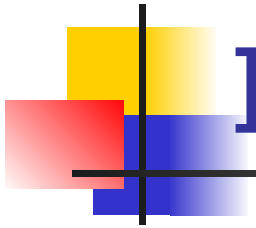
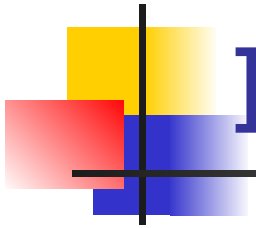
Procesos ETL
(Extracción, Transformación y Carga)

Tema 5



Indice



Indice

Introducción

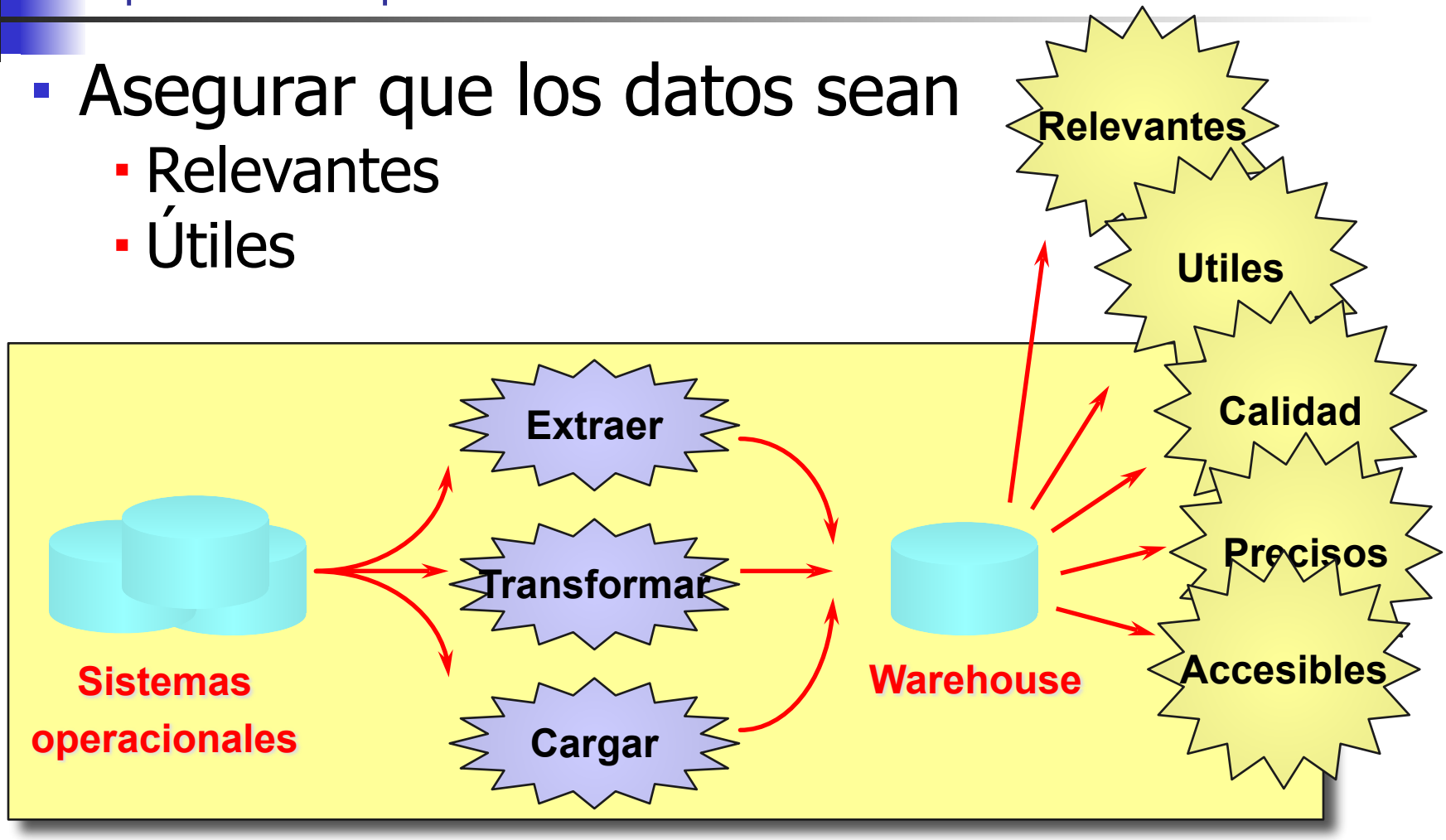
- Bill Inmon (90's): "Un almacén de datos es una colección de datos orientados por temas, integrados, no volátiles y variables en el tiempo en apoyo de la toma de decisiones estratégicas".
  - Datos procedentes de una gran variedad de fuentes

- ETL (Extraction-Transformation-Loading):
  - **Extracción** de datos desde fuentes de datos operacionales y heterogéneas,
  - **Transformación - limpieza**
    - Conversión de tipos, eliminación de valores nulos, …
  - **Carga - refresco** en el Almacén de datos

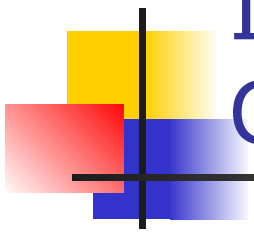
Introducción

Importancia de procesos ETL

- Asegurar que los datos sean
  - Relevantes
  - Útiles



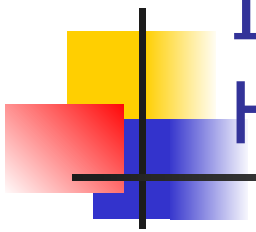
- **ETL tienen un coste elevado (tiempo, recursos)**



Introducción

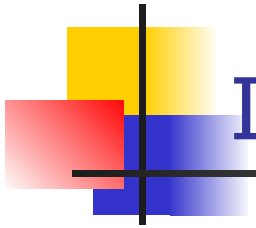
Consideraciones de diseño

- Definir política de calidad de datos con la empresa
- Documentar las fuentes de datos origen para comprenderlas
- Diseñar los procesos de limpieza
- Cuidado! Datos incorrectos -> Tomas de decisiones estratégicas erróneas
- Los procesos ETL suponen la fase más costosa en tiempo y recursos de los proyectos de almacenes de datos



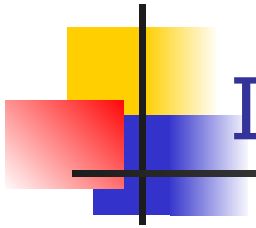
Introducción
Herramientas procesos ETL

- Lenguajes de programación (scripts)
- Herramientas especializadas
  - Integration Services (SQL Server)
  - …



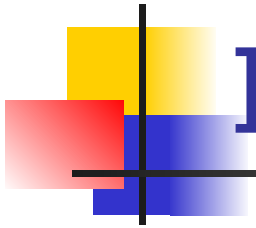
Introducción. Pasos detallados ETL

2. Seleccionar las fuentes para extraer datos
3. Transformar y limpiar los datos de las fuentes
4. Unir las fuentes
5. Seleccionar las estructuras destino a cargar datos (hechos, dimensiones, jerarquías,…)
6. Mapear los atributos de las fuentes en los destinos
7. Cargar los datos



Introducción. Pasos básicos

1. Extraer de las fuentes de datos
2. Transformar las fuentes
3. Cargar los datos

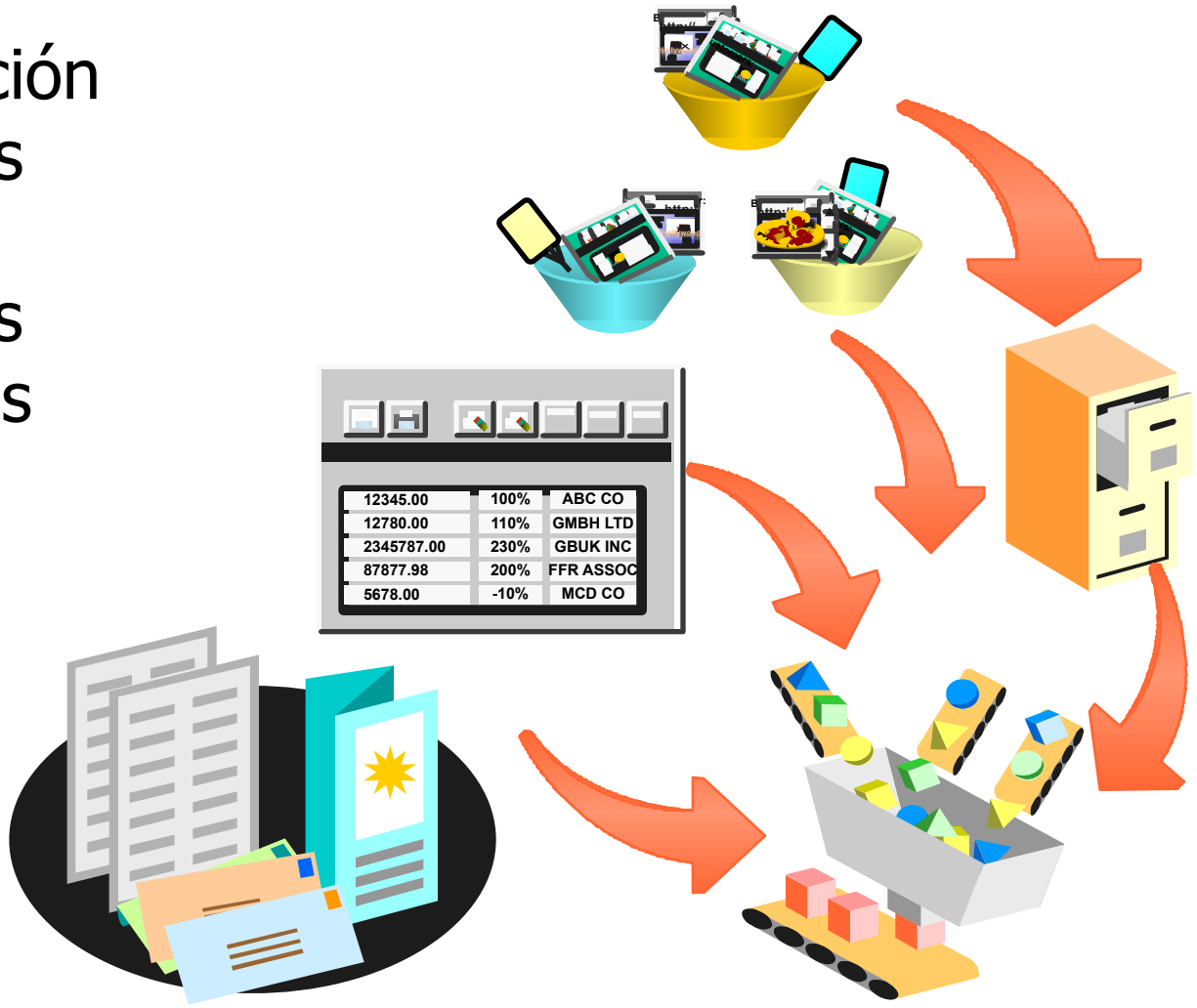


Indice

Extracción

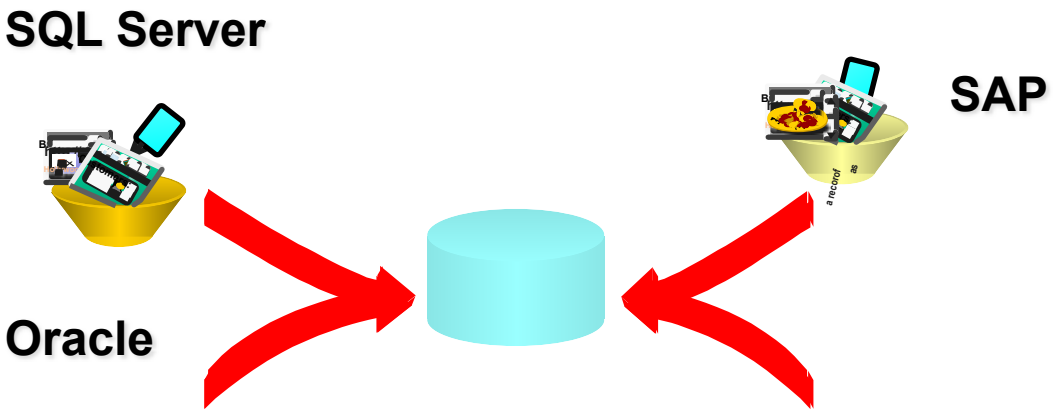
Fuentes de datos

- Producción
- Archivos planos
- Internas
- Externas



Extracción

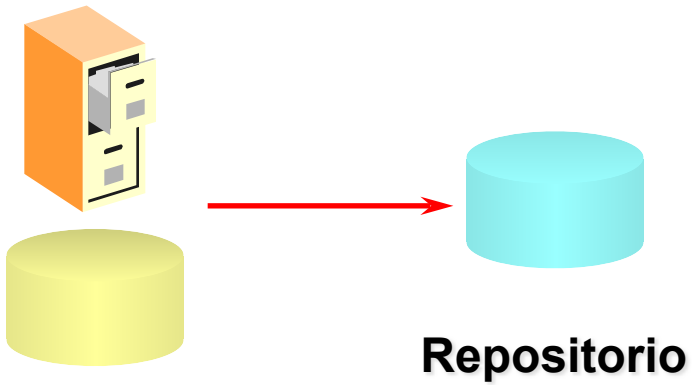
Fuentes de datos. Producción



- Sistemas operacionales (OLTP) de bases de datos

Extracción

Fuentes de datos. Archivos

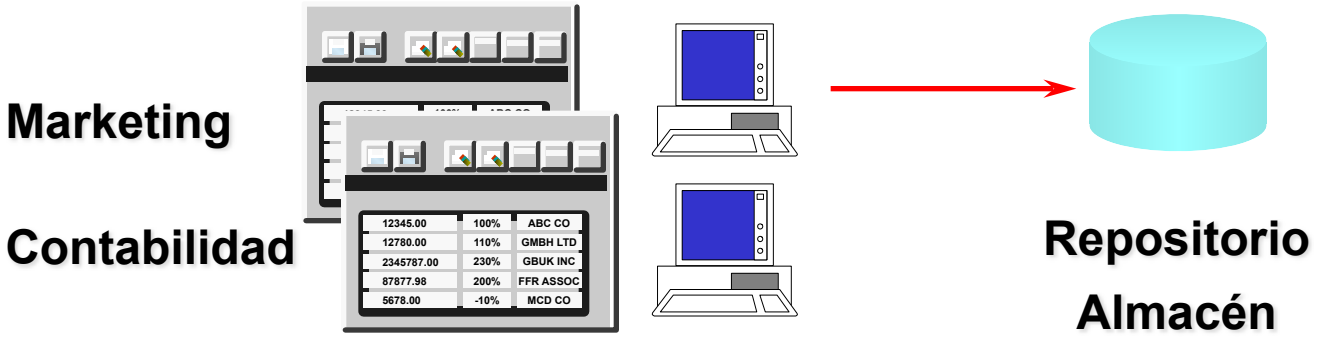


- Ficheros planos

Extracción

Fuentes de datos. Datos internos

- Información desde dentro de la organización

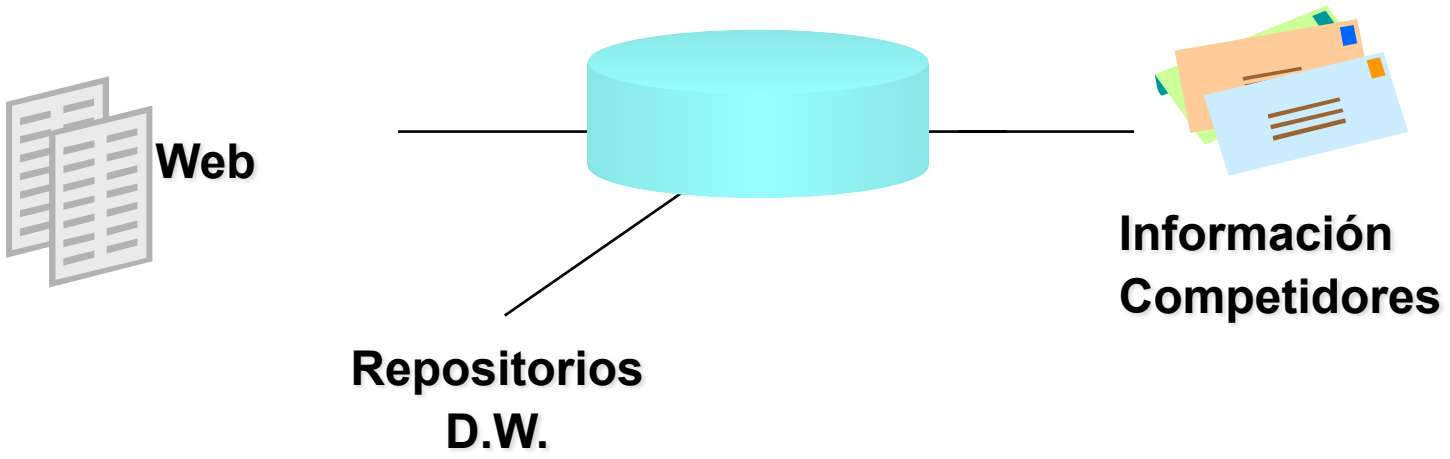


- Datos estructurados (Excel, …)
- No estructurados

Extracción

Fuentes de datos. Datos externos

- Información desde fuera de la organización

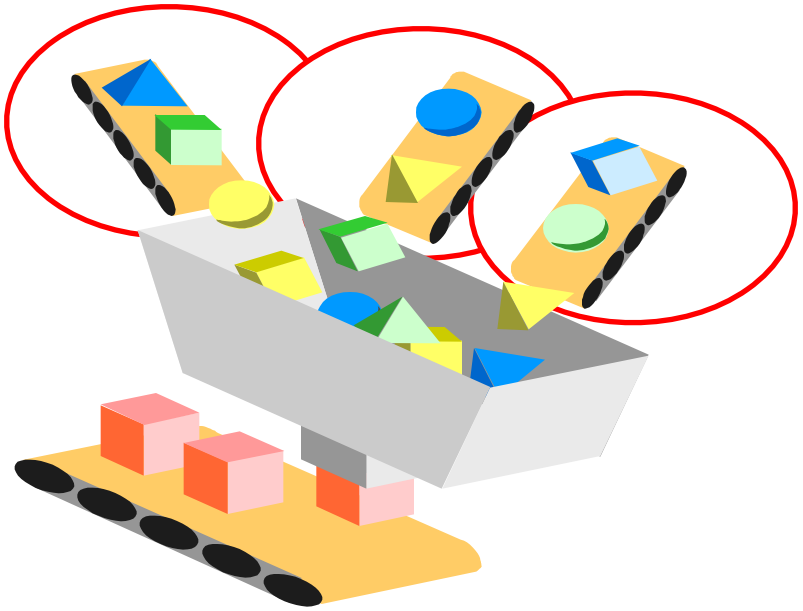


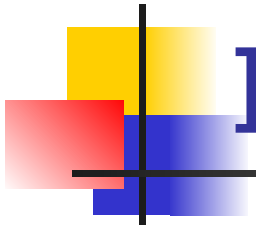
- Otras bases de datos
- Otros almacenes de datos
- …

Extracción

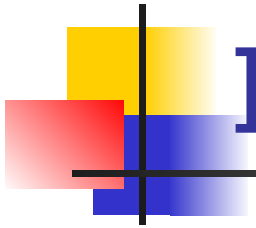
Técnicas de extracción

- Programas - C, PL/SQL …
- Herramientas
  - Limpieza de datos
  - Automatización
  - Coste inicial muy alto

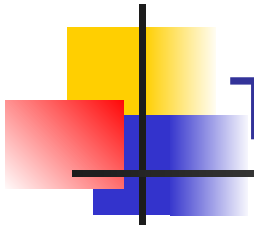




Indice



Indice



Transformación

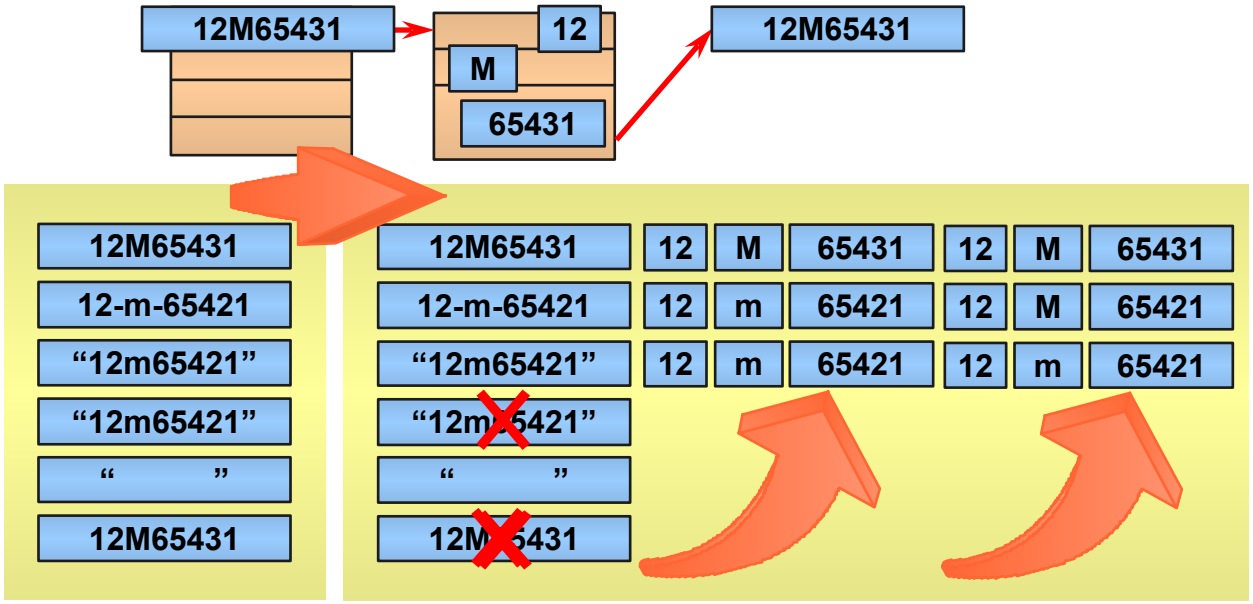
- Corrección de anomalías de los orígenes de datos
  - Limpieza de datos
- Ejemplo:
  - No tiene clave primaria
  - Diferentes nombres para el mismo cliente
  - Diferente formato para la dirección del mismo cliente
  - Diferente dirección para el mismo cliente

CLINUM	NOMBRE	DIRECCION
90328575	Telefonica SA	Calle 12, Madrid
90328575	Telefonica	Calle doce, Madrid
90238475	Telefonica I+D	Calle 12, CP 28080, Madrid
90233479	Telefonica Moviles	C/ 12, 28080 Madrid
90233489	Telefonica España	Avenida Blasco Ibañez, Valencia
90234889	Telefonica UK	Av. Vicente Blasco Ibanyez, Valencia
90345672	Telefonica Internacional	Av. Blasco Ibanyez, Valencia

Transformación

Algunas transformaciones comunes: Generadores de clave única

- Generador de clave única



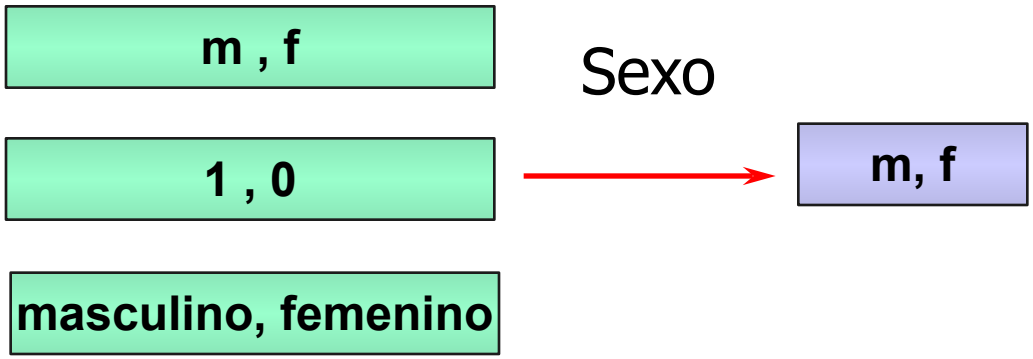
Código producto= 12M65431

Código país — **Zona Ventas** — **Número Producto**

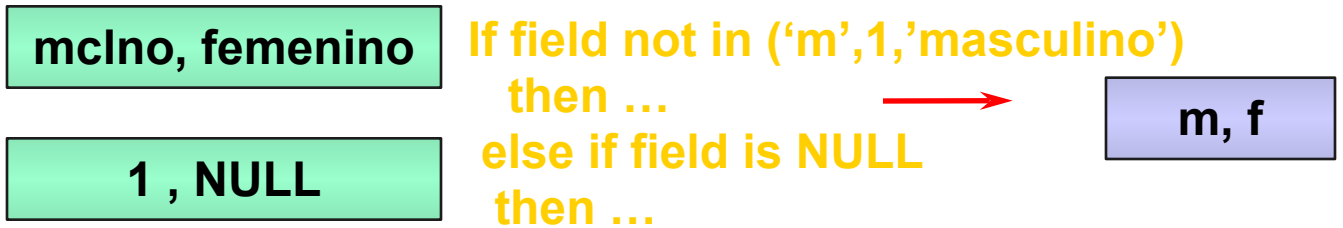
Transformación

Algunas transformaciones comunes: Codificación múltiple y detección de errores

- Codificación múltiple



- Detectar datos erróneos



Transformación

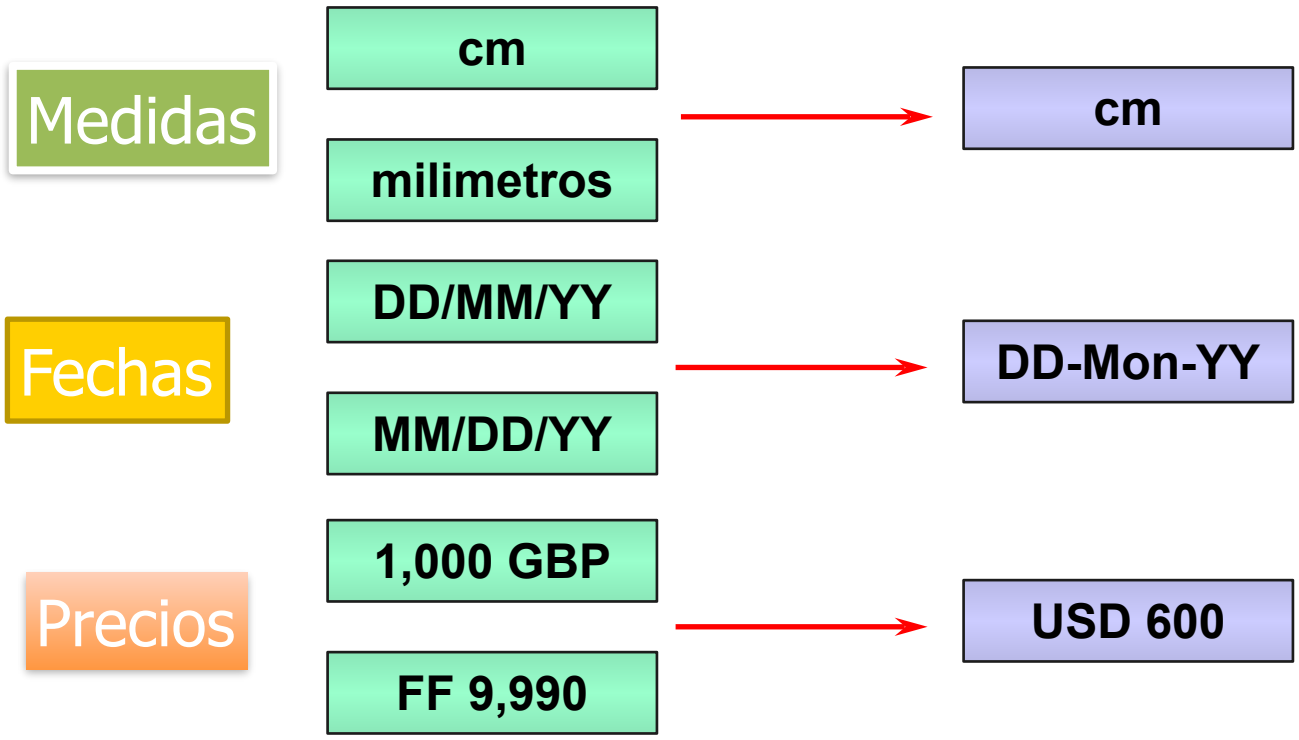
Anomalías de fuentes de datos

- Anomalías de instancias y codificado
- Inconsistencias de ortografía

CLINUM	NOMBRE	DIRECCION
90328575	Oracle Corp	100 NE 1st Street, Tampa
90328575	Oracle	100 NE. First St., Tampa
90238475	Oracle Services	100 North East 1st St., FLA
90233479	Oracle Limited	100 N.E. 1st St.
90233489	Oracle Computing	15 Main Road, Ft. Lauderdale
90234889	Oracle Corp. UK	15 Main Road, Ft. Lordadale, FLA
90345672	Oracle Corp UK Ltd	181 North Street, Key West, FLA

Transformación

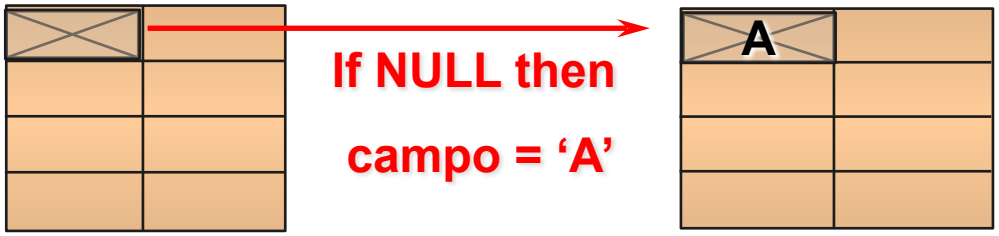
Algunas transformaciones comunes: Conversores



Transformación

Algunas transformaciones comunes: Filtrado

- NULL y valores que faltan

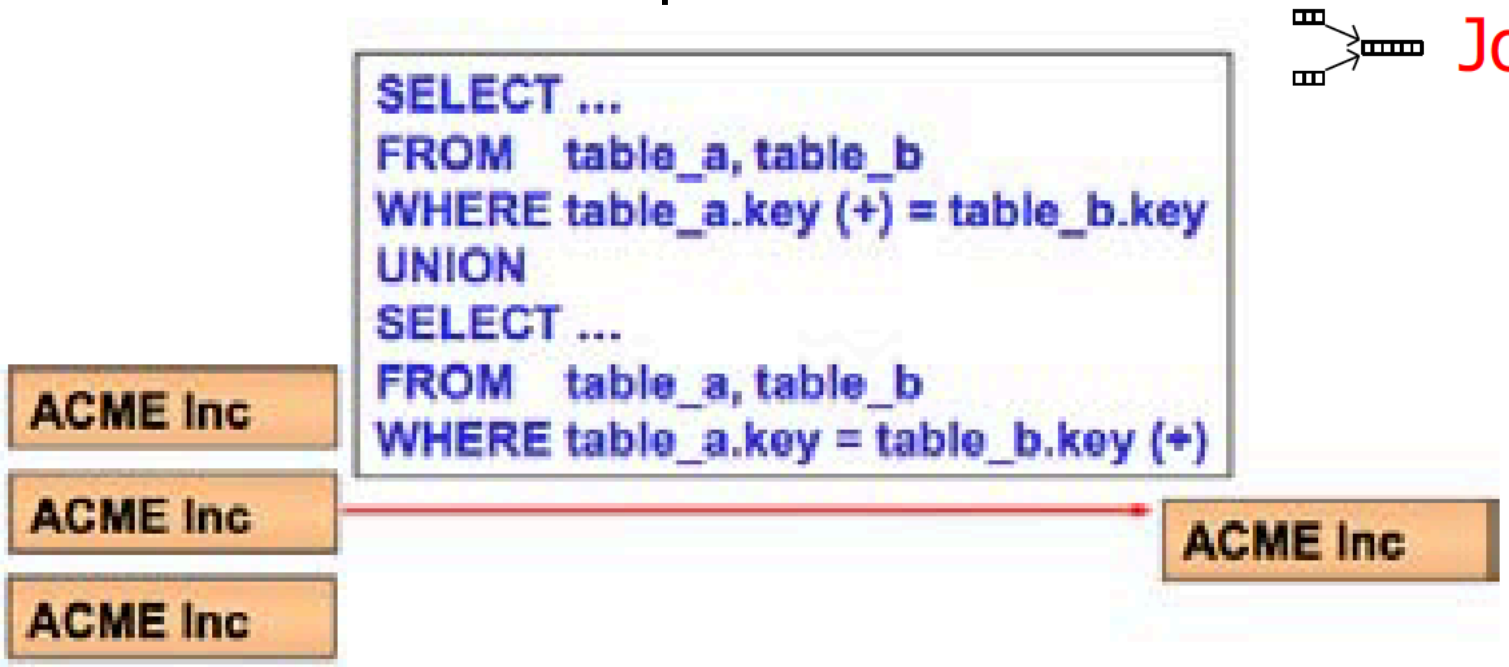


Filtrado: Operación que devuelve los datos que cumplen cierta condición

Transformación

Algunas transformaciones comunes: Union

Valores duplicados

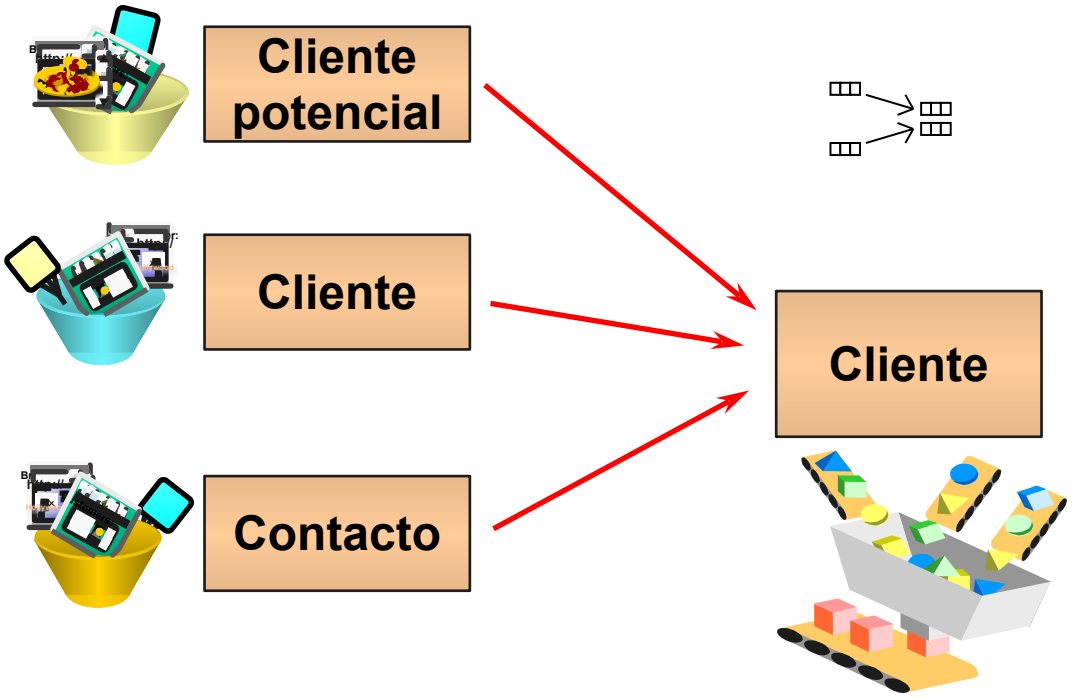


Diferentes filas de diferentes fuentes -> Una fila

Transformación

Algunas transformaciones comunes: Merge

- Atributos compatibles



Transformación

Algunas transformaciones comunes: Merge (fusión)

Fusiona diferentes datos de fuentes compatibles

#1	Venta	1/2/98	12:00:01 Pizza de jamón	$10.00
#2	Venta	1/2/98	12:00:02 Pizza de queso	$15.00
#3	Venta	1/2/98	12:00:02 Pizza de anchoas	$12.00
#4	Devol.	1/2/98	12:00:03 Pizza de anchoas	- $12.00
#5	Venta	1/2/98	12:00:04 Pizza de salchicha	$11.00

Valores de datos o claves artificiales

#dw1	Venta	1/2/98	12:00:01 Pizza de jamón	$10.00
#dw2	Venta	1/2/98	12:00:02 Pizza de queso	$15.00
#dw3	Venta	1/2/98	12:00:04 Pizza de salchicha	$11.00

Transformación

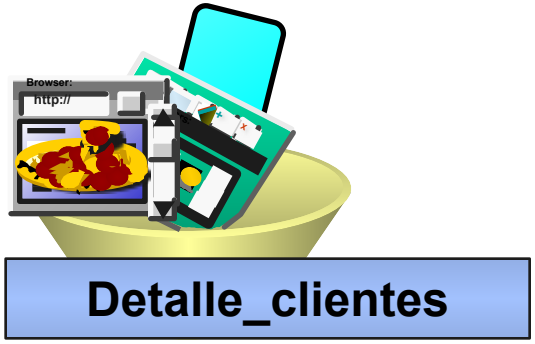
Algunas transformaciones comunes: Marcas de tiempo

- Marcado temporal de los datos para analizarlos a lo largo del tiempo: Hechos y dimensiones

Transformación

Algunas transformaciones comunes

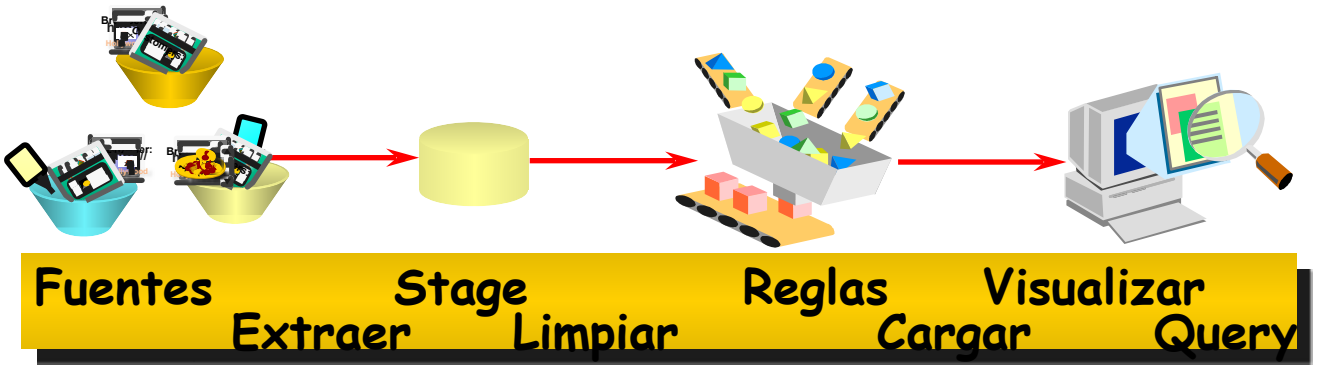
- Significado correcto de cada elemento



- **Evitar mala-interpretación**
- **Siempre documentar significado en METADATA**

Transformación

ETL. DIAGRAMA DE FLUJO DE PROCESOS



Transformación

Algunas transformaciones comunes. Ejemplo.

- Diferentes nombres y ortografía para el mismo miembro
- Diferente localización para el mismo miembro
- No hay clave única

	Nombre	Localización
Database 1	DIANNE ZIEFELD	N100
	HARRY H. ENFIELD	D589
	FRED AND SARA MULLEN	M300
Database 2	ZIEFLED, DIANNE	100
	ENFIELD, HARRY H	589
	MULLEN, SARA AND FRED	300