5. Herramientas ETL

5.1 Definición

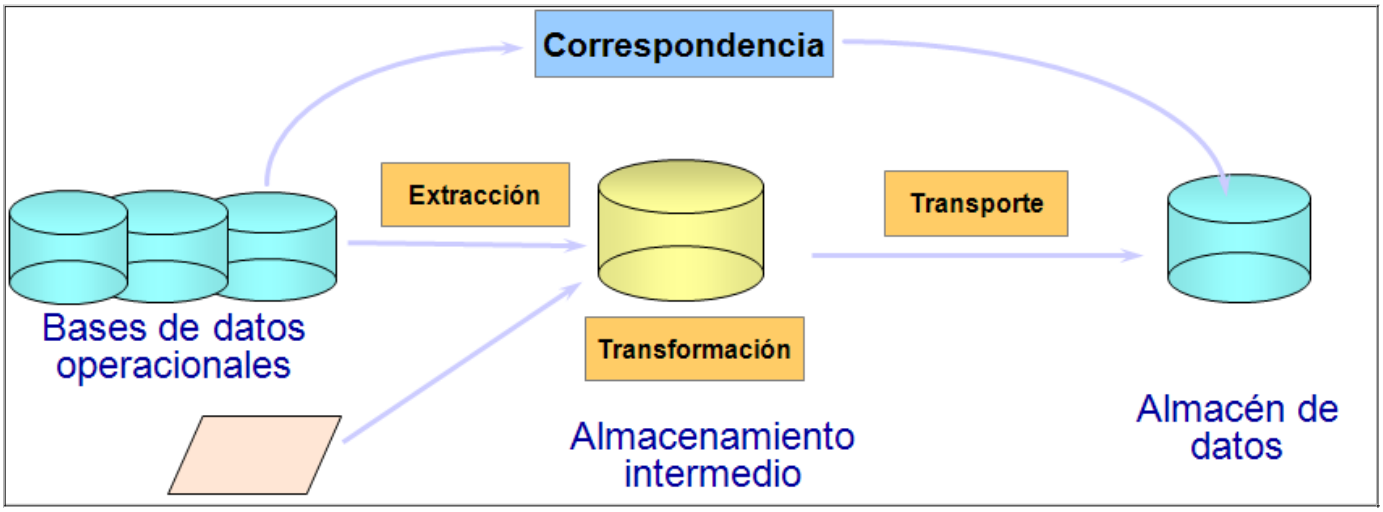
El sistema encargado del mantenimiento del almacén de datos es el Sistema E.T.L (Extracción - Transformación - Carga) o E.T.T (Extracción - Transformación -Transporte)

- La construcción del Sistema E.T.T es responsabilidad del equipo de desarrollo del almacén de datos.
- El Sistema E.T.T es construido específicamente para cada almacén de datos. Aproximadamente 50% del esfuerzo.
- En la construcción del E.T.T se pueden utilizar herramientas del mercado o programas diseñados específicamente.

.

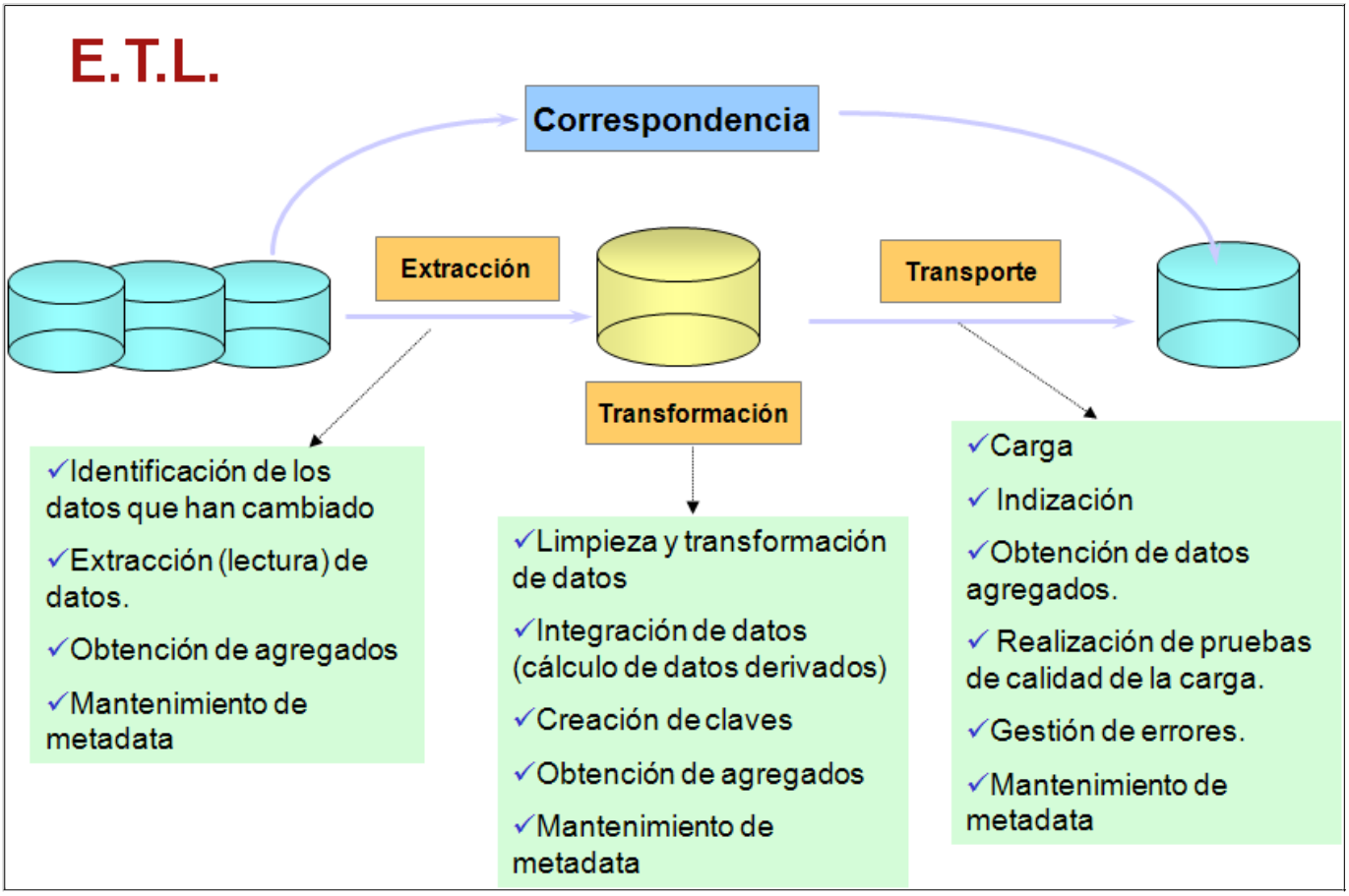
5.2 Funciones del Sistema E.T.T

- Carga inicial. (initial load)
- Mantenimiento o refresco periódico: inmediato, diario, semanal, mensual,... (refreshment)



El Almacenamiento intermedio permite:

- Realizar transformaciones sin paralizar las bases de datos operacionales y el almacén de datos.
- Almacenar metadatos.
- Facilitar la integración de fuentes externas.

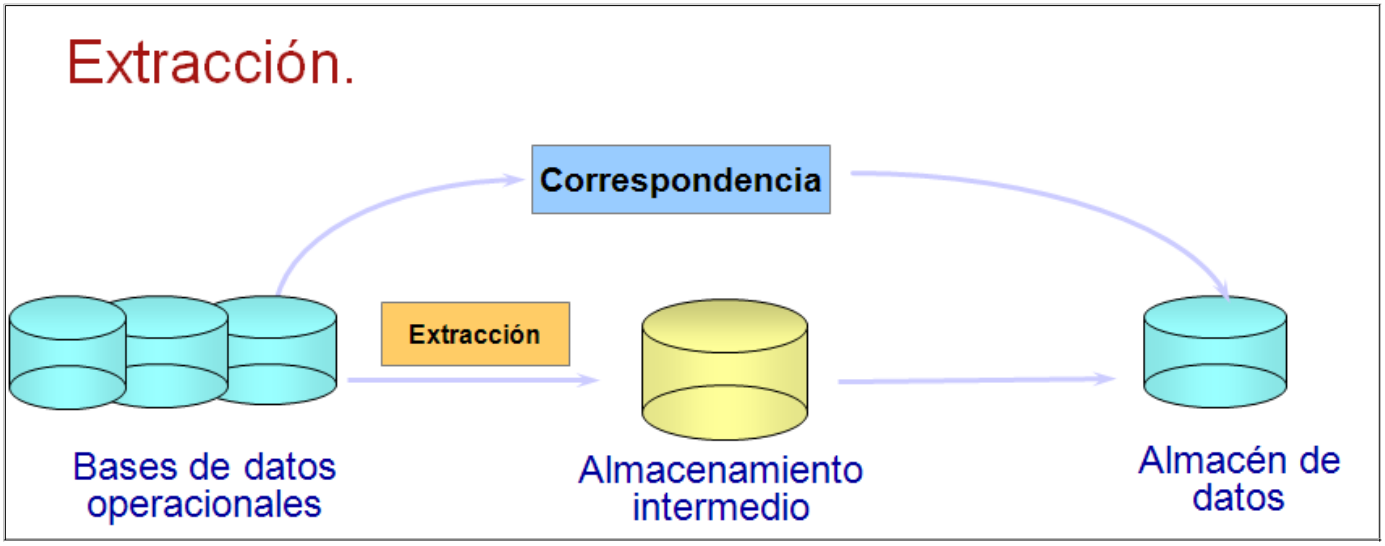


La “calidad de los datos” es la clave del éxito de un almacén de datos.

Definir una estrategia de calidad:

- actuación sobre los sistemas operacionales: modificar las reglas de integridad, los disparadores y las aplicaciones de los sistemas operacionales.
- documentación de las fuentes de datos.
- definición de un proceso de transformación.
- nombramiento de un responsable de calidad del sistema (Data Quality Manager).

5.3 Extracción



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

Extracción: lectura de datos del sistema operacional.

a) durante la carga inicial .

b) mantenimiento del DW.

Ejecución de la extracción:

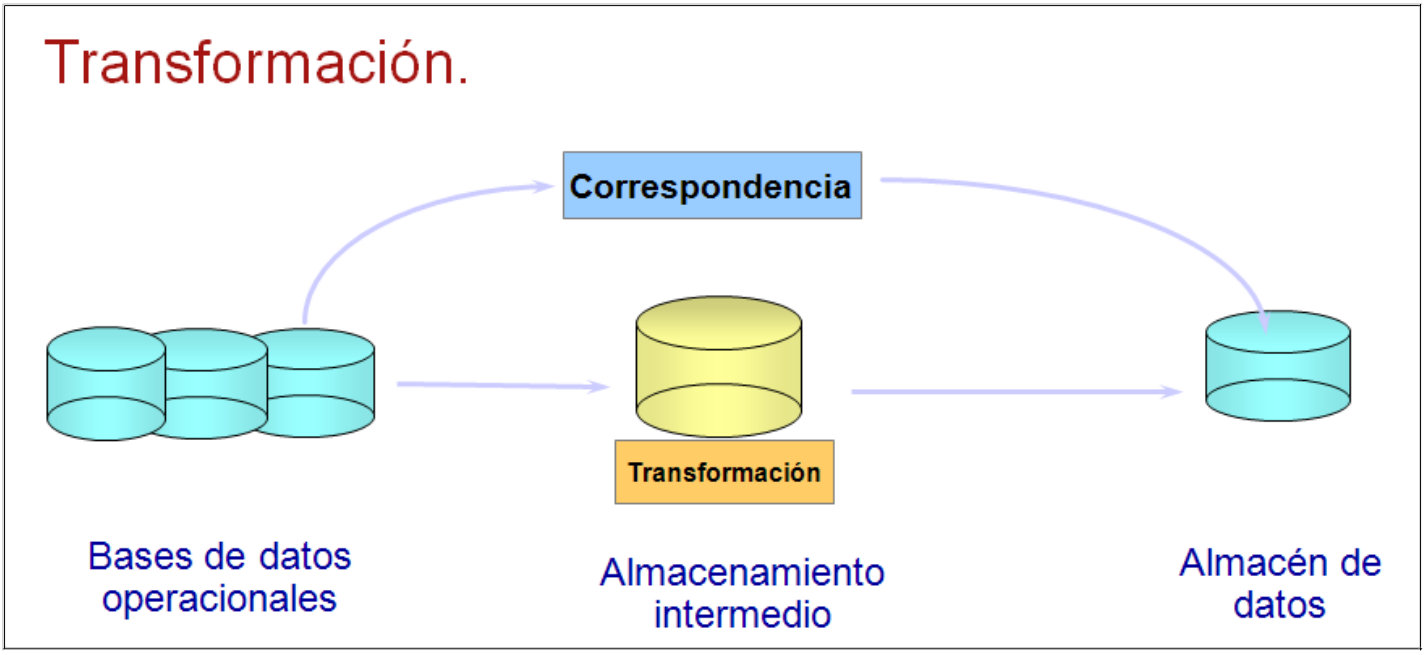
- a) si los datos operacionales están mantenidos en un SGBDR, la extracción de datos se puede reducir a consultas en SQL o rutinas programadas.
- b) si los datos operacionales están en un sistema propietario (no se conoce el formato de los datos) o en fuentes externas textuales, hipertextuales u hojas de cálculo, la extracción puede ser muy difícil y puede tener que realizarse a partir de informes o volcados de datos proporcionados por los propietarios que deberán ser procesados posteriormente.

Extracción: en el mantenimiento/refresco del DW. Antes de realizar la extracción es preciso Identificar los Cambios.

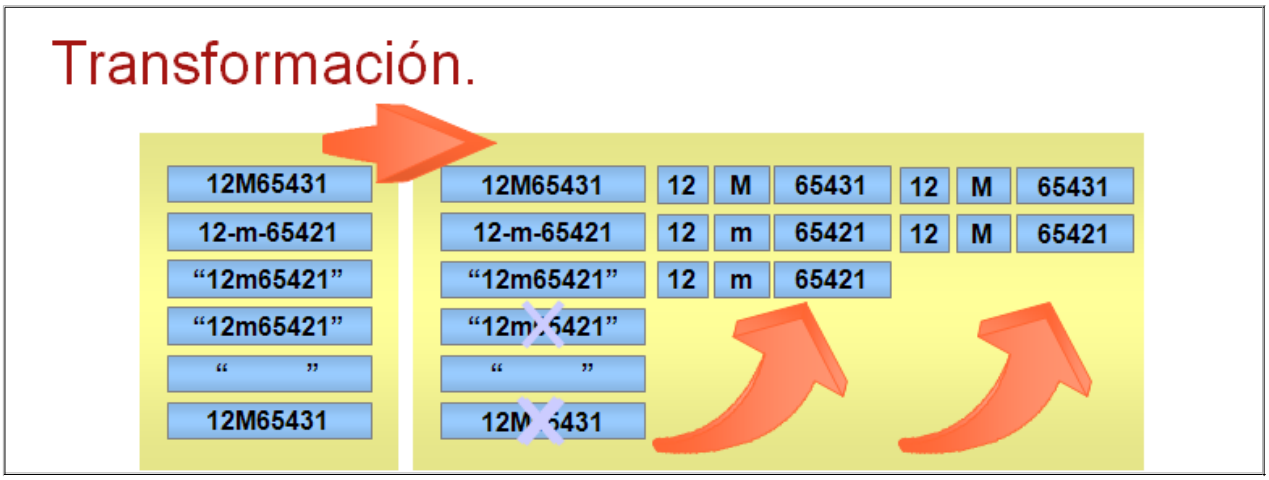
Identificación de Cambios.

- Identificar los datos operacionales (relevantes) que han sufrido una modificación desde la fecha del último mantenimiento.
- Métodos
  - Carga total: cada vez se empieza de cero.
  - Comparación de instancias de la base de datos operacional.
  - Uso de marcas de tiempo (time stamping) en los registros del sistema operacional.
  - Uso de disparadores en el sistema operacional.
  - Uso del fichero de log (gestión de transacciones) del sistema operacional.
  - Uso de técnicas mixtas.

5.4 Transformación



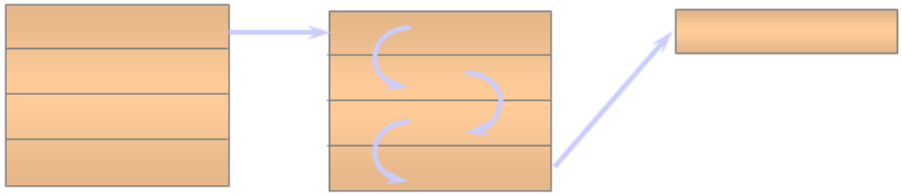
Transformar los datos extraídos de las fuentes operacionales: limpieza, estandarización. (cleansing)
Calcular los datos derivados: aplicar las leyes de derivación. (integration)



- En los datos operacionales existen anomalías: desarrollos independientes a lo largo del tiempo, fuentes heterogéneas, ..
- Eliminar anomalías:
  - Limpieza de datos: eliminar datos, corregir y completar datos, eliminar duplicados, ...
  - Estandarización: codificación, formatos, unidades de medida, ...

Transformación.

– Claves con estructura: descomponer en valores atómicos

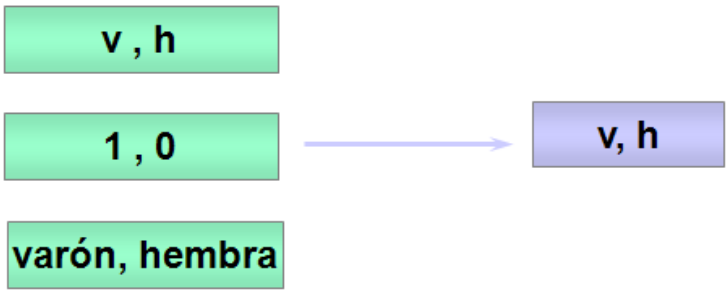


Código de producto = 12M65431345

código del país | **zona de ventas** | **número de producto** | **código de vendedor**

Transformación.

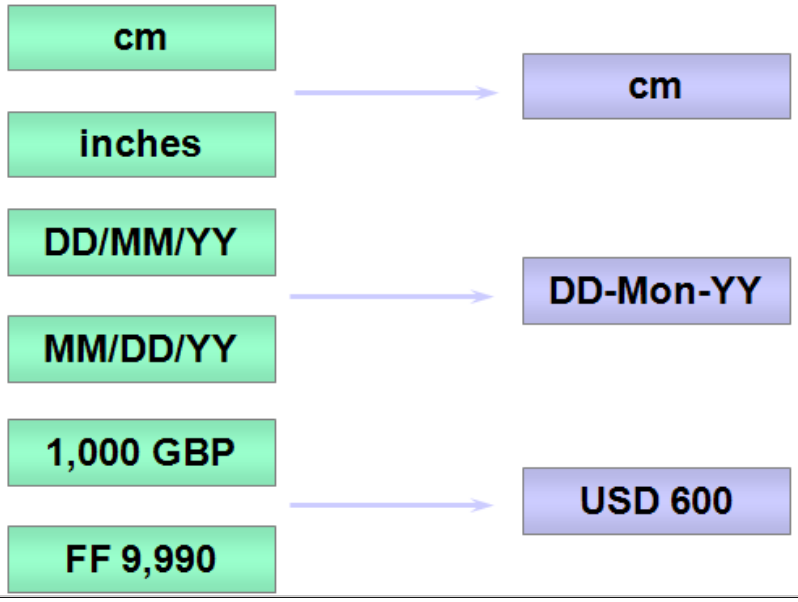
– Unificar codificaciones: existencia de codificaciones múltiples.



– Deben detectarse los valores erróneos.

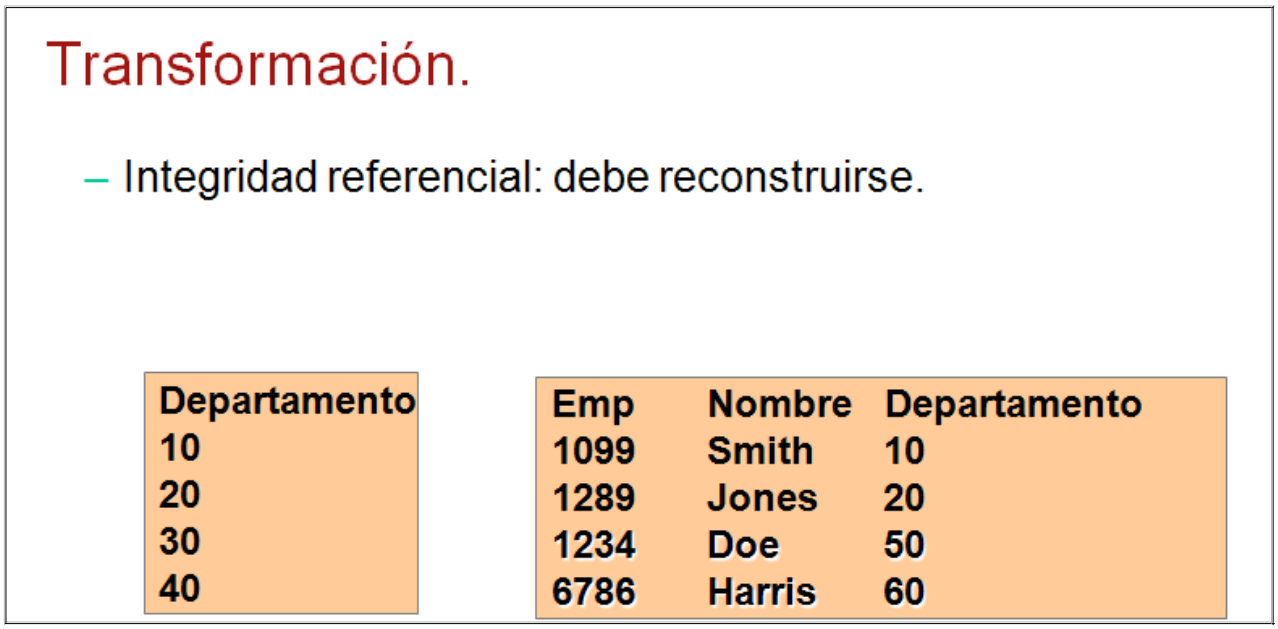
Transformación.

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



Transformación.

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



.

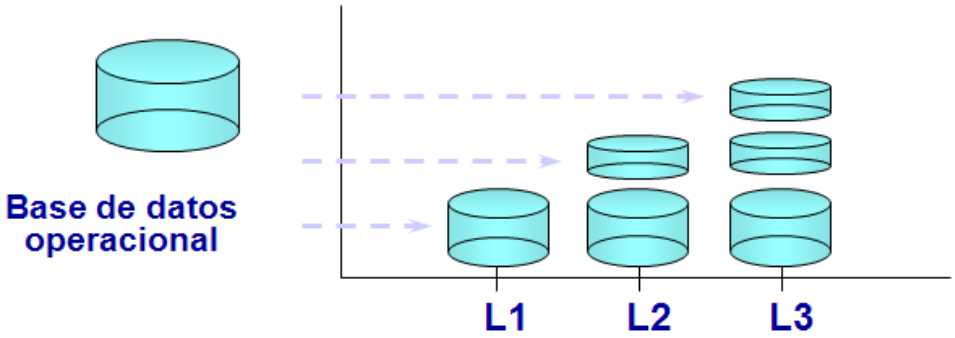
5.5 Carga

Load. (carga)

- La fase de Transporte consiste en mover los datos desde las fuentes operacionales o el almacenamiento intermedio hasta el almacén de datos y cargar los datos en las correspondientes estructuras de datos.
- La carga puede consumir mucho tiempo.
- En la carga inicial del DW se mueven grandes volúmenes de datos.

- En los mantenimientos periódicos del DW se mueven pequeños volúmenes de datos.
- La frecuencia del mantenimiento periódico está determinada por el gránulo del DW y los requisitos de los usuarios.

Carga. Creación y mantenimiento de un DW.

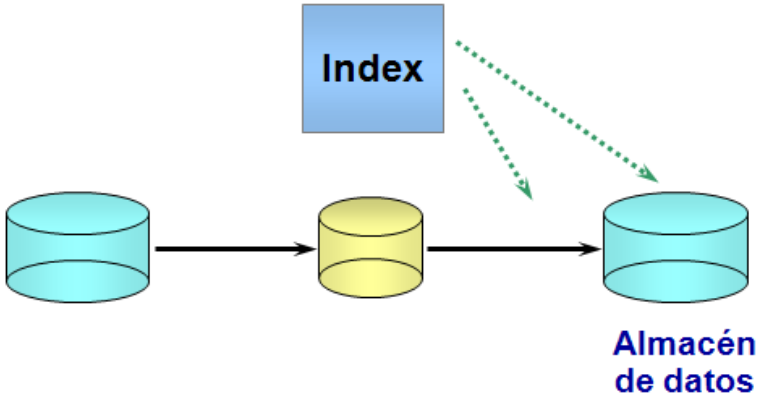


- Crear el AD (base de datos)
- En intervalos de tiempo fijos añadir cambios al AD. Se deben determinar las "ventanas de carga" más convenientes para no saturar la base de datos operacional.
- Ocasionalmente archivar o eliminar datos obsoletos que ya no interesan para el análisis.

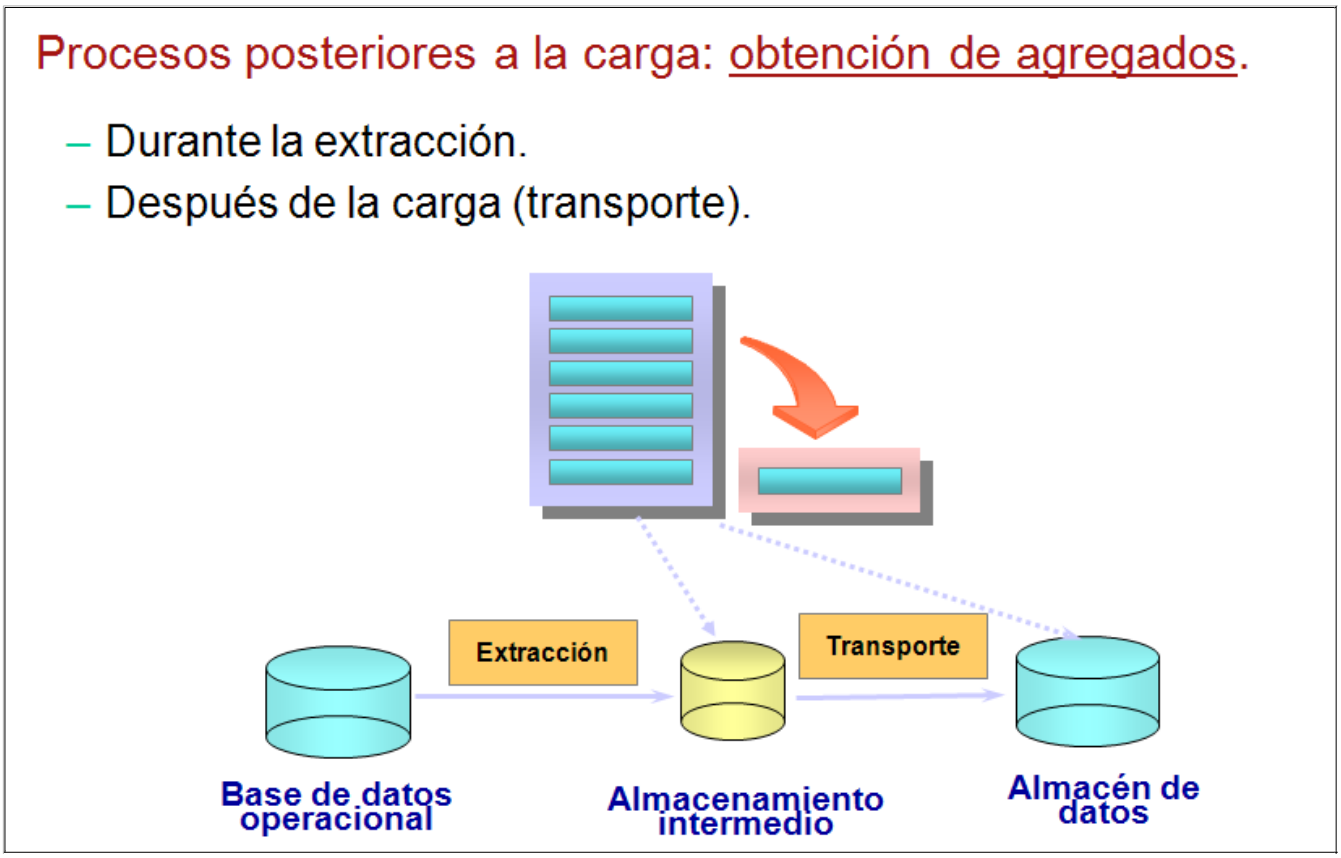
.

Procesos posteriores a la carga: indexación.

- Durante la carga:
  - carga con el índice habilitado
  - proceso tupla a tupla. (lento)
- Después de la carga:
  - carga con el índice deshabilitado
  - creación del índice (total o parcial). (rápido)



.



5.6 Herramientas

- Kettle (http://kettle.pentaho.org/)
- Talend (http://www.talend.com/)
- Clover (http://www.cloveretl.org/)