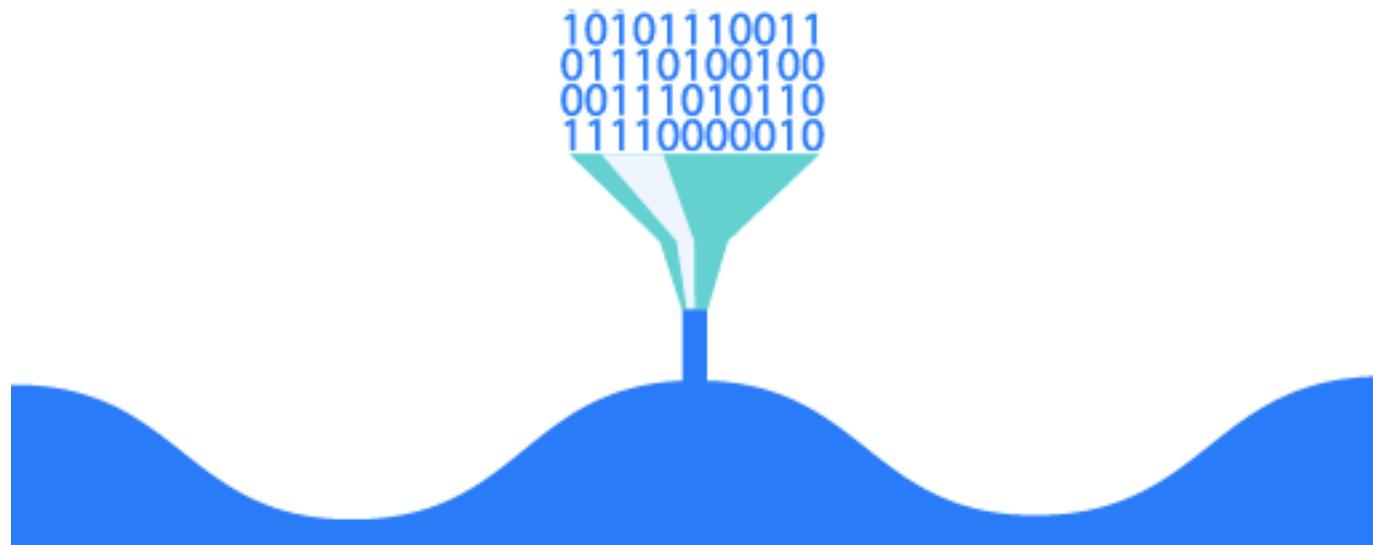


Lagos de datos

Data Lakes



Data Lakes

¿Por qué aparecen?

Data Lakes

Poder de la tecnología Big Data

+

Agilidad del ‘self-service’

Data Lake

“If you think of a datamart as a store of bottled water - cleansed and packaged and structured for easy consumption - the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the data lake can come to examine, dive in, or take samples.”

James Dixon, CTO, Pentaho

Data Lake

- Datos en su formato original
- Utilizados por diferentes tipos de usuarios

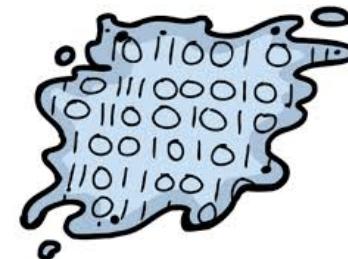


Etapas de madurez

- Data puddle (o charco de datos)
- Data pond (estanque de datos)
- Data lake
- Data ocean

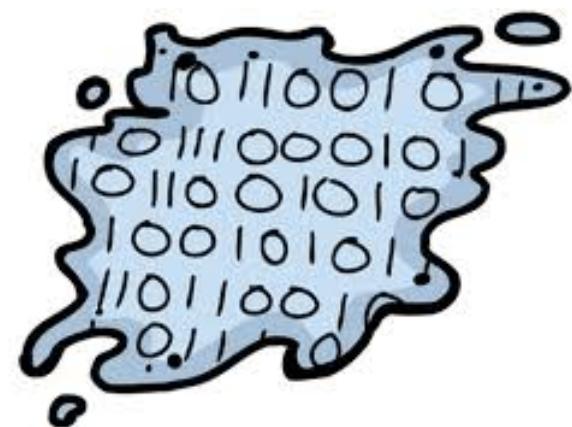
Data puddle

- Primer paso al adoptar tecnologías Big Data
- Es como un *datamart* con un único propósito o para un solo proyecto
- Usa tecnologías *big data*
- Datos son conocidos y entendidos



Data puddle

- Orientados a equipos pequeños
- Construidos en la nube
- Para proyectos muy específicos
- Alta participación de TI



Data pond

- Colección de *data puddles*
- Descarga de un DW
- Requiere participación de TI
- Sólo los datos que el proyecto requiera



Data Lake

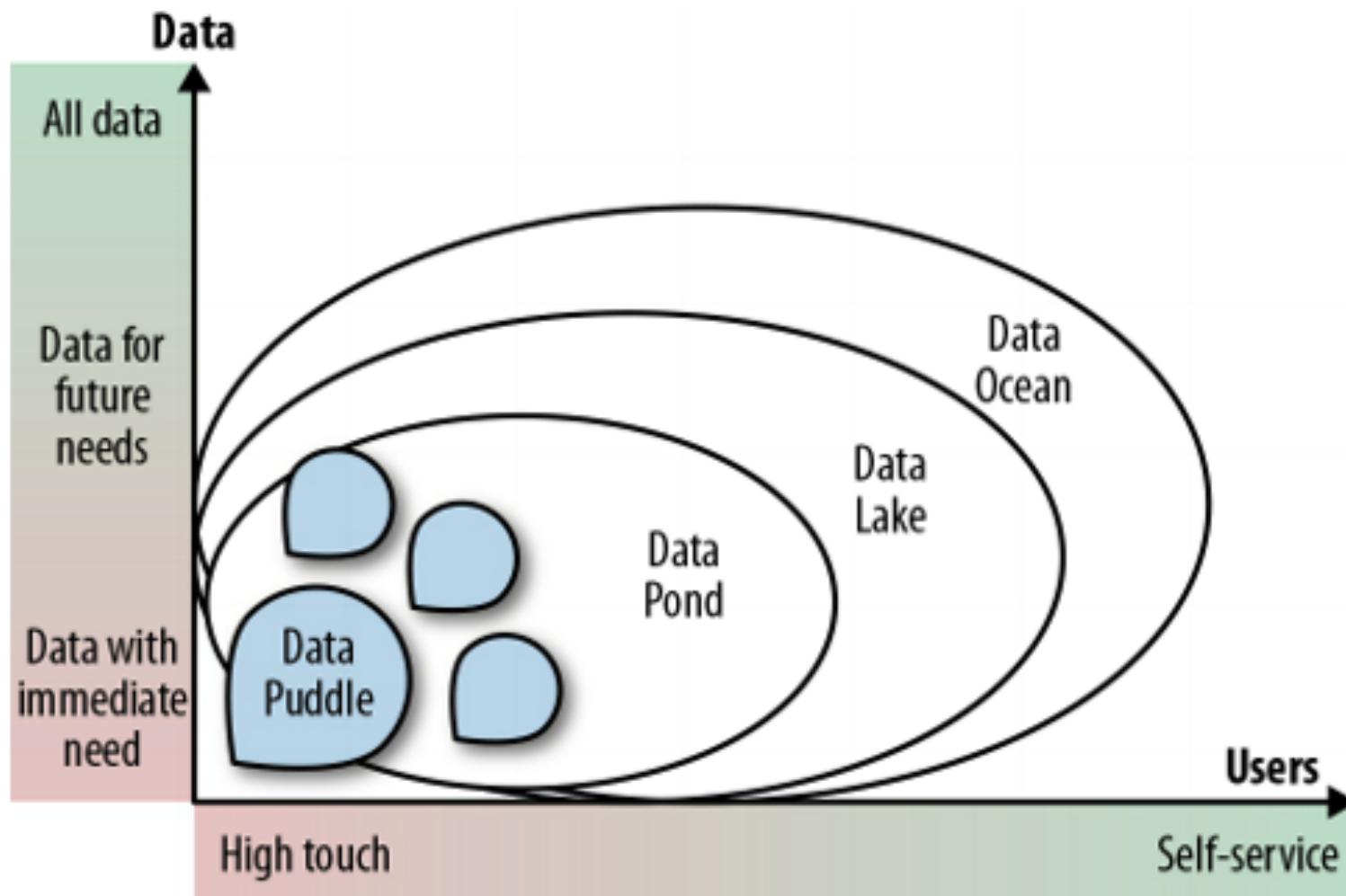
- Apoya el self-service para usuarios de negocio
- Contiene mayor cantidad de datos
- Baja participación de TI
- Análisis ad-hoc



Data Ocean

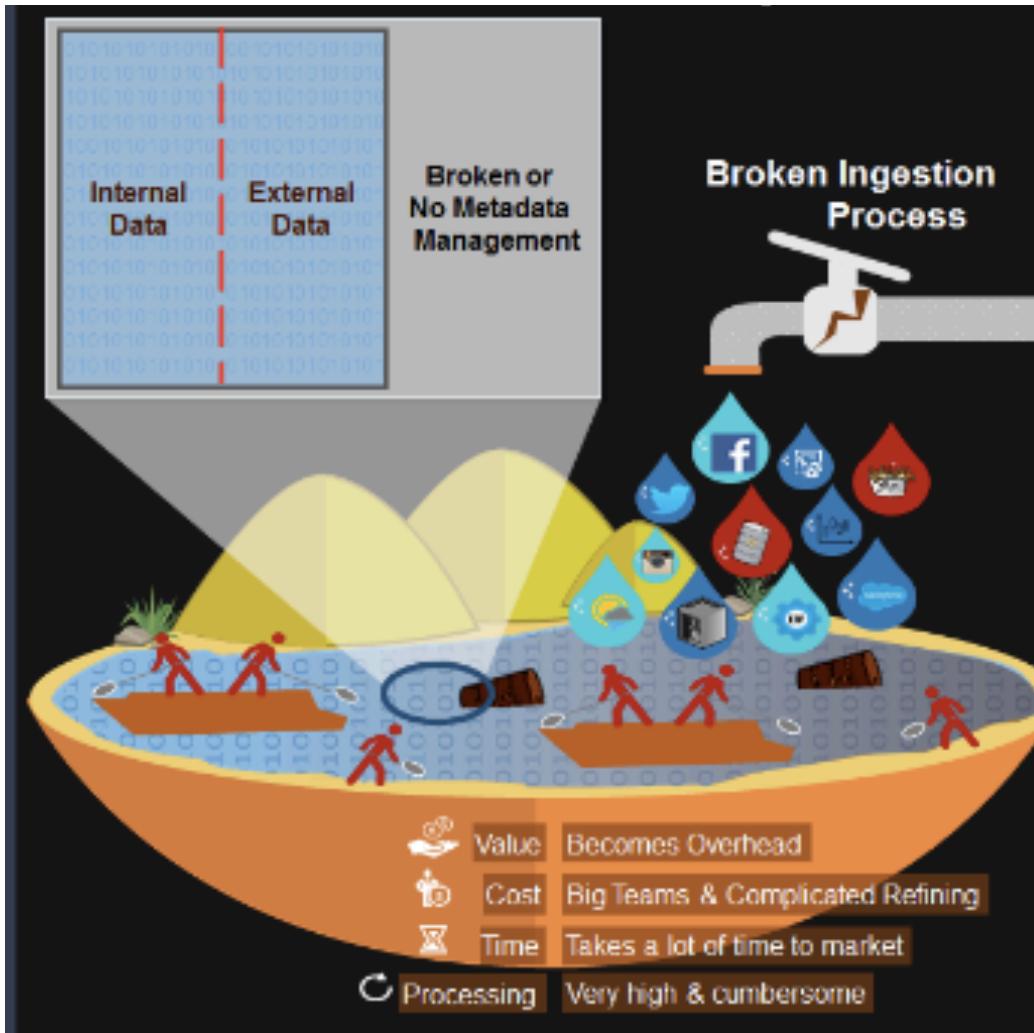
- Expande el auto servicio y la toma de decisiones basada en datos a toda la empresa
- No importa dónde estén los datos

Etapas de madurez





Data Swamp



Data Swamp

- Pantano de datos
- Un *data pond* con tantos datos como un *data lake* pero que no atrae a los usuarios.
- Datos no documentados

Diseño y construcción del lago de datos

Proceso

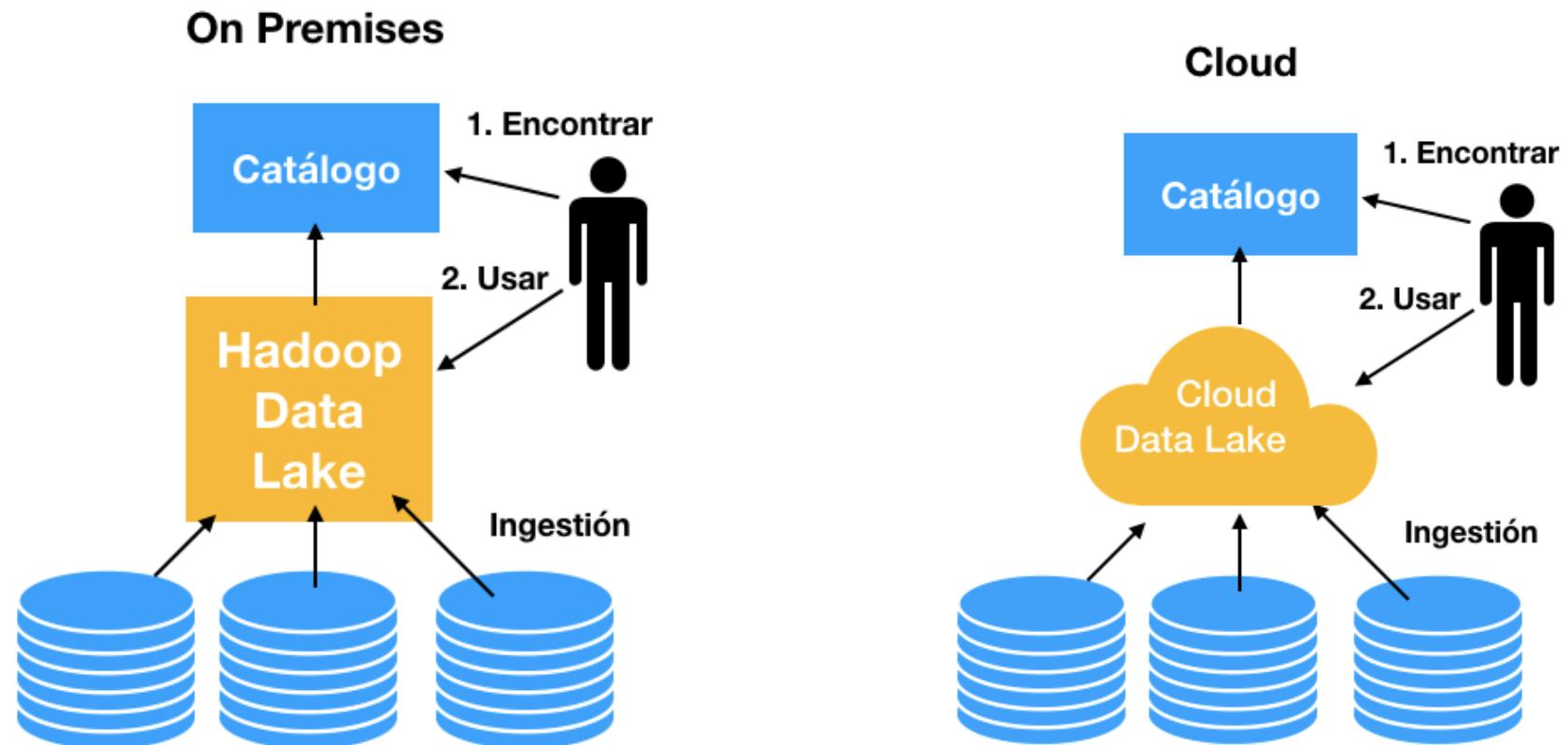
1. Poner en marcha la infraestructura
2. Organizar el *data lake*
3. Configurar el *data lake* para auto servicio
4. Abrir el *data lake* a los usuarios

1. La infraestructura

- Diferentes arquitecturas
- Inicialmente data lakes basados en Hadoop
- Data lakes 100% en la nube
- Enfoques híbridos
- Data lake lógico

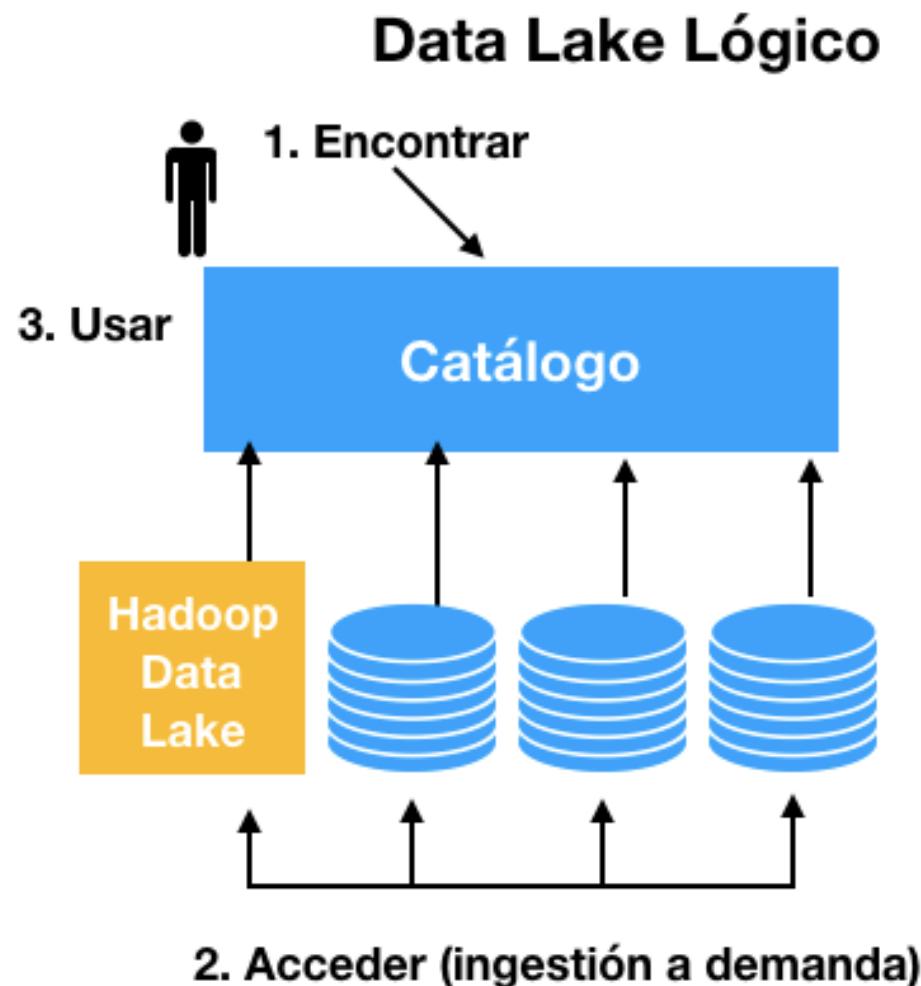
1. La infraestructura

Arquitecturas



1. La infraestructura

Arquitecturas



1. La infraestructura

El data lake lógico

- Alternativa al data lake centralizado
- Los datos están disponibles a través de un catálogo centralizado

2. Organizar el data lake

- Organizado por zonas
 - Landing zone*
 - Zona de producción (*gold zone*)
 - Zona de trabajo (*dev zone*)
 - organizada por proyecto, tema, etc.
 - Zona sensible

2. Organizar el data lake

- **Landing zone**
 - Zona de *staging*
 - Aquí se cargan los datos de fuentes externas
 - Estructura de folders que refleja su origen
 - /landing/twitter*
 - /landing/DW/DimProducto*

2. Organizar el data lake

- **Zona de producción (*gold zone*)**
 - Copia de la zona de landing pero con datos limpios, procesados, enriquecidos
 - Múltiples versiones de un mismo dato, según uso
 - Estructura de folder por sistema fuente

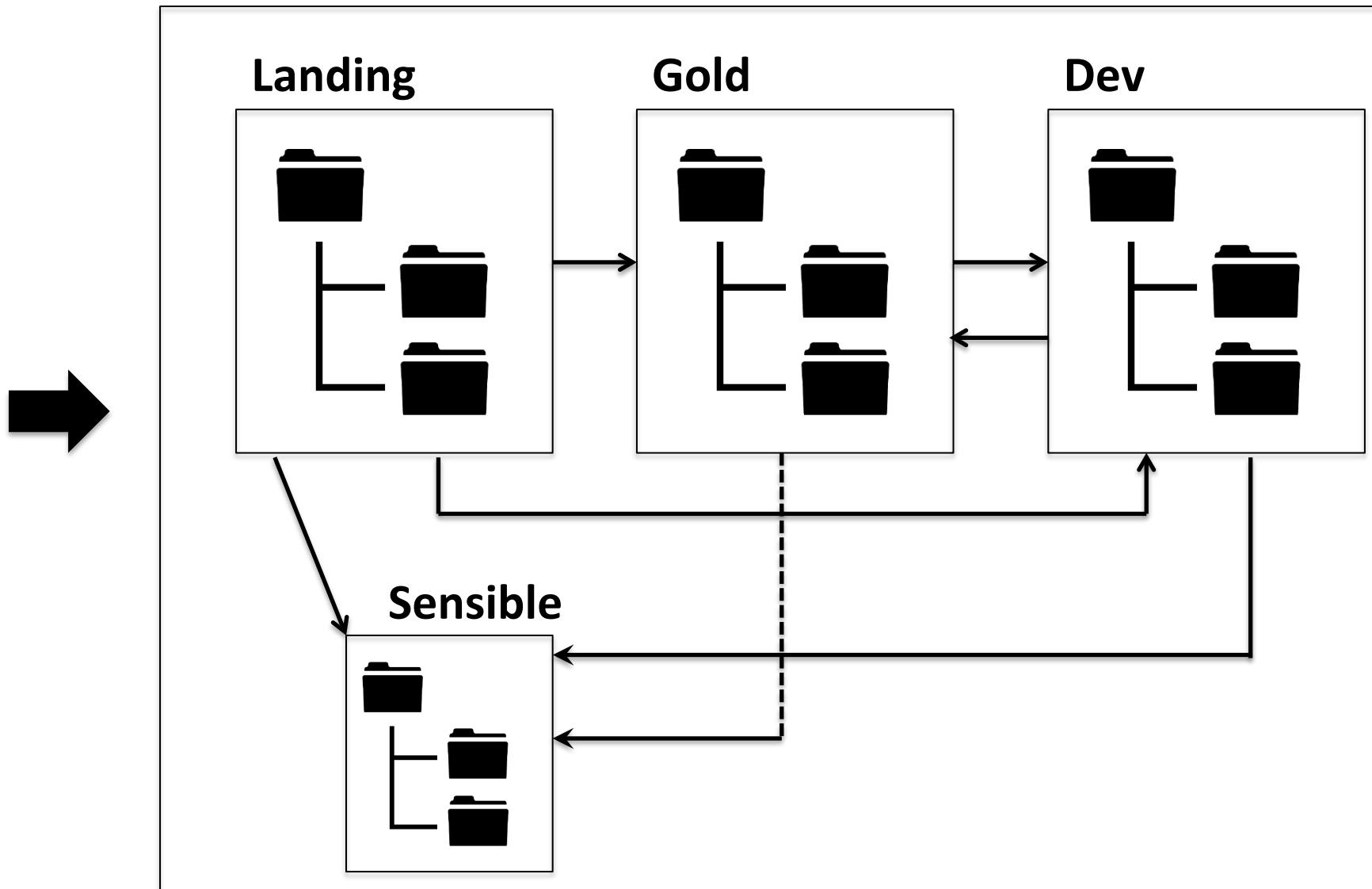
2. Organizar el data lake

- **Zona de trabajo (*dev zone*)**
 - Estructura refleja la estructura organizacional del negocio
 - Folders por proyecto, subfolders para reflejar detalles del proyecto
/Proyectos/RetencionClientes

2. Organizar el data lake

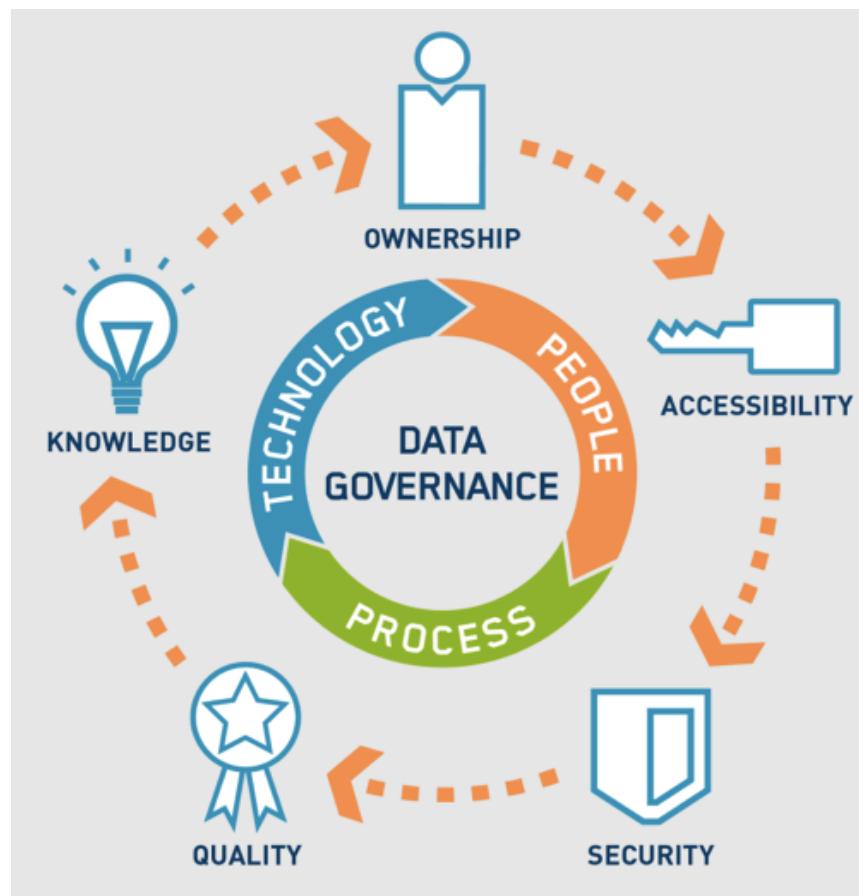
- **Zona sensible**
 - Mantener archivos con datos sensibles
 - Encriptados
 - Sólo personal autorizado (RH, data stewards)
 - Si los datos son requeridos para un proyecto ➔ anonimizarlos

2. Organizar el data lake



2. Organizar el data lake

La gobernanza de los datos



La gobernanza de los datos

Zona Sensible

- Data stewards
- Alta gobernanza
- Acceso restringido

Landing Zone

- Ingenieros de datos
- Gobernanza mínima
- No debe contener datos sensibles

Zona de trabajo

- Científicos de datos
- Gobernanza mínima
- No debe contener datos sensible

Gold Zone

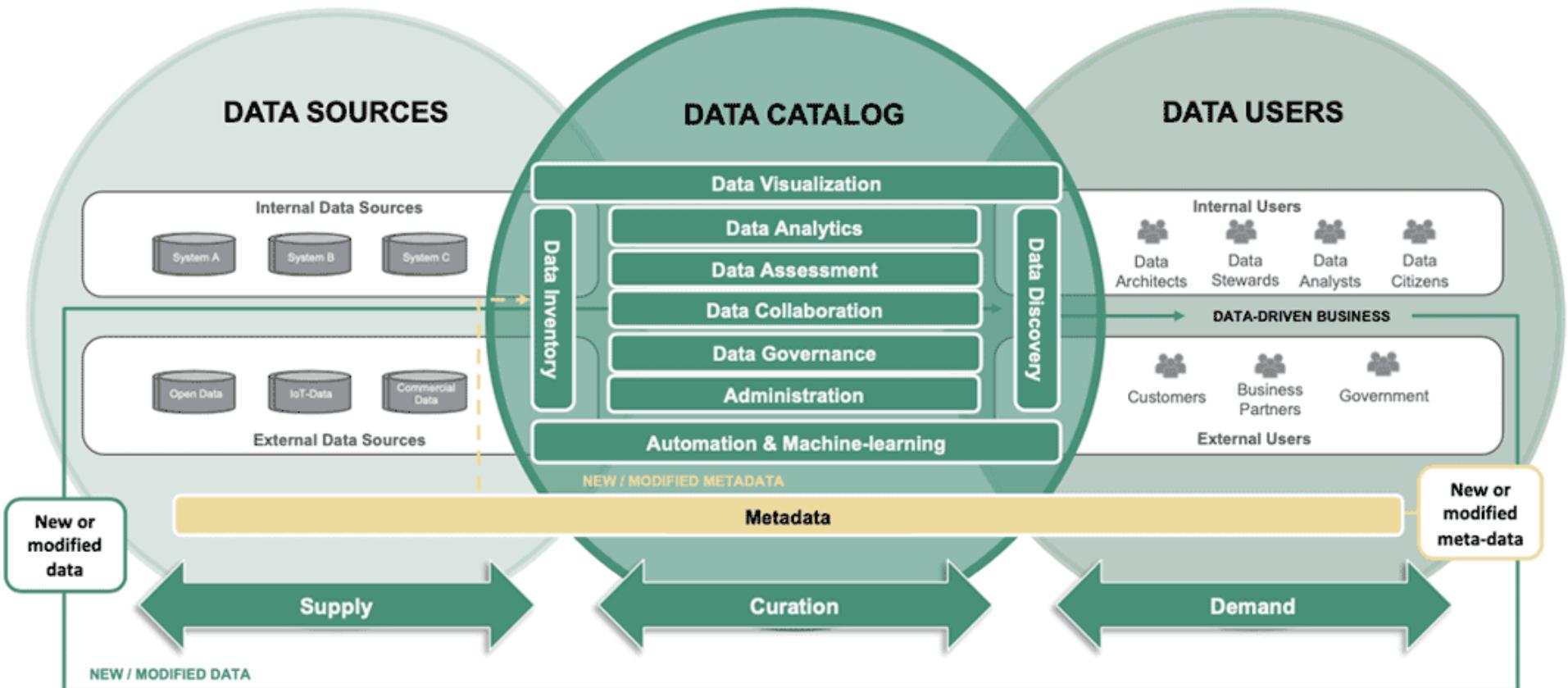
- Científicos de datos, analistas de negocio
- Alta gobernanza
- Datos curados, linaje y calidad de los dato

3. Configurar el *data lake* para auto servicio

- Tomar en cuenta el proceso que siguen los usuarios (futuros) en su trabajo



¿Cómo encontrar la información?



Data Catalog as an Integrated Platform for Bringing Data Supply and Demand Together

El catálogo

- La estructura de directorios no ofrece capacidad de búsqueda
- Los catálogos de datos resuelven el problema de la documentación de los datos
 - ¿ Cuál es el contenido de un archivo?
 - ¿De dónde proviene?

El catálogo

- Facilitar proceso de etiquetar datos
- Automatizar el proceso de catalogado

Metadata

- **Técnica**
 - Nombres de tablas, columnas, descripciones
 - Tipos de datos, tamaños de un campo, etc.
- **No técnica**
 - Reglas de negocio
 - Glosarios
 - Taxonomías
 - Ontologías

Metadata

- Metadata técnica
- Ejemplo

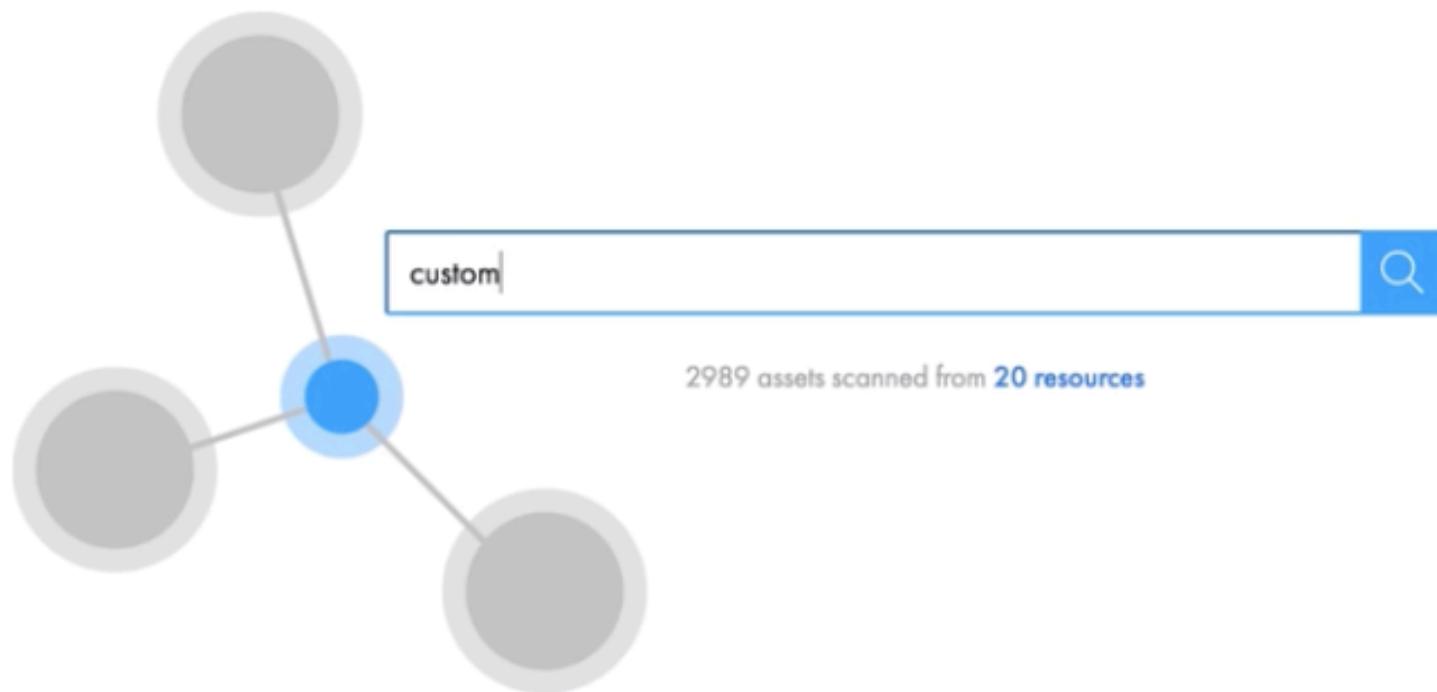
IdProducto	Año	Trim1	Trim2	Trim3	Trim4	Ene	Feb
ABC	2019	300	500	200	500	100	150
ABC	2018	200	400	100	700	50	75
XYZ	2019	600	500	800	200	200	300
...

Metadata

- **Metadata técnica**
- Ejemplo: metadata más compleja

IdProducto	Año	Período	Ventas
ABC	2019	Trim1	300
ABC	2018	Trim2	400
XYZ	2019	Trim3	800
...

El catálogo



El catálogo

Filter by

Search Results (1 - 20 of 113)

Show Details ↕

Resource Name

- All
- Hermes (26)
- BG_DEFAULT_RESOURCE (21)
- ACME_HDFS (17)
- HDFS_Resource (14)
- CRM (11)

Add Show All

Asset Type

- All
- Column (34)
- JSON Field (28)
- Term (20)
- Table (13)
- Data Domain (4)

Add Show All

Resource Type

- All
- Oracle (37)
- HDFS (31)
- Business Glossary (21)
- Hive (11)
- Tableau Server (8)

Add Show All

Resource	Asset Type	Resource Type	Last Updated	Business Description
Hermes	Resource	Oracle	Apr 05, 2017 05:11am	
	Topic: Customer Debt			Business Description: The primary enterprise dataset for customer information, com...
Customer	Term	Business Glossary	Apr 03, 2017 09:06pm	
	A party that receives or consumes products (goods or services) and has the ability to choose between different products and suppliers.			
acme_hive_customer	Table	Hive	Apr 27, 2017 09:47pm	Size: N/A
	Business Terms: Customer			Topic: Customer Segmentation
	Tag: Customer Churn Analysis			
CUST_TIER	Tier	Oracle	Mar 09, 2017 04:27pm	
	Asset Type: Column			
	CustomerTier LastName City FirstName			
	Inferred and accepted data domains.: CustomerTier			
CUST_NAME	Name	Oracle	Jan 11, 2017 03:57pm	
	Asset Type: Column			
	CustomerNames LastName City FirstName			

Algunas herramientas

- [Alation](#)
- [Informatica Data Catalog](#)
- [Apache Atlas](#)
- [AWS Glue](#)
- [IBM Watson Knowledge Catalog](#)

Acceso a los datos

- Considerar las características diferenciadoras del data lake

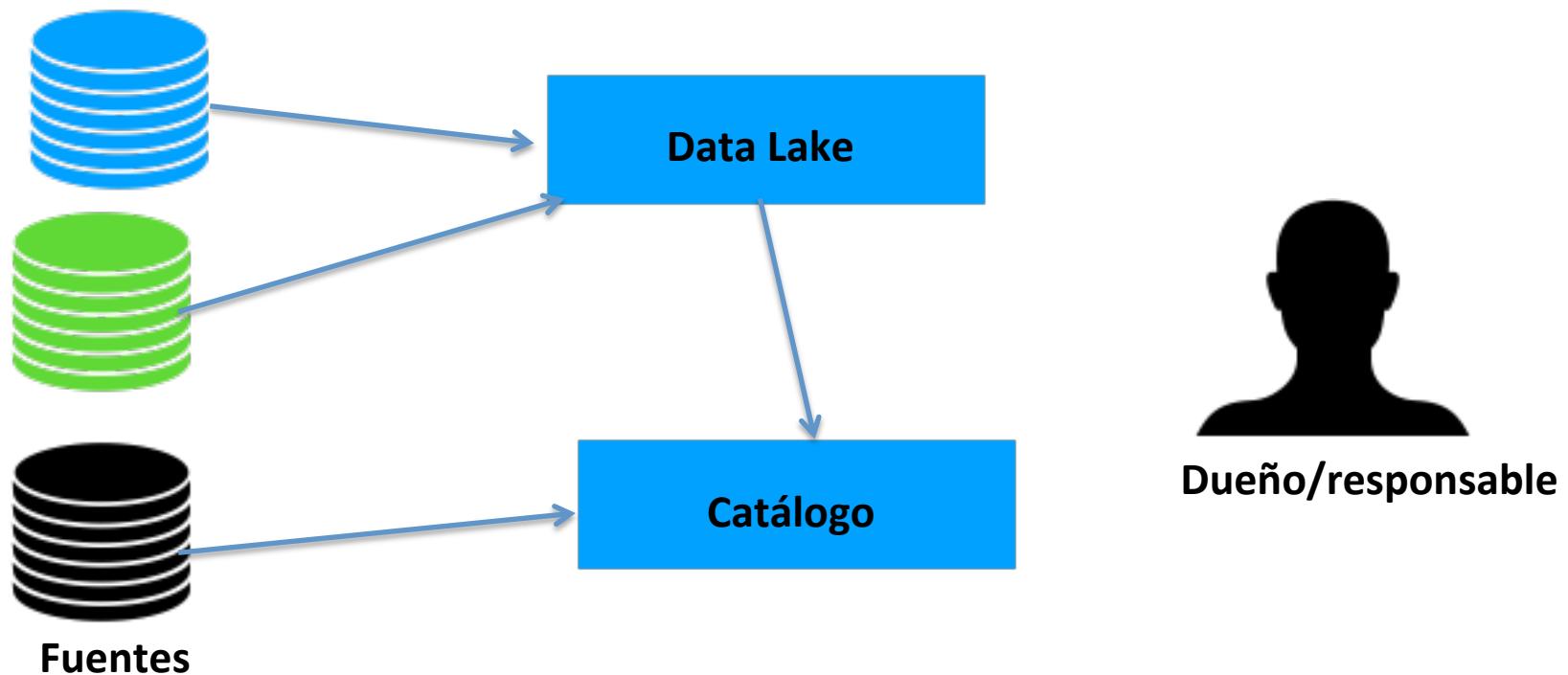


Acceso a los datos

- Diferentes enfoques:
 - Autorización por directorios
 - Políticas basadas en etiquetas
 - Gestión por autoservicio

Acceso a los datos

- El “dueño” o responsable de los datos publica en el catálogo.



Acceso a los datos

- Un analista los encuentra y solicita acceso.

