

Hola, bbdd



red.es



“El FSE invierte en tu futuro”
Fondo Social Europeo



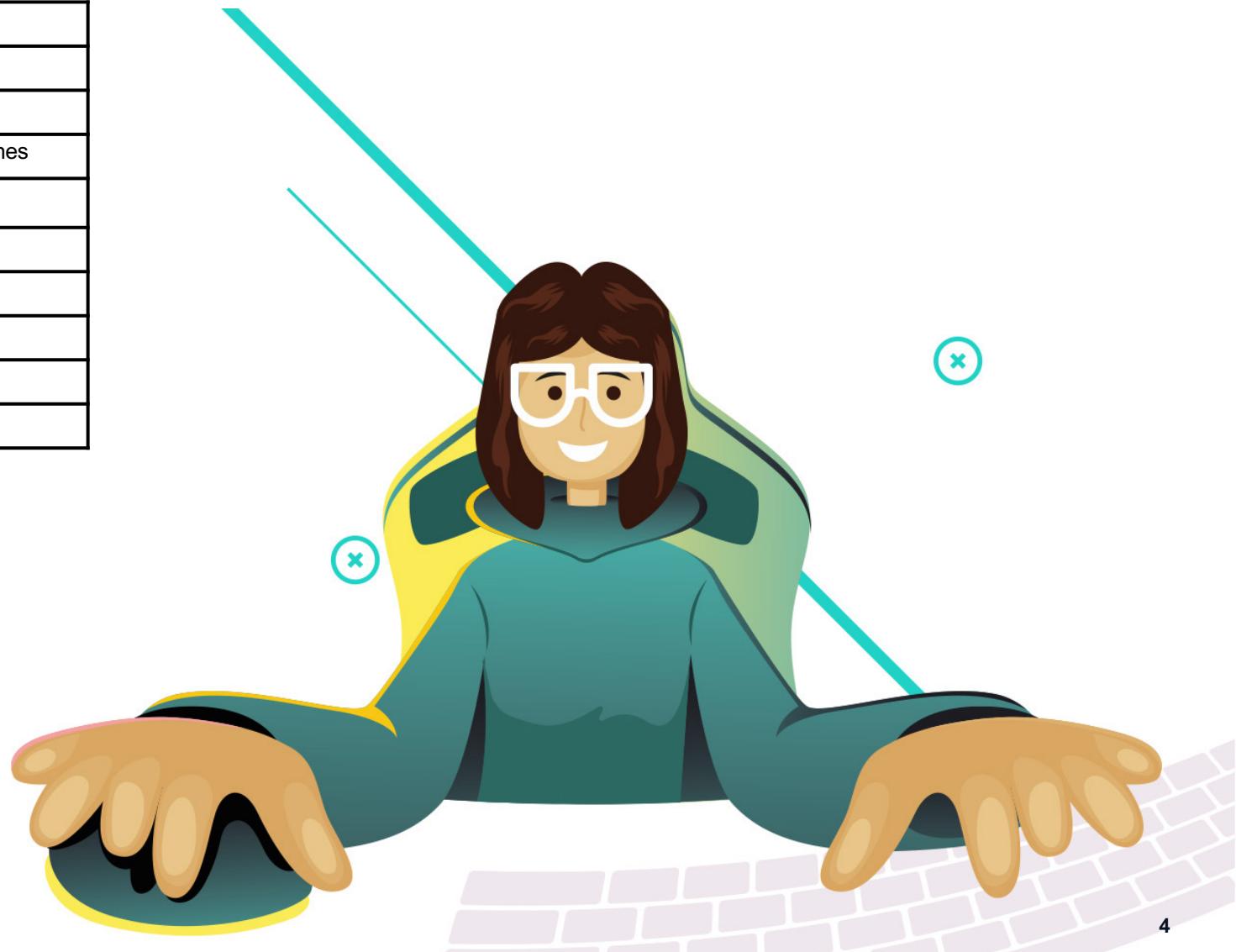
Índice

1. Análisis inicial de una base de datos
2. RECAP
3. ¿Big Data?
4. Ingesta y tratamiento

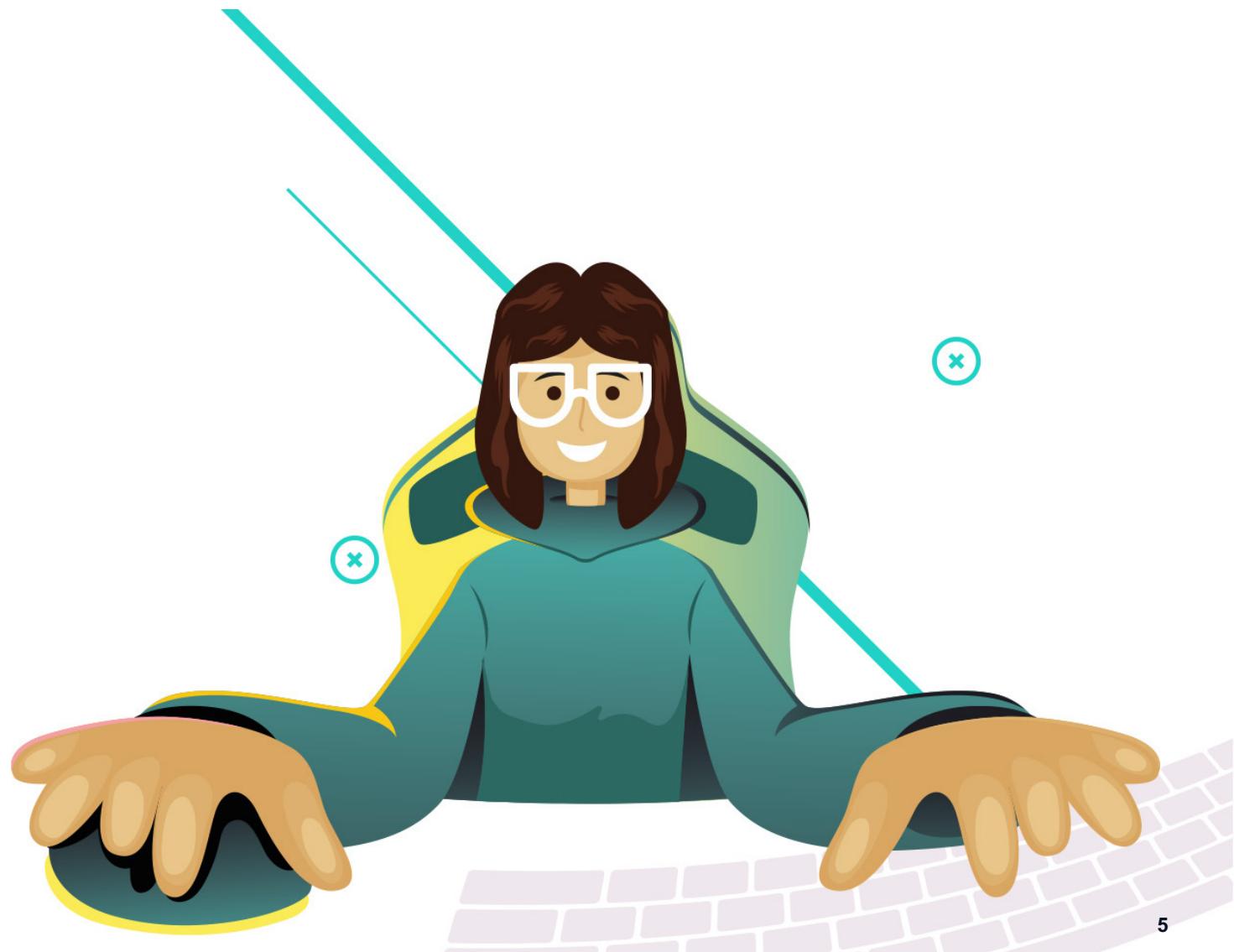
1. Análisis inicial de una bbdd[⊗]

¿Qué voy a aprender ahora?

Datos	1	¿Qué es un dato?
	2	Clases de datos
	3	Fuentes de datos
	4	¿Qué es imputar datos? Asunciones
Bases de datos	5	BDD vs Excel
	6	Bases de datos relacionales
	7	Primary keys y foreign keys
	8	De Excel a base de datos
	9	Intro a cruce de BDD - joins



IDENTIFICANDO DATOS



Identificando datos



¿QUÉ ES UN DATO?

Texto o string

Nombre del producto

Fecha y/u hora

Fecha de compra

Número o integer

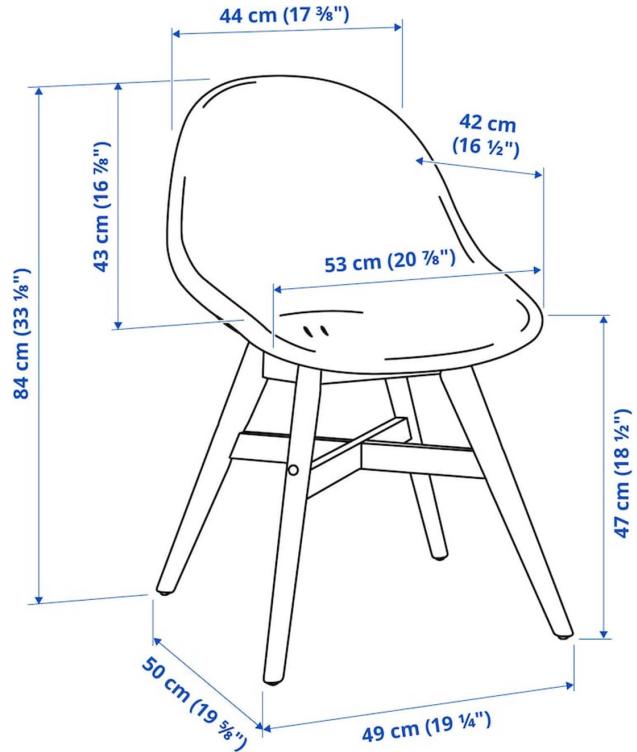
Cantidad que hay en stock

Boolean

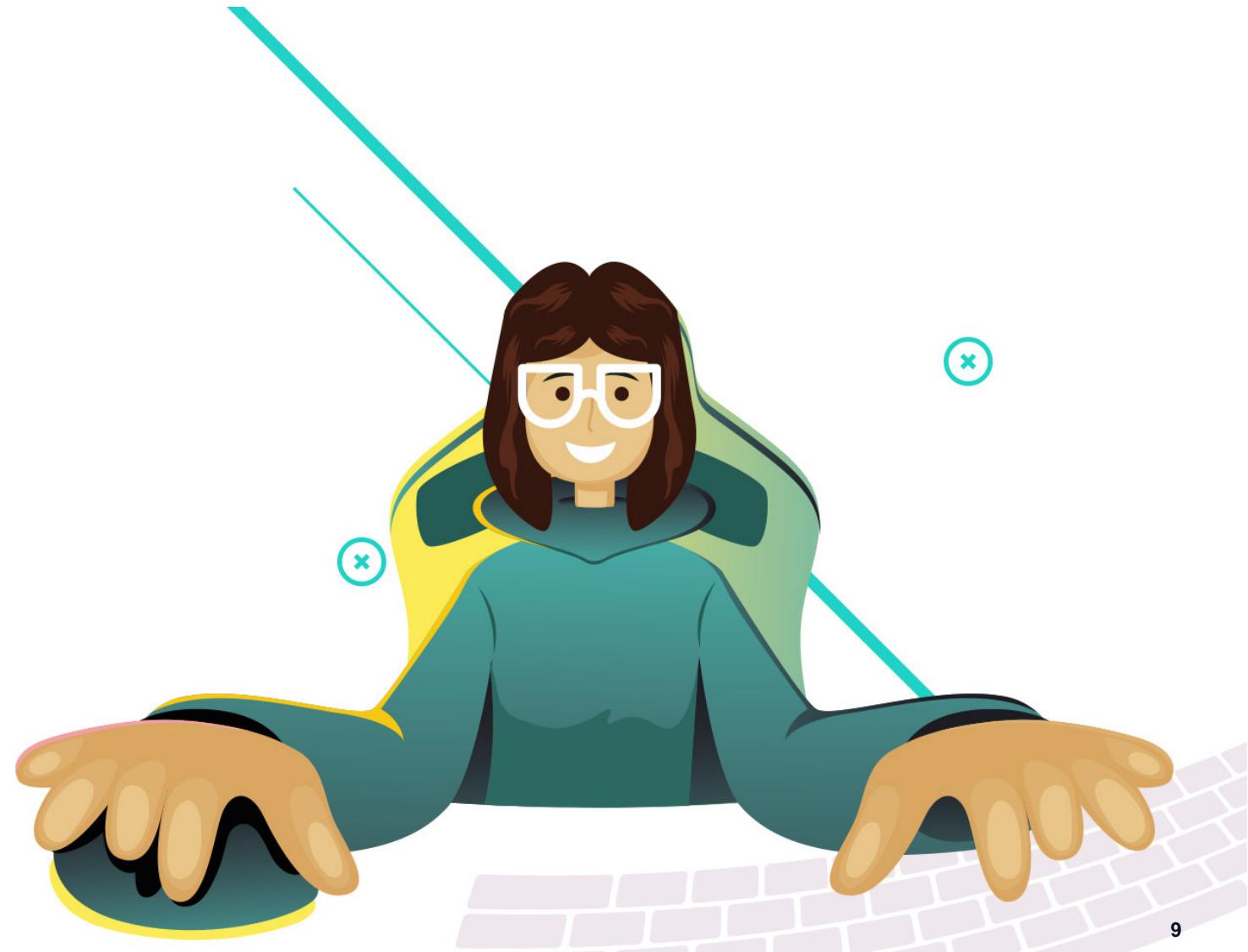
Sí/No, Verdadero/Falso, 1/0

**LAS FUENTES DE DATOS
PUEDEN ESTAR EN
CUALQUIER SITIO,
SÓLO HAY QUE SABER MIRAR**

Fuentes de datos



¿QUÉ ES UNA BASE DE DATOS?



[¿Qué es una base de datos?](#)

**Un sistema de información pragmático para
almacenar y organizar datos a los que vamos
a acceder de manera recurrente**



¿Qué es una base de datos?

¿ES EXCEL UNA BASE DE DATOS?

Excel cumple las premisas que acabamos de establecer

- Tiene un orden (páginas, columnas, filas...)
- Podemos acceder al archivo siempre que queramos
- Podemos analizar los datos

The screenshot shows a Microsoft Excel spreadsheet titled "AutoSave" and "Eddie Wang EW". The spreadsheet contains a table of employee data with the following columns: Name, Title, Department, Office Address, Start Date, Email, Phone, Birthday, and Image. The data includes various roles like Research Nurse, Software Consultant, and Geologist I, along with their respective department, office address, start date, email, phone number, birthday, and a placeholder image URL. The table has 28 rows of data, starting from row 1 and ending at row 28. The "Image" column contains URLs such as https://randomuser.me/api/portraits/men/89.jpg, https://randomuser.me/api/portraits/women/70.jpg, https://randomuser.me/api/portraits/women/45.jpg, https://randomuser.me/api/portraits/men/8.jpg, https://randomuser.me/api/portraits/women/17.jpg, https://randomuser.me/api/portraits/women/90.jpg, https://randomuser.me/api/portraits/men/26.jpg, https://randomuser.me/api/portraits/women/10.jpg, https://randomuser.me/api/portraits/women/7.jpg, https://randomuser.me/api/portraits/men/64.jpg, https://randomuser.me/api/portraits/women/40.jpg, https://randomuser.me/api/portraits/men/80.jpg, https://randomuser.me/api/portraits/men/10.jpg, https://randomuser.me/api/portraits/men/54.jpg, https://randomuser.me/api/portraits/women/19.jpg, https://randomuser.me/api/portraits/women/51.jpg, https://randomuser.me/api/portraits/men/28.jpg, https://randomuser.me/api/portraits/women/9.jpg, https://randomuser.me/api/portraits/men/76.jpg, https://randomuser.me/api/portraits/men/23.jpg, https://randomuser.me/api/portraits/women/28.jpg, https://randomuser.me/api/portraits/men/9.jpg, https://randomuser.me/api/portraits/men/52.jpg, and https://randomuser.me/api/portraits/men/48.jpg.

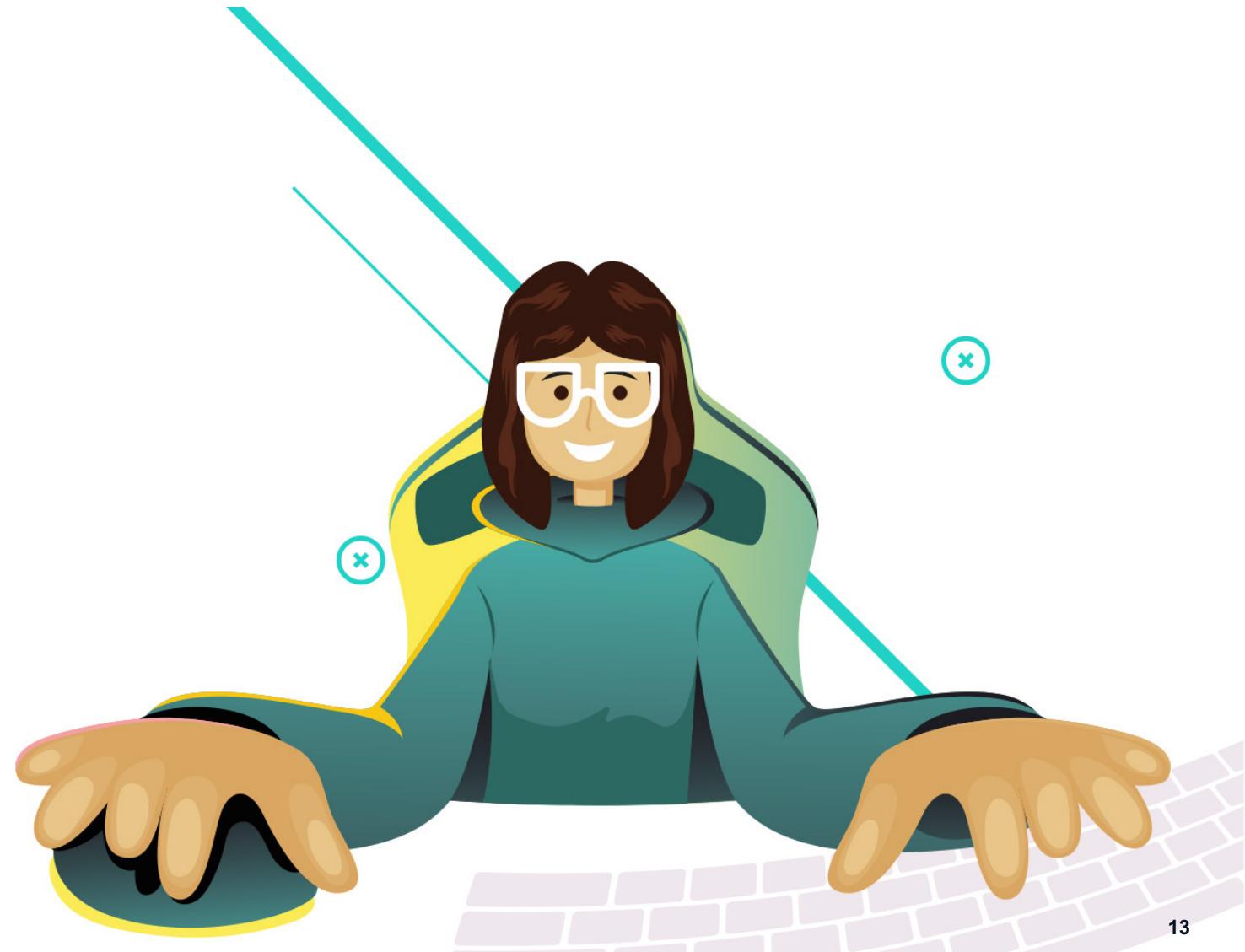
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Title	Department	Office Address	Start Date	Email	Phone	Birthday	Image				
2	Emmanuelle Garner	Research Nurse	Services	Bellevue, WA, 98004	5/18/2018	egarner@alexa.com	942-493-2803	11/18/1994	https://randomuser.me/api/portraits/men/89.jpg				
3	Nicolina Taff	Health Coach IV	Research and Dev	Bellevue, WA, 98004	8/7/2018	ntaff@jobs.org	776-430-4577	10/2/1995	https://randomuser.me/api/portraits/women/70.jpg				
4	Adrienne Leyborne	Recruiter	Legal	Bellevue, WA, 98004	12/28/2017	aleyborne@wsq.com	801-682-6424	7/6/1976	https://randomuser.me/api/portraits/women/45.jpg				
5	Tanner Casino	GIS Technical Architect	Business Development	Bellevue, WA, 98004	7/5/2016	tcasino@nordic.ru	293-662-7086	1/30/1967	https://randomuser.me/api/portraits/men/8.jpg				
6	Minnie Quilliam	Developer I	Legal	Bellevue, WA, 98004	11/4/2017	mquilliam@vulnart.com	853-234-0775	7/23/1987	https://randomuser.me/api/portraits/women/17.jpg				
7	Sawyer Livingstone	Administrative Officer	Business Development	Bellevue, WA, 98004	12/1/2017	slivingst@guardian.co.uk	400-329-3261	2/9/1979	https://randomuser.me/api/portraits/men/90.jpg				
8	Ursula Washington	Software Design manager	Human Resources	Seattle, WA, 98101	7/20/2017	uwashington@guardian.co.uk	554-303-3015	9/27/1975	https://randomuser.me/api/portraits/women/26.jpg				
9	Mora Colding	Software Consultant	Human Resources	Bellevue, WA, 98004	8/25/2017	mcolding7@datashow.com	591-828-4654	6/15/1970	https://randomuser.me/api/portraits/women/10.jpg				
10	Lorriy Lynell	Junior Executive	Engineering	Bellevue, WA, 98004	1/4/2017	lynell7@smugmug.com	857-813-2868	8/27/1967	https://randomuser.me/api/portraits/women/7.jpg				
11	Bartel Lamps	VP Sales	Marketing	Bellevue, WA, 98004	5/29/2017	hlamps@tipadevior.com	620-546-4759	9/3/1982	https://randomuser.me/api/portraits/women/64.jpg				
12	Alard Kobi	Geologist I	Sales	Bellevue, WA, 98004	3/13/2018	akobus@rambler.ru	304-195-3878	11/17/1976	https://randomuser.me/api/portraits/men/40.jpg				
13	Syvester Shrewsbury	Biostatistician IV	Legal	Bellevue, WA, 98004	7/11/2017	shreweb@tiny.cc	735-996-5448	12/10/1961	https://randomuser.me/api/portraits/men/80.jpg				
14	Brod Morsley	Research Associate	Business Development	Bellevue, WA, 98004	3/15/2019	bmorsley@tiny.cc	928-783-5846	8/28/1987	https://randomuser.me/api/portraits/men/10.jpg				
15	Costanza Fayer	Systems Administrator III	Services	Bellevue, WA, 98004	3/7/2017	cclayter@bigbang.com	285-245-3094	12/31/1983	https://randomuser.me/api/portraits/men/54.jpg				
16	Elspeth Schuch	Executive Secretary	Product	Seattle, WA, 98101	6/20/2017	elspethsch@bigbang.com	196-329-3775	3/4/1980	https://randomuser.me/api/portraits/women/21.jpg				
17	Hanni Lewellen	Senior Financial Analyst	Marketing	Seattle, WA, 98109	5/5/2017	hlewellen7@cybercure.org	555-680-1720	1/29/1980	https://randomuser.me/api/portraits/women/51.jpg				
18	Arlene Shoesmith	Paralegal	Legal	Seattle, WA, 98109	12/11/2016	eshoemsmith@wixpress.org	666-256-9424	5/22/1967	https://randomuser.me/api/portraits/women/43.jpg				
19	Ingamar D'Andrea	Developer IV	Research and Dev	Seattle, WA, 98109	4/5/2018	idandrea@ibm.com	554-819-6509	12/17/1979	https://randomuser.me/api/portraits/men/38.jpg				
20	Hedria Gut	Social Worker	Accounting	Seattle, WA, 98109	8/27/2017	hgut@infinity.com	185-927-0435	7/16/1976	https://randomuser.me/api/portraits/women/50.jpg				
21	Shelley Larchiere	Media Manager IV	Services	Seattle, WA, 98109	4/13/2019	slarchiere@va.gov	471-825-8911	10/14/1984	https://randomuser.me/api/portraits/women/22.jpg				
22	Zelma Baltoro	Chief Design Engineer	Product Management	Seattle, WA, 98109	12/25/2017	zbaltoro@zcombinator.com	443-596-4137	5/15/1976	https://randomuser.me/api/portraits/women/3.jpg				
23	Garey Pauter	Professor	Support	Seattle, WA, 98109	1/8/2018	gpauter@narod.ru	646-921-0513	10/25/1970	https://randomuser.me/api/portraits/men/76.jpg				
24	Leanne Auditor	Financial Representative IV	Business Development	Seattle, WA, 98109	10/20/2019	lrauditor@wixpress.org	597-888-1234	5/20/1980	https://randomuser.me/api/portraits/men/23.jpg				
25	Dunward Sharnoff	GIS Technical Architect	Sales	Seattle, WA, 98109	8/17/2018	dsharnoff@wixpress.com	255-547-0090	5/25/1973	https://randomuser.me/api/portraits/women/28.jpg				
26	Cyndy Vennar	Information Systems Manager	Accounting	Seattle, WA, 98109	10/25/2017	cvennar@dmz.org	733-905-2393	8/11/1997	https://randomuser.me/api/portraits/men/9.jpg				
27	Leonid Sinnock	Geologist IV	Support	Seattle, WA, 98109	4/11/2019	lsincock@vku.edu	428-564-7468	6/28/1983	https://randomuser.me/api/portraits/men/52.jpg				
28	Harland Cristofanini	Software Test Engineer II	Engineering	Seattle, WA, 98109	7/13/2016	hcristofanini@scicredaily.com	724-847-6747	6/15/1973	https://randomuser.me/api/portraits/men/48.jpg				

PERO TIENE GRANDES CARENCIAS EN OTROS ASPECTOS VITALES

- **Integridad y estructura:** es fácil cometer un error metiendo datos incorrectos/mal formateados
- **Almacenamiento y seguridad:** las reglas de acceso y edición son limitadas
- **Escalabilidad y eficiencia:** aunque son un buen punto de partida, pierden eficacia y rapidez a medida que crece el volumen de datos



LAS BASES DE DATOS RELACIONALES



¿Qué es una base de datos?

LAS BASES DE DATOS RELACIONALES AL RESCATE

Pero, ¿qué es una base de datos relacional?

Una base de datos relacional consiste en tablas con datos almacenados en filas y columnas **relacionadas entre sí**

Fila = Observación

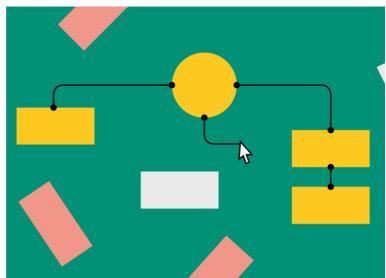
Celda = Valor

Columna = Variable

Nombre	Apellido	Nº Tel	NombreMascota	TipoMascota
Miguel	Lorenzo	689888777	Paco	Perro
Miguel	Lorenzo	689888777	Taco	Perro
Miguel	Lorenzo	689888777	Whiskas	Gato
José Carlos	Riesco	698777888	Romulo	Perro
María	Bortel	658475211	Klimt	Gato

¿Qué es una base de datos?

LAS BASES DE DATOS RELACIONALES AL RESCATE



- **Integridad y estructura:**
 - establece las relaciones entre tablas
 - valida los tipos de datos de las variables
- **Almacenamiento**
 - organiza los valores en tablas
 - construye índices de búsqueda
 - hace backups regularmente
- **Seguridad**
 - otorga permisos de acceso
 - establece credenciales de login
 - mantiene un log para auditoría
- **Escalabilidad**
 - capaz de guardar millones de datos
 - mantiene los niveles de accesibilidad y velocidad
 - análisis fácilmente replicables

Health policy

Alex Hern UK technology editor

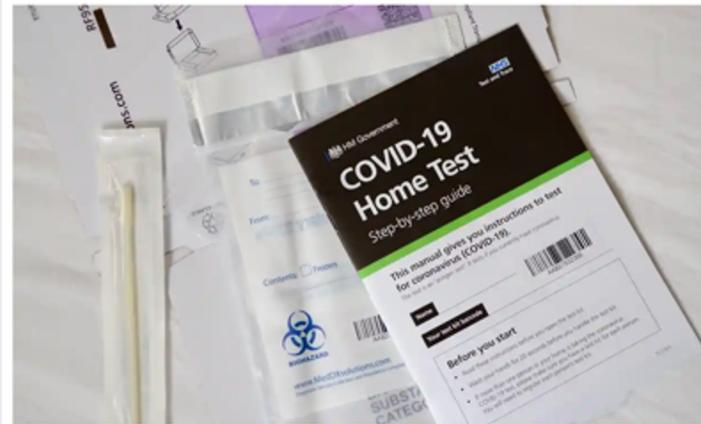
@alexhern
Tue 6 Oct 2020 08.21 BST



Covid: how Excel may have caused loss of 16,000 test results in England

Public Health England data error blamed on limitations of Microsoft spreadsheet

- Coronavirus - latest updates
- See all our coronavirus coverage



▲ More than 50,000 potentially infectious people may have been missed by contact tracers after 15,841 positive tests were left off the daily figures. Photograph: Simon Leigh/Alamy

A million-row limit on Microsoft's Excel spreadsheet software may have led to Public Health England misplacing nearly 16,000 Covid test results, it is understood.

The data error, which led to [15,841 positive tests being left off the official daily figures](#), means than 50,000 potentially infectious people may have been missed by contact tracers and not told to self-isolate.

¿Qué es una base de datos?

NORMALIZACIÓN

- **Integridad:**
 - cada tabla tiene una primary key (PK)
 - no hay PKs null
- **Eficiencia**
 - si quisiéramos añadir la dirección de Miguel, tendríamos que introducirla 1 vez y no 3

			Nombre	Apellido	NoTel	NombreMascota	TipoMascota	
IdDueño	Nombre	Apellido	NoTel			IdDueño	NombreMascota	TipoAnimal
PK1	Miguel	Lorenzo	689888777			Romulo	Pa co	Perro
				Maria			Gato	
2	José Carlos	Riesco	698777888	Bortel	658475211	Klimt	1 Ta co	Perro
3	María	Bortel	658475211				1 Whiskas	Gato
							2 Romulo	Perro
							3 Klimt	Gato

EL LENGUAJE DE LAS BASES DE DATOS (relacionales) ES SQL

- Ventajas de SQL (structured query language) vs Excel
- Analizar en Excel es seguir una receta (filtros, eliminación de columnas, eliminación de filas)
- Esto puede causar errores (ups, ¿he borrado esa fila?) y es difícil de replicar
- **SQL separa el análisis de los datos.**
- Con SQL *declaramos* lo que queremos, hacemos peticiones (querys)

NAME	TYPE	WEIGHT
bulbasaur	grass	15
charmander	fire	19
squirtle	water	20
pikachu	electric	13
oddish	grass	12
snorlax	normal	1014
mewtwo	psychic	269

EL LENGUAJE DE LAS BASES DE DATOS (relacionales) ES SQL

NAME	TYPE	WEIGHT
bulbasaur	grass	15
charmander	fire	19
squirtle	water	20
pikachu	electric	13
oddish	grass	12
snorlax	normal	1014
mewtwo	psychic	269

```
SELECT name, type  
FROM pokemon  
WHERE type = 'grass';
```

NAME	TYPE
bulbasaur	grass
oddish	grass

¿Qué es una base de datos?

GESTORES BASES DE DATOS RELACIONALES MÁS USADOS (I)

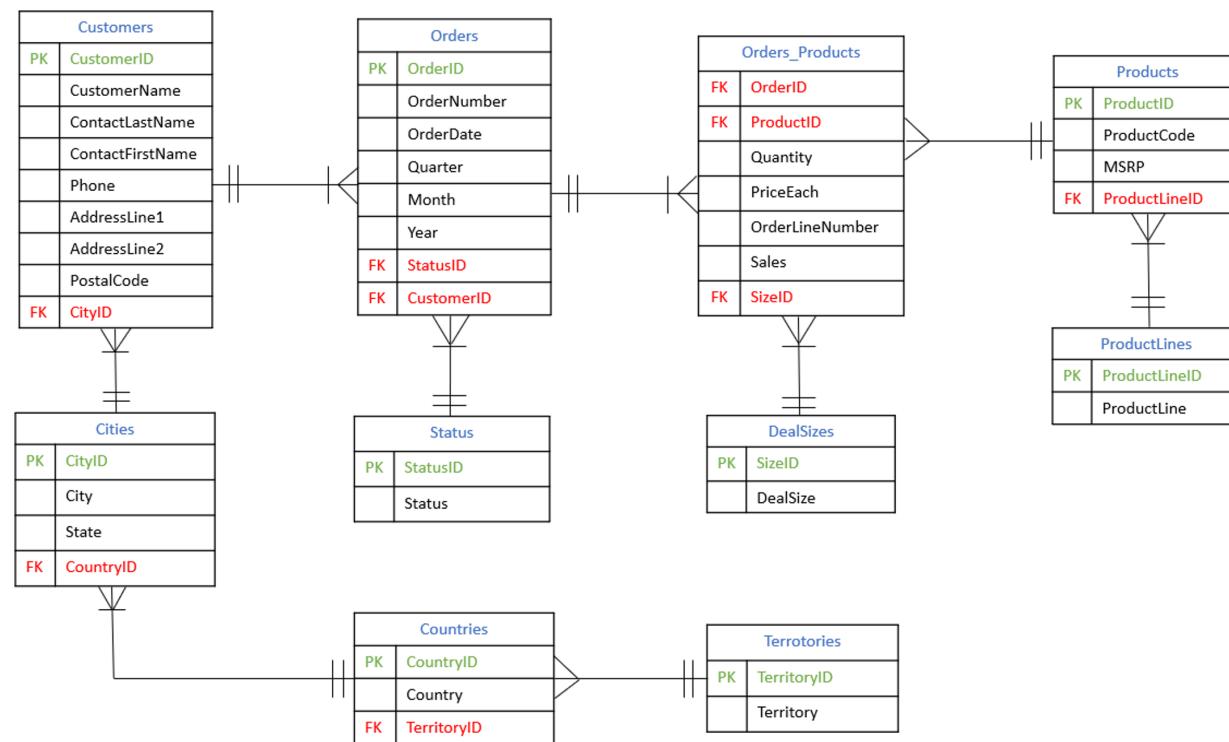


PRÁCTICA!

Con el Excel sales_data_sample:

- hacer un análisis descriptivo de las **variables** incluidas
- ¿qué nos gustaría saber con este dataset?
- echa un vistazo a los **tipos de datos** y formatea los que hagan falta
- ¿echas en falta alguna variable? ¿qué datos nos podrían venir bien?
- ¿podríamos utilizar alguna otra fuente de datos externa para enriquecer los datos?
- divide el dataset en **tablas** para crear una pseudo **base de datos relational**: ¿qué posibles tablas potenciales tendríamos?

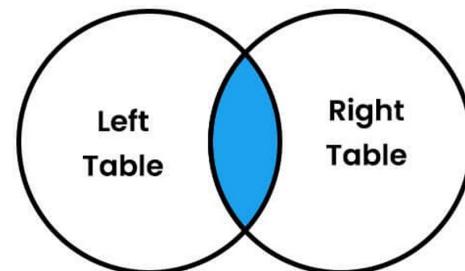
UN POSIBLE ESQUEMA FINAL



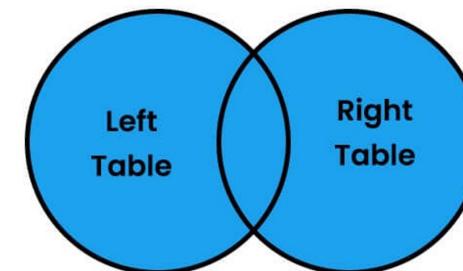
Ahora que tenemos nuestros datos organizados en tablas dentro de una base de datos relacional, ¿cómo podemos analizarlos conjuntamente?

Mediante **joins** -> combinan datos de dos o más tablas en una base de datos relacional.

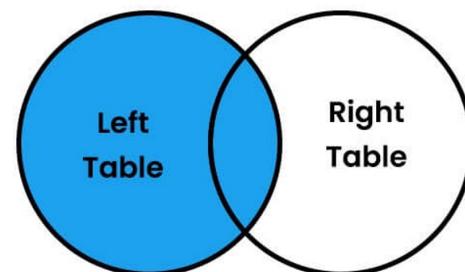
INNER JOIN



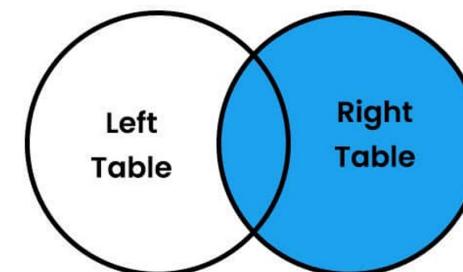
FULL JOIN



LEFT JOIN



RIGHT JOIN



HOLA, BBDD

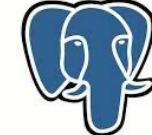
2. RECAP



ORACLE

MySQL

Microsoft
SQL Server™


PostgreSQL

 mongoDB®

 redis

 IBM
DB2


elasticsearch

3. ¿BIG DATA?



'Bird view'



¿Hoy qué vamos a empezar a hacer?



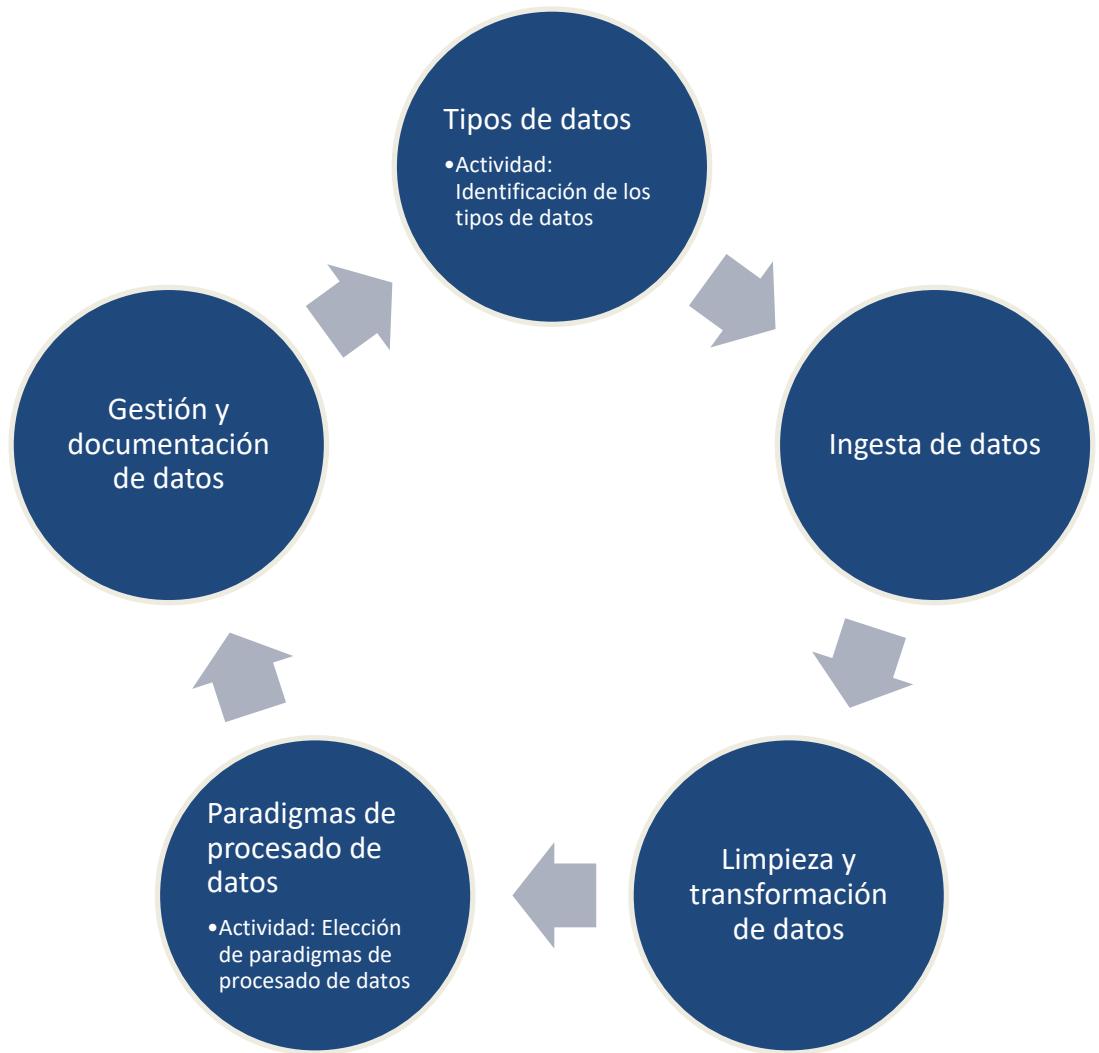
4. INGESTA Y TRATAMIENTO



INGESTA Y TRATAMIENTO DEL DATO

- Tipos de datos (ii).
- Ingesta de datos.
- Limpieza y transformación.
- Paradigmas de procesado.
- Gestión y documentación.

¿Qué es el tratamiento de datos?



¿Qué es la información?

Conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada, así como su comunicación y adquisición.

¿Qué es un dato?

Unidad mínima de información, generalmente fáctica, que puede representarse y almacenarse de forma digital para su tratamiento.

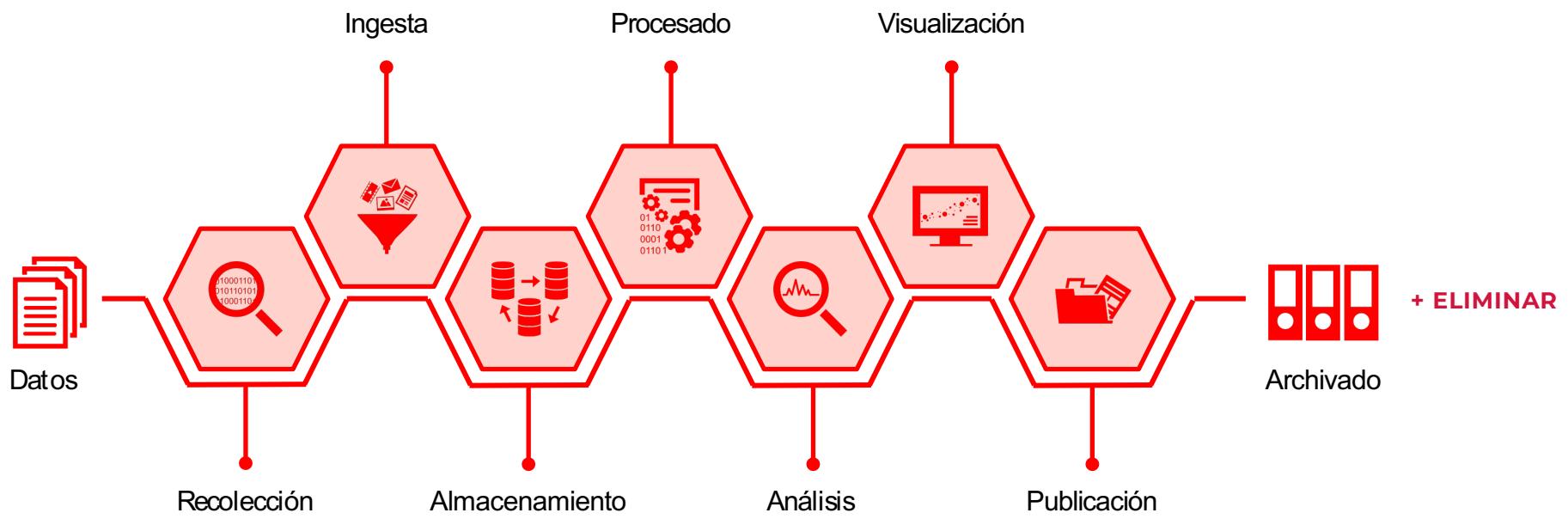
¿Qué es el tratamiento de datos?

Generalmente, un dato por sí mismo no proporciona información, o esta es mínima.

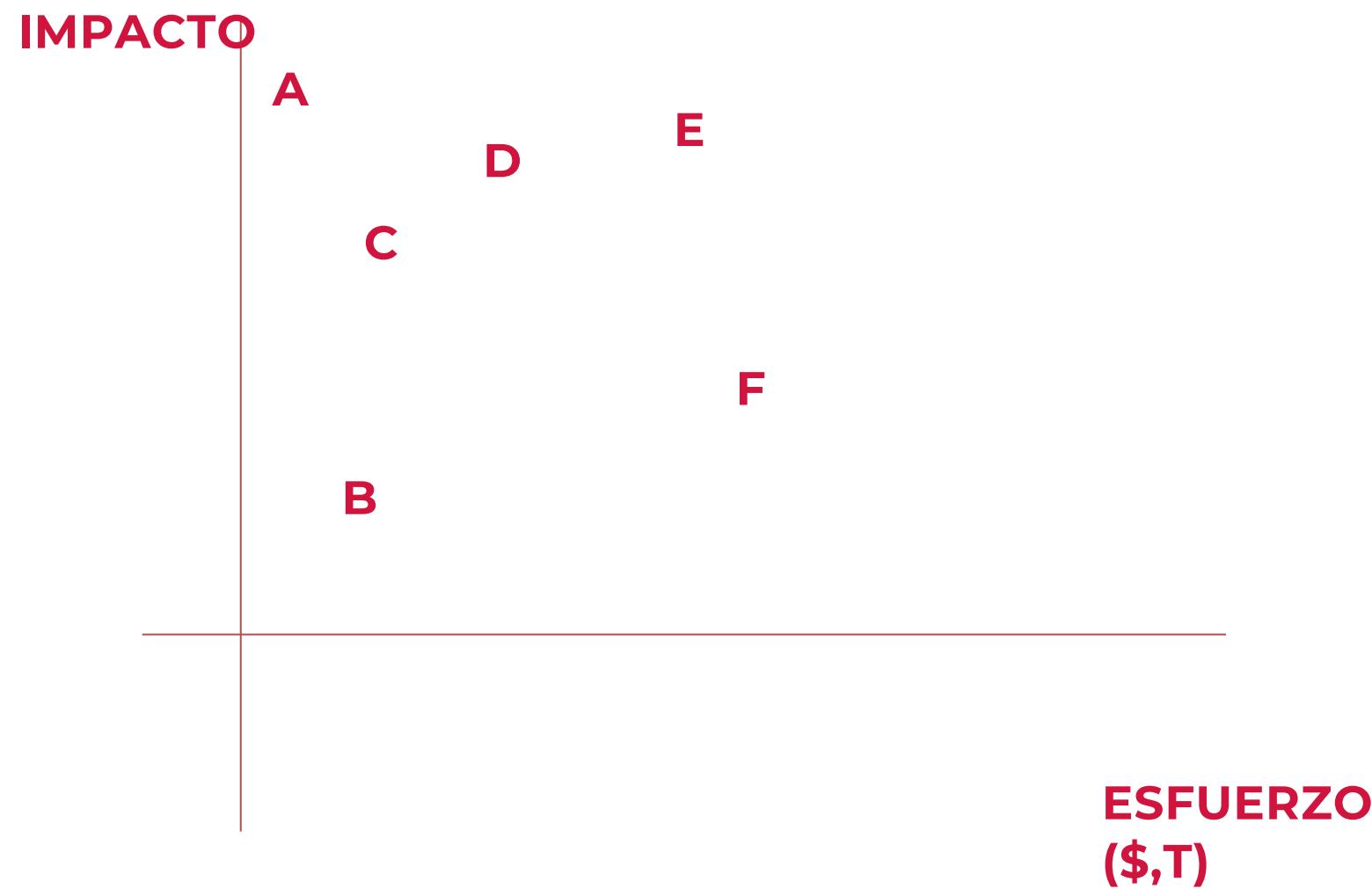
Por tanto, surge un interés en «extraer información» de un conjunto de datos, lo cuál generalmente requiere de algún tipo de proceso o tratamiento sobre los mismos.

Así, el tratamiento de datos es la serie de procesos a los que sometemos los datos para convertirlos en información relevante.

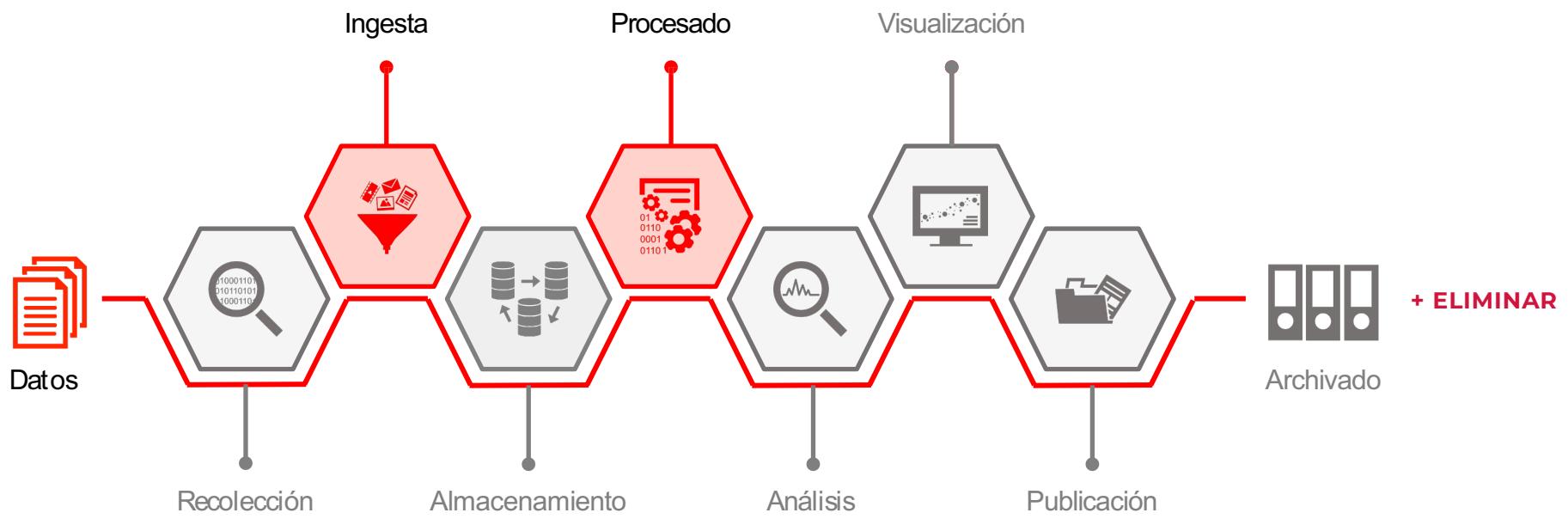
Ciclo de vida del dato



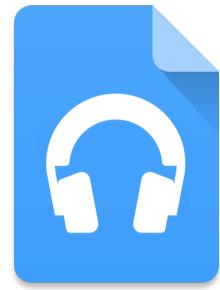
Priorización !



Ciclo de vida del dato



Principales tipos



Principales tipos

No estructurado



Semiestructurado



Estructurado



Datos estructurados

Son aquellos que poseen una estructura completamente definida, con un número de atributos (o columnas) fijos y con tipos de datos pre establecidos; por ejemplo: hojas de cálculo, bases de datos relacionales, etc.

	A	B	C	D	E	F	G	H	I
1	Código	Apellido	Nombre	CP	Ciudad	Fecha de Nacimiento	Departamento	Salario	Categoría
2	1	Alazart	Pedro	45720	Toledo	16/12/1976	Marketing	58.000,00	A1
3	2	Austria	Carolina	10001	Cáceres	04/05/1965	Producción	47.000,00	A2
4	3	Azcona	Pablo	46080	Valencia	12/02/1987	Ventas	25.000,00	C
5	4	Baamonde	Adán	47270	Valladolid	21/11/1969	Contabilidad	34.000,00	B
6	5	Ballesteros	Domingo	01006	Álava	19/02/1959	I+D	32.000,00	B
7	6	Batista	Clara	50100	Zaragoza	29/07/1981	Ventas	31.000,00	B
8	7	Bella	Inés	22050	Huesca	17/06/1990	Producción	21.000,00	C
9	8	Beltrán	Alberto	12002	Castellón	14/08/1978	Producción	61.000,00	A1
10	9	Bizet	Silvio	28003	Madrid	24/09/1975	Producción	29.000,00	C
11	10	Bollo	Adán	51126	Ceuta	13/11/1987	Ventas	26.000,00	C
12	11	Bon	Karina	13007	Ciudad Real	19/01/1975	IT	41.000,00	A2
13	12	Bonifaz	Luis	16100	Cuenca	11/03/1988	Marketing	43.000,00	A2
14	13	Boñar	Juan	08004	Barcelona	17/04/1989	Producción	24.000,00	C
15	14	Boveda	José Luis	15002	A Coruña	06/05/1976	IT	36.000,00	B
16	15	Bretón	María	08005	Barcelona	09/04/1991	Producción	28.000,00	C

Datos semiestructurados

Son aquellos que presentan cierta estructura, pero esta no es fija, pudiendo variar para diferentes registros.

xml

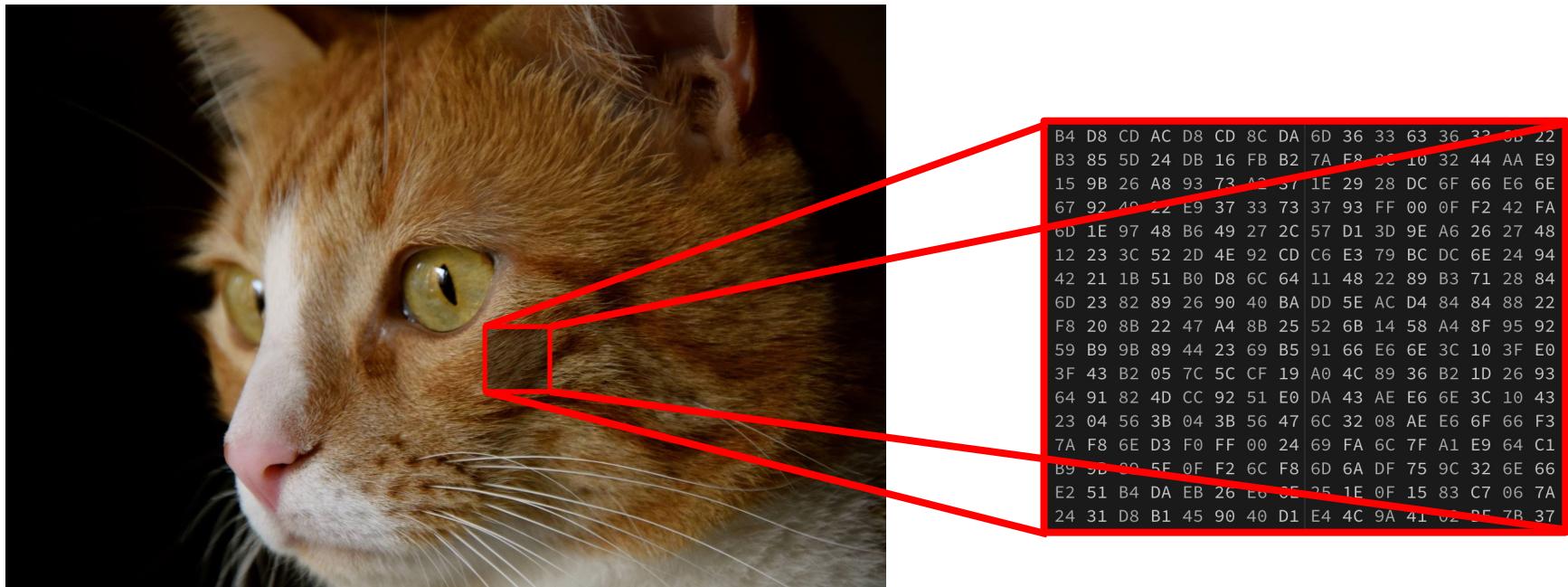
```
<?xml version="1.0" encoding="utf-8"?>
<Peliculas>
    <Pelicula ean="7509036232759">
        <Titulo>Lo que el viento se llevó</Titulo>
        <Año>1939</Año>
        <Director>Victor Fleming</Director>
        <Actores>
            <Actor>Clark Gable</Actor>
            <Actor>Olivia de Havilland</Actor>
        </Actores>
        <Productor>MGM</Productor>
    </Pelicula>
    <Pelicula ean="738572105723">
        <Titulo>Cinema Paradiso</Titulo>
        <Año>1988</Año>
        <Director>Giuseppe Tornatore</Director>
    </Pelicula>
</Peliculas>
```

json

```
{"películas": [
    {
        "ean": 7509036232759,
        "titulo": "Lo que el viento se llevó",
        "año": 1939,
        "director": "Victor Fleming",
        "actores": [
            "Clark Gable",
            "Olivia de Havilland",
        ],
        "productor": "MGM"
    },
    {
        "ean": 738572105723,
        "titulo": "Cinema Paradiso",
        "año": 1988,
        "director": "Giuseppe Tornatore"
    }
]} 
```

Datos no estructurados

Son aquellos que carecen de estructura clara o interpretable, por lo que su tratamiento digital acostumbra a ser más complejo, o requiere un mayor procesamiento.



Identificación de los tipos de datos

Vamos a iniciar un proyecto consistente en una filmoteca, para lo cuál queremos obtener y registrar un catálogo de productos audiovisuales, que incluirá películas, música y libros electrónicos.

Por el momento, únicamente estamos interesados en conocer las propiedades de estos productos de forma individual. Sin embargo, no existe un conjunto de propiedades fijo para cada tipo de medios (p. ej. dos películas distintas pueden tener un conjunto de atributos diferentes).

¿Aqué tipo de datos nos estamos enfrentando?

Identificación de los tipos de datos

Vamos a iniciar un proyecto consistente en una filmoteca, para lo cuál queremos obtener y registrar un catálogo de productos audiovisuales, que incluirá películas, música y libros electrónicos.

Por el momento, únicamente estamos interesados en conocer las propiedades de estos productos de forma individual. Sin embargo, no existe un conjunto de propiedades fijo para cada tipo de medios (p. ej. dos películas distintas pueden tener un conjunto de atributos diferentes).

¿Aqué tipo de datos nos estamos enfrentando?

Identificación de los tipos de datos

Vamos a iniciar un proyecto consistente en una filmoteca, para lo cuál queremos obtener y registrar un catálogo de productos audiovisuales, que incluirá películas, música y libros electrónicos.

Por el momento, únicamente estamos interesados en conocer las propiedades de estos productos de forma individual. Sin embargo, no existe un conjunto de propiedades fijo para cada tipo de medios (p. ej. dos películas distintas pueden tener un conjunto de atributos diferentes).

¿Aqué tipo de datos nos estamos enfrentando?

Datos semiestructurados

Identificación de los tipos de datos

Para dar valor a nuestra filmoteca, decidimos desarrollar una red social, donde los usuarios pueden valorar los contenidos, intercambiar opiniones, etc. Los mecanismos de interacción son bastante libres, por lo que los usuarios pueden decidir de forma opcional compartir contenido multimedia, indicar su ubicación, etc.

Queremos registrar estas interacciones de los usuarios para su posterior tratamiento.

¿Qué tipo de datos nos estamos enfrentando?

Identificación de los tipos de datos

Para dar valor a nuestra filmoteca, decidimos desarrollar una red social, donde los usuarios pueden valorar los contenidos, intercambiar opiniones, etc. Los mecanismos de interacción son bastante libres, por lo que los usuarios pueden decidir de forma opcional compartir contenido multimedia, indicar su ubicación, etc.

Queremos registrar estas interacciones de los usuarios para su posterior tratamiento.

¿Qué tipo de datos nos estamos enfrentando?

Datos semiestructurados / no estructurados

Identificación de los tipos de datos

En nuestra filmoteca, decidimos comenzar a incluir contenido audiovisual (más allá del catálogo). Idealmente, querremos ser capaces de tratar y procesar este contenido para poder extraer valor del mismo.

¿Aqué tipo de datos nos estamos enfrentando?

Identificación de los tipos de datos

En nuestra filmoteca, decidimos comenzar a incluir contenido audiovisual (más allá del catálogo). Idealmente, querremos ser capaces de tratar y procesar este contenido para poder extraer valor del mismo.

¿Aqué tipo de datos nos estamos enfrentando?

Datos no estructurados

Identificación de los tipos de datos

Finalmente, decidimos poner en alquiler parte del contenido que estamos incorporando en nuestra filmoteca.

Para ello, queremos registrar los clientes y las transacciones que estos realicen. Los clientes tendrán un identificador único, así como un correo electrónico, nombre, apellidos y contraseña. Las transacciones incluyen el identificador del cliente, el identificador del producto alquilado y el precio abonado por dicho alquiler.

¿A qué tipo de datos nos estamos enfrentando?

Identificación de los tipos de datos

Finalmente, decidimos poner en alquiler parte del contenido que estamos incorporando en nuestra filmoteca.

Para ello, queremos registrar los clientes y las transacciones que estos realicen. Los clientes tendrán un identificador único, así como un correo electrónico, nombre, apellidos y contraseña. Las transacciones incluyen el identificador del cliente, el identificador del producto alquilado y el precio abonado por dicho alquiler.

¿A qué tipo de datos nos estamos enfrentando?

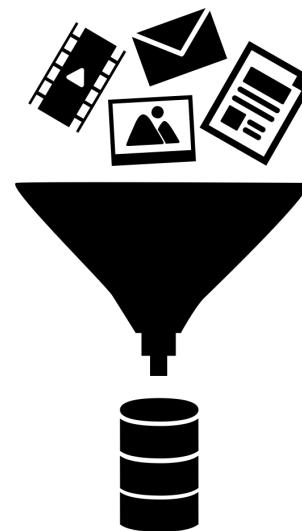
Datos estructurados

Ingesta y tratamiento del dato

Ingesta de datos

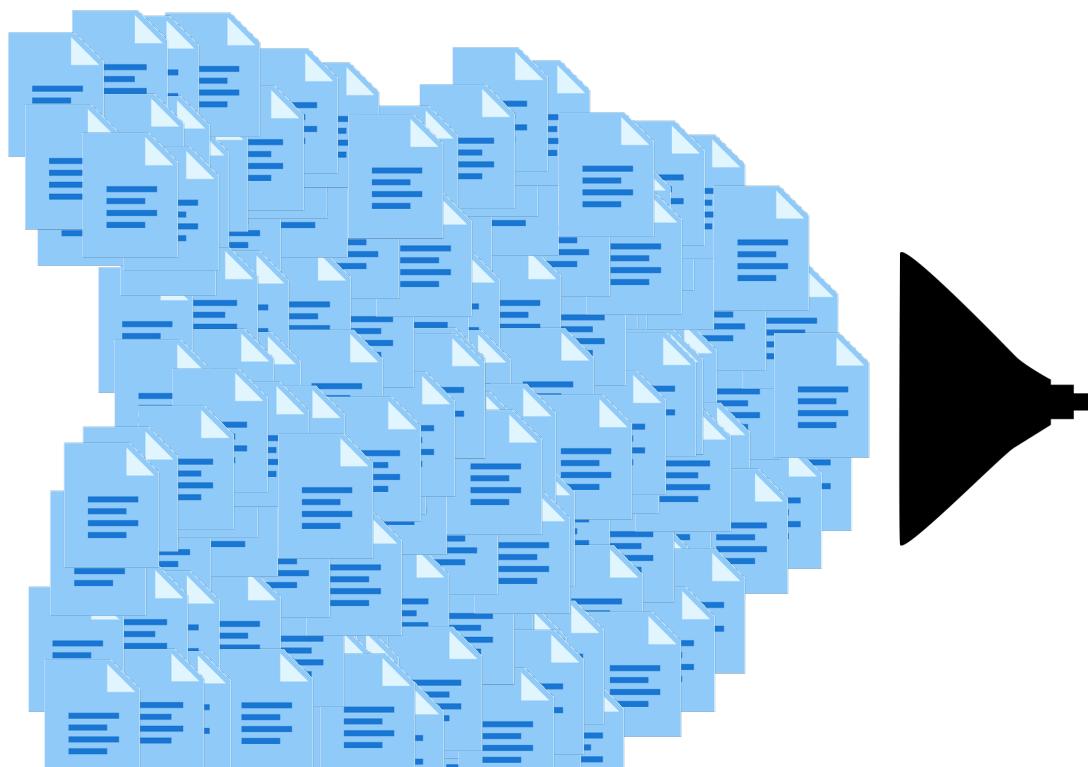
¿Qué es la ingesta de datos?

Ese proceso por el cual se recolectan datos de varias fuentes o bases de datos y se incorporan a un entorno unificado para su posterior procesamiento.



Las 3 Vs: Volumen

En ocasiones habrá que lidiar con el denominado «big data», es decir, cantidades enormes de datos cuyo volumen puede complicar el proceso de ingesta de datos, algo que puede aliviarse implementando soluciones y entornos distribuidos.



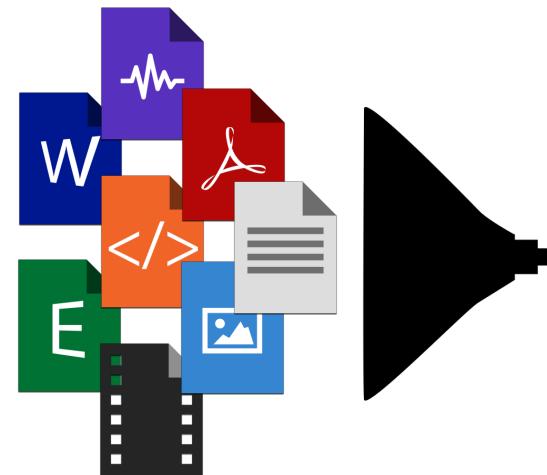
Las 3 Vs: Velocidad

En algunos casos, querremos ingerir y tratar datos que se generan a altas velocidades, lo cuál plantea nuevos retos para asegurarse de que todos estos datos se capturan correctamente.



Las 3 Vs: Variedad

Cuando los datos proceden de fuentes muy diversas, se complica el proceso de ingestión, pues se deben revisar constantemente las conexiones con estas fuentes y asegurar que los diversos datos se tratan convenientemente.



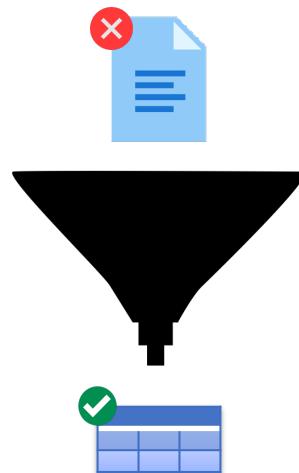
Ingesta y tratamiento del dato

Limpieza y transformación de datos

¿Por qué limpiar datos?

Durante la ingesta de datos, estos pueden venir con formatos diversos que puede resultar conveniente convertir o dotar de estructura.

Además, pueden contener errores o anomalías que deben ser corregidas.



Filtrado del dato

En ocasiones, podemos querer ignorar ciertos datos que no cumplen determinadas condiciones, o que no son relevantes para nuestro sistema.

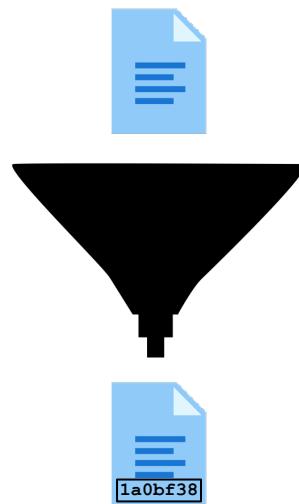
Ejemplo: si estamos recopilando «tuits» con noticias locales, podemos filtrar por geolocalización de las publicaciones.



Identificación del dato

Durante la ingesta, es importante dotar a los datos ingeridos de un identificador único (ya sea propio o dependiente de la fuente de datos).

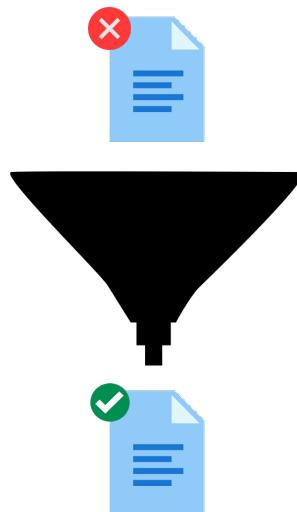
Ejemplo: podemos identificar los «tuits» con un identificador que nos proporciona Twitter, único para cada publicación.



Revisión del dato

Los datos deberían ajustarse al esquema (dominio) especificado, cumpliendo con las reglas de integridad y coherencia impuestas. Si lo hacen, pueden omitirse, subsanarse o marcarse como inválidos.

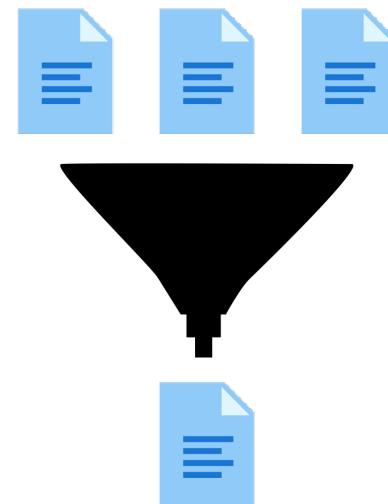
Ejemplo: por algún error, podría llegarnos un «tuit» con un número negativo de «retuits». De ser así, podríamos decidir sustituir este valor inválido por un cero.



Deduplicación del dato

Cuando incorporamos datos de varias fuentes (o varias consultas), es fácil que nos encontremos con datos duplicados, incluso si su estructura no es totalmente idéntica. En este caso, es conveniente eliminar los duplicados.

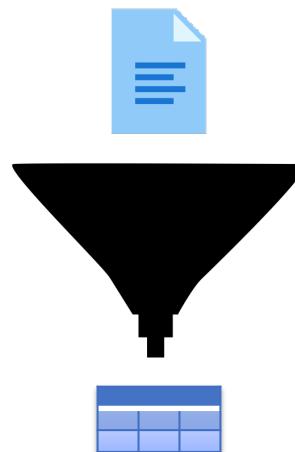
Ejemplo: si tenemos varias consultas activas recogiendo «tuits», una misma publicación podría ser devuelta por ambas consultas simultáneamente. En este caso, el identificador facilita la tarea de eliminar el registro duplicado.



Transformación del dato

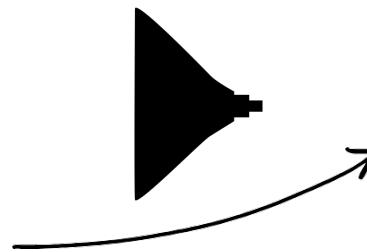
Cuando disponemos de datos estructurados o semiestructurados, puede ser conveniente transformar su estructura a una fija con el fin de unificar las diferentes fuentes de datos.

Ejemplo: en el caso de los «tuits», podemos decidir almacenar la información que nos interesa en formato JSON con un cierto esquema, obviando los campos que no son relevantes.



Estructuración del dato

Cuando se dispone de datos no estructurados, puede resultar interesante tratar de dotarlos de cierta estructura para facilitar su posterior análisis (en ocasiones, el «machine learning» puede ayudar a esto).



```
"labelAnnotations": [
  {
    "description": "Cat",
    "mid": "/m/01yrx",
    "score": 0.99598557,
    "topicality": 0.99598557
  },
  {
    "description": "Mammal",
    "mid": "/m/04rky",
    "score": 0.9890478,
    "topicality": 0.9890478
  },
  {
    "description": "Vertebrate",
    "mid": "/m/09686",
    "score": 0.9851104,
    "topicality": 0.9851104
  },
  {
    "description": "Whiskers",
    "mid": "/m/0117z4d"
  }
]
```

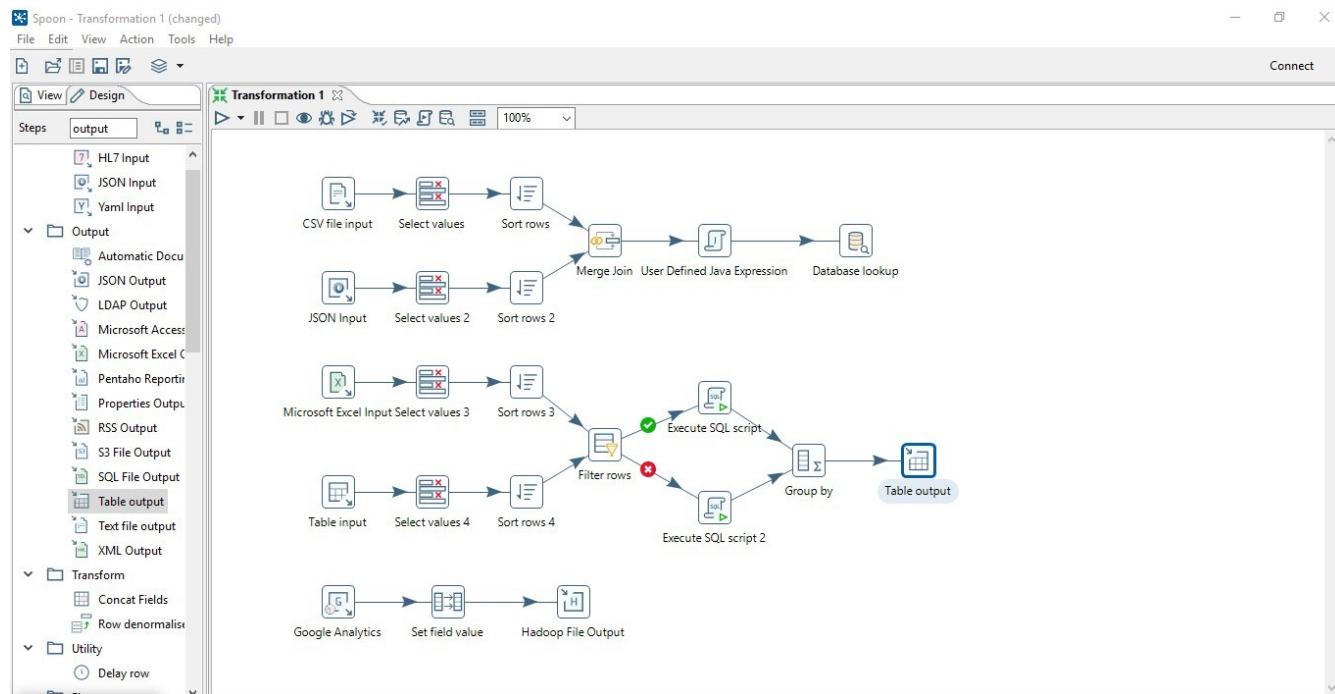
El proceso ETL

ETL son las siglas de **E**xtract — **T**ransform — **L**oad, y hace referencia al concepto de extraer datos de diferentes fuentes y transformarlos para posteriormente cargarlos en algún almacén o base de datos.

Como proceso, está muy relacionado con la ingesta de datos, aunque históricamente se ha denominado ETL al proceso de extracción de datos estructurados disponibles «por lotes».

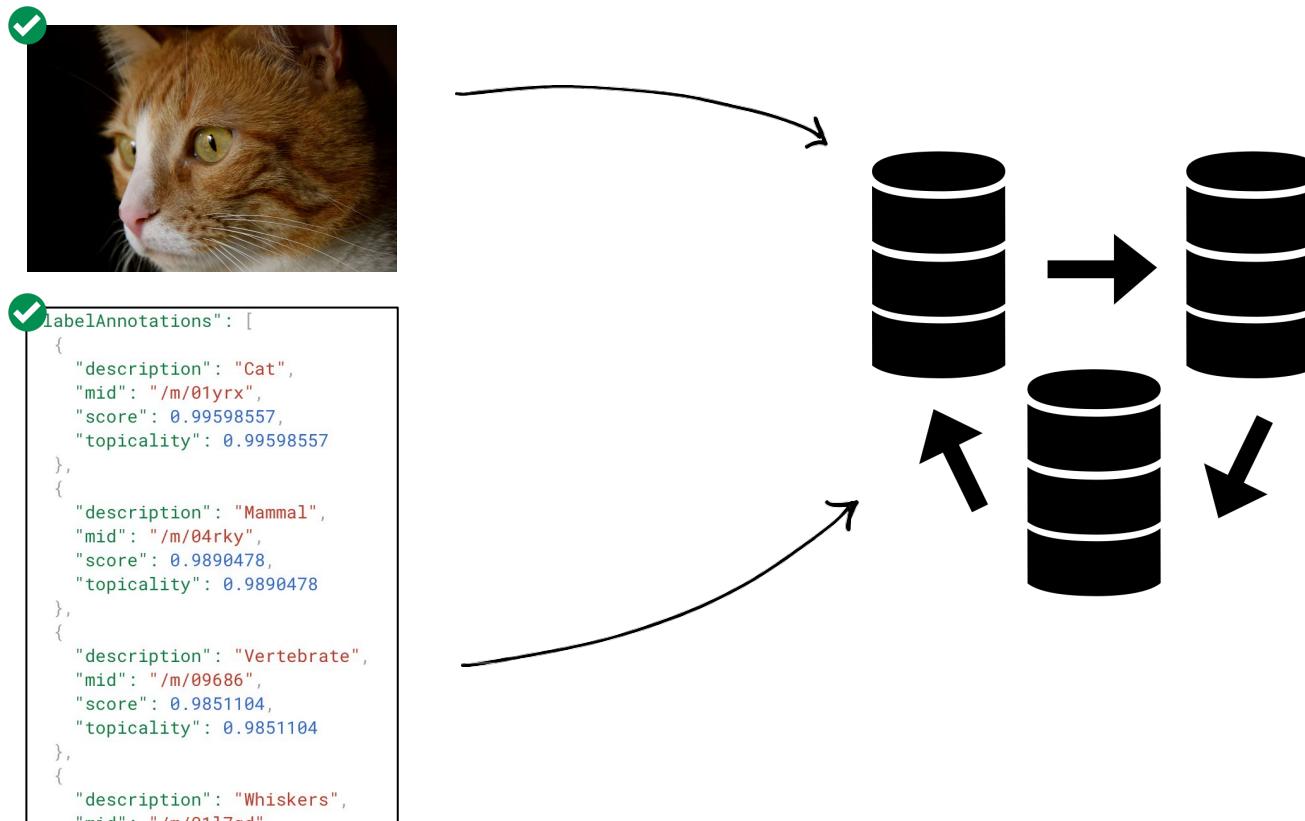
Herramientas ETL

Existen herramientas que permiten facilitar el proceso de ETL mediante el diseño de «pipelines» que indican los pasos a los que se someten los datos, tales como Talend o Pentaho.



Datos en crudo

Debido al abaratamiento de los costes de almacenamiento, puede ser interesante almacenar no solo el dato procesado, sino también el original o «crudo», por si fuera útil en el futuro.



ELT vs. ETL

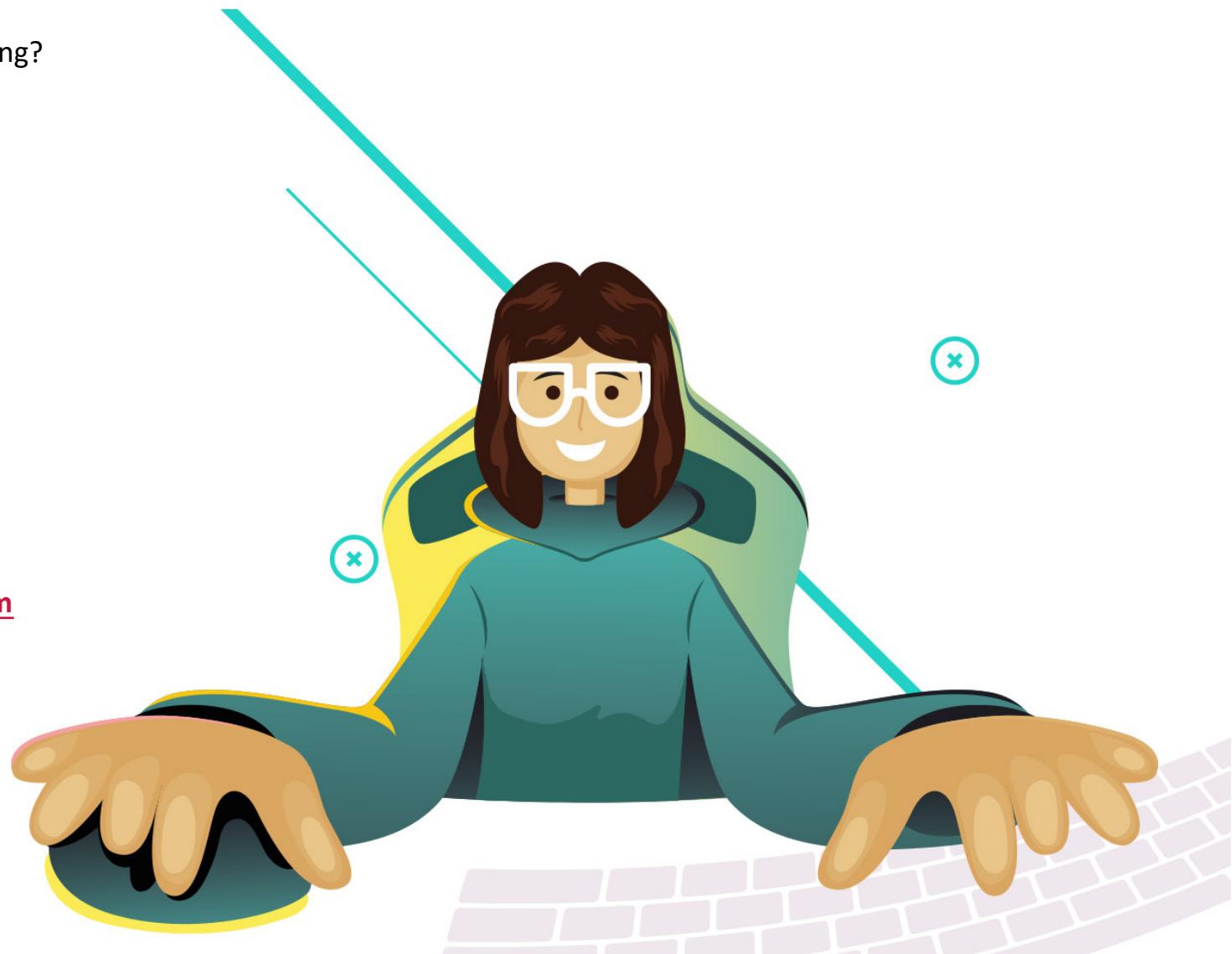
ELT es la filosofía **Extract — Load — Transform**, que plantea realizar la carga de los datos tras su extracción (en «crudo»), transformándolos cuando sea necesario y del modo que resulte más apropiado en cada momento.

Ejercicio 1

¿Qué es MapReduce? ¿Y Batch processing?

- ¿Cuándo nace MapReduce?
- ¿Qué es un clúster?
- ¿Qué es un NameNode?

DURACIÓN ESTIMADA: 1,5 horas
Plix, envíarme a julian@dragonlab.team

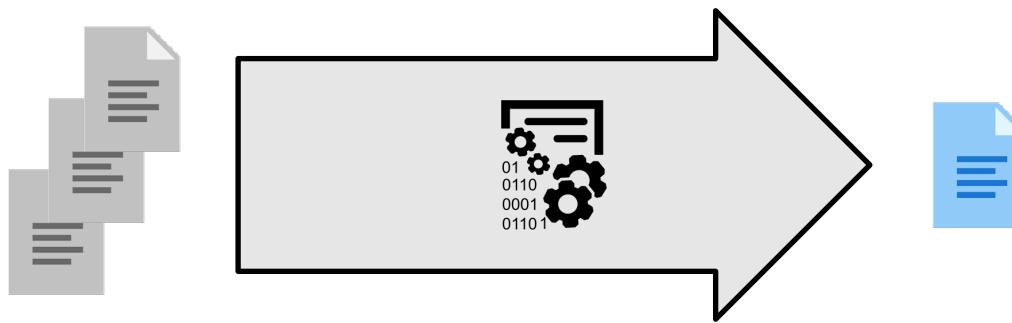


Ingesta y tratamiento del dato

Paradigmas de procesado de datos

Migración (lote único)

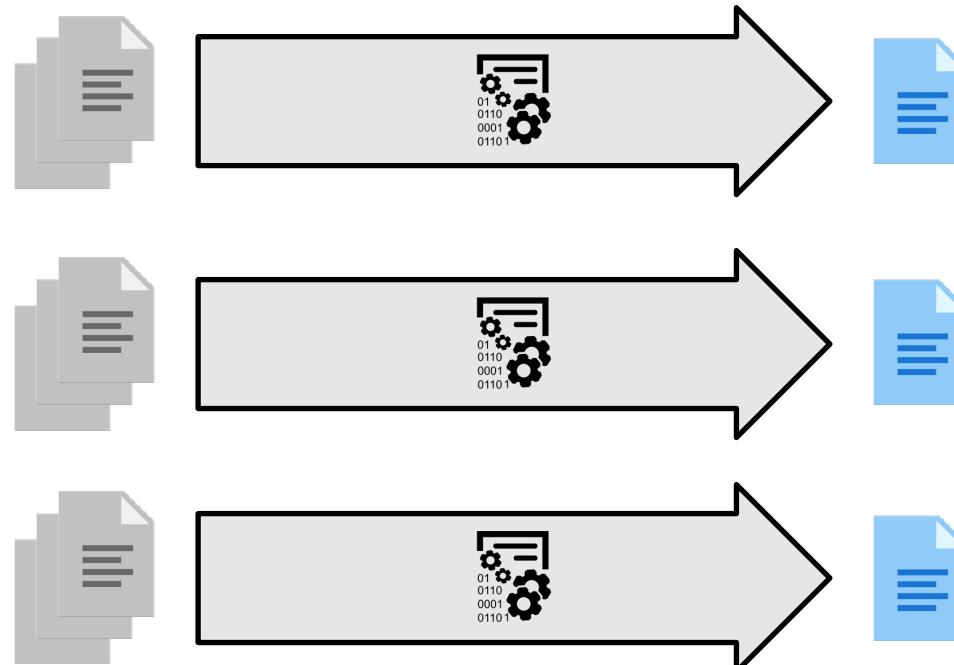
Es el proceso de realizar un procesado puntual para transformar unos datos en otros, o para tratarlos de algún modo.



Batch (por lotes)

En una aproximación *batch* o por lotes, grandes cantidades de datos se procesan de golpe, generalmente mediante algún enfoque distribuido.

Este proceso puede repetirse periódicamente, según se dispone de nuevos datos.



Batch: MapReduce

Es un paradigma de procesado de datos por lotes que permite transformar y agregar datos de forma distribuida entre diferentes servidores (o nodos).

Batch: MapReduce

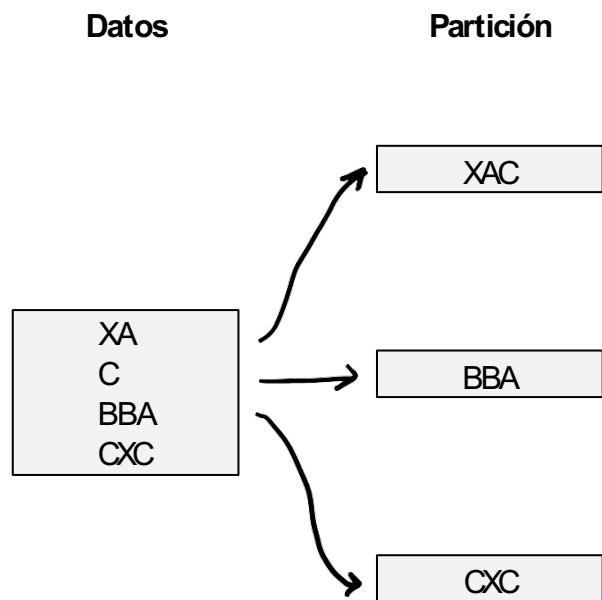
Es un paradigma de procesado de datos por lotes que permite transformar y agregar datos de forma distribuida entre diferentes servidores (o nodos).

Datos

XA
C
BBA
CXC

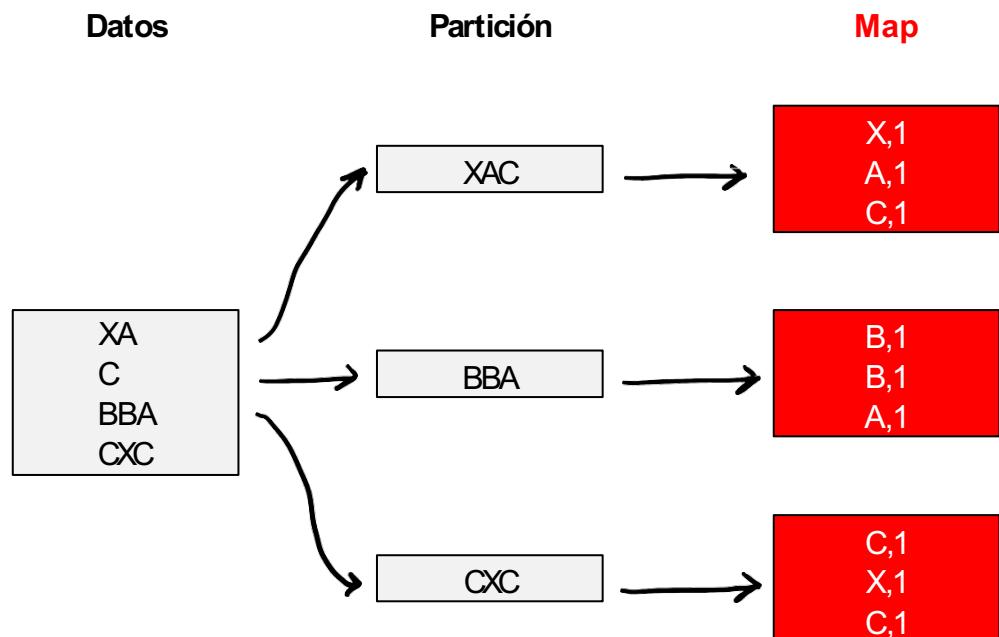
Batch: MapReduce

Es un paradigma de procesado de datos por lotes que permite transformar y agregar datos de forma distribuida entre diferentes servidores (o nodos).



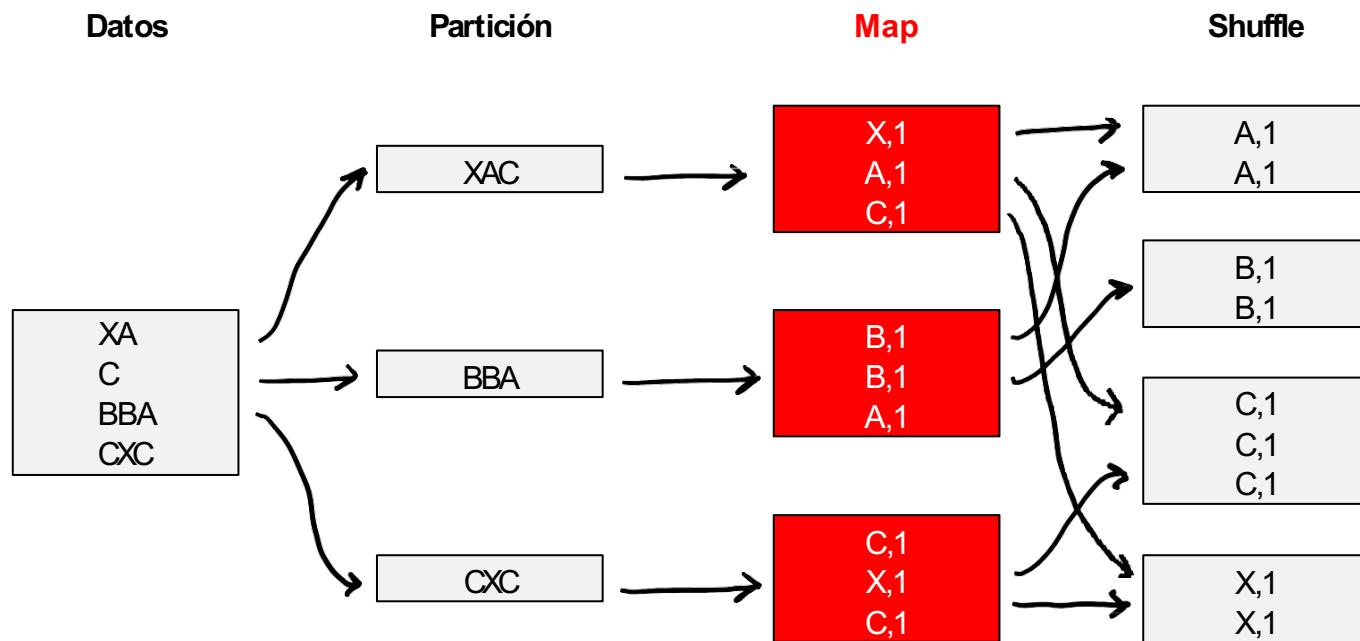
Batch: MapReduce

Es un paradigma de procesado de datos por lotes que permite transformar y agregar datos de forma distribuida entre diferentes servidores (o nodos).



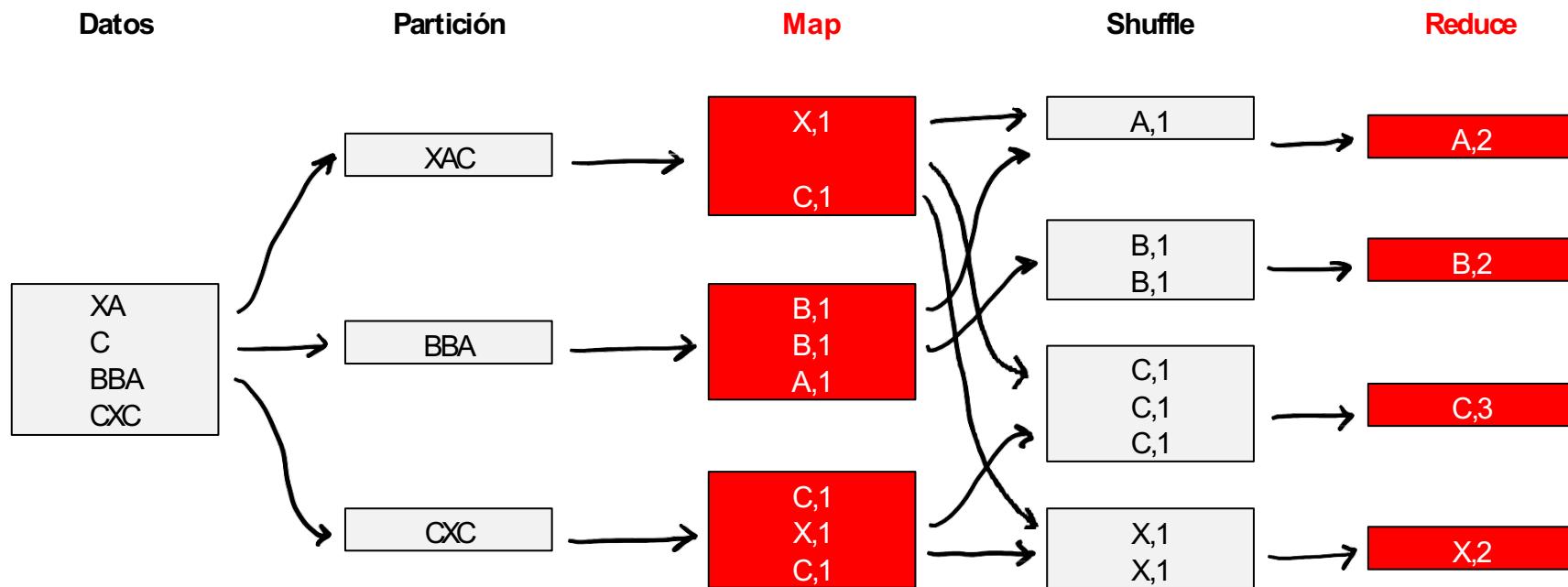
Batch: MapReduce

Es un paradigma de procesado de datos por lotes que permite transformar y agregar datos de forma distribuida entre diferentes servidores (o nodos).



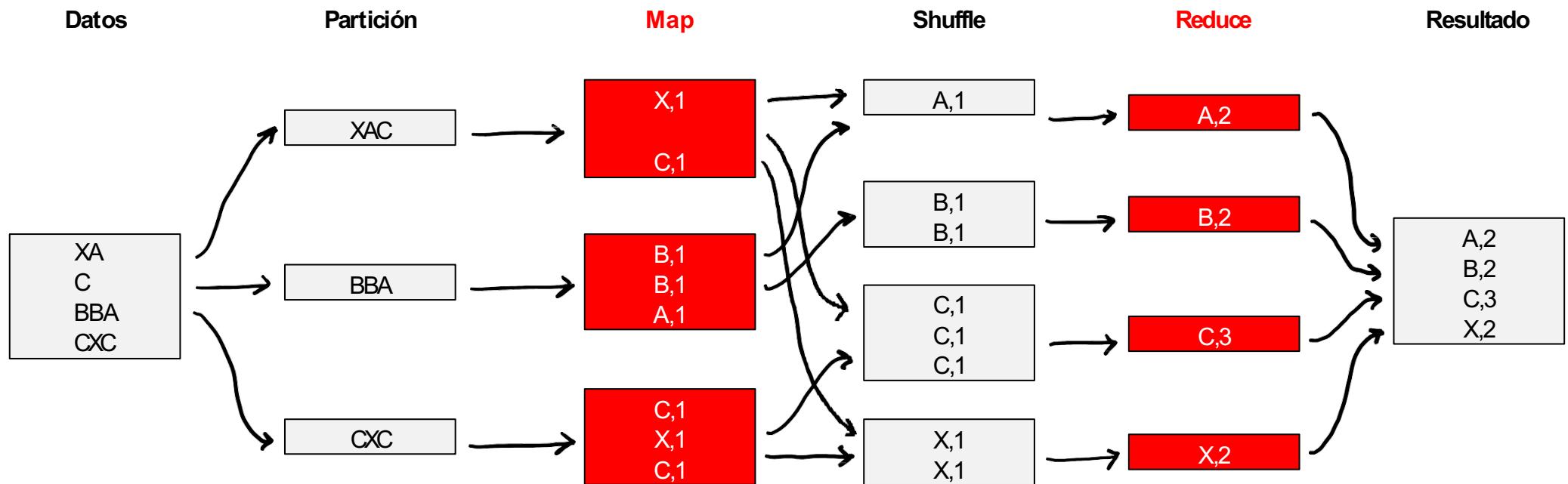
Batch: MapReduce

Es un paradigma de procesado de datos por lotes que permite transformar y agregar datos de forma distribuida entre diferentes servidores (o nodos).



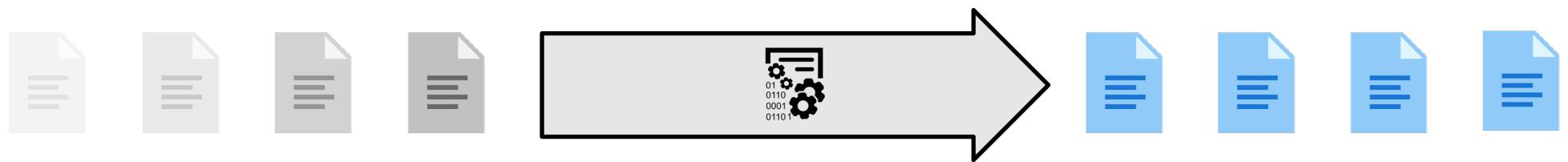
Batch: MapReduce

Es un paradigma de procesado de datos por lotes que permite transformar y agregar datos de forma distribuida entre diferentes servidores (o nodos).



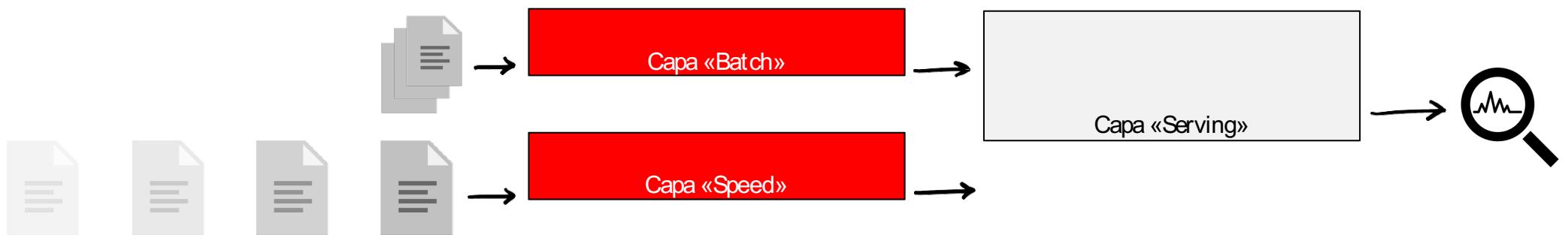
Streaming (tiempo real)

En una aproximación en streaming, los datos llegan de forma continua y a gran velocidad, y este flujo de datos se va transformando y procesando a medida que va llegando al sistema.



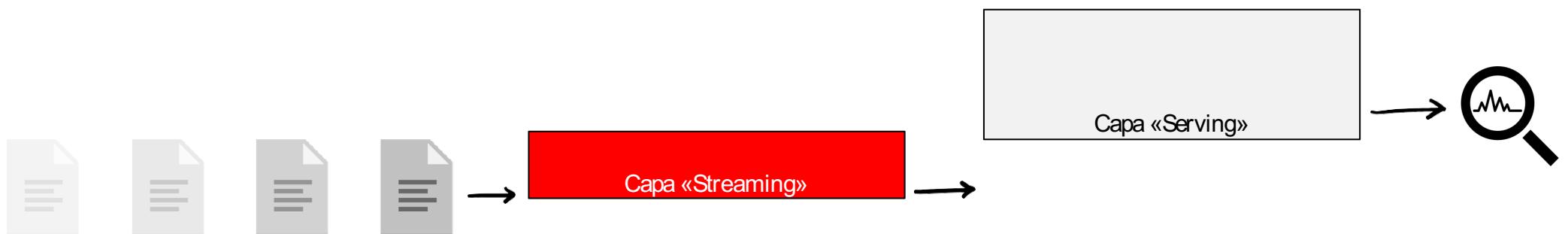
Streaming: Lambda

Es una arquitectura que permite el procesado de datos tanto por lotes como en tiempo real, permitiendo que ambas capas se retroalimenten, y proporcionando una salida conjunta.



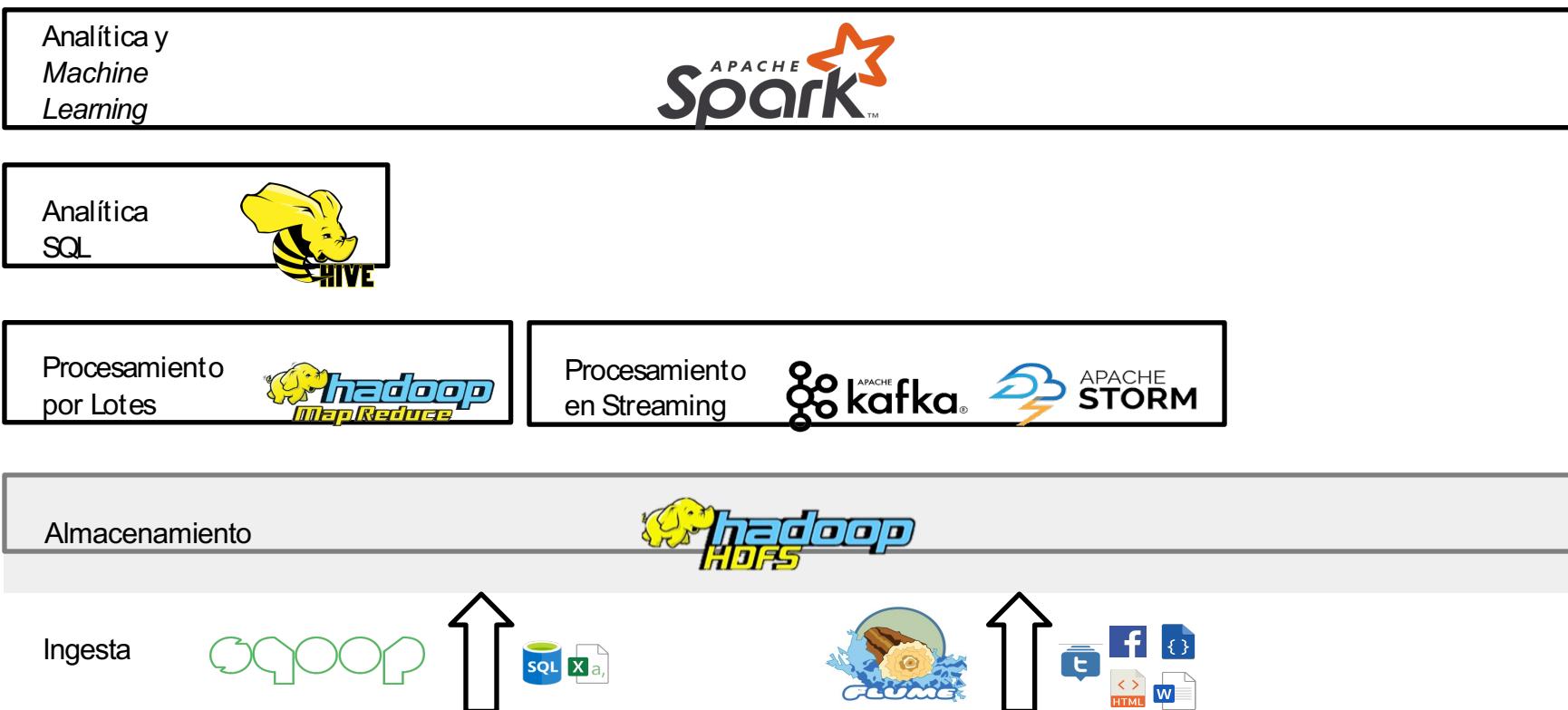
Streaming: Kappa

Simplifica la arquitectura Lambda eliminando la capa de procesamiento por lotes, permitiendo un mejor aprovechamiento de los recursos, que puede ser destinado íntegramente al procesado el tiempo real.



El ecosistema de Hadoop

Hadoop es un ecosistema consistente en multitud de herramientas para almacenamiento, procesado, análisis y gestión integral de «Big Data», inspirado en tecnologías presentadas por Google a principios de los 2000.



Elección de paradigmas de procesado de datos

Para nuestro catálogo audiovisual, disponemos de un sistema que de forma constante está registrando toda la información sobre las visitas que recibe nuestro catálogo online, incluyendo la dirección IP del visitante, los datos de su navegador web, la página que está visitando, si el servidor ha devuelto algún error, etc.

Al finalizar el año, queremos obtener un reporte de los contenidos más visitados de nuestro catálogo, reportando también los principales países de origen de los visitantes.

¿Qué paradigma de procesado de datos resulta más adecuado?

Elección de paradigmas de procesado de datos

Para nuestro catálogo audiovisual, disponemos de un sistema que de forma constante está registrando toda la información sobre las visitas que recibe nuestro catálogo online, incluyendo la dirección IP del visitante, los datos de su navegador web, la página que está visitando, si el servidor ha devuelto algún error, etc.

Al finalizar el año, queremos obtener un reporte de los contenidos más visitados de nuestro catálogo, reportando también los principales países de origen de los visitantes.

¿Qué paradigma de procesado de datos resulta más adecuado?

- > Procesamiento por lotes (batch)

Elección de paradigmas de procesado de datos

Cuando el catálogo comienza a tener suficientes visitas, sospechamos que algunos usuarios malintencionados están tratando de acceder a los recursos audiovisuales sin autorización; es decir, tratan de hacer ingeniería inversa de la web para intentar descargar los ficheros multimedia (archivos de vídeo y archivos de música) sin pagar el coste del alquiler.

Tras investigar la cuestión, desarrollamos un modelo de «machine learning» que permite identificar comportamientos anómalos, y decidimos emplearlo para que, en caso de que llegue un nuevo usuario malintencionado a nuestro sistema, impedirle el acceso de inmediato.

¿Qué paradigma de procesado de datos resulta más adecuado?

Elección de paradigmas de procesado de datos

Cuando el catálogo comienza a tener suficientes visitas, sospechamos que algunos usuarios malintencionados están tratando de acceder a los recursos audiovisuales sin autorización; es decir, tratan de hacer ingeniería inversa de la web para intentar descargar los ficheros multimedia (archivos de vídeo y archivos de música) sin pagar el coste del alquiler.

Tras investigar la cuestión, desarrollamos un modelo de «machine learning» que permite identificar comportamientos anómalos, y decidimos emplearlo para que, en caso de que llegue un nuevo usuario malintencionado a nuestro sistema, impedirle el acceso de inmediato.

¿Qué paradigma de procesado de datos resulta más adecuado?

Procesamiento en tiempo real (streaming)

Elección de paradigmas de procesado de datos

Pronto nos damos cuenta de que nuestro sistema de detección de anomalías resulta en demasiados falsos positivos, es decir, impide el acceso a muchos usuarios que no tienen malas intenciones, causándoles un perjuicio y empeorando la imagen de nuestro catálogo.

Para solventarlo, decidimos emplear los registros que nuestro sistema ha ido almacenando con el fin de reentrenar y refinar nuestro modelo de «machine learning».

¿Qué paradigma de procesado de datos resulta más adecuado?

Elección de paradigmas de procesado de datos

Pronto nos damos cuenta de que nuestro sistema de detección de anomalías resulta en demasiados falsos positivos, es decir, impide el acceso a muchos usuarios que no tienen malas intenciones, causándoles un perjuicio y empeorando la imagen de nuestro catálogo.

Para solventarlo, decidimos emplear los registros que nuestro sistema ha ido almacenando con el fin de reentrenar y refinar nuestro modelo de «machine learning».

¿Qué paradigma de procesado de datos resulta más adecuado?

- > Procesamiento por lotes (batch)

Elección de paradigmas de procesado de datos

Cuando implementamos la primera versión de nuestro catálogo, decidimos emplear una base de datos relacional (SQL) con el fin de almacenar los registros, dada la popularidad de esta tecnología.

Sin embargo, con el paso del tiempo, nos hemos dado cuenta de que la flexibilidad que nos proporciona es insuficiente, y optamos por diseñar un nuevo esquema más flexible, basado en documentos JSON para almacenar la información.

Apartir de ahora, nuestros nuevos ítems del catálogo se almacenarán con la nueva estructura. No obstante, es importante adaptar todos nuestros datos a este nuevo esquema.

¿Qué paradigma de procesado de datos resulta más adecuado?

Elección de paradigmas de procesado de datos

Cuando implementamos la primera versión de nuestro catálogo, decidimos emplear una base de datos relacional (SQL) con el fin de almacenar los registros, dada la popularidad de esta tecnología.

Sin embargo, con el paso del tiempo, nos hemos dado cuenta de que la flexibilidad que nos proporciona es insuficiente, y optamos por diseñar un nuevo esquema más flexible, basado en documentos JSON para almacenar la información.

Apartir de ahora, nuestros nuevos ítems del catálogo se almacenarán con la nueva estructura. No obstante, es importante adaptar todos nuestros datos a este nuevo esquema.

¿Qué paradigma de procesado de datos resulta más adecuado?

> Migración

Ingesta y tratamiento del dato

Gestión y documentación de datos

¿Por qué documentar?

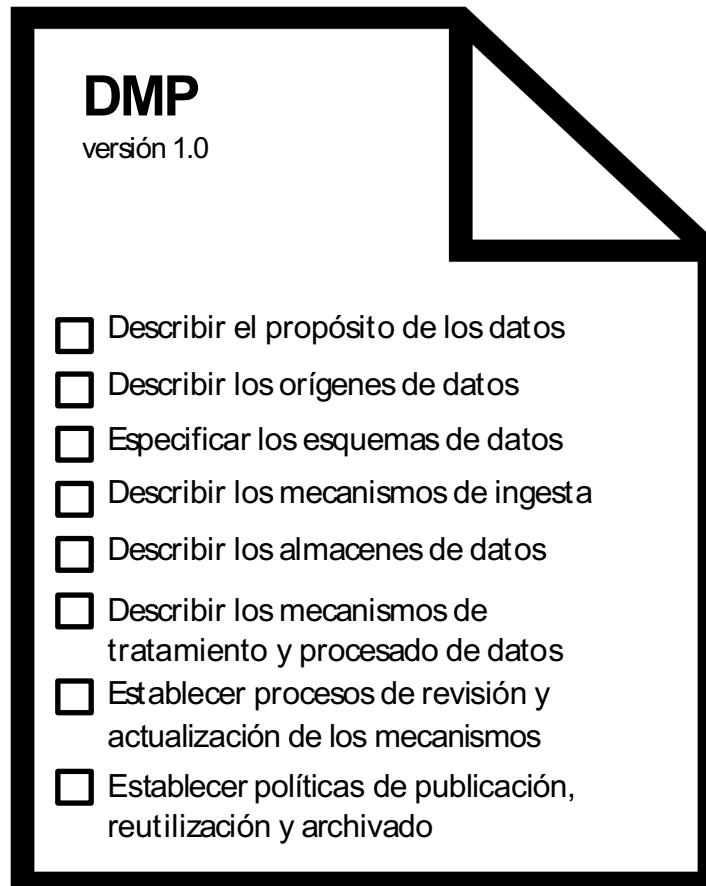
En demasiados casos, las compañías no tienen documentación actualizada sobre sus fuentes de datos, mecanismos de ingesta y procesado, recursos de almacenamiento, etc.

Es útil disponer de un **documento de gestión de los datos**.

Ventajas:

- > Facilita la comprensión de los procesos de datos a los nuevos empleados de la compañía, y permite la continuidad en caso de renovación del equipo de gestión de datos.
- > Permite la revisión sistemática de los procesos de ingesta y tratamiento de datos, asegurándose de que las fuentes de datos permanecen correctamente conectadas y la ingesta se realiza satisfactoriamente.
- > Facilita la auditoría de los datos, tanto interna como externa, ya sea con fines técnicos o regulatorios.
- > Facilita la revisión de mecanismos de almacenamiento y procesado para asegurarnos de que siempre responden a las necesidades de la compañía.

Contenidos de un Data Management Plan





GOBIERNO
DE ESPAÑA

VICEPRESIDENCIA
PRIMERA DEL GOBIERNO
MINISTERIO
DE ASUNTOS ECONÓMICOS
Y TRANSFORMACIÓN DIGITAL

SECRETARÍA DE ESTADO
DE DIGITALIZACIÓN
E INTELIGENCIA ARTIFICIAL

red.es

Centro de
Referencia Nacional
en Comercio Electrónico
y Marketing
CRN
Digital



UNIÓN EUROPEA

"El FSE invierte en tu futuro"

Fondo Social Europeo


Barrabés

 The Valley