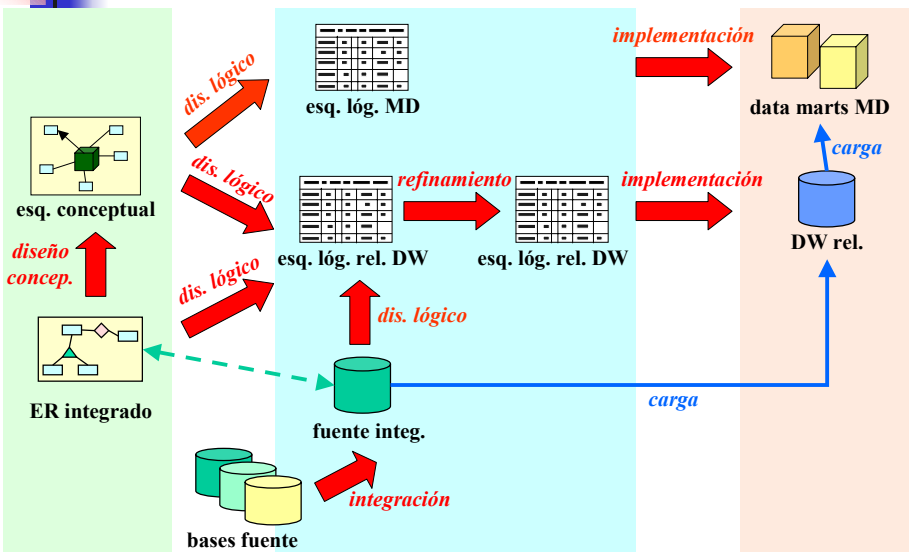
Proceso de Diseño

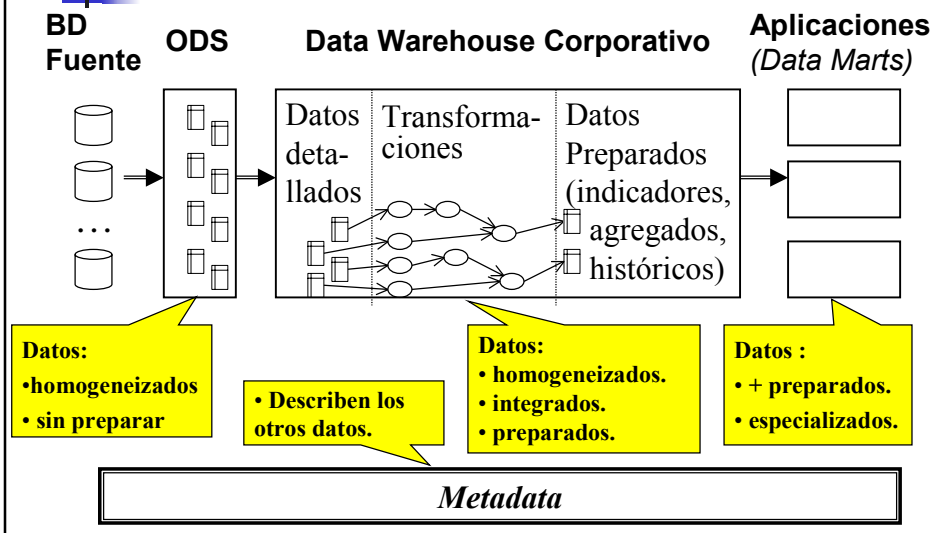


Carga y Mantenimiento de DW

Plan

- **Contexto - Ciclo de vida de un DW**
  - Carga inicial
  - Problemática del proceso de actualización
- **Carga inicial**
- **Herramientas ETL**
  - Laboratorio
- **Conclusión**

Estructura del Data Warehouse



Ciclo de vida de un DW

- **3 grandes etapas**
  - diseño
  - carga inicial
  - refresque

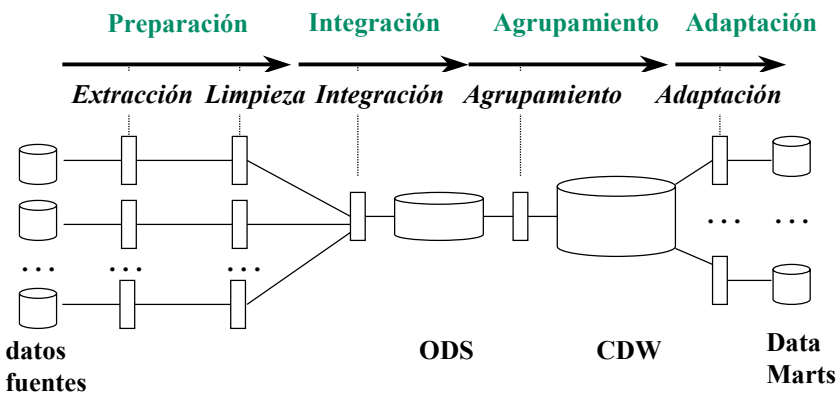
Ciclo de vida de un DW / Diseño

- **Etapa *diseño***
  - Consiste en la definición de:
    - esquema del DW y de los DMs
    - extractores de fuentes
    - limpiadores de datos
    - integradores de datos
  - El resultado es un conjunto de especificaciones formales o semi-formales que alimentan la metadata usada por el sistema y las aplicaciones dw.

Ciclo de vida DW / Carga Inicial

- **Etapa *carga inicial***
  - Consiste en la generación inicial del contenido del dw.
  - 4 actividades:
    - preparación
    - integración
    - agrupamiento (high level aggregation)
    - adaptación (customization)

Carga inicial



Carga inicial

- *Preparación* se realiza para c/fuente y consiste en
  - la *extracción* de datos
  - la *limpieza* de datos
  - el *almacenamiento* de datos
- *Integración de datos* consiste en
  - la *reconciliación de* datos provenientes de fuentes heterogéneas
  - la generación de las relaciones (vistas de base) del ODS.

Carga inicial

- *Agrupamiento* consiste en la generación de las "*vistas agrupadas, resumidas"* a partir de las vistas de base.
- *Adaptación* consiste en la generación y especialización de las "*vistas usuario"* que definen a los data marts.
- Esta descomposición en 4 pasos es llevada a la implementación de diferentes maneras en los productos y en los trabajos de investigación.

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).
- Cambios que ocurren en las fuentes
  - Esquema y datos
- Pocos trabajos sobre impacto en el depósito de datos debido a cambios en los requerimientos.

Actualización

- **Etapa *actualización* tiene un flujo de datos similar a la etapa de *carga*.**
- **Sin embargo, el proceso de actualización:**
  - captura los cambios diferencia que ocurren en las fuentes
  - propaga dichos cambios a lo largo de la jerarquía de depósitos

Carga inicial y Actualización

- **Diferencias**
  - Período de disponibilidad requerida de las fuentes
    - Carga inicial: un período largo
    - Actualización: período / no sobrecargue 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
    - Mayor en refresque que en carga inicial

Parámetros del proceso de actualización

- **Parámetros estáticos y dinámicos.**
- **Estáticos**
  - Requerimientos de las aplicaciones
    - ej.: "frescura" de los datos, tiempos de cálculo de consultas y de vistas, modo de actualización (historia, sobreescritura, ...).
  - Restricciones de las fuentes
    - ej.: períodos de disponibilidad, frecuencia de cambios
  - Restricciones del sistema de dw
    - ej.: limite de espacio, limites de funcionalidades
  - Estos parámetros pueden evolucionar llevando a reconfigurar la arquitectura del dw y cambiar la estrategia de actualización.

Parámetros del proceso de actualización

- **Dinámicos**
  - volumen de cambios en las fuentes
  - "perfiles" de consultas

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 el refresque 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 el refresque sin detener demasiado el despacho de consultas.

Dificultades en la actualización

- **La carga transaccional.**
  - *Actualización de un DW* puede involucrar transacciones pesadas de carga y acceso.

    ⇒ uso de arquitecturas paralelas + compresión para transmisión + transacciones de larga duración.
  - *Refresque de un DM* puede involucrar transacciones que acceden muchos datos, realizan muchos cálculos para resumir y actualizan pocos datos en el DM.

    ⇒ problema porque se debe actualizar en una cierta ventana de tiempo.

Problema de la actualización

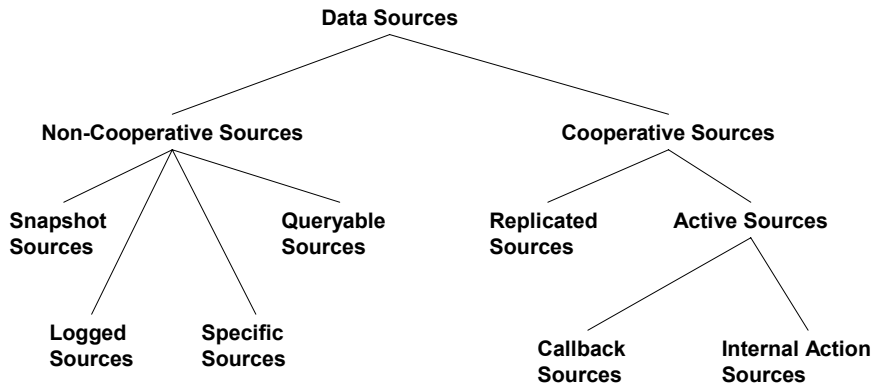
- El *problema de la actualización de dws* puede ser visto como la definición de un *proceso de construcción incremental de dws*.
- La incrementalidad aparece en distintos niveles
  - extracción
  - integración
  - carga

Problema de la actualización

- La *extracción* debe poder encargarse de registrar los cambios ocurridos en una fuente. Esta tarea requiere:
  - la detección de cambios en las fuentes
  - la extracción de los cambios, y
  - el registro de los cambios.
- *Wrappers* (uno en cada fuente)
  - Funcionalidad típica: Traducir datos de la fuente a modelo de datos común
  - En DW: Detectar y extraer cambios

Problema de la actualización

- **Una clasificación de las fuentes**



Problema de la actualización

- **La *integración* debe ser incremental.**
  - La limpieza debe ser incremental.
  - Determinar las operaciones a aplicar sobre el ODS.
  - Determinar los datos que deben ser cambiados en el dw.
    - Determinar información de otras fuentes para calcular el nuevo dato del dw.

Problema de la actualización

- **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 refresque se aplican.