

UNIVERSIDAD  
COMPLUTENSE  
DE MADRID

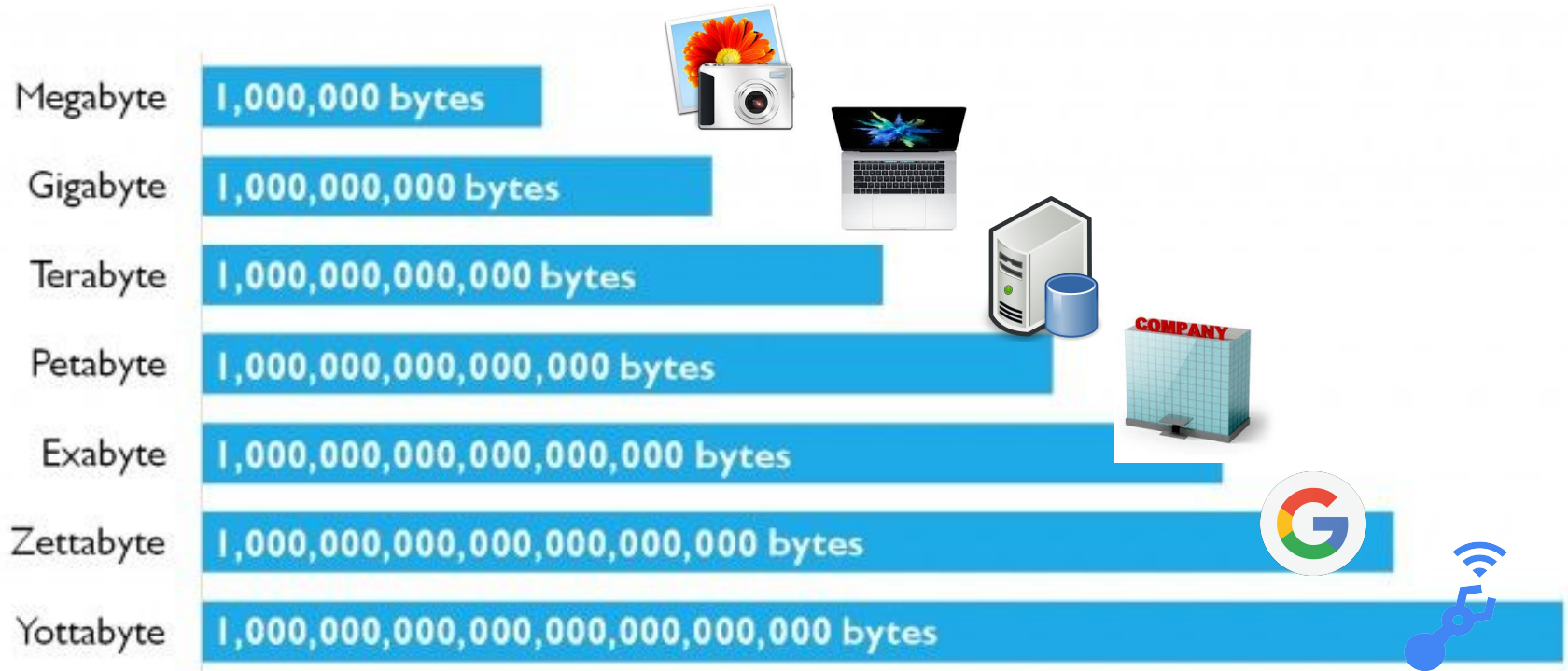




Almacenar datos en la nube.

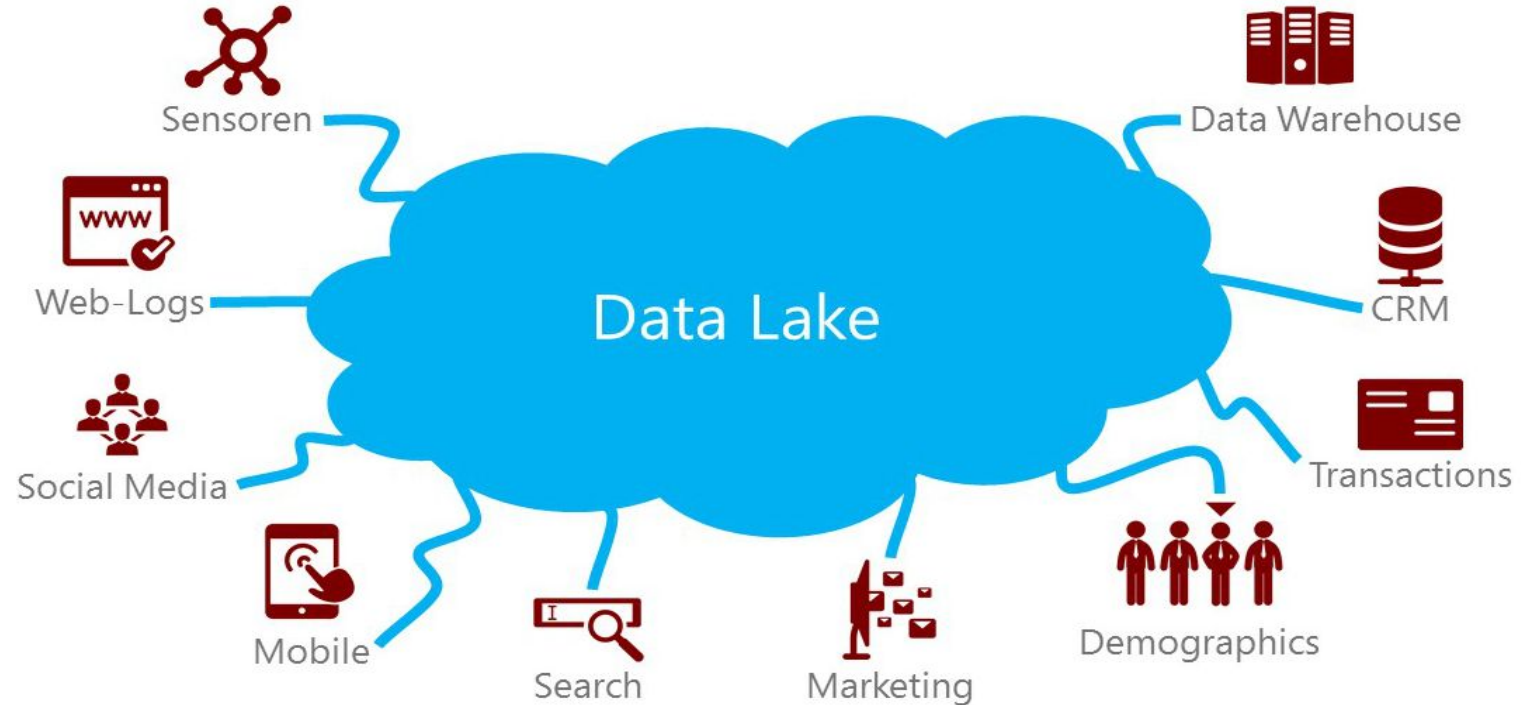
# Google Cloud Data Lake

## Volúmenes de Datos



# Google Cloud Data Lake

Un lugar para todos los datos - Data Lake



# Google Cloud Data Lake

## Data Lake Procesos



**Ingesta**

**Procesar**

**Almacenar**

**Analizar  
&  
Enriquecer**

**Visualizar**

- Almacenar datos en bruto u optimizada
- Admite datos estructurados y no estructurados
- Agnóstico de las herramientas de procesamiento, cada usuario elige
- Soporte para diferentes velocidades de datos
- Soporte para el modelo de seguridad de extremo a extremo
- Ofrece herramientas de auditoría para rastrear el uso de datos
- Proporciona diferentes temperaturas de datos

- Escala el almacenamiento a cualquier volumen, ¡incluso PB!
- Procese cualquier volumen en movimiento o en reposo
- Maximiza la utilidad de los datos
- Preocupaciones separadas entre los productores, procesamiento y consumidores
- Agilidad para acceder a los datos
- A prueba de desastres, sin pérdida de datos

# Google Cloud Storage (GCS)

## Globally Scalable Object Storage



# Google Cloud Data Lake

## Regional

Tus datos se almacenan en una región específica con replicación en zonas de disponibilidad en esa región. Es bueno para ubicar el cómputo y el almacenamiento para un alto rendimiento.

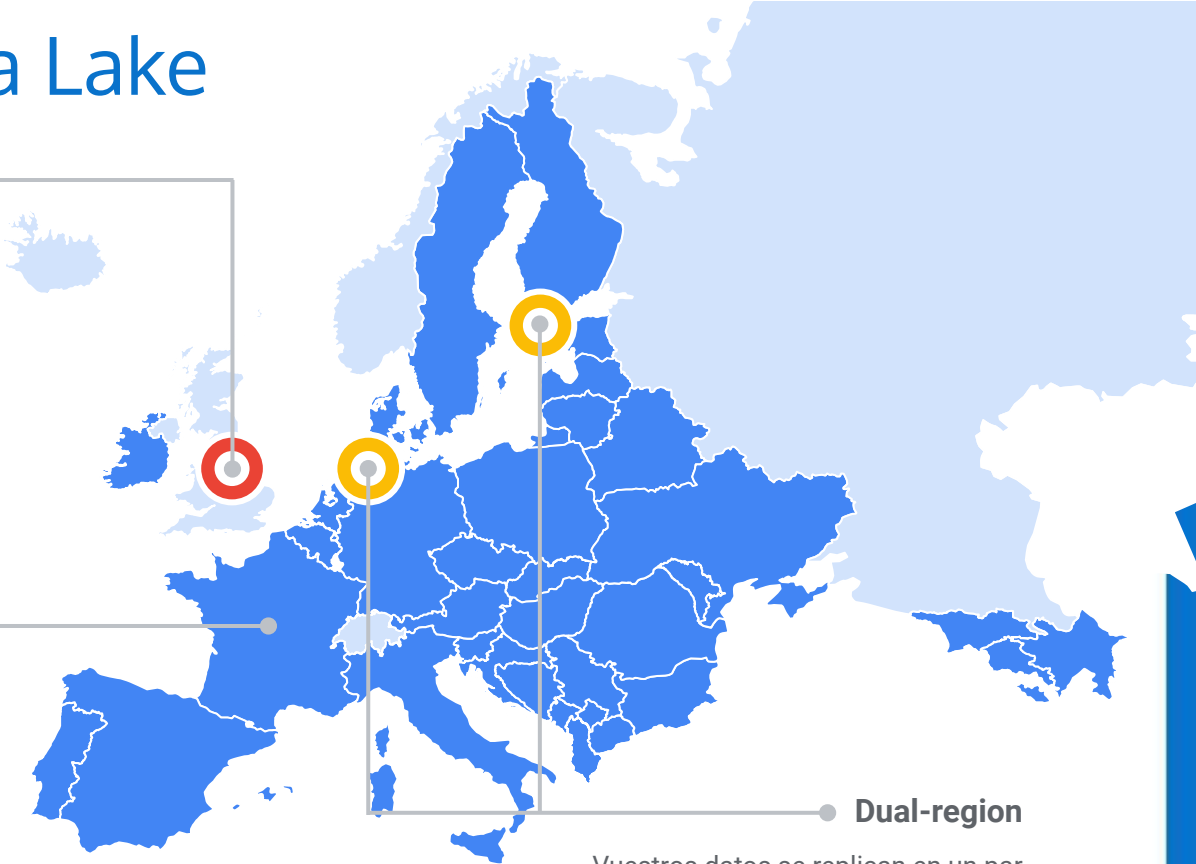
## Multi-Region

Tus datos se distribuyen de forma redundante en EE. UU, Europa o Asia. Es bueno para servir contenido a usuarios finales y cuando necesitemos una conmutación por error automática.

## Dual-region

Vuestros datos se replican en un par específico de regiones. Es bueno para cuando se necesita computación y almacenamiento juntos y failover automático.

## Elegir el tipo de almacenamiento



# Google Cloud Data Lake

## Google Cloud Data Lake Cuatro tipos de Almacenamiento



### Coldline

Cold  
99% SLA  
Milliseconds

Archive  
Source file backup  
Disaster recovery



### Nearline

Infrequent Access  
99% SLA  
Milliseconds

Backup  
Long-tail content  
Rarely accessed docs



### Regional

Regional  
99.9% SLA  
Milliseconds

Transcoding  
Data Analytics  
General compute



### Multi-Regional

Geo-redundant  
99.95% SLA  
Milliseconds

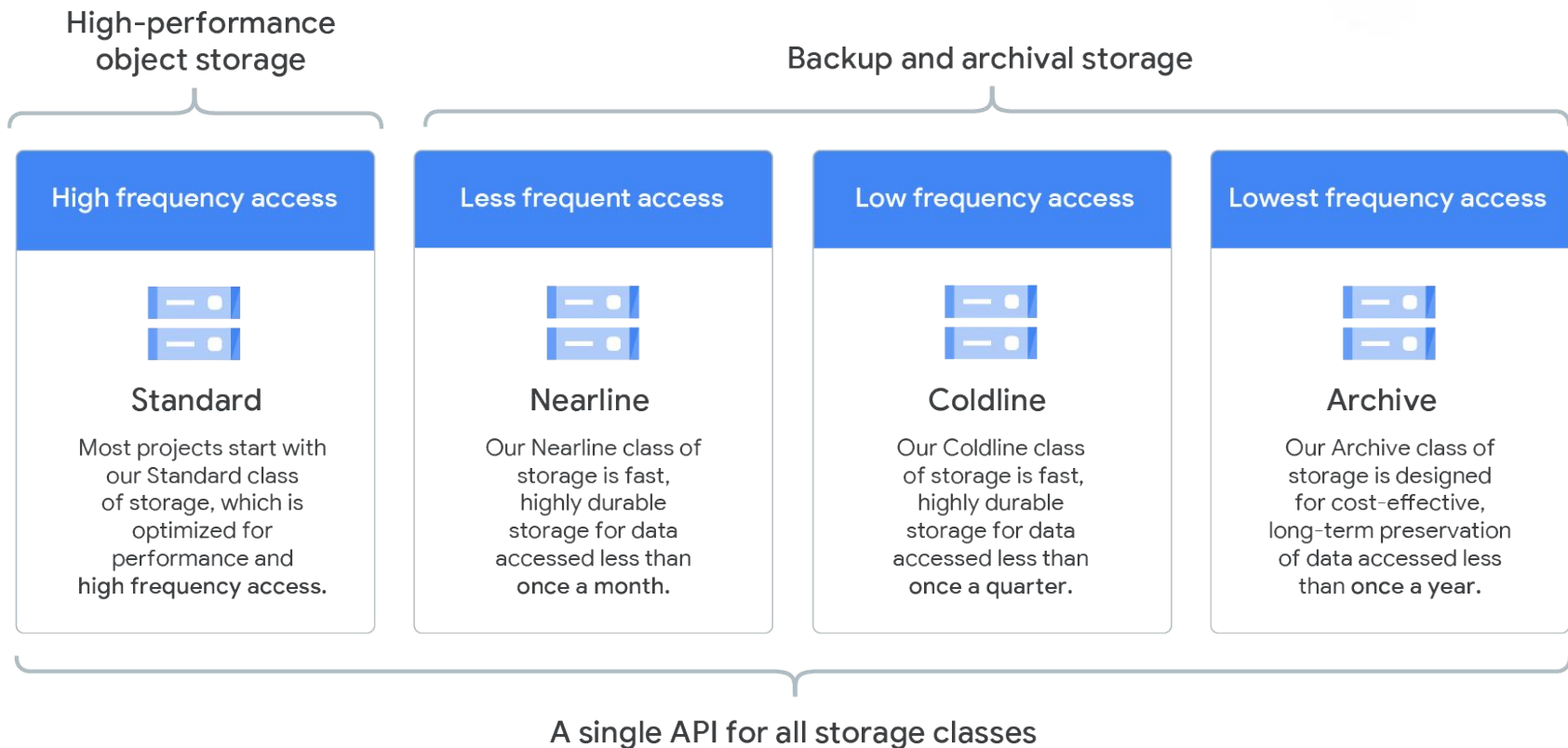
Video  
Multimedia  
Business continuity

Retrieval Frequency





# Google Cloud Data Lake



# Tipos de Almacenamiento



## GOOGLE CLOUD STORAGE

Exabytes, propósito general, escalado automático



## DISCOS PERSISTENTES

SSD/HDD Discos Persistentes  
Alto rendimiento, replicados



## ALMACENAMIENTO LOCAL

Local SSD (NVMe) muy rápidos  
conectados físicamente via PCI



## FILESTORE

Alta disponibilidad, POSIX compliant compartidos a través  
de decenas de miles de nodos



## SOLUCIONES de TERCEROS

Soluciones de terceros como NetApp, Elastifile, DDN, etc.  
Mover petabytes a GCS usando el Data Transfer Appliance

Data  
Storage



# GCS Object Storage

Cloud Storage

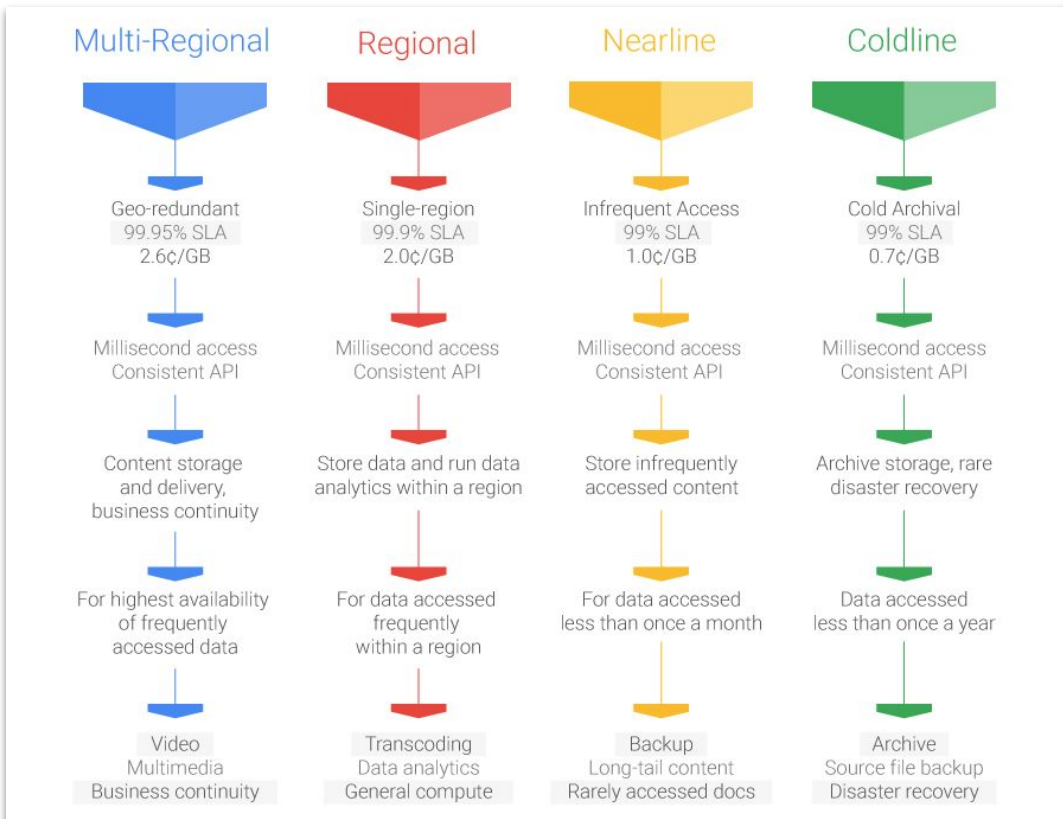
vs.

Discos Persistentes:

- Escala a exabytes
- Accesible desde cualquier sitio
- Mayor latencia que los discos persistentes
- Versionado
- Más barato

Google Cloud

## Google Cloud Storage Opciones



# Disco

## Persistent Disk

- Disponible en Red
- Tamaño configurable (hasta 64 TB)
- Disco Magnético o SSD
- Independiente de la VM
  - Snapshots globales
  - Lectura desde múltiples VM
- Mayor tamaño == mayor velocidad

## Local SSD

- Mínima latencia
- 375 GB por disco (hasta 8x)
- Atado al ciclo de vida de la VM

¡Siempre cifrado!



## Create a disk

Name ?

disk-1

Description (Optional)

Local SSD scratch disk (maximum 8)

Max disk size: 375 GB

SSD persistent disk

Max disk size: 65,536 GB

Standard persistent disk

Max disk size: 65,536 GB

Size (GB) ? (Optional)

Estimated performance ?

Operation Type

Read

Write

Sustained random IOPS limit

Sustained throughput limit (MB/s)

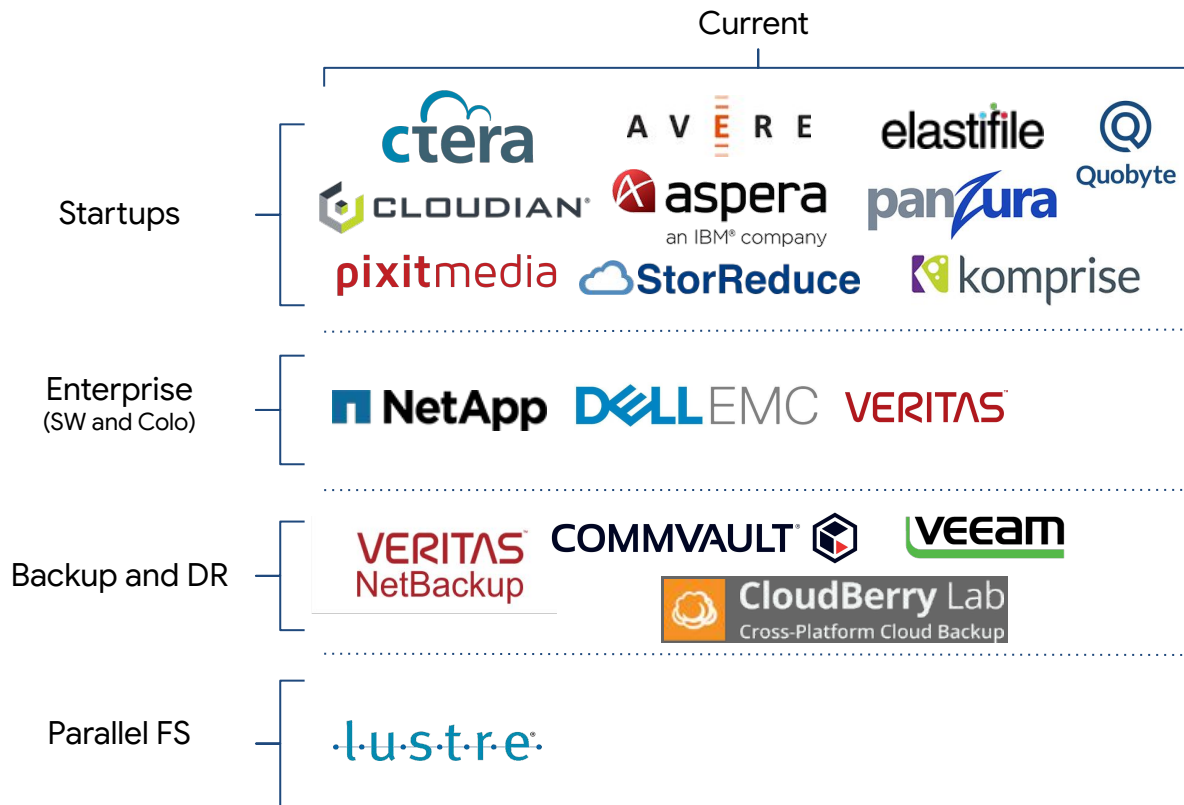


# Opciones de Almacenamiento

	Local SSD Fastest, Attached, Ephemeral	Persistent Disk: SSD Fast, Persistent, Durable, Remote	Persistent Disk: HDD Cheapest, Persistent, Durable, Remote
Target scenarios	<ul style="list-style-type: none"> <li>- <b>High-performance scratch space. Frequently accessed data.</b></li> <li>- Excellent for scientific workloads, especially when combined with fast compute VMs like GPU instances</li> </ul>	<ul style="list-style-type: none"> <li>- Latency sensitive applications and files.</li> <li>- High performance database and enterprise applications</li> <li>- Databases</li> </ul>	<ul style="list-style-type: none"> <li>- Large data processing workloads</li> <li>- Latency incentive tasks with lots of data: Genomics processing, video transcoding in GCE</li> </ul>
Features	<ul style="list-style-type: none"> <li>- Ephemeral storage</li> <li>- Highest-performance</li> <li>- IOPS: 680k read / 360k write</li> <li>- Max Throughput: 2.5 GB/s read, 1.4GB/s write</li> </ul>	<ul style="list-style-type: none"> <li>- Persistent storage</li> <li>- Performance sensitive</li> <li>- IOPS: up to 40k read / 30k write</li> <li>- <b>Max Throughput: 1200 MB/s read, 400 MB/s write</b></li> </ul>	<ul style="list-style-type: none"> <li>- Persistent storage</li> <li>- IOPS: 3k read / 15k write</li> <li>- <b>Max Throughput: 180 MB/s read, 120 MB/s write</b></li> </ul>
	Encryption 3TB per instance. 375 GB per partition, up to 8 partitions	Encryption, Snapshots 64 TB, Disk Size sets performance (Attach larger VMs for max SSD performance)	
Price / GB	\$0.08	\$ 0.17	\$0.04

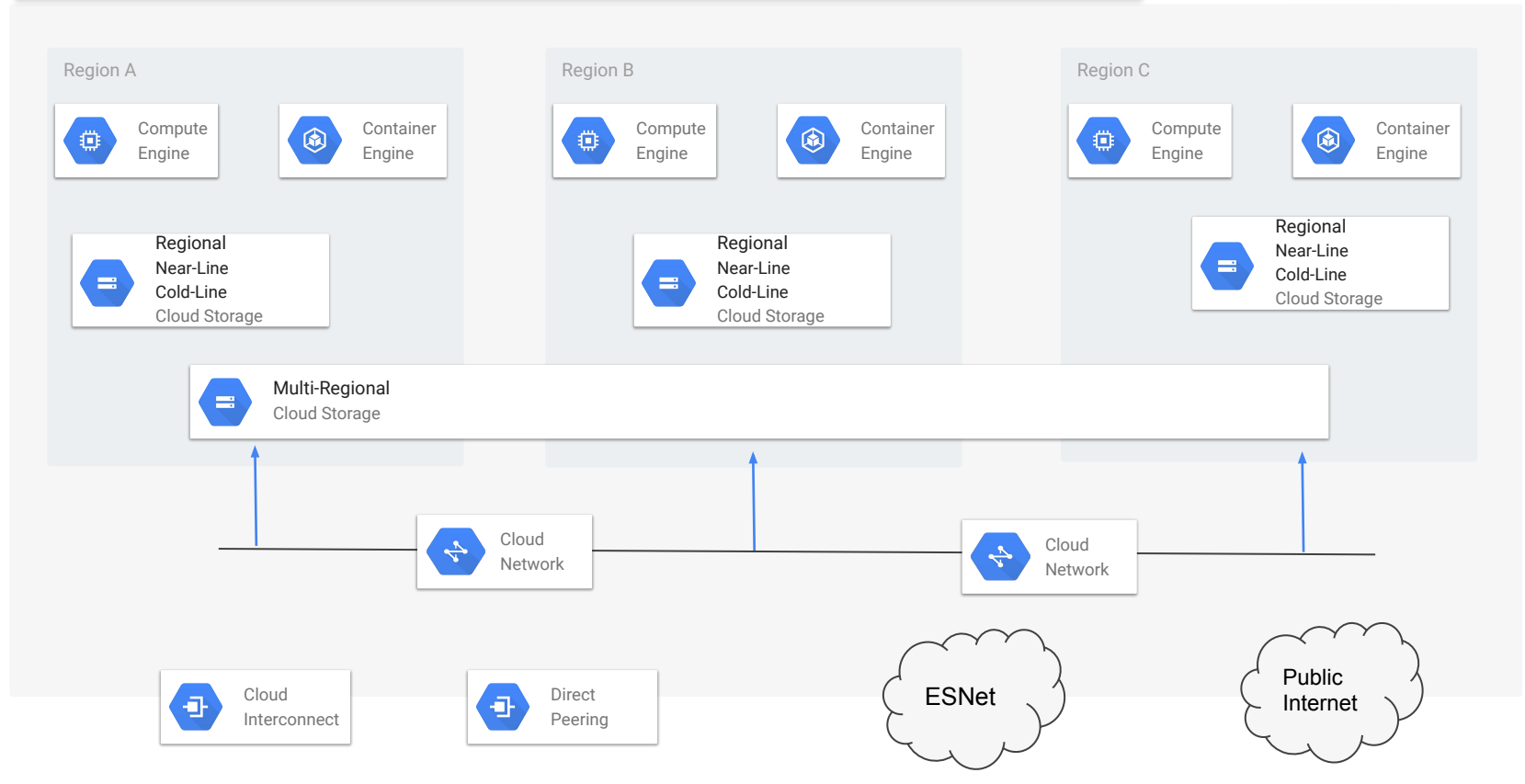


# Herramientas de Terceros



# Google Cloud Data Lake

Architecture: Storing Atlas Data in Google Cloud Storage

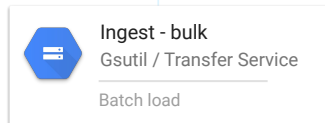
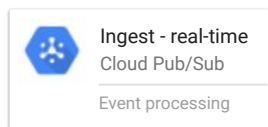


# Cloud Storage - Analytics Pipeline

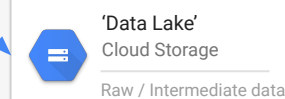
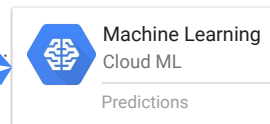
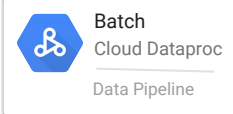
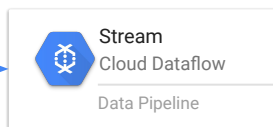
Análisis de cliente omnicanal (por ejemplo, compras, juegos, telecomunicaciones, finserv)



## Ingestar

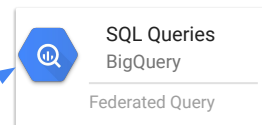


## Analizar

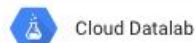


Cloud Storage: Regional  
Storage / Analytics

## Servir



Interactive Data Exploration



BI Tools




Tiempo  
Real  
(hot)

Batch  
(cold)



# GCS Object Storage

Standard		Nearline	Coldline	Archive
Multi-region Para enviar contenido de forma global		Para datos que se accede mensualmente	Para datos que se accede anualmente	Almacenamiento para datos a largo plazo, regulatorio
 Streaming videos	 Video transcoding	 Serving rarely accessed docs	 Serve rarely used data	 Regulatory archives
 Images	 Genomics	 Backup	 Movie archive	 Tape replacement
 Websites	 General data analytics & compute		 Disaster recovery	
 Documents				



# MUCHAS GRACIAS

UNIVERSIDAD  
COMPLUTENSE  
DE MADRID

