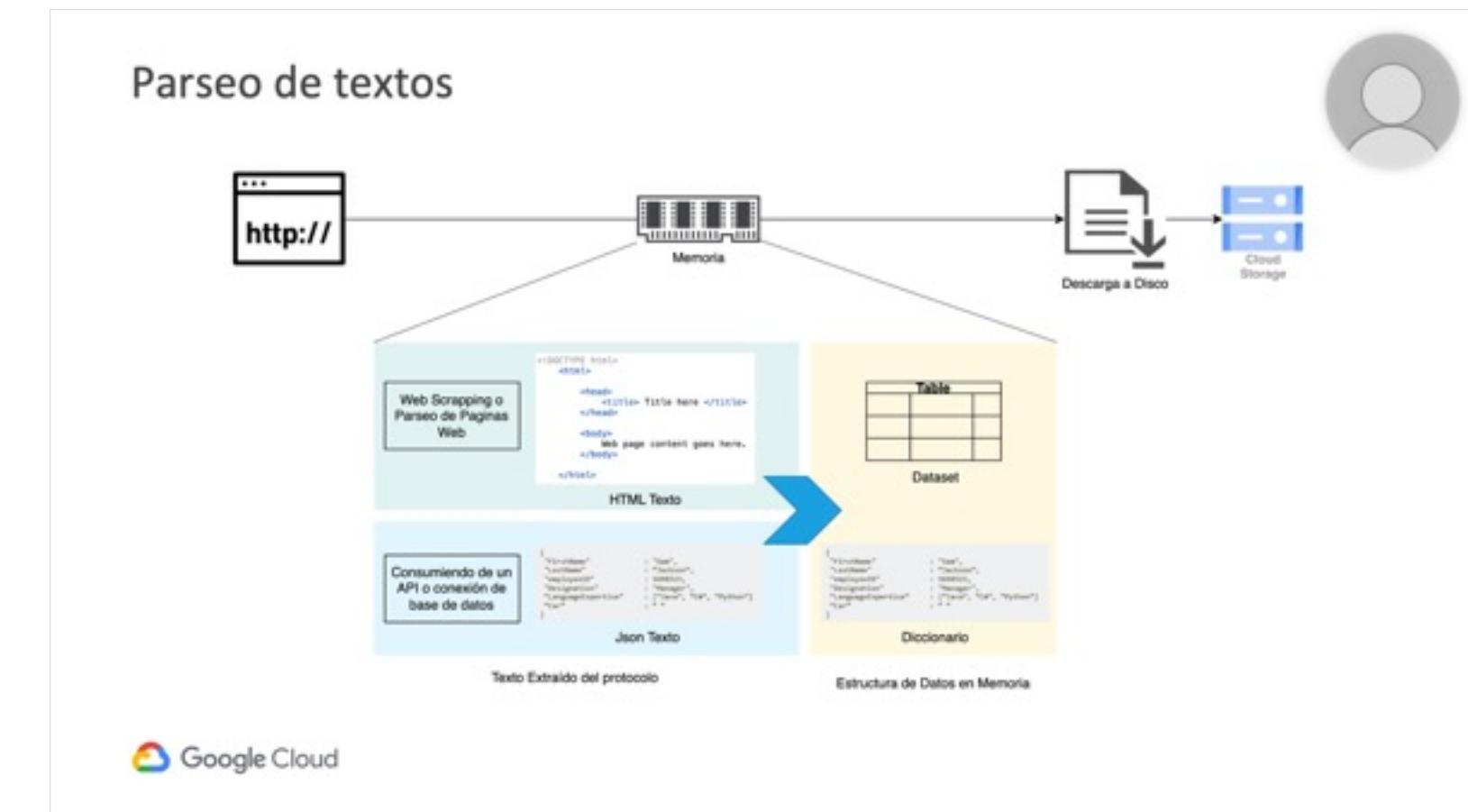
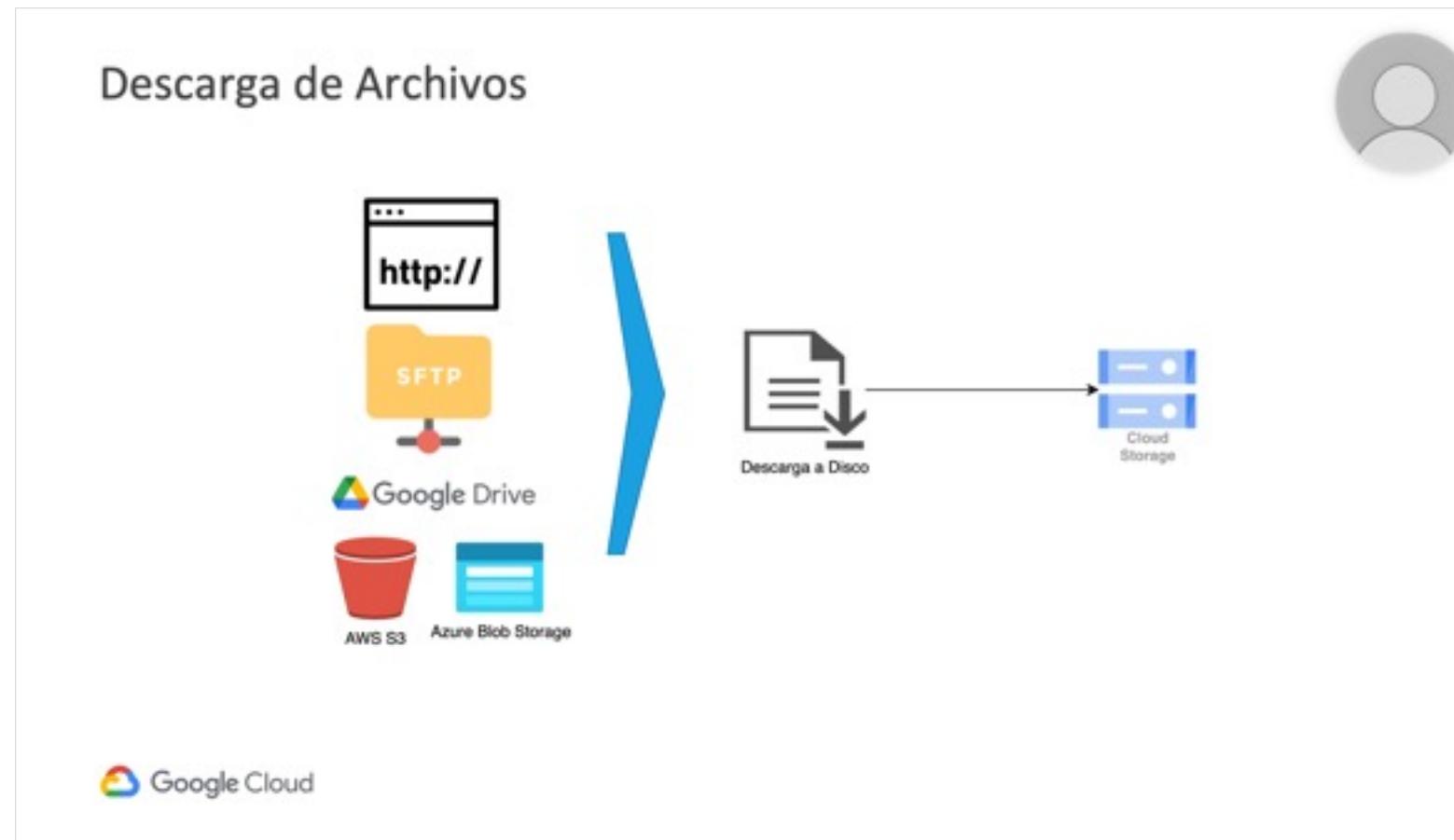
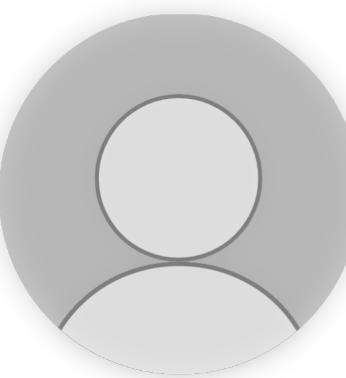
	Casos de uso	Consideraciones
<b>gsutil &amp; GCP Console</b>	<ul style="list-style-type: none"><li>Herramientas de línea de comandos para administrar plataformas de almacenamiento de GCP como Cloud Storage</li></ul>	<ul style="list-style-type: none"><li>Recomendado para 10TB o menos</li><li>Muy bueno para mover datos on-premise</li><li>Limitado por el ancho de banda de red</li></ul>
<b>Cloud Storage Transfer</b>	<ul style="list-style-type: none"><li>Importación rápida de data online</li></ul>	<ul style="list-style-type: none"><li>Bueno para datos de escala de 10 TB + o PB</li><li>Monitoreo de más de 100 millones de objetos</li><li>Programador 24 horas / día</li></ul>
<b>Transfer Appliance</b>	<ul style="list-style-type: none"><li>Transfiera y envíe de forma segura sus datos a una instalación de carga de Google</li></ul>	<ul style="list-style-type: none"><li>Sus datos residen en ubicaciones en las que Transfer Appliance está disponible.</li><li>Llevaría más de una semana cargar sus datos a través de la red.</li></ul>
<b>Pub/Sub</b>	<ul style="list-style-type: none"><li>Mensajería asincrónica que desacopla a remitentes y receptores</li></ul>	<ul style="list-style-type: none"><li>Los mensajes se entregan casi en tiempo real, y se da preferencia a entregar primero los mensajes más antiguos.</li></ul>



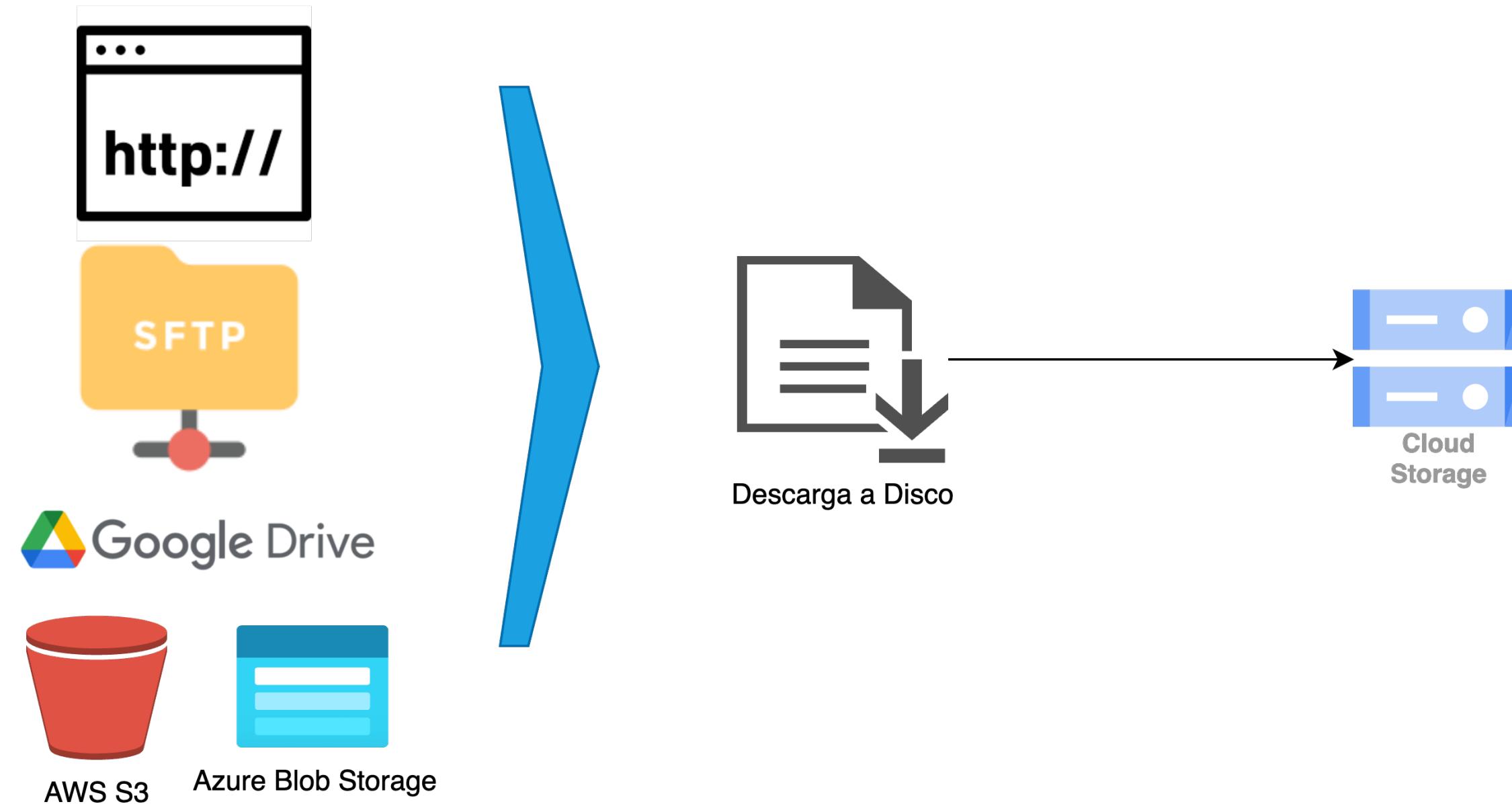
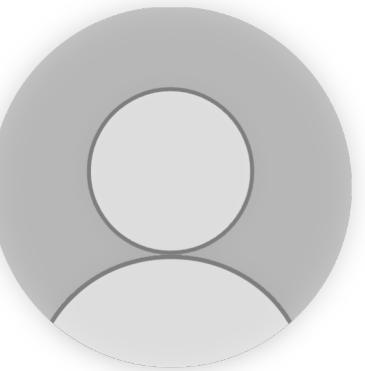
# Usando el sdk de Python para importar datos a Cloud Storage



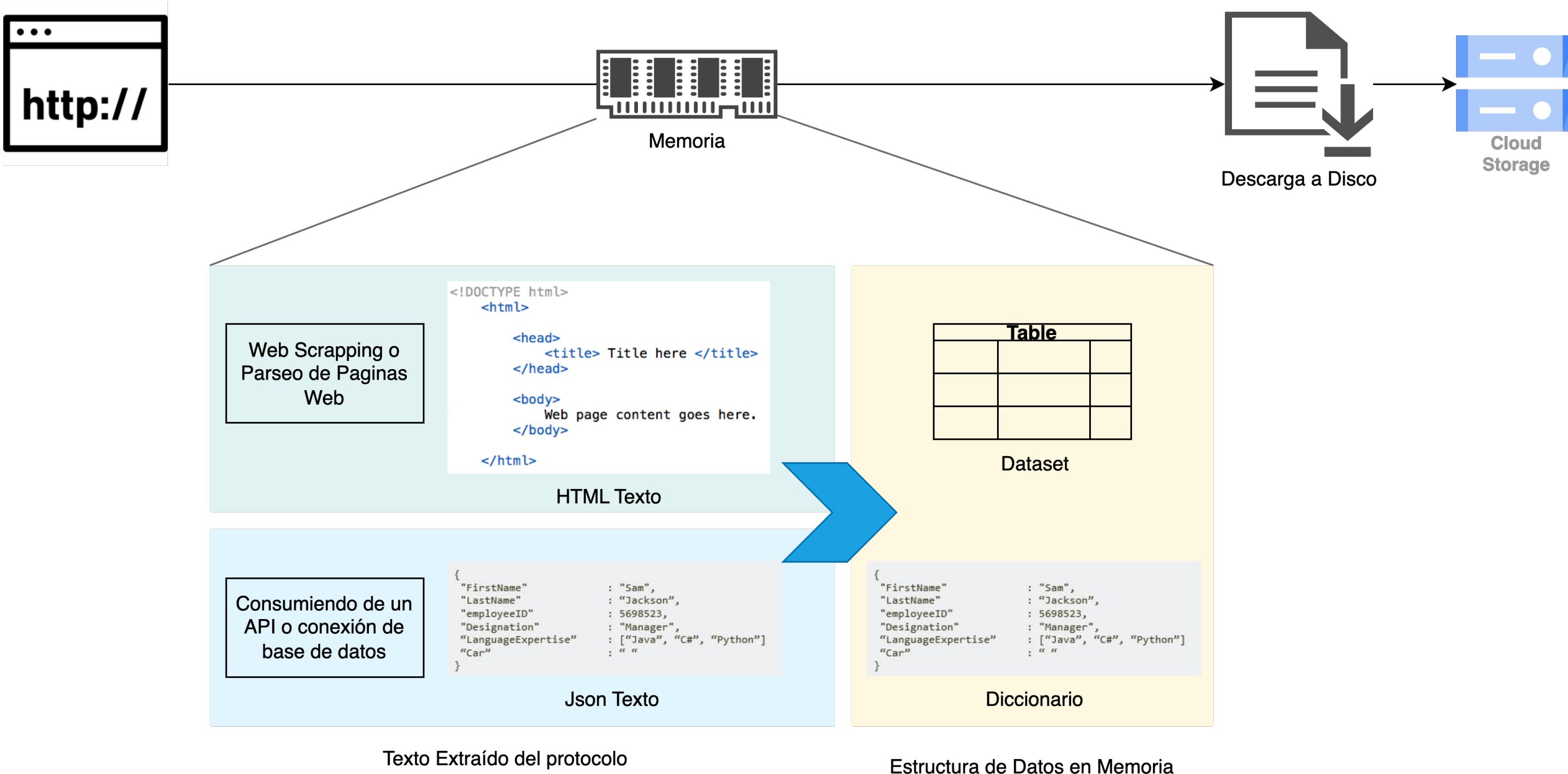
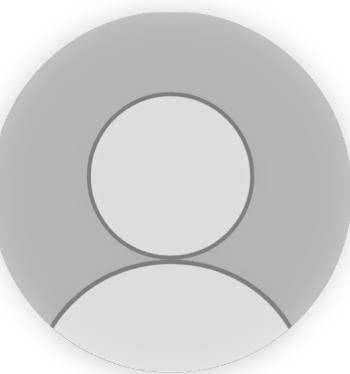
# Patrones de Ingesta a Cloud Storage



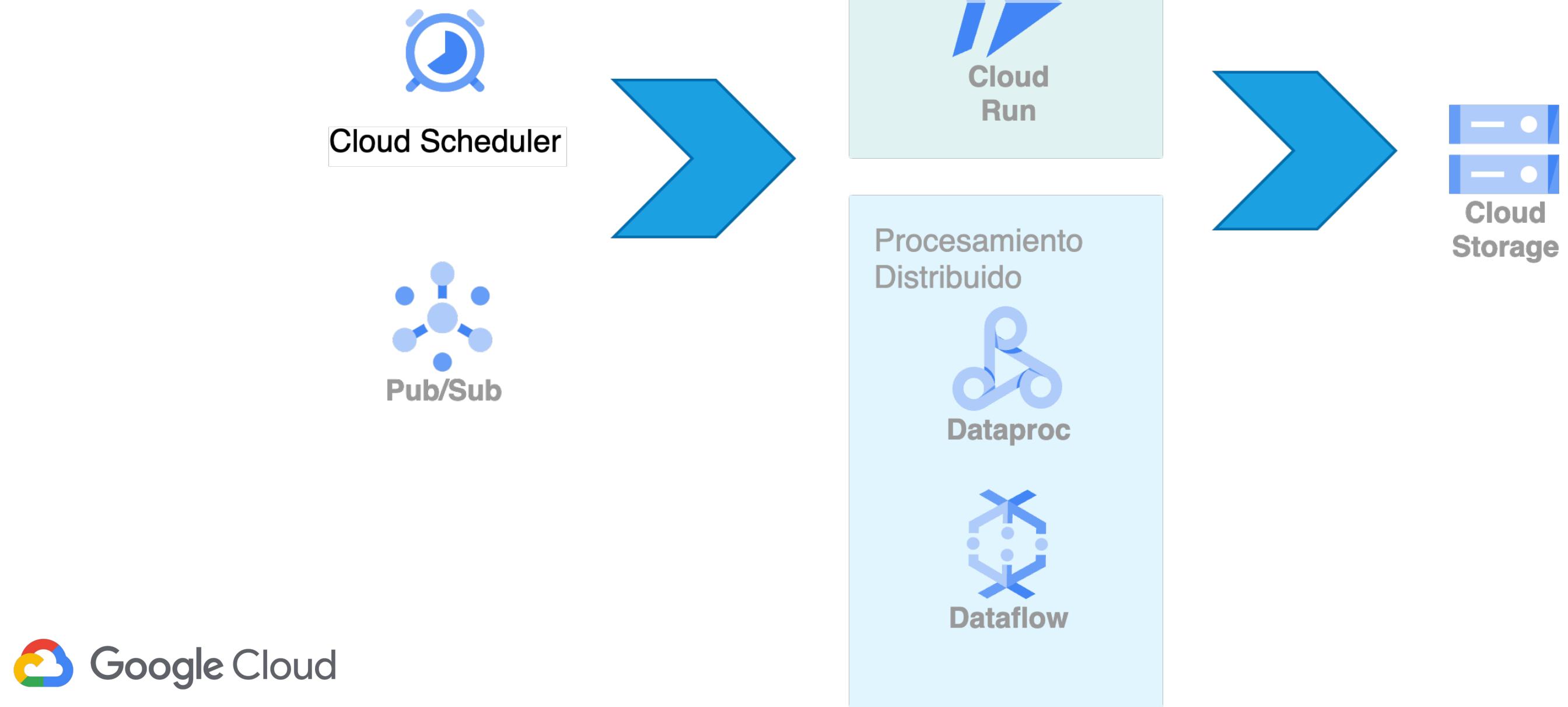
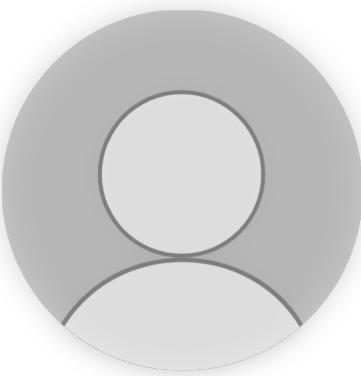
# Descarga de Archivos



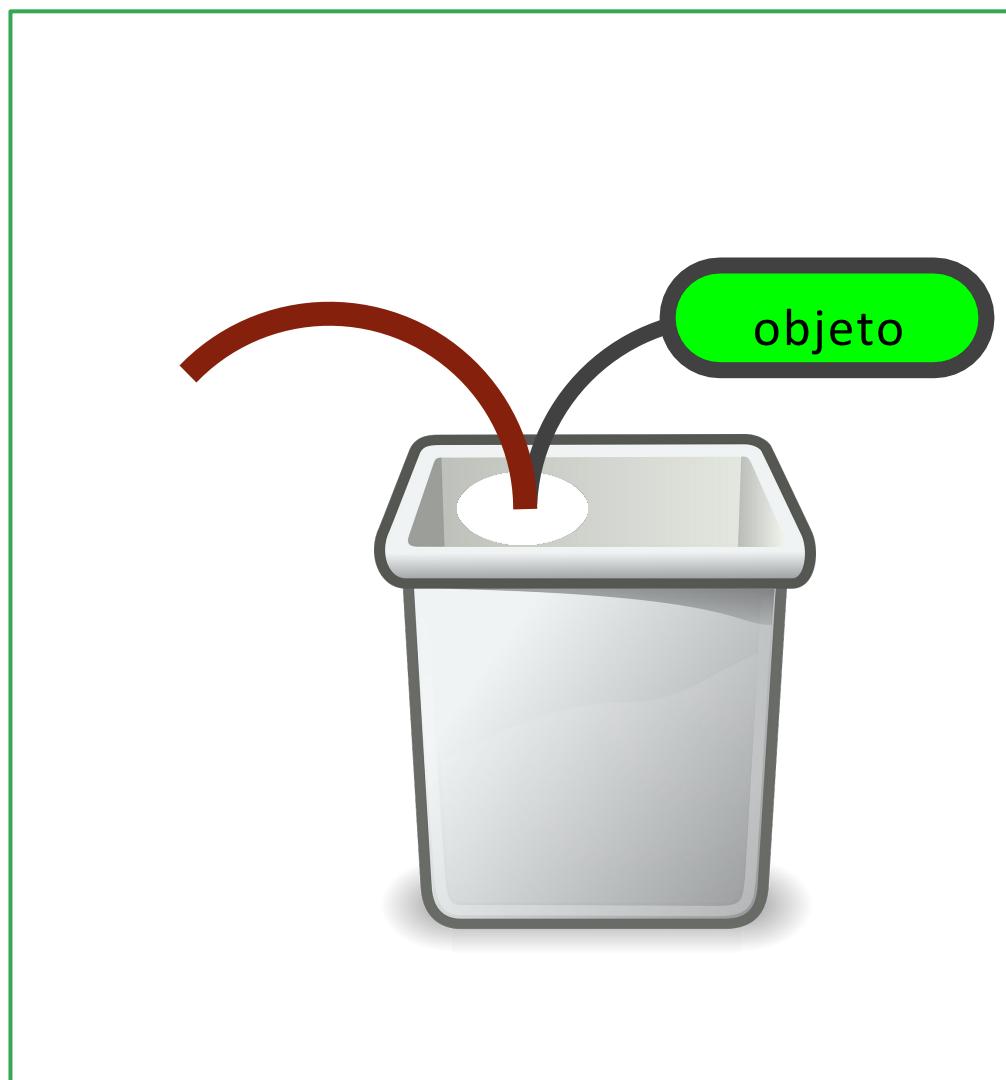
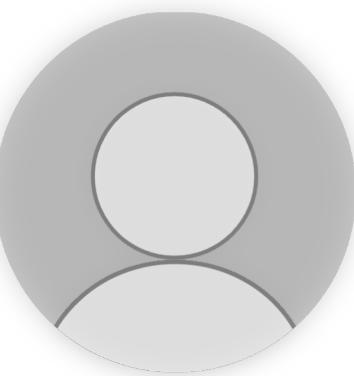
# Parseo de textos



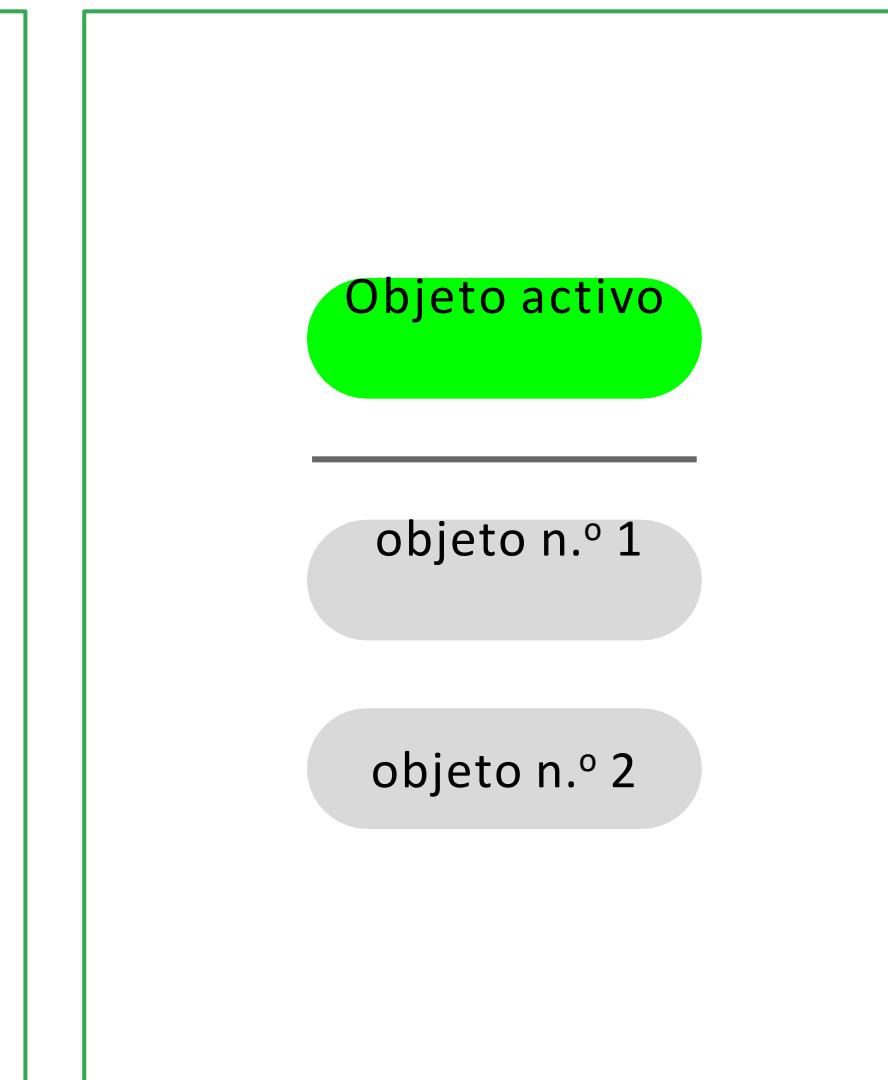
# Patrones de Arquitectura para ingestá a Cloud Storage



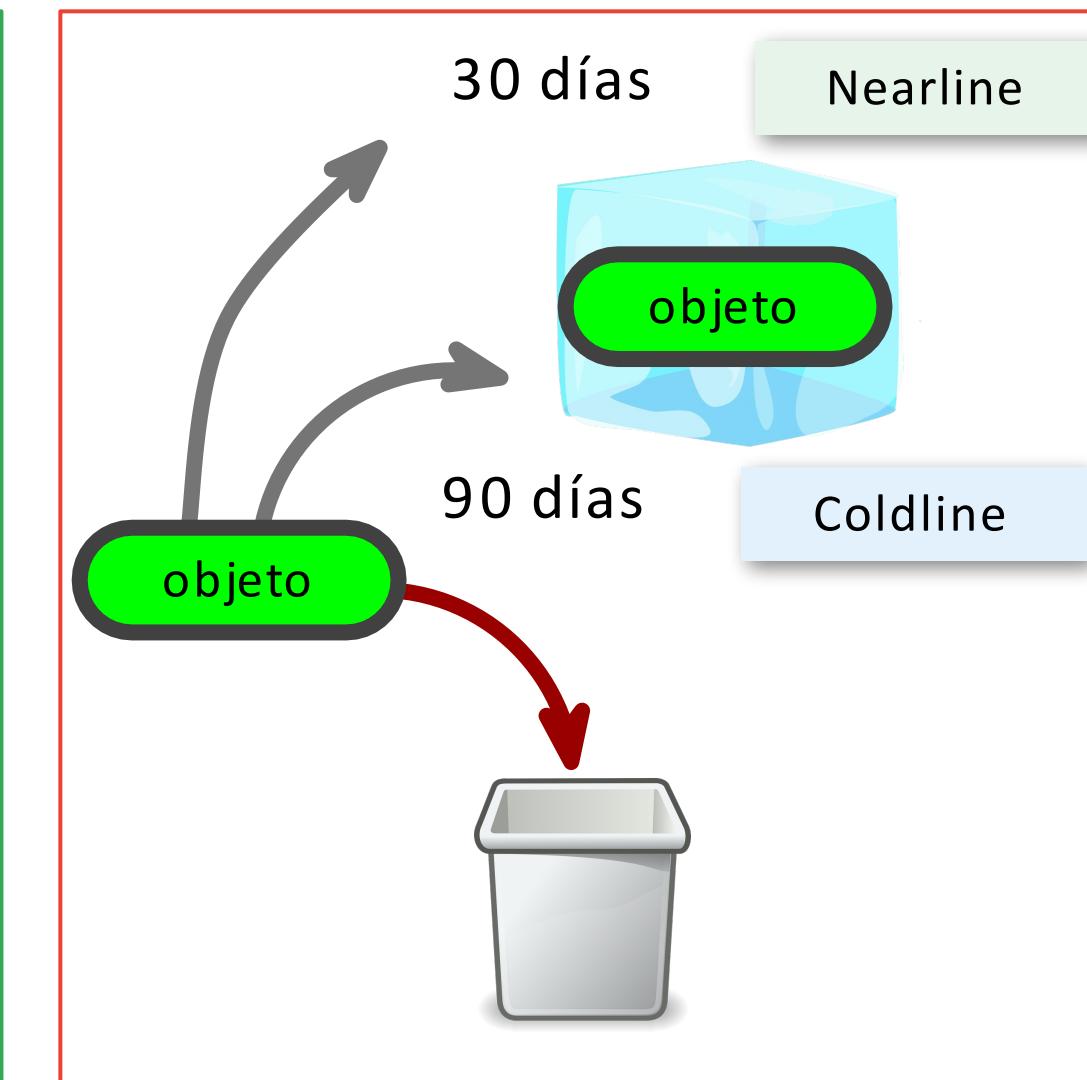
# Cloud Storage tiene varias características de administración de objetos



Política de retención



Control de versiones



Administración del ciclo de vida

# Arquitectura de administración de datos moderna serverless (sin servidores)

