

Sistemas transaccionales: operativa -> soportan los procesos diarios que tiene la empresa, son los que generan data, son el origen de los datos, usan db OLTP

Sistemas de soporte de decisiones: interactúan con los sistemas transaccionales a través de los datos para la toma de decisiones, usan db OLAP

Por esto, ambos sistemas se complementan

Semana 2 – Integración de datos

En la actualidad, se pueden integrar los sistemas transaccionales con los sistemas de tomas de decisiones con un ETL

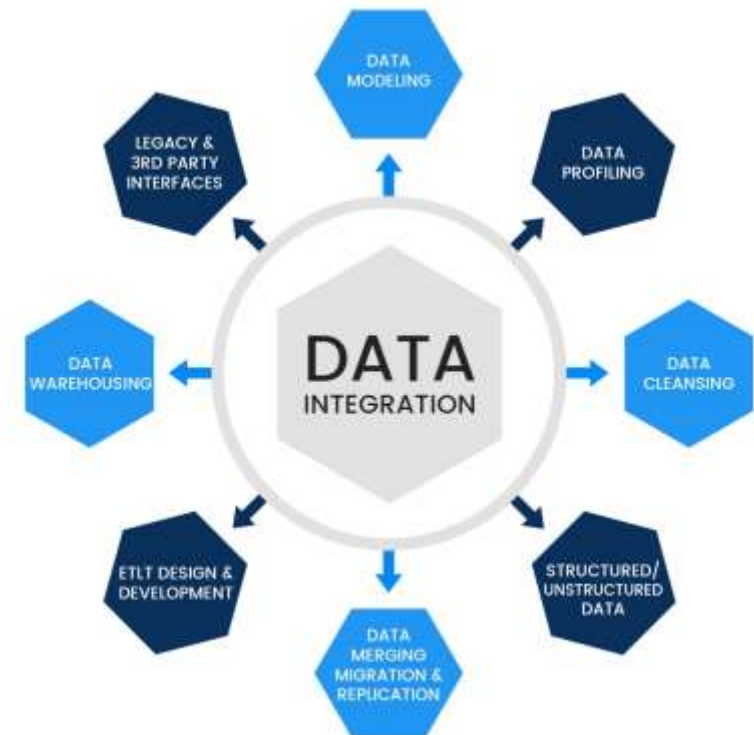
¿Qué es?

- La integración de datos es un proceso que involucra combinar datos de diferentes fuentes y permitir su consumo desde una sola interface.



¿Cómo le ayuda al negocio?

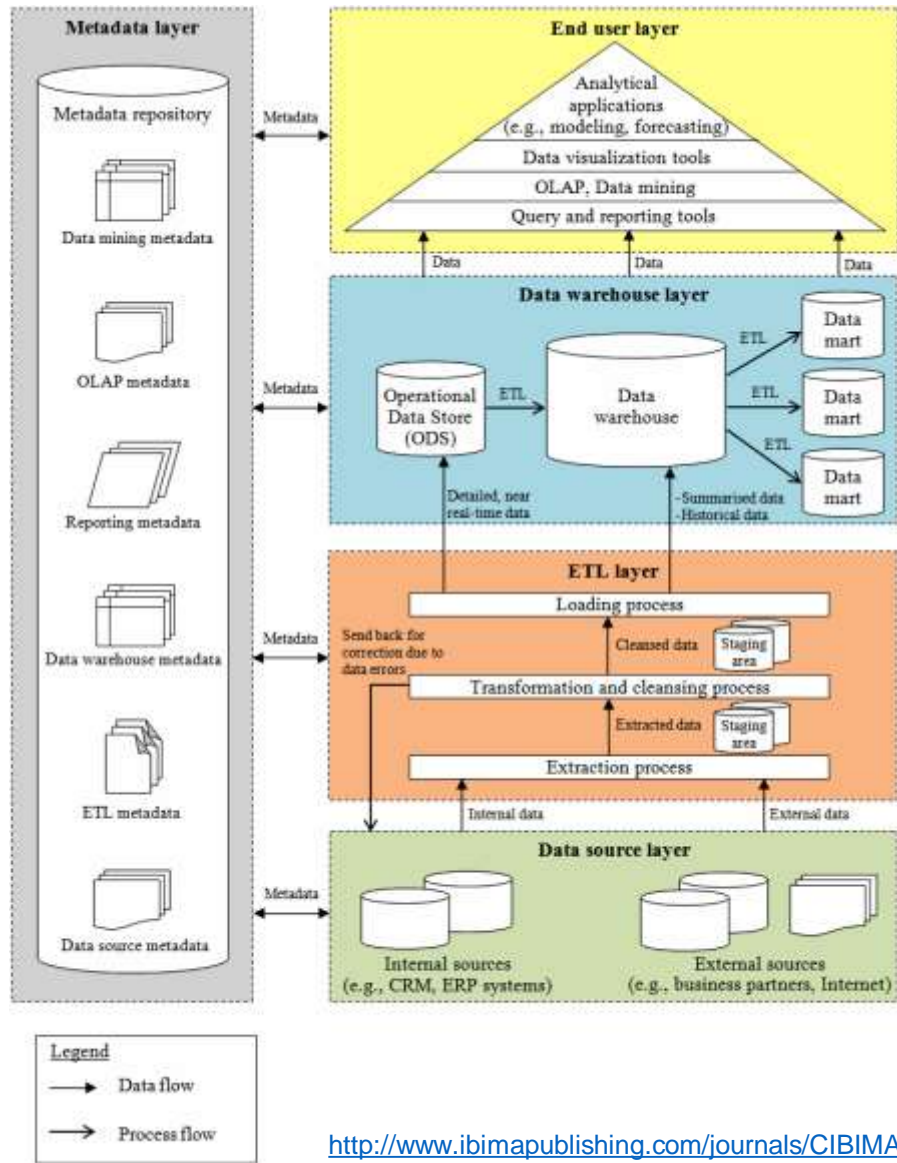
- Nos puede ayudar a descubrir espacios de eficiencia.
- Permite incrementar la competitividad.
- Mejora el proceso de toma de decisiones.



¿Qué es un ETL?

- Proceso para transferir, formatear, limpiar y cargar datos desde aplicaciones productivas (ERP, CRM, BDR, archivos) a los sistemas de BI (Data Marts, BDD, OLAP).
- Su construcción es responsabilidad del equipo TI.
- Se construye para cada BI/DW.
- Implica entre 40% al 60% del esfuerzo en la implementación de un DW.
- Se utiliza herramientas y programas diseñados ad hoc
- Extracción: Desde las diferentes fuentes de datos.
- Transformación: Convertir los datos al formato que se necesita en el proyecto analítico.
- Carga: Subir los datos a la base de datos analítica destino.

¿Cómo encaja en una aplicación de BI?

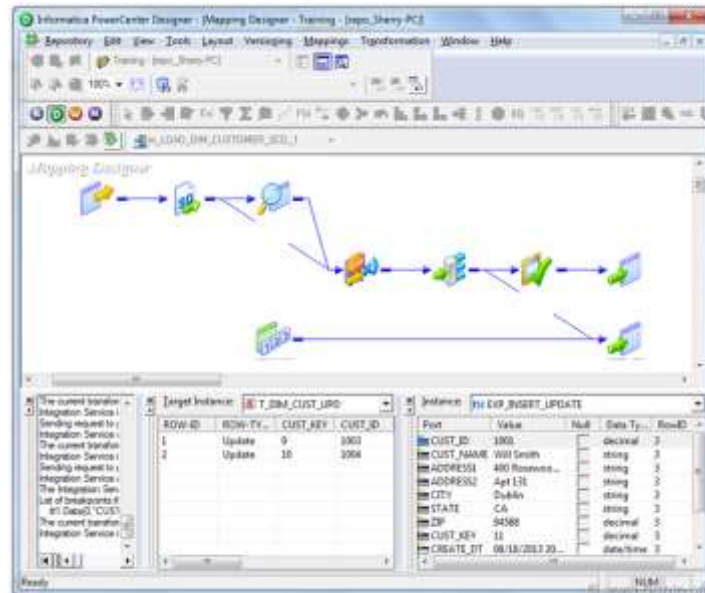
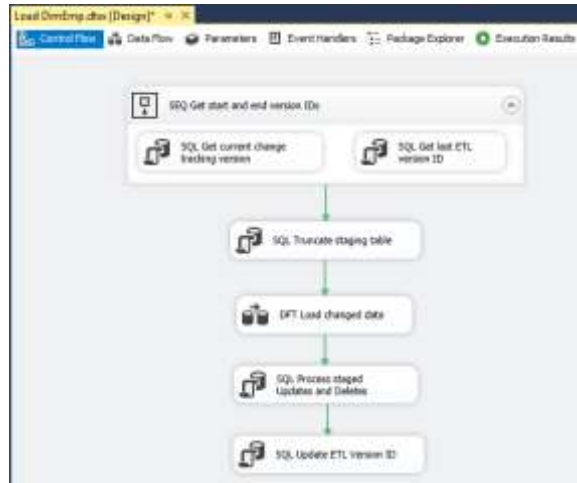


¿Qué es lo complicado en un ETL?

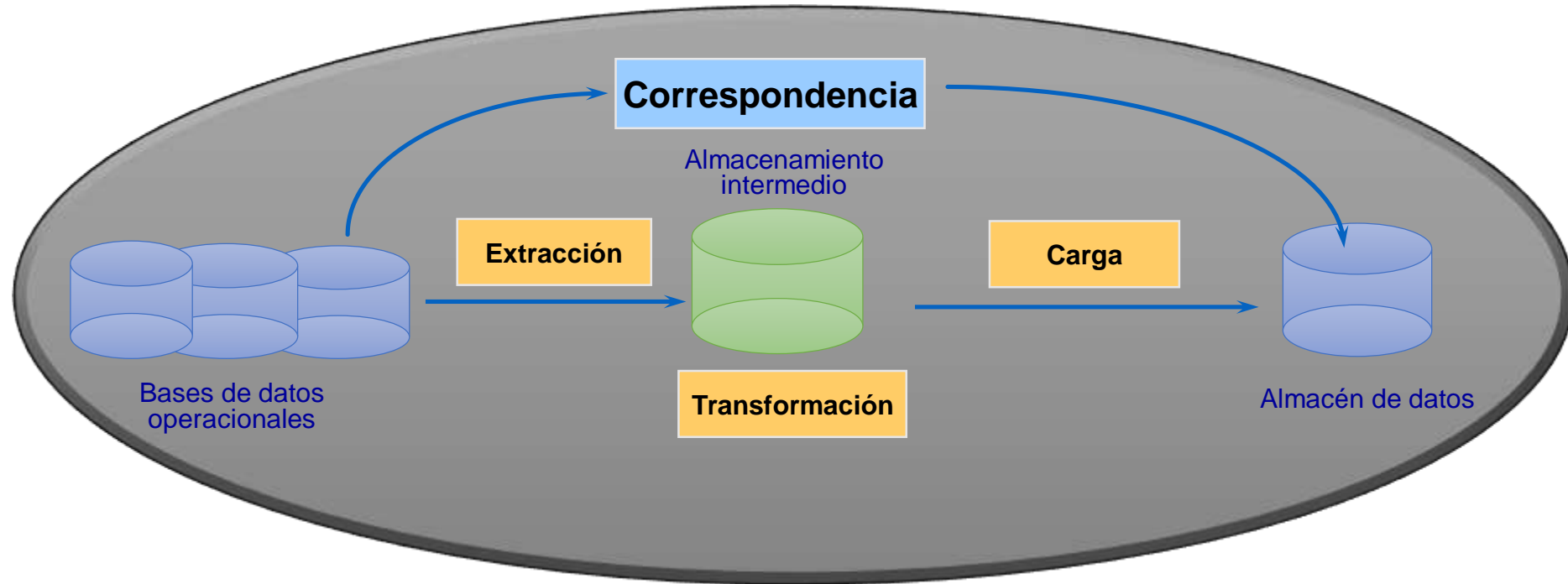
- Múltiples tecnologías se pueden tener muchos sistemas transaccionales al mismo tiempo
- Reportes obsoletos.
- No existe Metadata. 🗨️
- Diferentes algoritmos de calculo. Diferentes formatos para guardar la data en las distintas db
- Diferentes niveles de extracción.
- Diferentes niveles de detalle (granularidad).
- Diferentes nombres de campos de datos.
- Diferentes significados de campos de datos.
- Perdida de información.
- No exista reglas de corrección de datos.
- No exista capacidad de Drill Down.



Ejemplos

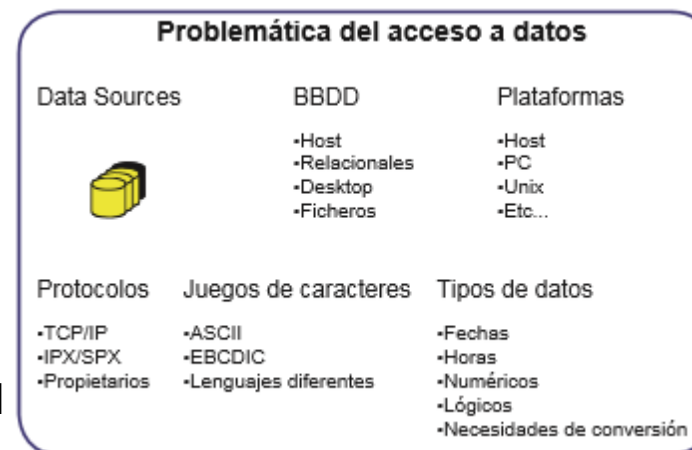


El proceso ETL



Extracción de datos

- Permite extraer los datos desde los sistemas de origen.
- Cada organización puede usar software diferente con datos y formatos
- Los formatos de las fuentes se encuentran en:
 - bases de datos relacionales
 - archivos planos
 - bases de datos no relacionales
- La extracción convierte los datos a un formato preparado para iniciar el
- Se analiza los datos extraídos y se verifica si cumplen estructura esperada o rechazados.



Extracción - Calidad de Datos

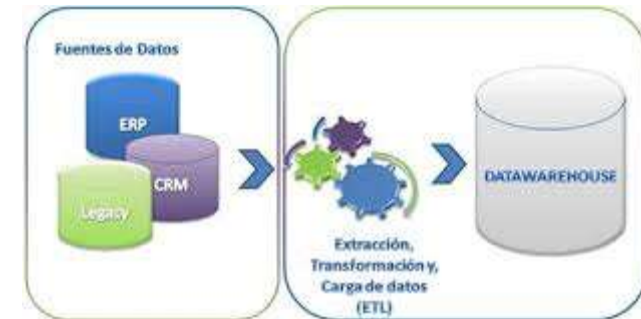
- La “calidad de los datos” es la clave del éxito de un almacén de datos.
- Definir una estrategia de calidad:
 - Actuar sobre los sistemas operacionales:
 - Modificar las reglas de integridad
 - Usar disparadores.
 - Documentar las fuentes de datos.
 - Definir procesos de transformación.
 - Nombrar un responsable de calidad de datos



Extracción de datos

- Requerimiento importante:

- Que el proceso de extracción cause mínimo impacto en el sistema origen. Si son muchos datos, el sistema origen puede reducir su desempeño o colapsar, evitando su uso cotidiano.

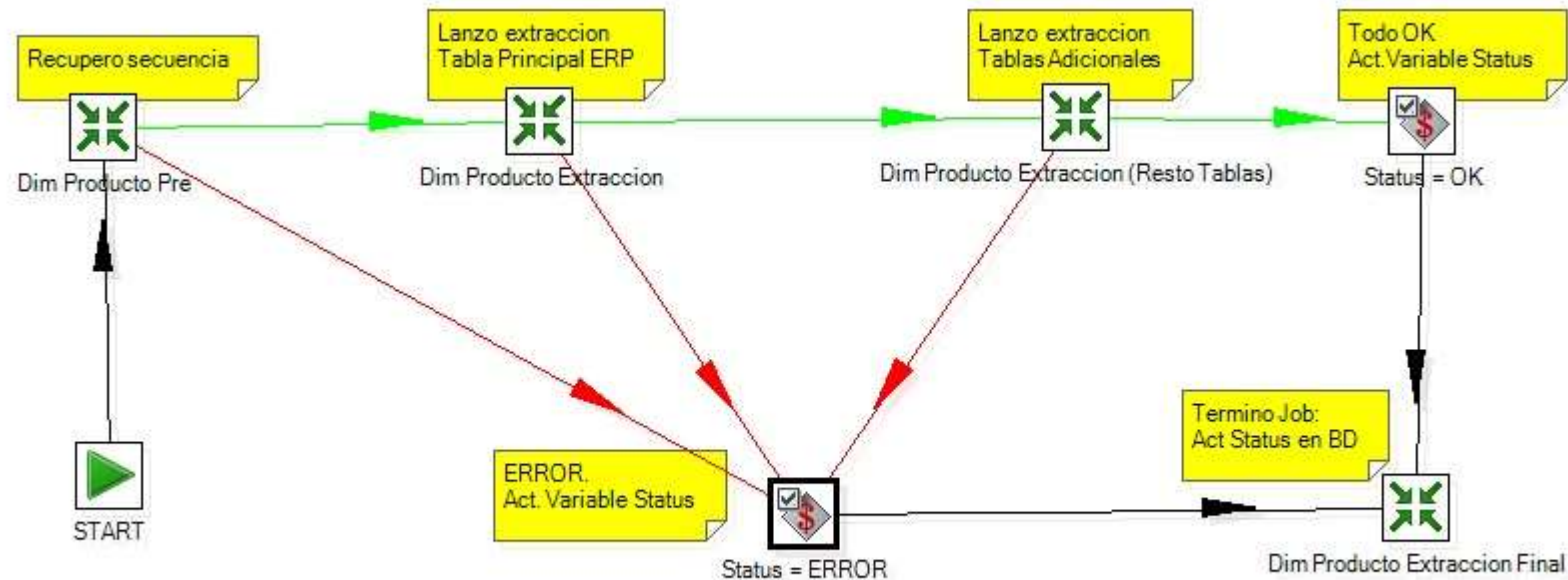


- Por esta razón:

- En sistemas grandes las operaciones de extracción suelen programarse en horarios o días donde el impacto sea nulo o mínimo.

Extracción de datos

- Programas diseñados para extraer datos de las fuentes.
- Herramientas: data migration tools, wrappers, ...



Extracción de datos

- Lectura de datos del sistema OLTP:
 - Durante la carga inicial .
 - Mantenimiento del BI/DW
- Ejecución de la extracción:
 - DBMS. Preparar consultas SQL o rutinas programadas.
 - Sistemas propietarios. Lectura y pre-procesamiento.
 - Texto libre. Lectura y pre-procesamiento.
 - Hipertextuales. Lectura, formateo y pre-procesamiento.
 - Hojas de cálculo. Lectura, formateo y pre-procesamiento.



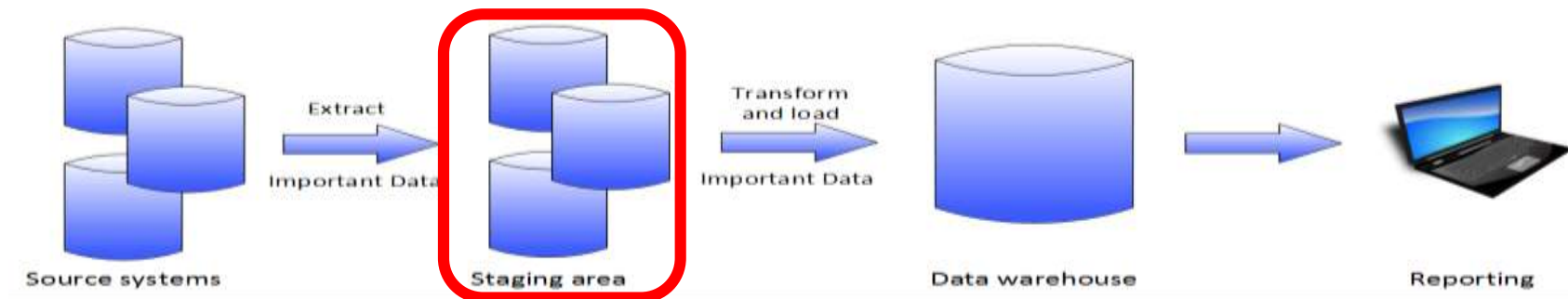
Identificación de cambios

- Identificar los cambios antes de extraerlos.
- Identificar los datos operacionales que se han modificado desde la fecha del último mantenimiento.
- Métodos
 1. Carga total. cada vez se empieza de cero.
 2. Comparación de instancias de la base de datos operacional.
 3. Uso de marcas de tiempo (time stamping) en los registros del sistema operacional.
 4. Uso de disparadores en el sistema operacional.
 5. Uso del fichero de log (gestión de transacciones) del sistema operacional.
 6. Uso de técnicas mixtas.

Data Staging Area

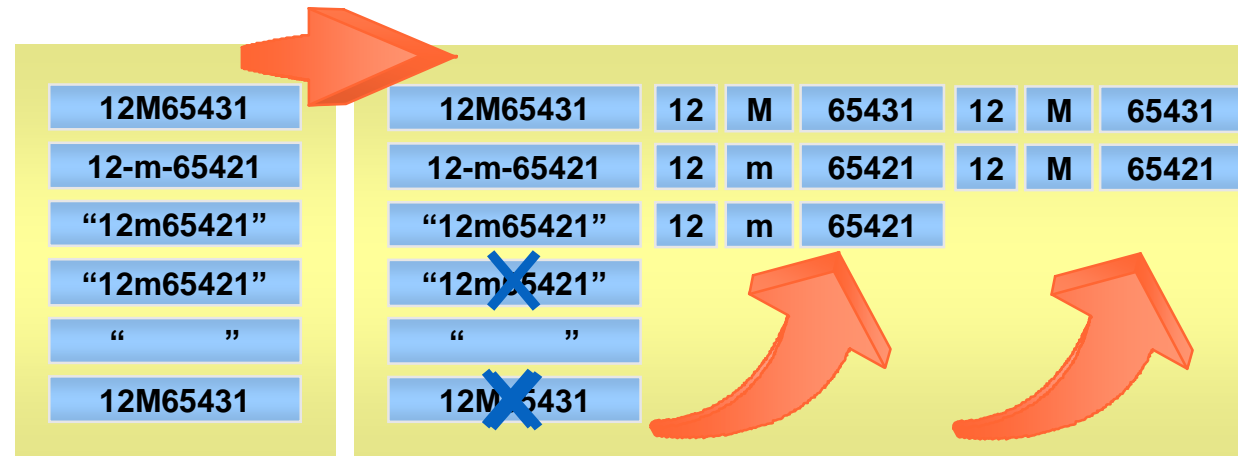
Espacio intermedio para transformar la data y convertilos en el formato que necesite el data warehouse para poder almacenarlos

- El Área de almacenamiento intermedio o Data Staging Area (DSA) es una ubicación temporal donde se copian los datos de los sistemas de origen. En las arquitecturas de almacenamiento de datos se requiere principalmente un área de ensayo por razones de tiempo. En resumen, todos los datos requeridos deben estar disponibles antes de que los datos puedan integrarse en el Almacén de Datos.
- Debido a ciclos de negocios variables, ciclos de procesamiento de datos, limitaciones de hardware y recursos de red y factores geográficos, no es factible extraer todos los datos de todas las bases de datos operativas al mismo tiempo.



Transformación - Cleansing (Limpieza)

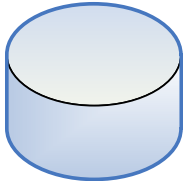
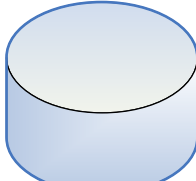
- En los datos operacionales de fuentes heterogéneas existen anomalías desarrolladas independientemente a lo largo del tiempo, por lo que es necesario eliminarlas:
 - **Limpieza de datos**: eliminar datos, corregir y completar datos, eliminar duplicados.
 - **Estandarización**: codificación, formatos, unidades de medida.



Transformación - Múltiple codificación

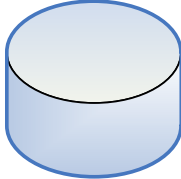
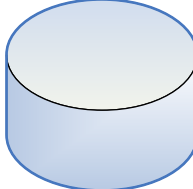
Mismos datos que están almacenados en diferentes db de diferentes maneras, hay que estandarizarlos

- Codificación y descripción del genero del individuo.
- Almacenado de diferentes maneras: “M” y “F”, “1” y “0”, “Hombre” y “Mujer” ó “Masculino” y “Femenino.”
- En la transformación, habrá que elegir una convención única para el DW, que puede ser “M” y “F” y transformar los datos.

 Operacional	 Data Warehouse
Aplicación A: M y F	M - F
Aplicación B: 1 y 0	
Aplicación C: Masculino y Femenino	

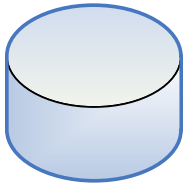
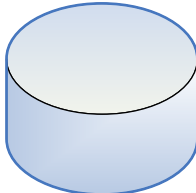
Transformación - Unidades de medida

- Los diferentes departamentos de la empresa pueden tener distintas unidades de medidas, según el origen del sistema OLTP. Un ejemplo es hablar de litro, centímetros cúbicos o hectolitros.
- Habrá que elegir una única unidad de medida que sea útil para el DW y transformar los datos.

 Operacional	 Data Warehouse
Aplicación A: litros	Litros
Aplicación B: cm ³	
Aplicación C: Hectolitros	

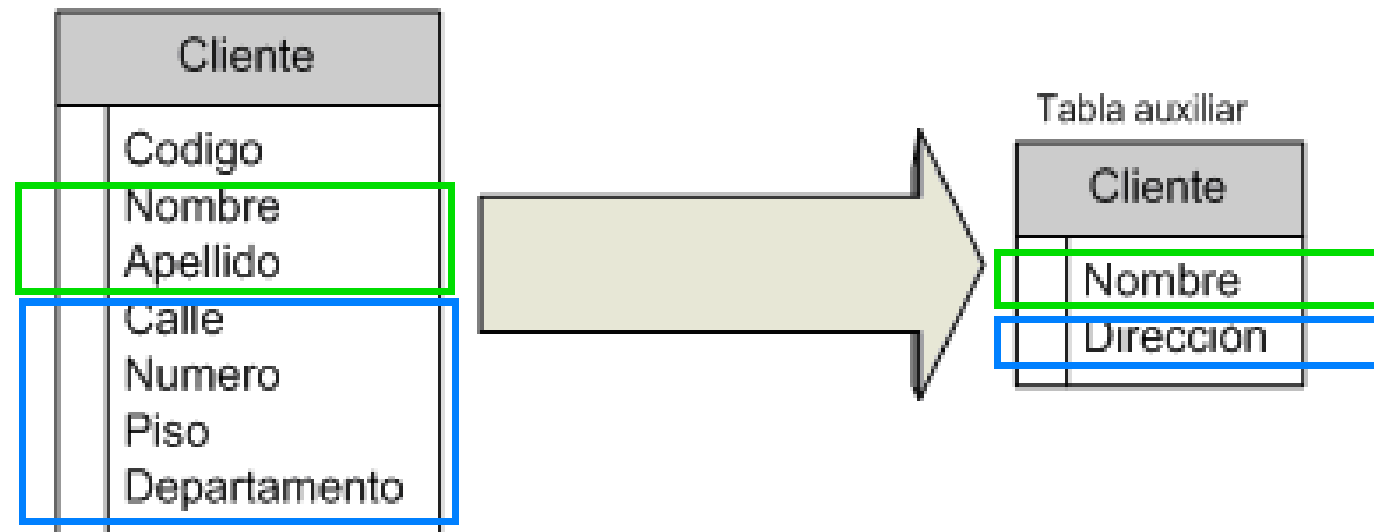
Transformación - Formatos

- Los formatos de fecha que encontramos en los diferentes sistemas operacionales pueden estar almacenados en multiples formatos.
- Las fechas pueden estar almacenadas como yyyy/mm/dd, mm/dd/yyyy ó dd/mm/yyyy.
- En el desarrollo del sistema DW, debemos elegir alguna de ellas y realizar la transformación correspondiente.

 Operacional	 Data Warehouse
Aplicación A: yyyy/mm/dd	dd/mm/yyyy
Aplicación B: mm/dd/yyyy	
Aplicación C: dd/mm/yyyy	

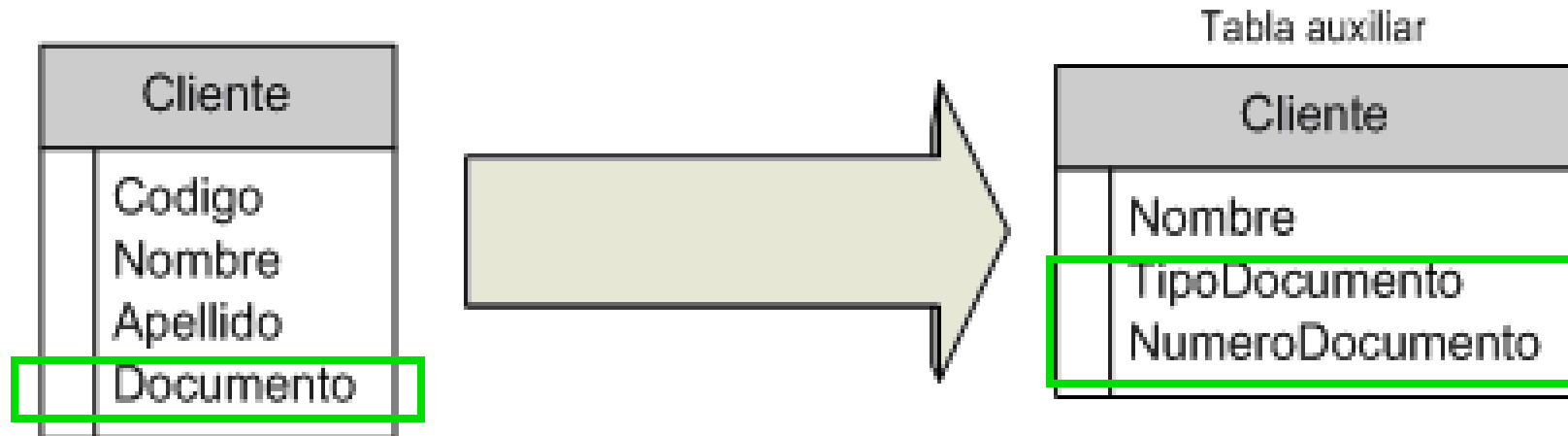
Transformación - Varias columnas en una

- Los datos de una persona, como dirección pueden almacenarse en diferentes campos de la misma tabla (Calle, Número, Piso y Departamento).
- En un sistema DW, es posible que los almacenemos en una única columna.
- Lo mismo puede suceder con el Nombre y Apellido.



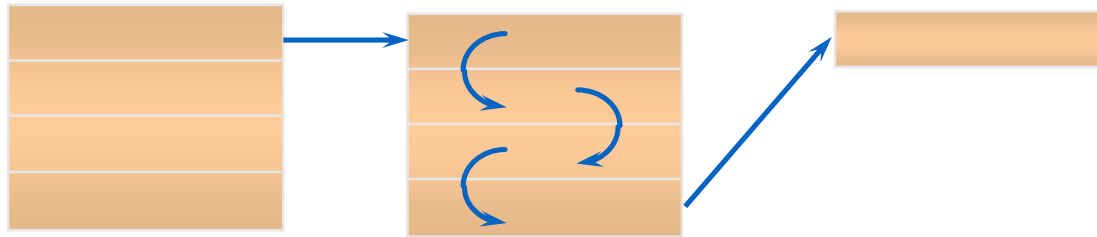
Transformación - Una columna en varias

- Los sistemas antiguos solían colocar el tipo y número de documento en el mismo campo de la tabla.
- En un DW, es posible que necesitemos colocar el tipo de documento en un campo y el número de documento en otro.



Transformación - Claves con estructura

- Claves con estructura: descomponer en valores atómicos



Código de producto = 12M65431345

código
del país

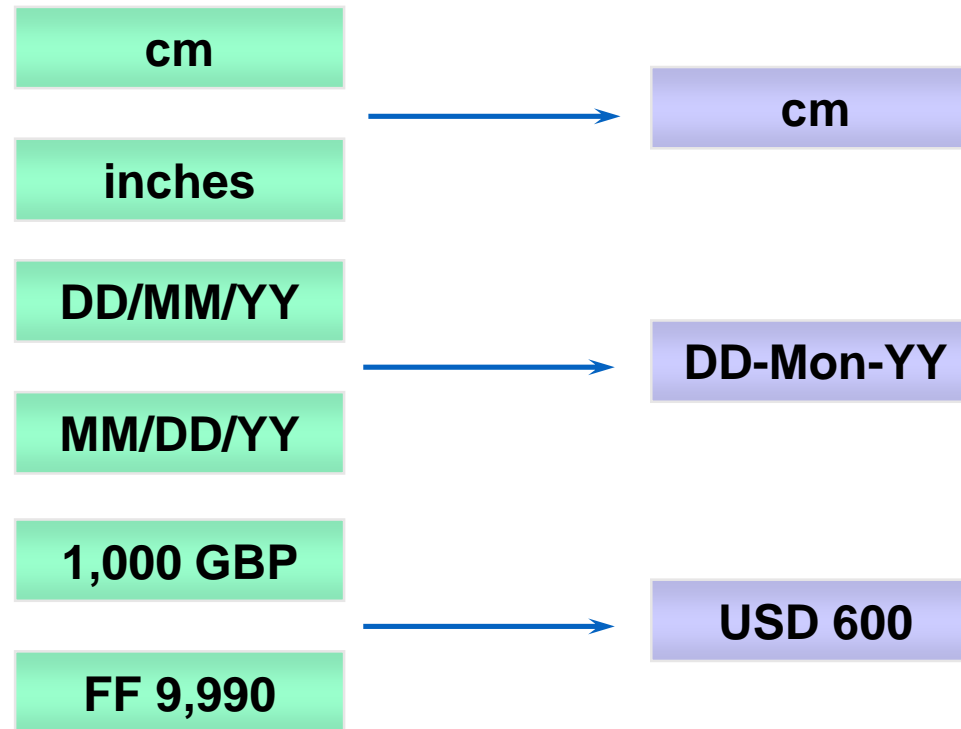
zona de
ventas

número de
producto

código de
vendedor

Transformación - Estándares

- Unificar estándares
- Unidades de medida, unidades de tiempo, moneda,...



Transformación - Duplicados

- Valores duplicados: deben ser eliminados.
- SQL restricciones en el SGBDR



Transformación - Integridad referencial

- Integridad referencial: debe reconstruirse.

Departamento	Emp	Nombre	Departamento
10	1099	Smith	10
20	1289	Jones	20
30	1234	Doe	50
40	6786	Harris	60

Transformación - Creación de claves

- Las claves primarias de las tablas no deben tener algún significado más allá de representar un identificador único.

#1	Venta	1/2/98	12:00:01 Ham Pizza	\$10.00
#2	Venta	1/2/98	12:00:02 Cheese Pizza	\$15.00
#3	Venta	1/2/98	12:00:02 Anchovy Pizza	\$12.00
#4	Devolución	1/2/98	12:00:03 Anchovy Pizza	- \$12.00
#5	Venta	1/2/98	12:00:04 Sausage Pizza	\$11.00

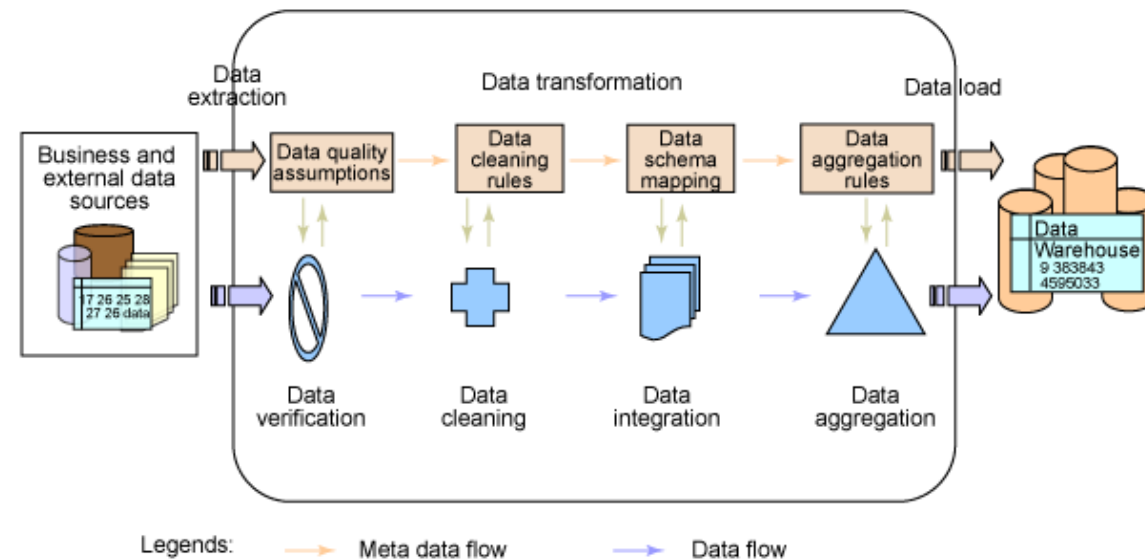


Claves sin significado

#dw1	Venta	1/2/98	12:00:01 Ham Pizza	\$10.00
#dw2	Venta	1/2/98	12:00:02 Cheese Pizza	\$15.00
#dw3	Venta	1/2/98	12:00:04 Sausage Pizza	\$11.00

Carga

- Es el momento en el cual los datos transformados son cargados en el sistema de destino.
- Consiste en mover los datos desde las fuentes operacionales o el almacenamiento intermedio hasta el almacén de datos y cargar los datos en las tablas estructuras de datos.
- Puede consumir mucho tiempo.



Carga



Existen dos etapas del proceso de carga:

Carga inicial. Mueve grandes volúmenes de datos.

Mantenimiento periódico. Mueve pequeños volúmenes de datos.



Existen dos formas de desarrollar el proceso de carga:

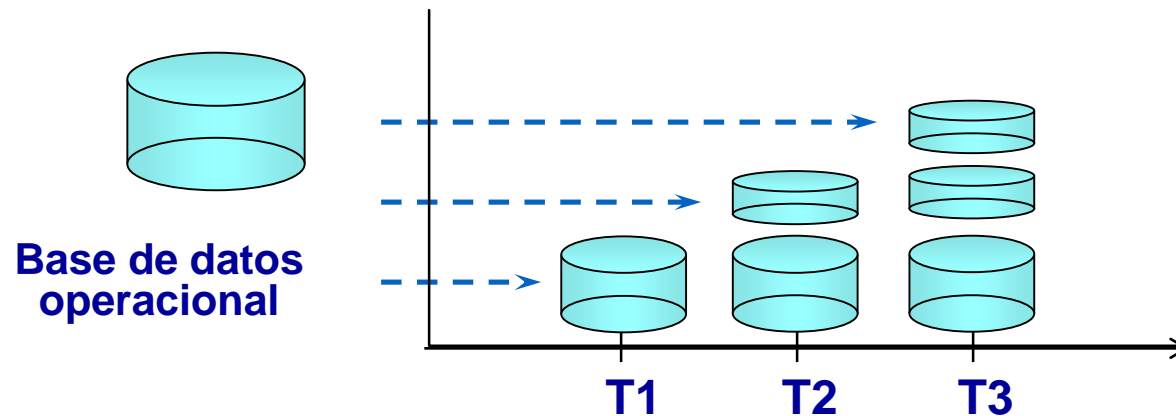
TAL (Trunc and Load): Limpia el repositorio de datos y carga la información de nueva cuenta.

Incremental: Se utiliza cuando se carga información nueva o información que necesita ser actualizada.

Frecuencia de mantenimiento

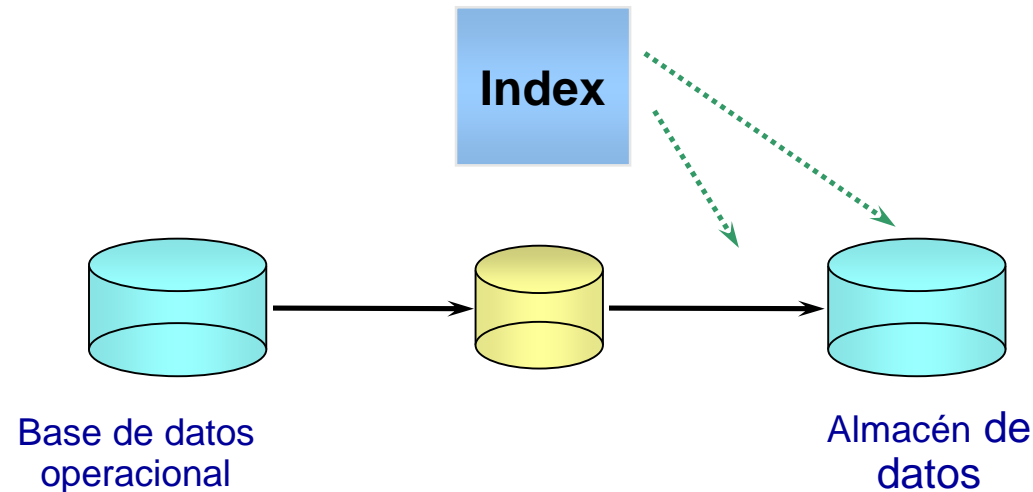
- La frecuencia del mantenimiento periódico está determinada por la granularidad de las dimensiones y los requisitos de los usuarios.
- Se deben determinar las “ventanas de carga” más convenientes para no saturar la base de datos operacional.
- Ocasionalmente archivar o eliminar datos obsoletos que ya no interesan para el análisis.

Hay que darle mantenimiento a la data de acuerdo a la frecuencia de cambio de esta



Indización

- Determinar cómo se van a generar los índices:
- Generación del índice **durante la carga**.
 - carga con el índice habilitado
 - proceso tupla a tupla. (lento)
- Generación del índice **después de la carga**:
 - carga con el índice deshabilitado
 - creación del índice (total o parcial). (rápido)



Integridad Referencial

- Deshabilitar Integridad referencial.
- Para agilizar el proceso de transferencia de datos.
- Cada vez que se inserta un registro no se efectúan las validaciones de integridad referencial.
- Al finalizar el proceso se habilita la integridad referencial.

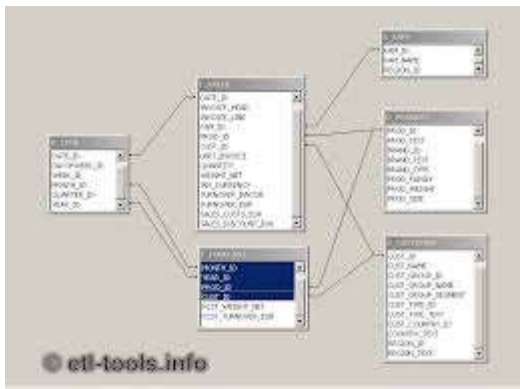
La **integridad referencial** garantiza que una entidad (fila o registro) siempre se relaciona con otras entidades válidas (que existen en la BD).

Dichos datos serán correctos, sin repeticiones innecesarias, datos perdidos y relaciones mal resueltas.

CLIENTE				CUENTA	
NOMBRE	CEDULA	CUENTA	CIUDAD	CUENTA	SALDO
CANO	7.245.310	C-101	CALI	C-101	50.000
PEREZ	1.352.851	C-121	PASTO	C-121	120.000
TORO	9.874.115	C-203	BOGOTA	C-203	70.000
LOPEZ	9.765.398	C-302	BUGA	C-302	98.000
SERNA	2.458.698	C-109	TADO	C-209	42.000
VEGA	4.111.119	C-230	LIMA	C-109	108.500
CANO	7.245.310	C-209	CALI	C-230	59.000
PEREZ	1.352.851	C-209	PASTO		

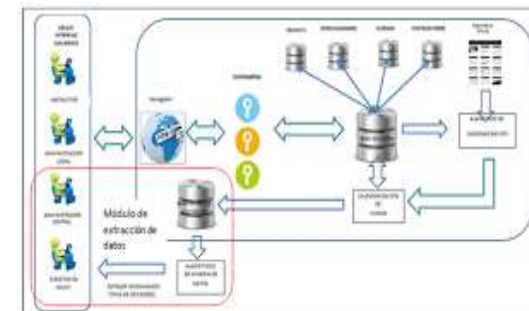
Herramientas de ETL – Principales funcionalidades

- **Diseño gráfico:** Entorno que permite a los desarrolladores establecer la relación entre las fuentes de datos, las transformaciones, los procesos y las tareas para desarrollar la carga. Los diseños se deben almacenar en un repositorio Metadata.
- **Gestión del Metadata:** Proveer un repositorio donde definir, documentar y gestionar la información del proceso ETL y su ejecución. El Metadata debería ser accesible también desde otras aplicaciones.
- **Extracción:** Extracción de la información mediante conectores, como ODBC65, SQL nativos de los distintos motores de bases de datos o ficheros planos. Los conectores deberían acceder al Metadata para determinar qué información extraer y cómo.



```
metadata[1] - Notepad
File Edit Format Help

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0
[transitional//EN">
<html>
<head>
<title>CMS200 Login information</title>
<meta name="keywords" content="CMS200: content
management; solution; business users; developers;
benefits">
</head>
<body>
</body>
</html>
```



Herramientas de ETL – Principales funcionalidades

- **Transformación:** Deberían proveer de librerías de transformación que permitan a los desarrolladores transformar los datos origen en los destino con las nuevas estructuras y crear las tablas de agregación para mejorar el rendimiento.
- **Carga:** Utilizar adaptadores para poder insertar o modificar los datos en el datawarehouse. • **Servicios de transporte:** Las herramientas ETL utilizan las redes y sus protocolos (por ejemplo: FTP, File Transport Protocol) para mover los datos entre las distintas fuentes y los sistemas destino.

