Bases de datos Masivas

Introducción a Data Warehouse
Proceso de ETL

Banchero, Santiago

Agosto de 2016

Introducción a Data Warehouse

¿Qué es un Data Warehouse?

+ Definido de muchas maneras diferentes, pero no rigurosamente.

  - Una base de datos de apoyo a las decisiones, que **se mantiene separada** de la base de datos operativa de la organización.
  - Apoyar el procesamiento de información, proporcionando una plataforma sólida de datos históricos consolidados para el análisis.

Data warehousing: Es el proceso de construir y usar un DW.

[Han and Kamber, 2006]

Introducción a Data Warehouse

¿Qué es un Data Warehouse?

"A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, *and* **nonvolatile** *collection of data in support of management's decision-making process."*

W. H. Inmon[1]

[1] https://en.wikipedia.org/wiki/Bill_Inmon

Data Warehouse — Subject-Oriented

- Organizado en torno a grandes temas, como: clientes, productos, ventas (Otros ejemplos...)
- Centrándose en el modelado y análisis de los datos para los tomadores de decisiones, **no en las operaciones diarias** o **procesamiento de transacciones**.
- Provee una visión **simple y concisa** sobre cuestiones temáticas particulares por **exclusión de los datos que no son útiles en el proceso de apoyo a las decisiones**.

Data Warehouse — Integrated

- Construido por la integración de múltiples y heterogéneas fuentes de datos

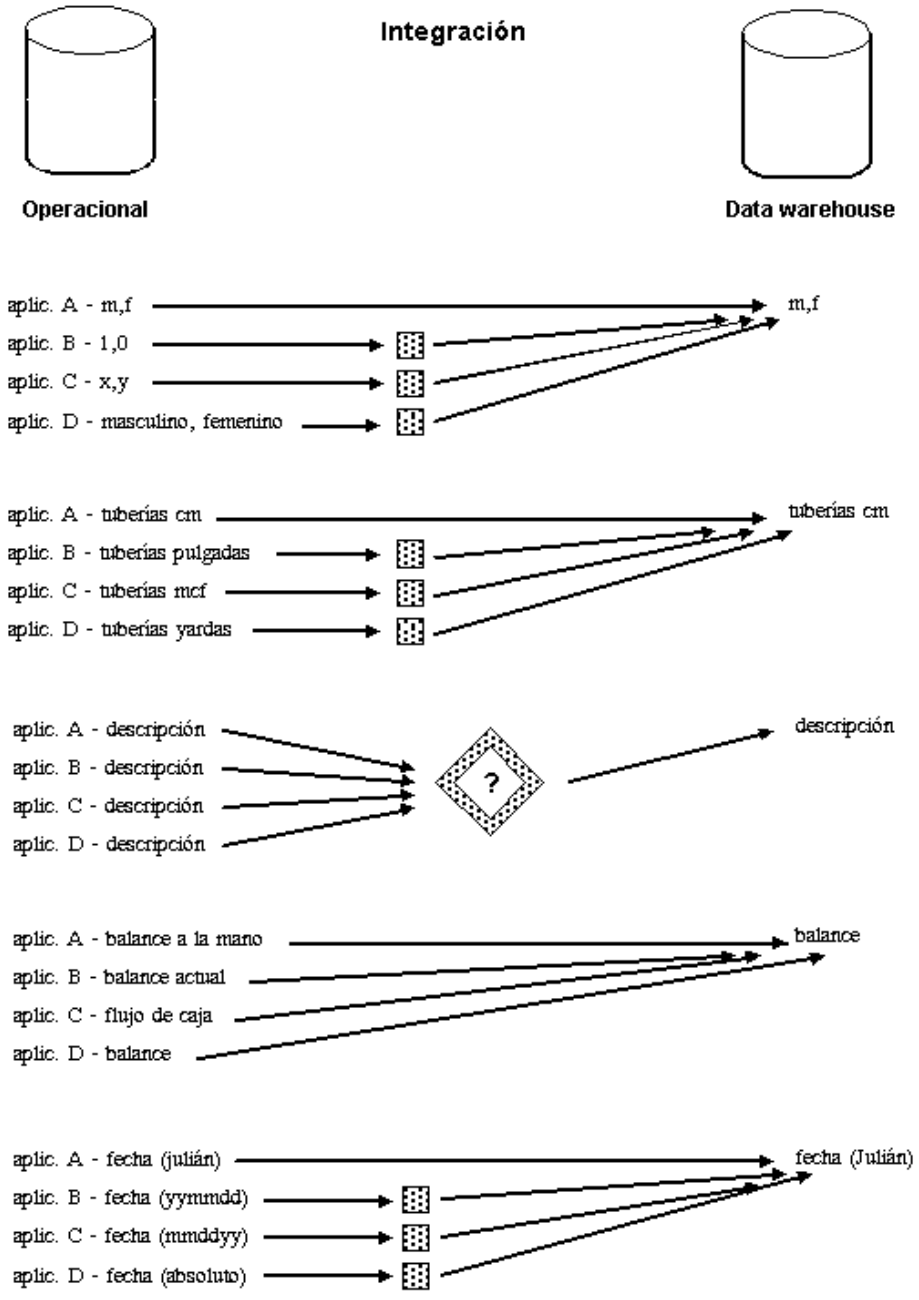
  + Bases de datos relacionales, archivos planos, XML, hojas de cálculo, etc.

- Se aplican técnicas de integración y de limpieza de datos.

  + Garantizar la coherencia en las convenciones de nomenclatura, las estructuras de codificación, medidas de atributos, etc.; entre las diferentes fuentes de datos

  + Todas las conversiones se realizan cuando los datos son movidos al DW.

Data Warehouse — Integrated



Data Warehouse — Time Variant

- Un de los objetivos es descubrir las tendencias en los negocios, el análisis va a requerir de grandes cantidades de datos.
- El horizonte de tiempo en el DW es significativamente más largo que el de los sistemas de bases de datos operacionales. **Los requisitos de performance** exigen que los datos históricos sean trasladados a un archivo.
  - DB transaccionales: datos con valores actuales, recientes.
  - Los datos en el DW: proveen información de una perspectiva histórica. (Ej. 2,3,..,10 años)
- Cada clave en la estructura del DW
  - Contiene un elemento de tiempo, explicito o implícito.
  - Pero una clave en datos operacionales, pueden o no tener un "elemento tiempo" asociado

Data Warehouse — Time Variant

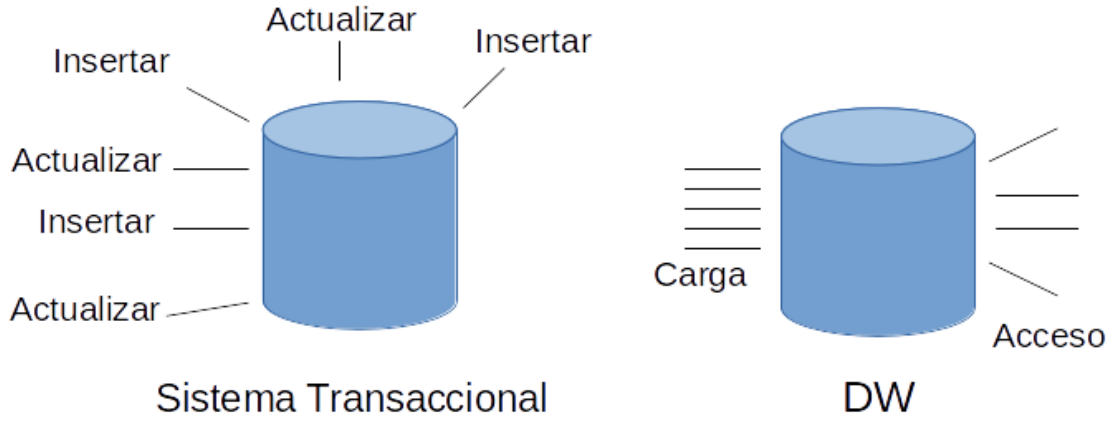
La **información es útil sólo cuando es estable**

Los datos operacionales cambian sobre una base momento a momento.

La perspectiva más grande, esencial para el análisis y la toma de decisiones, requiere una base de datos estable.

Data Warehouse — Nonvolatile

- Se trata de un almacenamiento físicamente separado, de datos transformados desde el ambiente operativo.
- La actualización de los datos **no se produce en el entorno data warehouse**.
  - No se requieren mecanismos de control de concurrencia, recuperación o proceso de transacciones.
  - Requiere solo dos operaciones:
    - La carga inicial de los datos
    - Acceso a los datos



OLTP y el DW

Los sistemas transaccionales tradicionales (**OLTP** - *On Line Transaction Processing*) son inapropiados para el soporte a las decisiones.

Las Tecnologías de Data Warehouse se han convertido en una importante herramienta para integrar fuentes de datos heterogéneas y darle lugar a los sistemas de **OLAP** (*On Line Analytic Processing*)

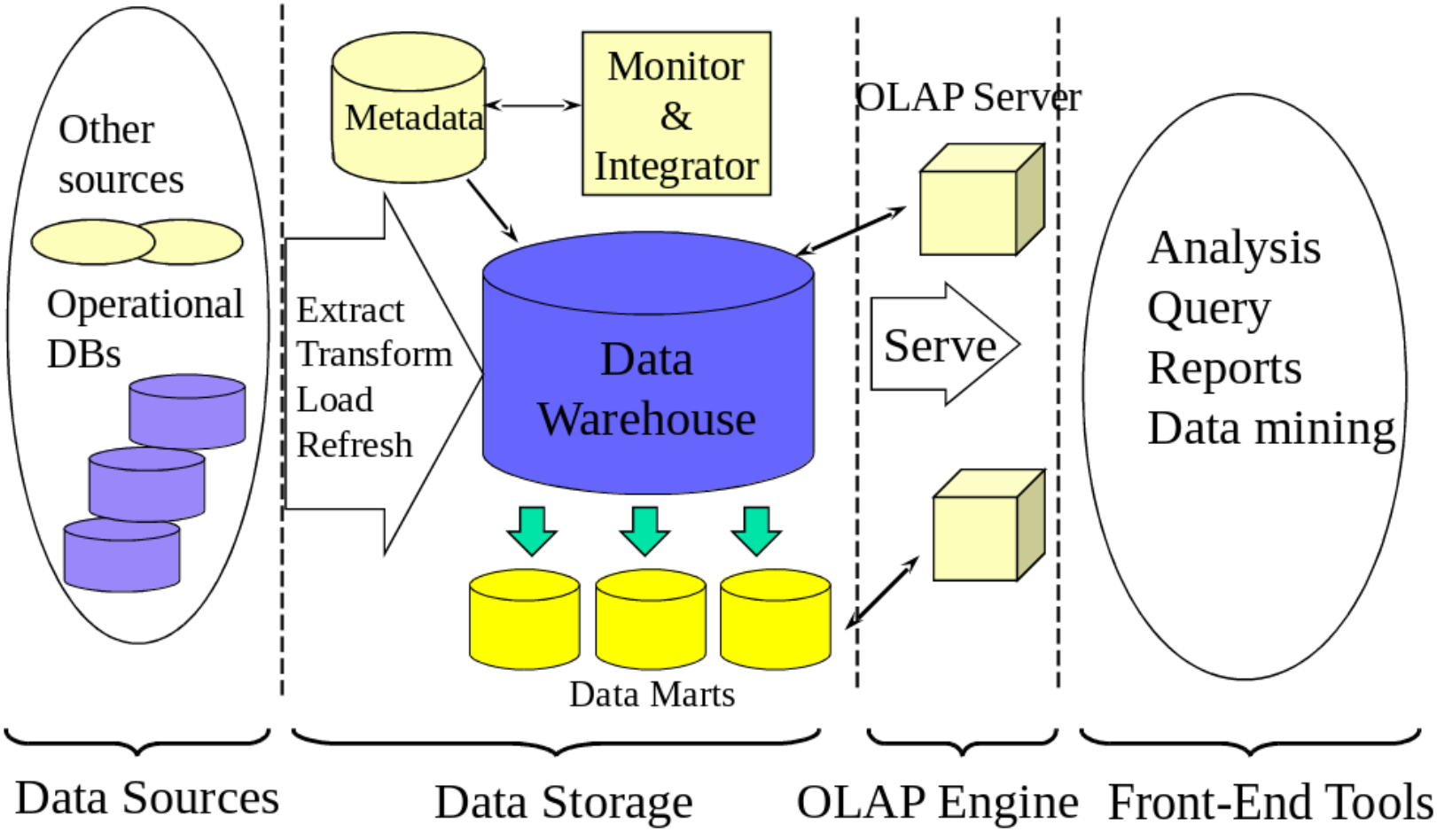
Diferencias entre OLTP y OLAP

Característica	OLTP (Relational)	OLAP(Multidimensional)
Tipo de información	atomizada	resumida
Nivel de agregación	Un registro por unidad de tiempo	muchos registros
Propósito	Orientada a procesos	Orientada a tema
Tamaño BBDD	GigaBytes	Giga a TeraBytes
Origen Datos	Interno	Interno y Externo
Actualización	On-Line	Batch
Periodos	Actual	Histórico
Consultas	Predecibles	Ad Hoc
Actividad	Operacional	Analítica

Why a Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

Data Warehouse: A Multi-Tiered Architecture

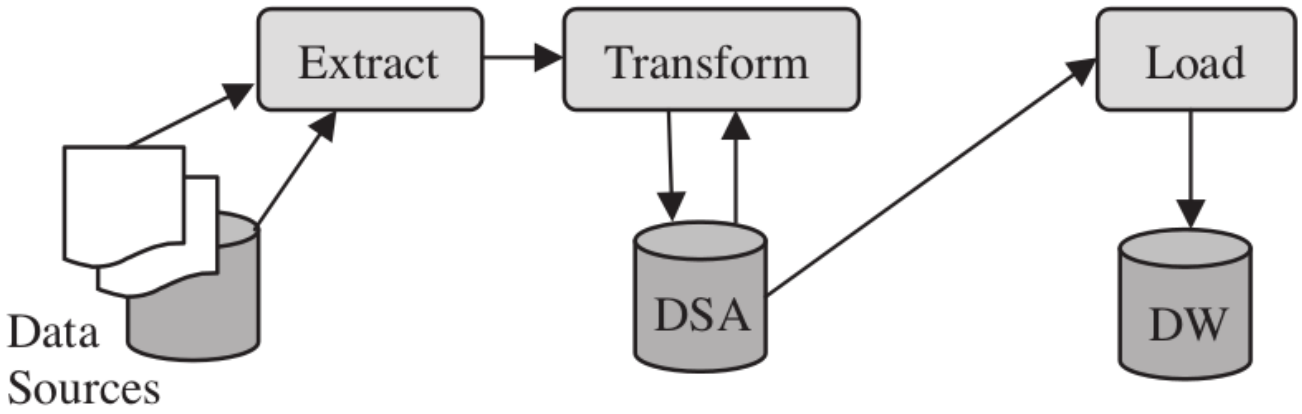


Extraction, Transformation, and Loading (ETL)

Las herramientas de *Extraction–transformation–loading* (**ETL**) son piezas de software responsables de la extracción de datos desde varias fuentes, su limpieza, puesta a punto, re formateo, integración e inserción en un Data Warehouse.

Construir el proceso de ETL es una de las grandes tareas de la implementación de un data warehouse.

La construcción de un data warehouse requiere enfocarse en entender tres cuestiones: las fuentes de datos, quienes son los destinatarios y cómo mapear esos datos (proceso de ETL)



Extraction, Transformation, and Loading (ETL)

Durante el proceso de ETL los datos **son extraidos desde bases de datos OLTP**, transformados para que coincidan con el DW esquema y luego cargados en el DW.

Muchos data warehouses además incorporan datos desde sistemas **no-OLTP**, tales como archivos de texto sistemas de herencia y hojas de cálculo.

ETL es a menudo una compleja combinación de **procesos y tecnologías** que consumen una porción significativa del **esfuerzo de desarrollo** y requiere la habilidad de análisis de negocio, diseño de bases de datos y desarrollo de aplicaciones.

El proceso de ETL **no es una tarea de una sola vez**.

Como las fuentes de datos cambian, los data warehouse deben ser actualizados periodicamente.

Extraction, Transformation, and Loading (ETL)

Extracción de datos
Obtener datos de múltiples, heterogéneos y fuentes externas

Limpieza de datos
Detectar errores en los datos y rectificarlos cuando sea posible

Transformación de datos
Convertir datos de formato heredado o acogida al formato de almacén

Carga
Clasificar, resumir, consolidar, calcular puntos de vista, comprobar la integridad, y construir índices del y particiones

Refrescar
Propagar las actualizaciones de las fuentes de datos para el almacén

Extraction

- El primer paso en un escenario de ETL es la extracción de datos
- Cada una de las fuentes de datos tiene sus propias características que se necesitan manejar en orden para extraer los datos de forma efectiva.
- El proceso debe integrar eficazmente los sistemas que tienen diferentes plataformas, como:
  - los diferentes sistemas de gestión de bases de datos,
  - sistemas operativos diferentes y
  - diferentes protocolos de comunicación

Extraction

Durante el proceso de extracción de datos de diferentes fuentes, el equipo de ETL debe ser consciente de:

- (a) el uso de controladores que conectan a las fuentes de bases de datos,
- (b) **comprender la estructuras** de datos de las fuentes, y
- (c) **saber cómo manejar las fuentes de diferente naturaleza** tales como mainframes.

Extraction

The extraction process consists of two phases, initial extraction, and changed data extraction.[2]

- **In the initial extraction**, it is the first time to get the data from the different operational sources to be loaded into the data warehouse
- **The incremental extraction** is called changed data capture (CDC) where the ETL processes refresh the DW with the modified and added data in the source systems since the last extraction

This **process is periodic according to the refresh cycle and business needs**. It also captures only changed data since the last extraction by using many techniques as audit columns, database log, system date, or delta technique.

[2] Libro de Kimball

Transformation

The second step in any ETL scenario is data transformation.

The transformation step tends to make some **cleaning** and **conforming** on the incoming data to gain accurate data which is correct, complete, consistent, and unambiguous.

This process includes data cleaning, transformation, and integration.

It defines the granularity of fact tables, the dimension tables, DW schema (stare or snowflake), derived facts, slowly changing dimensions, factless fact tables.

All transformation rules and the resulting schemas are described in the metadata repository.

Load

Loading data to the target multidimensional structure is the final ETL step.

In this step, extracted and transformed data **is written into the dimensional structures** actually accessed by the end users and application systems.

Loading step includes both **loading dimension tables** and **loading fact tables**.

Referencias

Han, J. and Kamber, M. (2006).
Data mining: Concepts and techniques, 2nd ed.
http://hanj.cs.illinois.edu/bk2/.
Accedido: 2015-08-01.