

# Los Anillos de Seguridad:

## Salvaguardando el Poder de tu Data Lake en

AWS

Alvaro Garcia  
@alvarongg  
01 – 07 - 2023



# AGENDA

- Introducción a los Data Lakes en AWS
- Arquitectura de un Data Lake Serverless en AWS
- Amenazas y riesgos de seguridad en un Data Lake
- Los Anillos de poder para gobernar Data Lakes Serverless
- Consideraciones finales
- Q&A



# ME PRESENTO:



**Alvaro Garcia - @alvarongg** 

- Data Lead en Cloudbase.
- +1 década trabajando en data.
- Entusiasta de la seguridad informática.
- Uruguayo.



- Introducción a los Data Lakes en AWS



“El recurso más valioso del mundo ya no es el petróleo, sino los datos.”

-The Economist, 2017

<https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>

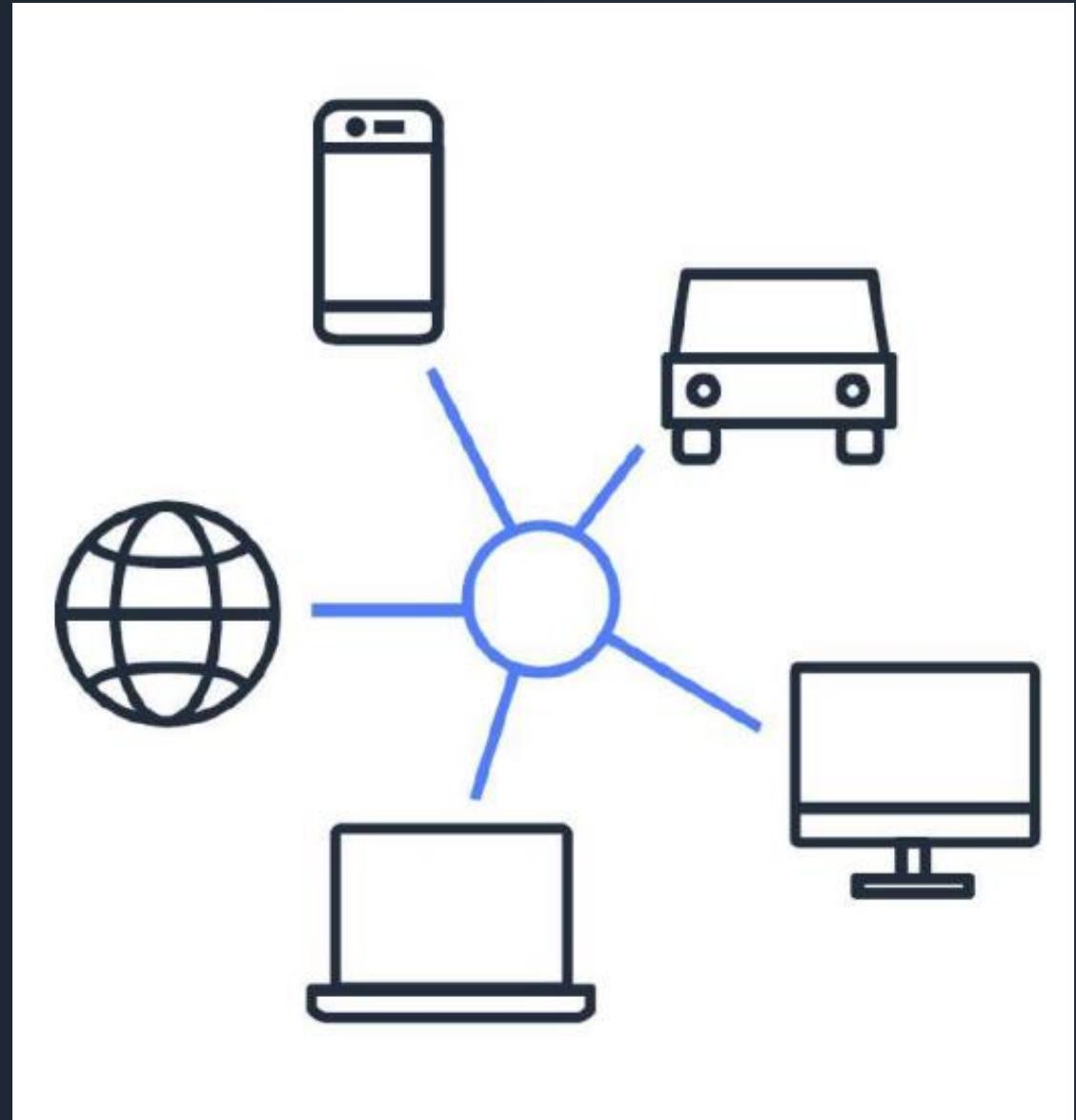


# Definición de un Data Lake

**Datos capturados en formato digital**

- Personas
- Eventos
- Elementos
- Transacciones

**Los datos de los clientes se convierten  
en información**



# Qué es Big Data?



Cuando la cantidad de datos es tan grande y crece tan rápidamente que almacenarlos y administrarlos se convierte en un desafío.



Tiene las siguientes características: volumen, velocidad, variedad y veracidad.



# Las cuatro “V” del Big Data

## Volumen

Escala del tamaño de los datos

## Velocidad

Velocidad a la que se crean y crecen los datos

## Variedad

Variedad de formatos y orígenes de datos

## Veracidad

Incertidumbre de los datos

# Flujo de los Datos



**Recopilar y conservar  
todos los datos**

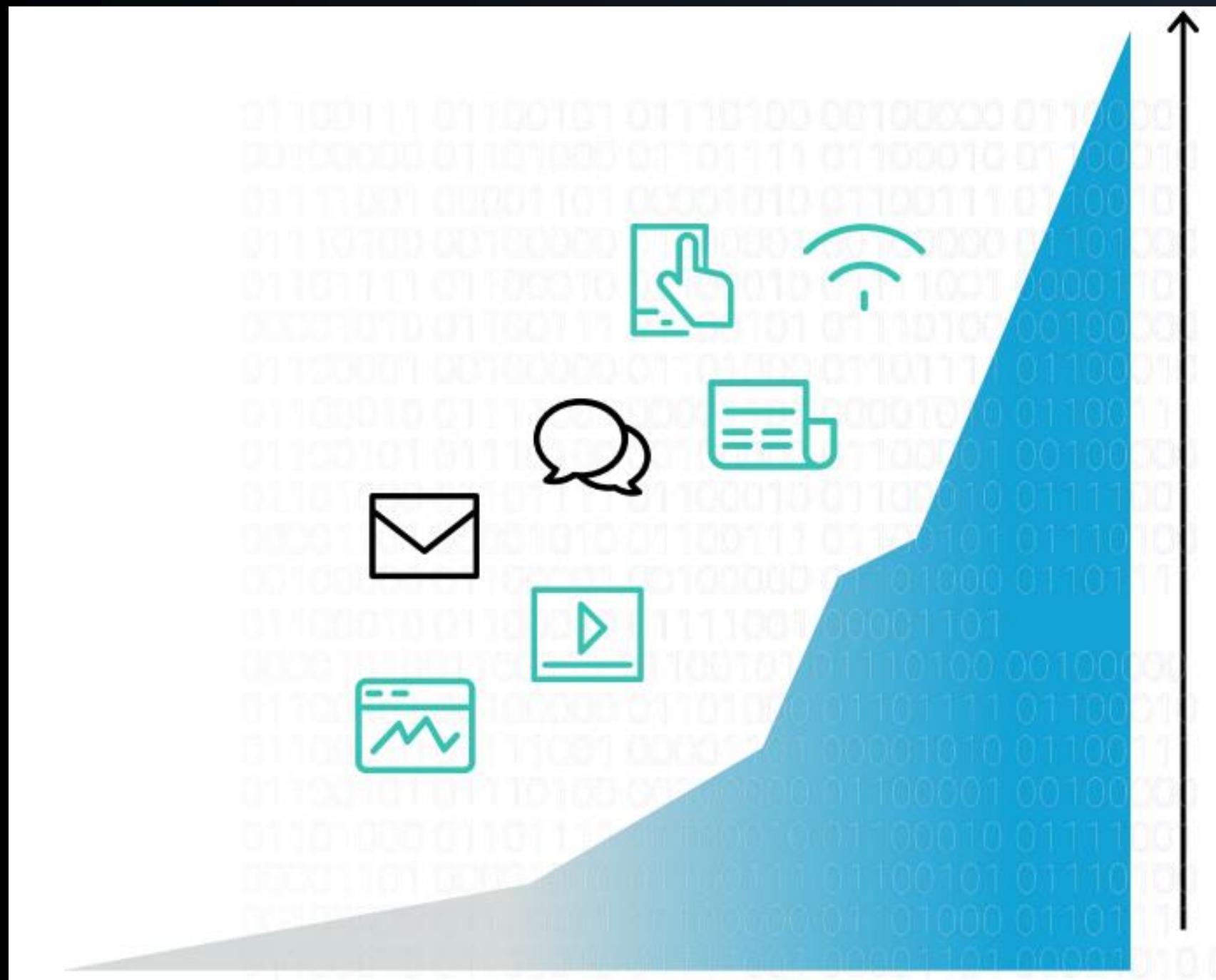


**Hacer que los datos estén  
a disposición de los  
usuarios**



**Proporcionar a los usuarios  
tecnologías de  
procesamiento de datos**

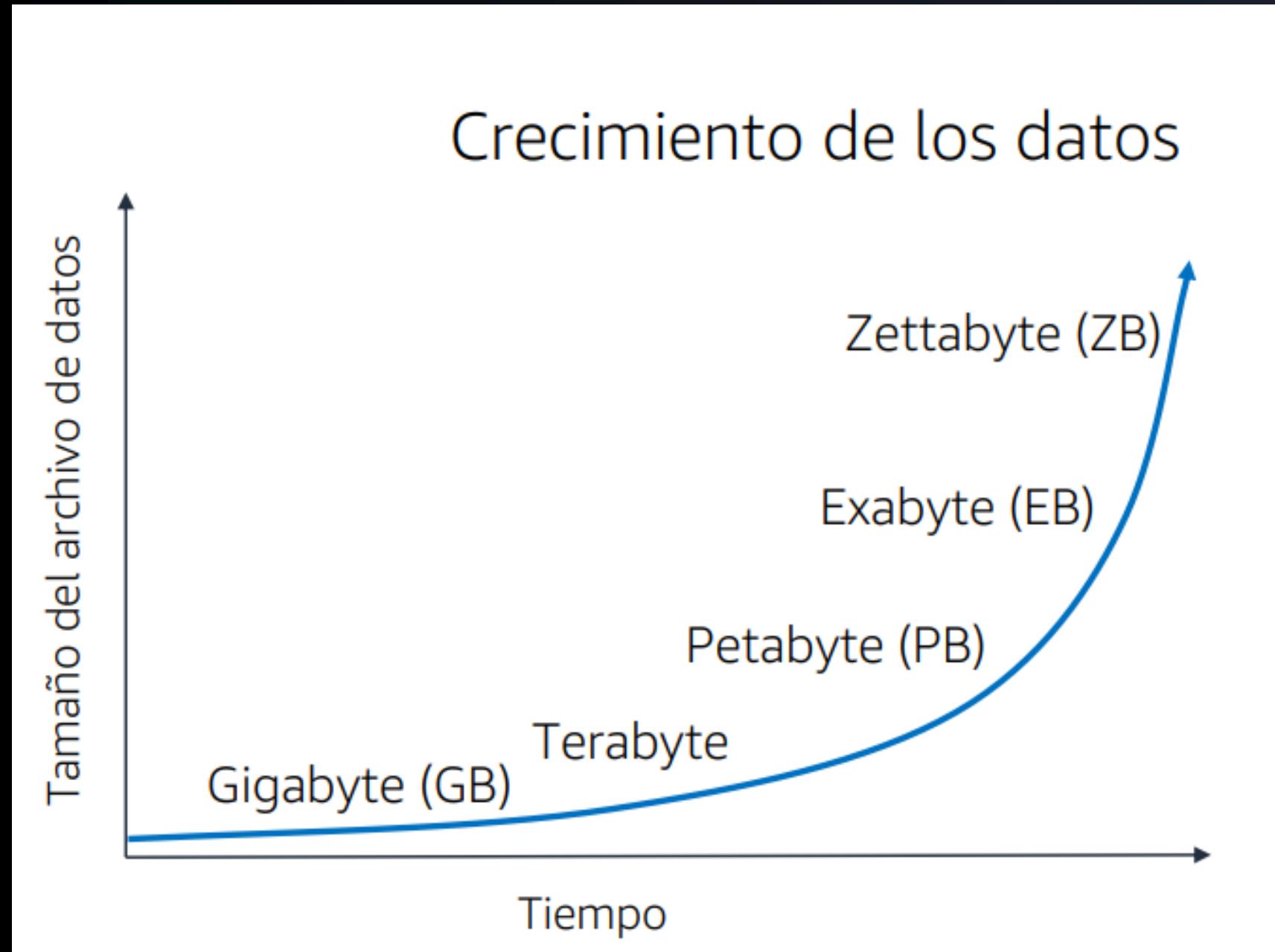
# Datos en constante crecimiento



Hay más datos de los que la gente piensa

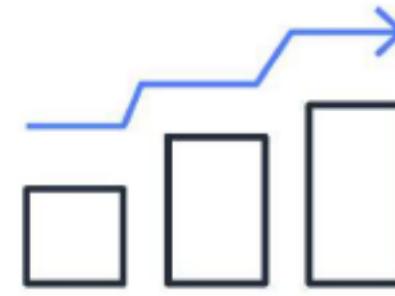
Datos	Plataformas de datos deben
Crecen <b>&gt;10 veces</b> cada 5 años	Escalar <b>x 1000</b>

# Los datos están creciendo más rápido que nunca



IDC predice que la esfera de datos global crecerá de **33 ZBs en 2018** a **175 ZBs en 2025**

# Desafíos en la administración de un Data Lake



Están creciendo exponencialmente



Vienen de nuevas fuentes



Son cada vez más diversos



Utilizan muchas personas

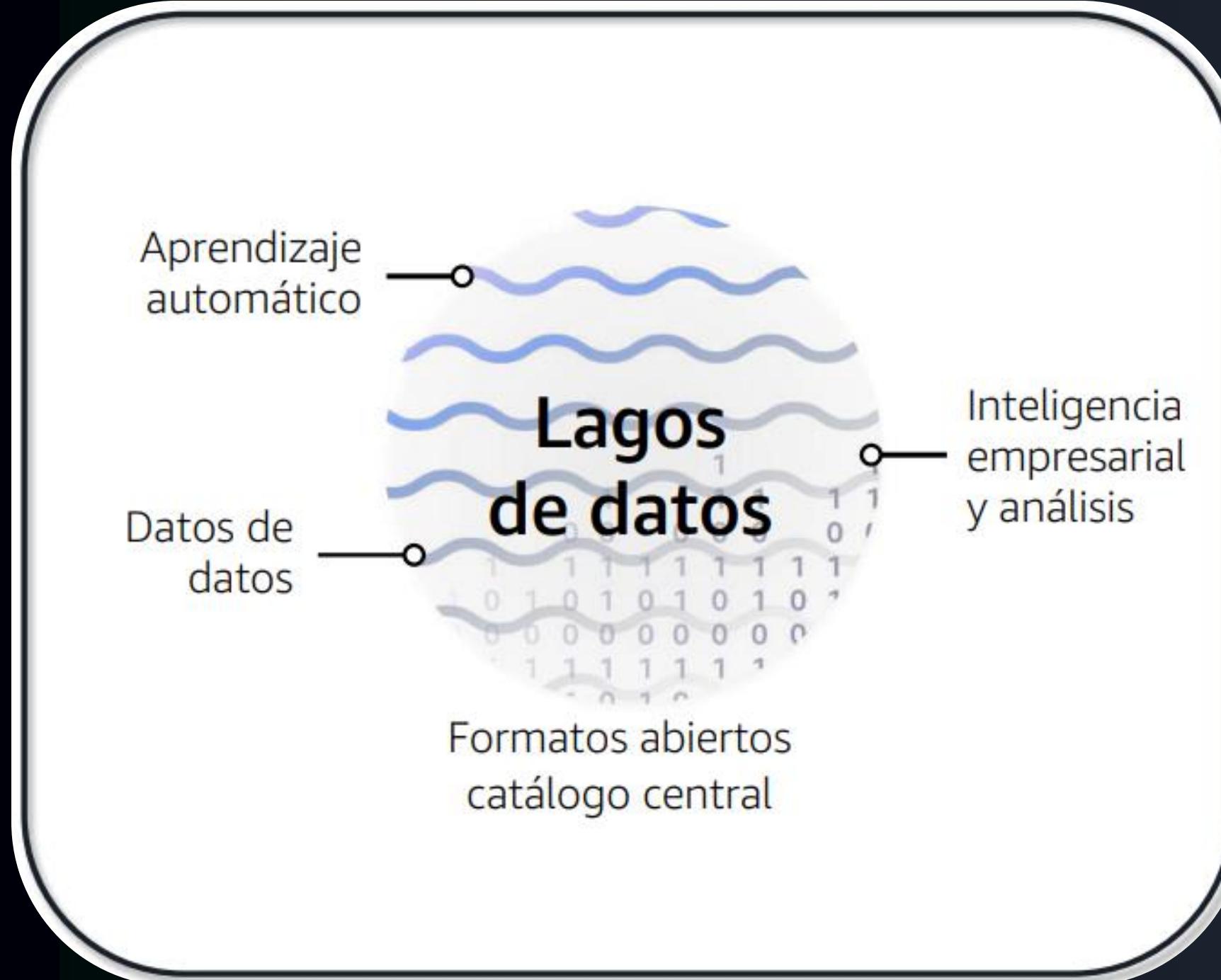


Se analizan con muchas aplicaciones

- Arquitectura de un Data Lake en AWS



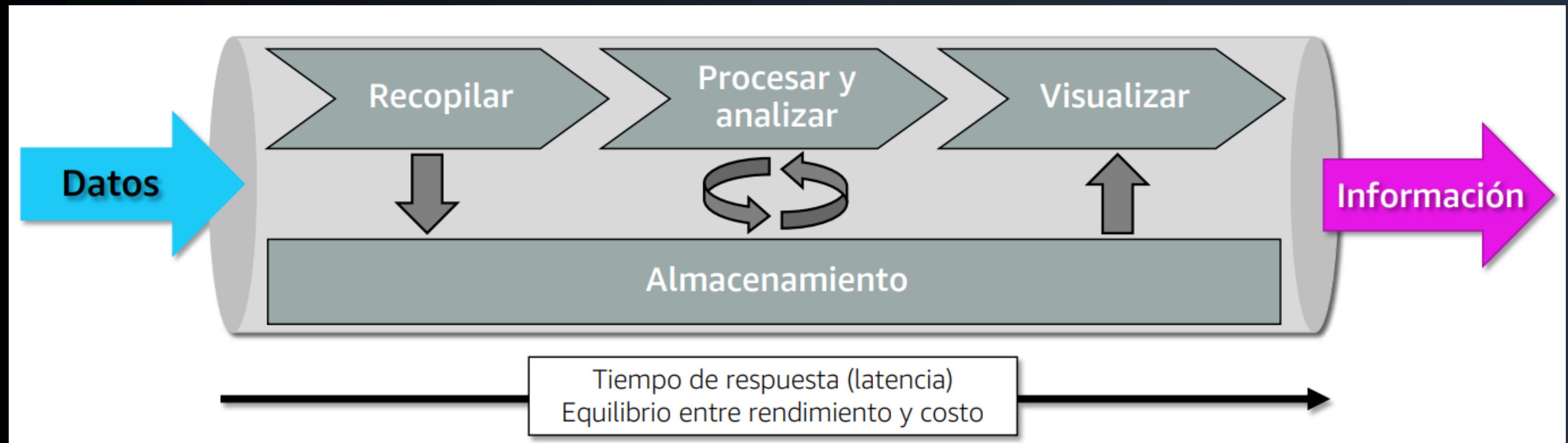
# Una arquitectura de datos moderna



## Beneficios:

- Desplazarse a un único almacén; un lago de datos en la nube.
- Almacenar datos de forma segura en formatos estándar.
- Crecer a cualquier escala, con costos bajos.
- Analizar los datos de diversas formas.
- Democratizar el acceso a los datos y el análisis.

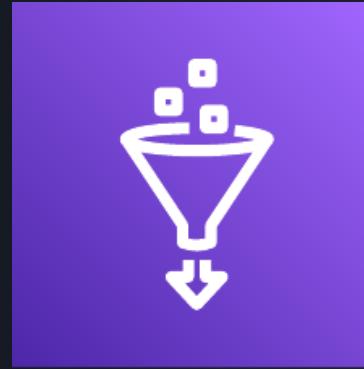
# Canalización de análisis de datos



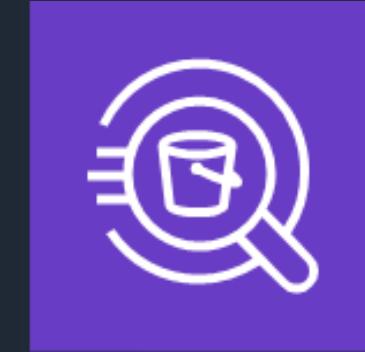
# Principales servicios de un Data Lake en AWS



Amazon S3



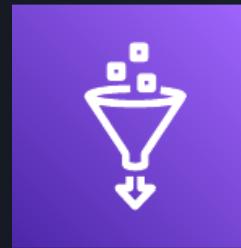
AWS  
Glue



Amazon  
Athena



Amazon  
S3 Glacier



AWS Glue  
Jobs

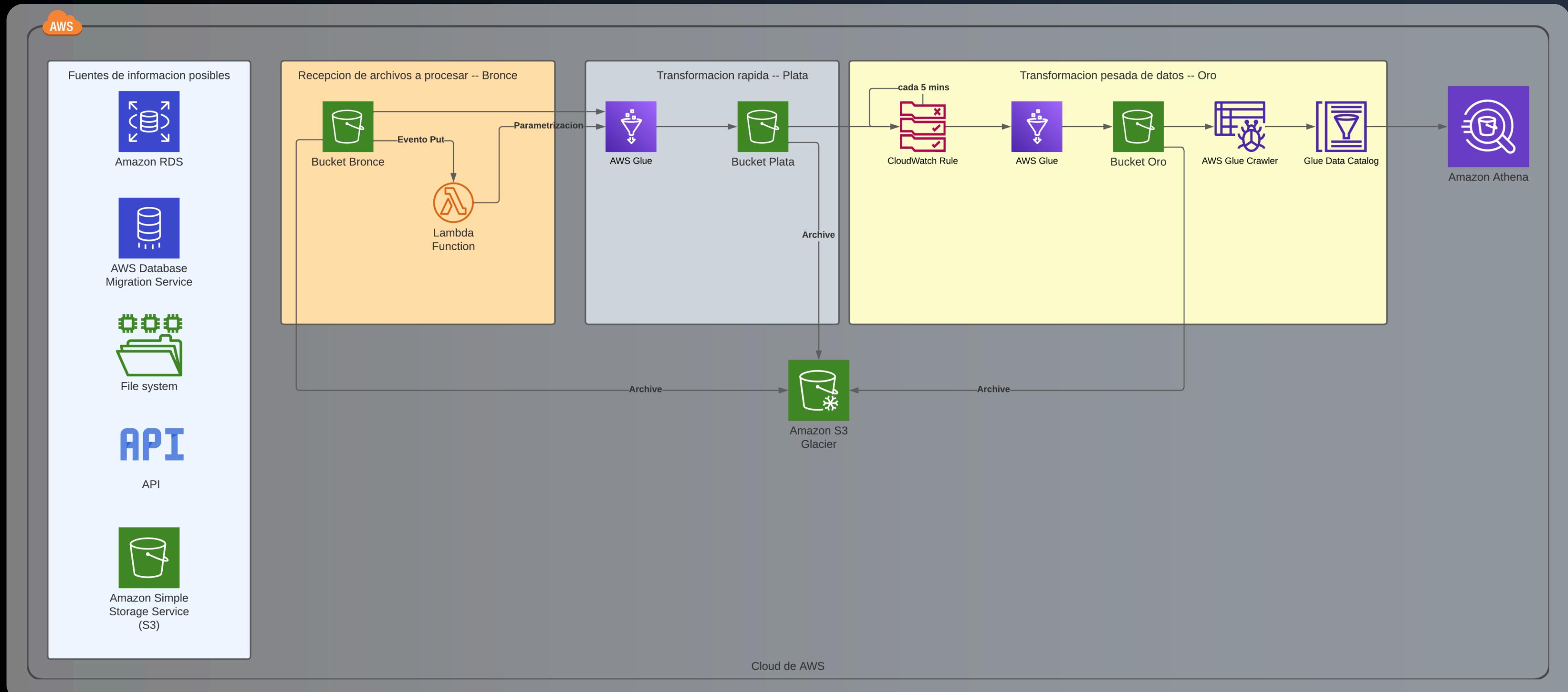


AWS Glue  
Crawlers



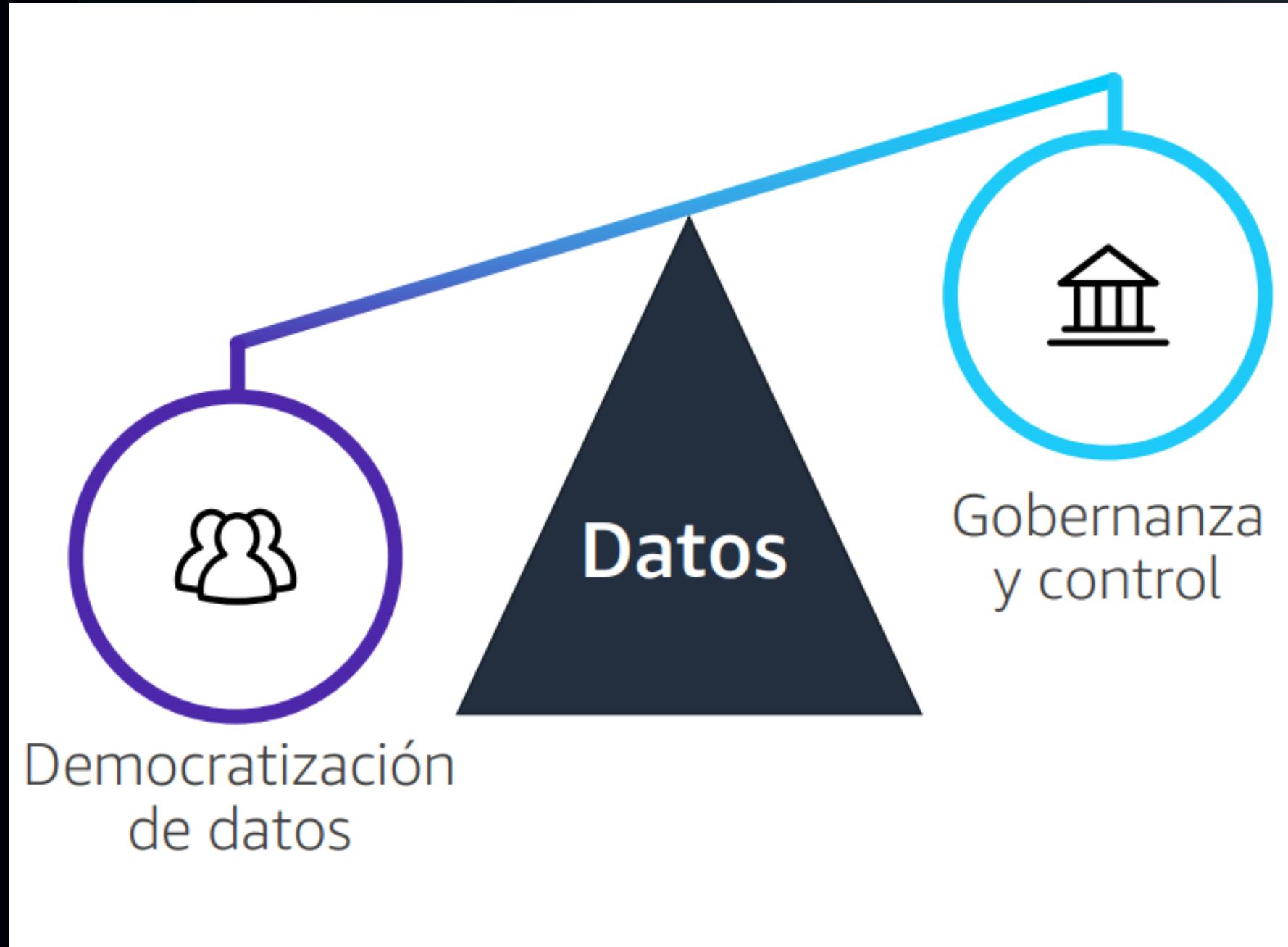
AWS Glue  
Catalog

# Arquitectura Base de un Data Lake



- Amenazas y riesgos de seguridad en un DL

# Gobernanza de datos



Hay más personas trabajando con datos que **nunca**

---

Proporcionar **acceso democratizado** a los datos y, al mismo tiempo, hacer cumplir **la gobernanza de datos** es el verdadero desafío.

# Desafíos de gobernanza en los lagos de datos

- Protección de datos
- Auditoría del uso de datos
- Administración del acceso a datos
- Protección de información confidenciales y PII
- Mantenimiento de los reglamentos y los mandatos



# Amenazas y riesgos de seguridad en un DL

- Brechas de seguridad y accesos no autorizados.
- Ataques de inyección y manipulación de datos.
- Riesgos de privacidad y cumplimiento normativo.
- Amenazas internas y privilegios excesivos.
- Ataques de denegación de servicio (DDoS).
- Fugas de datos y pérdida de información.
- Amenazas emergentes y ataques sofisticados.



- Los Anillos de poder para gobernar  
Data Lakes Serverless



# Los Anillos de poder para gobernar Data Lakes Serverless

- Modelo basado en capas.
- Tienen funciones claras.
- 100% serverless.





# Anillo Único: Datos

- Nuestro objetivo a proteger.
- Separación por unidad de negocio.
- Priorización de Vistas sobre Tablas.
- Testeo de nuestro código ETL.



**AWS Glue**



**Amazon  
Athena**



**Amazon S3**



# Amazon S3

- Separar la información en buckets por unidad de negocio
- Asignar políticas a los buckets:
  - Restricciones por VPC, HTTPS, filtrado por IP, KMS Keys
- Restricciones utilizando Tags y Condiciones:
  - "Condition": {"StringEquals": {"S3:ResourceTag/GPRD": "True"}}
- Activar Encriptación y Versionado.
- Aplicar ACL's.
- Recordar que AWS ya nos provee cifrado por defecto.





# AWS Glue

- Utilizar los ACL's sobre los grupos (Crawler & Catalog ).
- Replicar la misma separación por unidad de negocios que en S3.
- Modularizar el código para aumentar la mantenibilidad.
- Evitar librerías externas a las que trae Glue de base.
- Testear el código (PyTest).
- Activar la encriptación de Bookmarks.
- Validar tamaño de archivos/particiones.



# Amazon Athena

- Utilizar los ACL's sobre los grupos.
- Replicar la misma separación por unidad de negocios que en S3.
- Priorizar el uso de vistas por sobre tablas:

```
CREATE VIEW V_Ventas_Frutas AS
SELECT sum(v.monto)
FROM ventas AS v
INNER JOIN productos as p
ON v.sku = p.sku
WHERE p.categoría = "Frutas"
```



# Anillo Azul : Roles y Permisos

- Agrupamiento de roles por BU.
- Siempre access Keys en Secretos.
- Privilegios mínimos y específicos.



AWS Identity and  
Access  
Management  
(IAM)



AWS Secrets  
Manager



Amazon  
Athena



Amazon S3



AWS Lake  
Formation



AWS Glue



# AWS Identity and Access Management (IAM)

- Identificar los actores y roles del flujo de datos.
- Siempre utilizar los IAM Permission Boundaries para limitar los privilegios .
- Planificar los niveles de acceso de cada grupo con la mayor granularidad posible:
  - Un analista no necesita escribir en un bucket de S3.
    - "glue:GetTable" / "glue:GetPartition"
  - Un ingeniero de datos debe escribir en producción.
    - S3:putObject,S3:GetObject,S3:DeleteObject --  
> arn:partition:service:region:account-id:bucket\_dev/\*

# Identificar los actores del flujo de datos



- **Curadores de datos :**
  - Acceso a los datos crudos y finales.
  - Muy pocos usuarios (1 por Unidad de negocio).
  - Es quien debe asignar los permisos a los demás usuarios.
- **Mineros de datos (Data Engineers y Data Scientist) :**
  - Acceso a datasets en general de ambientes no productivos.
  - Los accesos se limitan por clase y por tiempo.
  - No deben de tener la posibilidad de extraer los datos fuera de AWS.
- **Usuarios de Negocio (Analistas o Ejecutivos):**
  - Acceso mínimo y granular a la información.
  - Nivel de acceso por Rol y Unidad de negocio.
  - El acceso de estos usuarios puede llegar a los datos productivos.



## Secret Manager

- Aprovechar este servicio para trabajar con diferentes ambientes.
- Muy importante para subir esquemas de dataset que necesitemos validar.
- Reduce la necesidad de tener archivos perdidos en un repositorio.



## Lake Formation

- Nos permite asignar y quitar permisos al nivel de datos que necesitemos.
- Ayuda a mantener el DL sin información duplicada.
- Encriptación de datos por defecto.



# Anillo Blanco : Cifrado y ofuscación

- Cifrado en tránsito y en reposo.
- Detección de PII.
- Ofuscación de datos en dev/test.
- Normas de compliance.



AWS  
KMS



Amazon Macie



AWS Identity and  
Access  
Management  
(IAM)



AWS CloudHSM



AWS Secrets  
Manager



AWS Glue



Amazon  
Athena



Amazon S3

A dark, atmospheric scene of a castle at night. The castle has multiple towers and glowing windows, with a path leading up to it. The ground is covered in fallen leaves.

# Encriptación de datos en reposo



## KMS

- Serverless – Gestionado por AWS.
- Costo basado en la demanda.
- Norma FIPs 140-2 nivel 2.
- Free Tier (20k peticiones/mes).



## Cloud HSM

- Requiere aprovisionamiento – Hardware Específico.
- La gestión las keys corre por nuestra cuenta.
- Costo por hora.
- Norma FIPs 140-2 nivel 3.

**En resumen:** Si el data lake va a manejar información personal debemos utilizar Cloud HSM , de lo contrario podemos utilizar KSM



# Detección de datos sensibles



# Amazon Macie

- Activarlo sobre los Buckets necesarios.
- Nos ayuda a separar y tagear los datos sensibles.
- Podemos crear nuestros clasificadores mediante REGEX o palabras clave.
- Podemos procesar los findings como queramos (SNS, S3, Security Hub, etc).
- Tambien nos avisa de buckets mal securizados.
- Free Trial de 30 Dias.



# Anillo Rojo : Control , Monitorización y Auditoria

- Control de eventos de seguridad.
- Auditoria de logs.
- Cumplimiento de Compliance.
- Observabilidad de la plataforma.





# AWS SNS

- Excelente herramienta de Observabilidad.
- Manejo de eventos y notificaciones.
- Podemos disparar un flujo específico para mitigar un error o informarlo de manera interna y externa sin exponer ningún dato sensible.
- Permite desacoplar el desarrollo.



# Amazon CloudWatch

- No utilizar los logs por defecto.
- Tratar de siempre separar los logs al nivel de proceso.
- Manejar la verbosidad con la que se escriben.
- Revisar que los tags de los logs esten acordes al proceso que los genera.
- Utilizar las alarmas y metricas para detectar gastos innecesarios.
- Podemos hacer consultas de los logs utilizando CloudWatch Insight.



# Amazon CloudTrail

- Procesar todos los logs provenientes de:
  - API calls a KMS / CloudHSM
  - S3 calls



# Consideraciones Generales

# Consideraciones Generales

- Los servicios mostrados son los mínimos que necesitamos para construir un data lake.
- Siempre es recomendable acompañar el despliegue con alguna herramienta de automatización (CDK + GitHub Action = ❤️ )
- Lo más importante siempre es acompañar el caso de uso de la manera clara y simple.

# Q&A

Alvaro Garcia



¡ Muchas Gracias !

