

#### Sumario

- Datos no estructurados
- Datos semi estructurados
- Datos estructurados
- Anonimización de los datos



## Tipos de datos

Diversos criterios para diferenciar los tipos de datos

- Según su estructura
- Según su contenido
- Según su procedencia



## Tipos de datos según su estructura

Estructurados

• Semi-estructurados

No estructurados

SQL		
JSON		
XML		
TEXTO PLANO		



#### Datos no estructurados

- La información no tiene un modelo predefinido
- Gran volumen de información, típicamente en texto plano
- El contenido agrega diferentes tipos de datos en un mismo registro
- Los repositorios de datos contienen grandes cantidades de datos no estructurados
- ¡Pero mi documento tiene una estructura interna!
  - necesita el conocimiento de un experto para entender la semántica entre las entidades de datos



### Muchos ejemplos de datos no estructurados

- Imágenes satelitales
- Datos científicos: datos meteorológicos, oceanográficos, tránsito de vehículos
- Imágenes y vídeos
- Volcados de documentación empresarial: informes, resultados de encuestas, correos electrónicos, memorándums corporativos
- Datos provenientes de redes sociales: volcados de canales, mensajes de texto
- Contenido de páginas web, archivos de log



#### Datos semi-estructurados

- La información tiene una estructura basada en categorías o etiquetas para separar elementos distintos
- Los archivos de datos con algún formato de lenguaje mark-up como XML o JSON son típicamente semiestructurados
- Los datos están almacenados en registros y son fácilmente identificables
- Los repositorios de datos semi-estructurados pueden contener ambigüedades e inconsistencias



## Ejemplo de datos semi-estructurados

```
nombre:"Pedro",
apellido:"Perez",
id_pedido: "1234",
total:"12.34" },
nombre:"Petra",
apellido: "Lopez",
id_pedido : "5678",
total: "99.76" }]
```

Esquema flexible

No se gestionan posibles errores ni incoherencias

#### Datos estructurados

- Los datos pertenecen a un modelo de datos abstracto
  - Organiza los elementos principales en entidades
  - Define las relaciones entre las entidades
- Cada elemento del conjunto de datos tiene asignado un contenido semántico en función de su relación con el resto de entidades
- Se ha realizado un proceso previo de modelización y formalización sobre los datos
- Suelen formar parte de bases de datos relacionales: su gestión se realiza mediante SGBD como mariaDB o MySQL



### Ejemplo de datos estructurados

Los datos están estructurados en tablas donde se almacenan entidades y atributos

Nombre	Apellido	Id_pedido	Total
Pedro	Perez	1234	12.34
Petra	Lopez	5678	99.76

TABLA PEDIDOS

### Anonimización de los datos

- Proceso de eliminar las referencias cualquier identificación personal de un conjunto de datos (PII: Información de Identificación Personal)
- Este proceso se aplica sobre un conjunto de datos para conservar la privacidad de los individuos a los que se hace referencia
- Es necesario aplicar este proceso cuando se quieren hacer públicos datos con identificadores como nombres, direcciones o códigos postales
- En general, se debe intentar eliminar cualquier información que pueda identificar al individuo



## Tipos de anonimización

- Eliminación de datos
- Cifrado de los datos sensibles: depende del uso de una buena clave para el cifrado y el coste del proceso
- Enmascaramiento de los datos sensibles: se pretende mantener la estructura de la fuente de datos original. Los tipos se conservan, sus valores no



## La anonimización no asegura la protección

- La anonimización no proporciona protección absoluta
- Cualquier conjunto de datos suficientemente extenso puede ser desanonimizado usando información del contexto
- Con datos sensibles, como la salud, hay que limitar los campos más sensibles respecto a la privacidad y aplicar técnicas más complejas





# Tipos de datos



