Sistemas de Información para el Análisis de GVDatos

Instituto de Computación - Facultad de Ingeniería

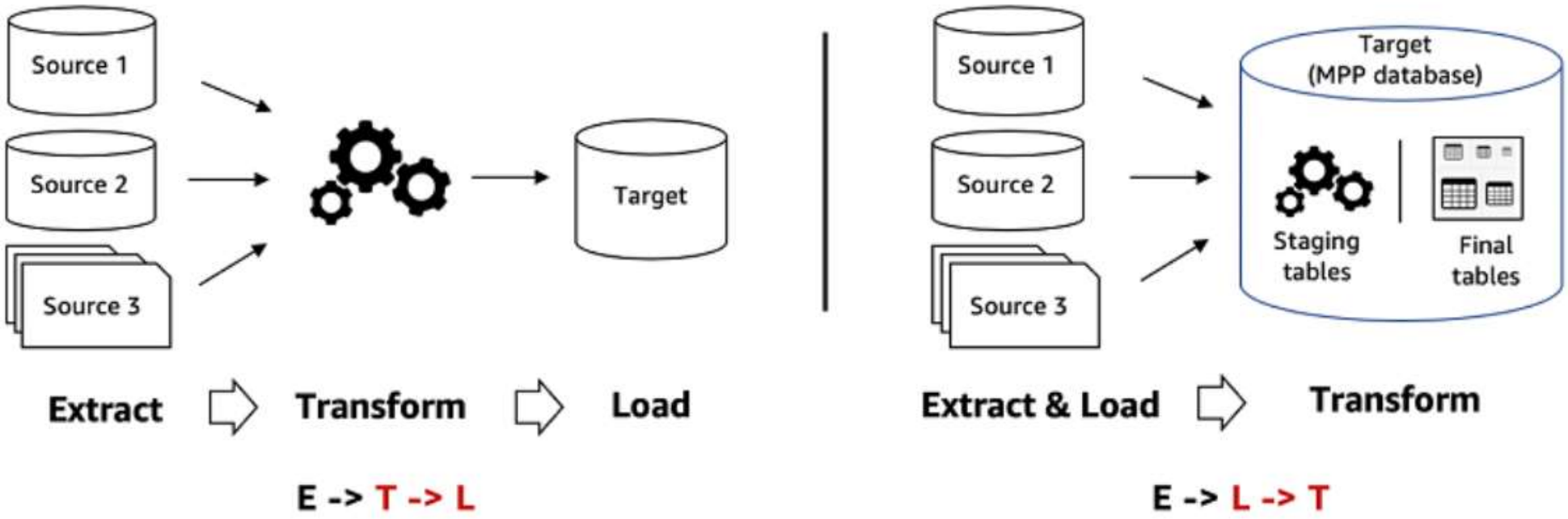
Mayo 2024

Carga y Actualización

Temario

- ETL y ELT
- Introducción
- Tipos de operaciones
- Flujos: datos y control
- Ciclo de vida de un DW
- Herramientas ETL
- Conclusiones

ETL y ELT



AWS Redshift. https://aws.amazon.com/es/blogs/big-data/etl-and-elt-design-patterns-for-lake-house-architecture-using-amazon-redshift-part-1/

ETL y ELT

- ETL
  - Transforma los datos antes de cargarlos en el DW con herramientas especializadas.
  - DW clásico
  - DW con alta exigencia en calidad de datos
- ELT
  - Carga los datos en el DW y luego utiliza SQL y el poder del procesamiento paralelo masivo para hacer las transformaciones dentro del DW.
  - Altamente optimizado y escalable.
  - DW con reqs de altos niveles de actualización de los datos
  - DW donde se puede resignar calidad de datos

ETL: visión general

- Objetivo del proceso de ETL:
  - Mantener cargado el DW y DMs con los datos correspondientes.
- Estructura general de los procesos ETL:
  - Operaciones de manipulación de datos que se realizan con un cierto orden y comunicando entradas y salidas.
  - El DW y DMs se cargan inicialmente, y luego se mantienen actualizados.
  - Normalmente involucra volúmenes de datos mucho mayores a los habituales en operaciones R/W OLTP.

ETL: aspectos principales

- Operaciones básicas para el proceso ETL
  - Identificación, Extracción, Filtrado, Integración, Transformación.
- Flujos de datos y control.
  - Los procesos ETL tienen ambos flujos.
- Ciclo de vida del DW.
  - Diseño, Carga Inicial, Actualización.

Operaciones

- Identificación de fuentes de datos
- Extracción de datos de las BDs fuente
- Integración de varias estructuras fuente en una destino
- Filtrado de datos (selección o proyección)
- Transformación de datos de entrada en otros de salida
- Control de calidad o limpieza (cleaning).
- Historización (agregar una marca de tiempo a un dato)
- Cálculos (generan una nueva variable calculada)

Flujos de datos y control

- Ejecución de un proceso ETL.
  - Ejecución de operaciones en un cierto orden:

    ➔ Flujo de control.

    - Una operación debe realizarse en forma previa/posterior a otra.
    - No implica que se conecten Salida→Entrada.

  - Conexiones entre operaciones (Salida→Entrada).

    ➔ Flujo de datos.

    - La entrada de una operación es la salida de otras.

Ciclo de vida de un DW

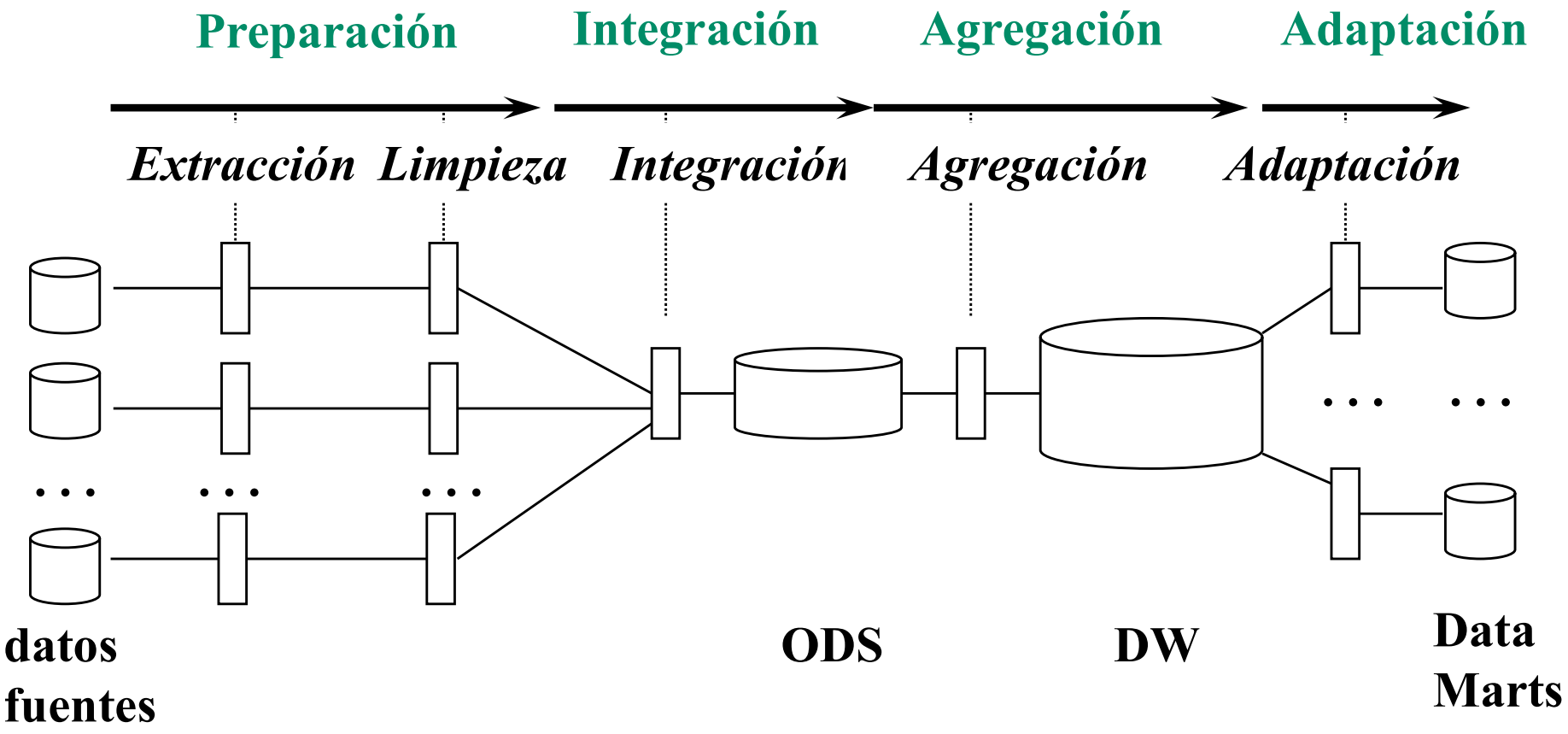
- 3 grandes etapas:
  - Diseño.
    - De las estructuras de datos.
    - De los procesos de carga / actualización.
  - Carga inicial.
    - Inicializa estructuras (DW, DMs) con una primera instancia.
  - Actualización, o Mantenimiento (refresh).
    - Mantiene actualizadas las estructuras con instancias:
      - Válidas, correctas.
      - Actualizadas con respecto a la realidad (BDs fuente).

Carga Inicial – enfoque [Jar2000]

- Etapa *carga inicial*
  - Consiste en la generación inicial del contenido del DW y DMs.
  - 4 actividades:
    - preparación
    - integración
    - agregación
    - adaptación

[Jar2000] M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis. "Fundamentals of Data Warehouses". Springer-Verlag, 2000.

Carga inicial



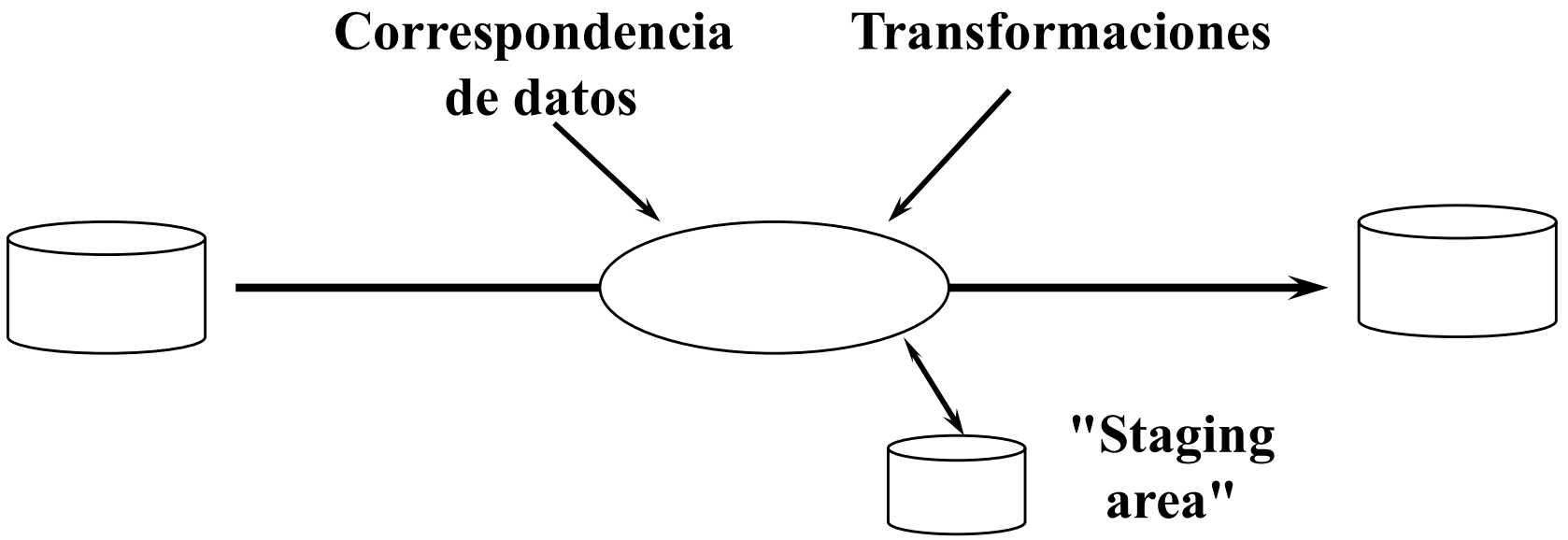
Carga inicial

- *Preparación* se realiza para c/fuente y consiste en
  - la *extracción* de datos
  - la *limpieza* de datos
- *Integración de datos* consiste en
  - la *reconciliación de* datos provenientes de fuentes heterogéneas
  - *estructuración* para el ODS o directamente para el DW

Carga inicial

- *Agregación* consiste en la generación de las "vistas agrupadas, resumidas" a partir de las tablas base.
- *Adaptación* consiste en la generación y especialización para el dominio de análisis de cada data mart.
- Esta descomposición en 4 pasos es llevada a la implementación de diferentes maneras en los productos y en los trabajos de investigación.

Carga inicial – enfoque gral



- Extracción de datos
- Transformación (Limpieza)
- Herramientas de ETL

Extracción de datos

- Involucra *técnicas* para la extracción de información en las fuentes.
  - Programas específicos (ej.: C, PL/SQL)
  - Herramientas ETL.
- Desde el punto de vista de *arquitectura*, el enfoque utilizado consiste en asociar *una componente por c/ fuente*.
  - Se le suele llamar *wrapper*.
  - Función:
    - básica: Traducir datos de la fuente a modelo de datos común.
    - en contexto DW: (básica) + detectar y extraer cambios de interés ocurridos en las fuentes y propagarlos.

Transformación

- La *limpieza* de datos constituye *uno* de los procesos dentro de la *transformación de datos* para la construcción de un DW.
  - La transformación de datos involucra:
    - cambios en las estructuras de representación de los datos
    - limpieza
    - integración de diferentes valores y estructuras de datos
    - resumen y agrupamiento de datos
- En el laboratorio se experimenta la programación de *transformaciones* usando una herramienta específica: Kettle.

Limpieza de datos

- *"Data cleaning*" ("*data cleansing*")
- Presente en la mayoría de los procesos de migración de datos
- Su objetivo es mejorar la *calidad de los datos* obtenidos al final de la migración
  - Calidad de datos como juicio sobre la condición o el estado de los datos
  - El nivel de calidad es definido según los requerimientos de las aplicaciones

Ejemplos de datos "sucios"

- Diferentes formatos de datos para el mismo atributo.
  - Ej.: la información sobre el departamento en un atributo dirección puede aparecer bajo las siguientes formas:
    - abreviación
    - nombre
    - un código
- Conflicto entre la descripción del atributo y los valores.
  - Ej.: un atributo nombre-empresa que en algunos casos contenga la dirección.

Ejemplos de datos "sucios"

- Atributos de texto libre pueden ocultar información importante.
  - Ej.: algunas etiquetas como "CP" dentro de nombres y direcciones, "Fax: ", ...
- Valores faltantes que deben ser asignados de acuerdo al esquema destino.
  - Datos incompletos.

Ejemplos de datos "sucios"

- Valores inconsistentes para la misma entidad.
  - Ej.: El mismo cliente en dos fuentes distintas tiene diferente dirección
- Información duplicada
  - Información sobre la misma entidad aparece duplicada, con claves u otros atributos diferentes.
  - Esta situación puede ocurrir tanto trabajando con una o varias fuentes origen.
  - Problema conocido como: *Object Identification, Entity Resolution, Entity Matching, etc.*

Funcionalidades de ayuda

- Las herramientas de migración gral. y orientadas a DW ofrecen funcionalidades para ayudar a resolver los problemas anteriores:
  - Funciones de conversión y de normalización
  - Limpieza para casos y dominios específicos
  - Algoritmos de correspondencias entre campos equivalentes de fuentes diferentes.

Conversión y normalización

- Conversión: convertir tipos y representación de los datos
- Normalización: usar un formato común para todos los datos pertenecientes al mismo tipo para permitir la comparación entre campos.
  - Ej.: Strings a mayúsculas o a minúsculas
    Fechas en formato "dd/mm/yyyy"
- Otros tipos de normalización pueden ser orientadas a comparar campos equivalentes.
  - Ej.: Corregir guiones que separan palabras.

Limpieza para dominios específicos

- Ejemplo: Nombres y direcciones
- Las técnicas utilizan metainformación.
  - Tablas para buscar datos válidos (ej.: códigos postales)
  - Diccionarios para buscar sinónimos y abreviaciones (e.g. "Apto", "Apt.", "Apartamento").

Herramientas ETL

Extraction, Transformation and Loading

- Características generales
  - Objetivo principal
    - *facilitar* el desarrollo de aplicaciones que *migran* datos aplicando *transformaciones.*
  - En este tipo de aplicaciones, los objetos típicos a definir:
    - conexiones
    - estructuras de los depósitos de datos
    - correspondencias y transformaciones entre los depósitos
    - excepciones
    - planificaciones de las transformaciones

Características generales

- Las herramientas ETL son *ambientes especializados* que permiten la definición y manipulación de objetos típicos en aplicaciones de intercambios de datos.
  - Facilidades para la modificación y mantenimiento de las aplicaciones.
- En estas herramientas, el data warehouse y/o los data marts son vistos como depósitos adonde migrar datos transformados.

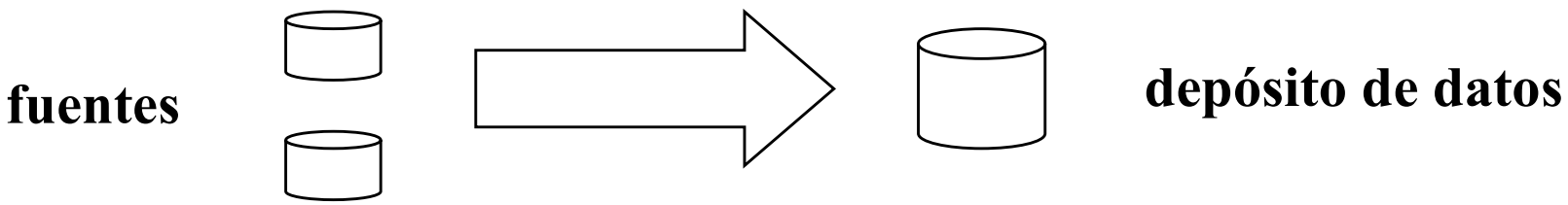
Características generales

- En general, ETLs *NO* ofrecen funcionalidades específicas para:
  - la captura de cambios en los datos,
  - la integración de esquemas y datos
- ETLs son "pobres" en cuanto al manejo de excepciones.
  - No significa que no se puedan manejar sino que su manejo suele ser "engorroso".

Ambientes especializados

- Editores gráficos para definición y planificación de procesos de carga.
- Lenguajes de programación para definir las transformaciones.
  - Proveen el motor de ejecución de los programas escritos en estos lenguajes.
  - Ofrecen funciones predefinidas y permiten el agregado de funciones definidas por el usuario.
- Mecanismos para el control del flujo de los procesos.

Ciclo de vida de DW: Actualización



- La *actualización* en sistemas de dw trata el problema de cómo reflejar los *cambios que ocurren en las fuentes* a partir de las cuales el depósito ha sido definido.
- En inglés, *Refreshment Process*.

Actualización

- Concepto de "frescura" (freshness)
  - No se refiere necesariamente a los datos más actuales.
  - "Frescura" requerida por las aplicaciones (los usuarios).
    - Diferencia aceptable entre en el grado de actualización de los datos del DW con respecto a la realidad.
- Cambios que ocurren en las fuentes
  - Fundamentalmente datos (manejado por herramientas)
  - Eventualmente esquema (hay que ajustar diseño DW y/o proceso de carga)
- Pocos estudios sobre impacto en el DW debido a cambios en los requerimientos.

Actualización

- La *actualización* tiene un flujo de datos similar a la etapa de *carga*.
- Sin embargo, el proceso de actualización:
  - Tiene en cuenta los cambios que ocurren en las fuentes.
  - Propaga dichos cambios al DW o DMs.
  - Se ejecuta:
    - Con una frecuencia planificada.
    - En «ventanas de tiempo » establecidas.

Diferencias Carga inicial - Actualización

- Período de disponibilidad requerida de las fuentes
  - Carga inicial: un período largo
  - Actualización: no debe interferir con las aplicaciones que usan a las fuentes.
- Restricciones sobre el tiempo de respuesta
  - Carga inicial: el tiempo de respuesta se mezcla con la duración del proyecto.
  - Actualización: depende de los requerimientos.
- Paralelismo en la etapa de preparación
  - La sincronización del paralelismo está dada por la integración.

Dificultades en la actualización

- El volumen de datos almacenados en un DW
  - Los cambios deben propagarse a los distintos niveles de la jerarquia de depósitos de datos.
    - Datos de interés y también datos de los niveles intermedios.
- Concurrencia entre la actualización y el procesamiento de consultas del DW
  - Escenarios donde esta concurrencia es necesaria:
    - Período corto o inexistente en que no hay consultas.
    - Nivel de "frescura" de los datos.
  - La dificultad radica en realizar la actualización sin detener demasiado el despacho de consultas.

Actualización: Extracción

- La *extracción* debe encargarse de identificar y propagar los cambios ocurridos en las fuentes.

- 2 posibles enfoques
  - Extracción periódica
    - Siempre se extrae lo nuevo generado en un período fijo (por ej, se extrae una vez por semana).
  - Extracción cada vez que hay cambios
    - Esta tarea requiere la detección de cambios en las fuentes (realizada por módulos llamados wrappers)

Actualización: Integración

- La *integración* debe ser incremental.
  - Cuando ingresan nuevos datos se deben poder integrar con los ya existentes.
  - La limpieza debe ser incremental.
  - Determinar los datos que deben ser cambiados en el dw.
    - Determinar información de otras fuentes para calcular el nuevo dato del dw.

Actualización: Carga

- La *carga* debe ser incremental.
  - Las transacciones de actualización deben ser sincronizadas de manera que las vistas accedidas por las consultas se encuentren en un estado "consistente".
  - Planificar el momento en que las transacciones de actualización se aplican.

Conclusiones

- Carga y actualización de DW y DMs:
  - Involucra múltiples operaciones de transformaciones de datos
  - Se tienen flujos de datos y de control
  - Requiere un diseño específico
- Operaciones muy variadas:
  - Desde extracción, hasta control de calidad
  - Integración de datos
- No hay estandarización
  - Ni siquiera en un modelo conceptual
- Las herramientas de ETL
  - Facilitan diseño y mantenimiento de los procesos de carga y actualización