



# Bases de datos Masivas

## Introducción a Data Warehouse

### Proceso de ETL

Banchero, Santiago

Agosto de 2017

# 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?

*“Un data warehouse es una colección de datos integrada, que **varia en el tiempo, no volátil y orientada a un tema**; para asistir a sectores gerenciales en el proceso de toma de decisiones.”*

W. H. Inmon<sup>1</sup>

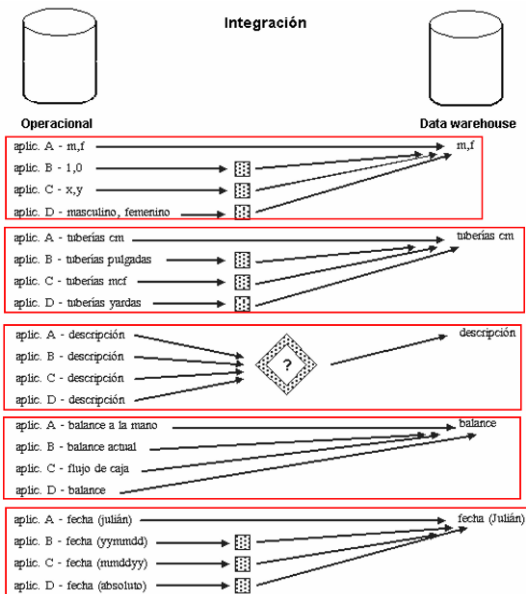
---

<sup>1</sup>[https://en.wikipedia.org/wiki/Bill\\_Inmon](https://en.wikipedia.org/wiki/Bill_Inmon)

- 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.**

- 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 — Integrada



# Data Warehouse — Variante en el tiempo

- Un de los objetivos es realizar análisis de tendencias, este 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, explícito o implícito.
  - Pero una clave en datos operacionales, pueden o no tener un “elemento tiempo” asociado

## 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 — No Volátil

- 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**.

Por lo tanto:

- 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



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

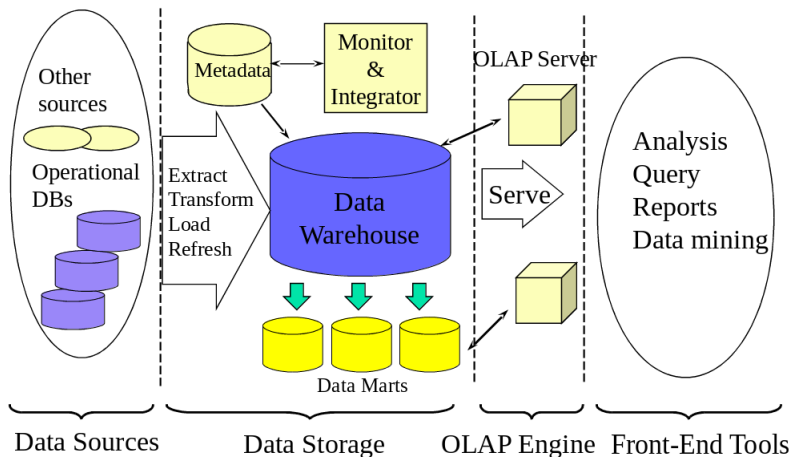
Característica	OLTP (Relational)	OLAP (Multidimensional)
Tipo de información	atomizada	Resumida
Nivel de agregación	Un registro por unidad de tiempo	Muchos registros
Propósito	Orientada a procesos	Orientada a tema
Tamaño BBDD	GigaBytes	Giga a TeraBytes
Origen Datos	Interno	Interno y Externo
Actualización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

# ¿Por qué un Data Warehouse separado?

- Alto rendimiento en ambos sistemas
  - **DBMS** Son optimizados para tareas de **OLTP**: métodos de acceso, indexación, control de concurrencia, recuperación ante fallas, etc.
  - **DW** Es optimizado para **OLAP**: Consultas complejas, vistas multidimensionales, consolidación, etc.
- Funciones diferentes y datos diferentes
  - **Datos faltantes**: El soporte a las toma de decisiones requiere datos históricos que las DB operacionales generalmente no mantienen.
  - **Consolidación de datos**: El soporte a las toma de decisiones requiere consolidación (agregación, resúmenes) de fuentes de datos heterogéneas.
  - **Calidad de datos**: diferentes fuentes de datos generalmente utilizan inconsistentes representaciones de datos, códigos y formatos que deben ser reconciliados.

Cabe destacar que cada vez existen más y más sistemas que realizan OLAP en bases de datos relacionales :D.

# 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 extraídos desde bases de datos OLTP**, transformados para que coincidan con el DW esquema y luego cargados en el DW.

# Extraction, Transformation, and Loading (ETL)

Durante el proceso de ETL los datos **son extraídos 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, hojas de cálculo, entre otros.



# Extraction, Transformation, and Loading (ETL)

Durante el proceso de ETL los datos **son extraídos 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, hojas de cálculo, entre otros.

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.

# Extraction, Transformation, and Loading (ETL)

Durante el proceso de ETL los datos **son extraídos 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, hojas de cálculo, entre otros.

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 periódicamente.

# Extraction, Transformation, and Loading (ETL)

Durante el proceso de ETL los datos **son extraídos 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, hojas de cálculo, entre otros.

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 periódicamente.

# 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

- 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.

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

- ❶ el uso de controladores que conectan a las fuentes de bases de datos,
- ❷ **comprender la estructuras** de datos de las fuentes, y
- ❸ **saber cómo manejar las fuentes de diferente naturaleza** tales como mainframes.

El proceso de extracción consta de dos fases, la **Extracción inicial** y **Extracción de datos modificados**.<sup>2</sup>

- En la **extracción inicial**, es la primera vez que se obtienen los datos de las diferentes fuentes operativas que se cargan en el DW.
- La **extracción incremental** se llama captura de datos modificados (CDC) donde los procesos ETL refrescan el DW con los datos modificados y añadidos en los sistemas fuente desde la última extracción.

Este proceso es **periódico según el ciclo de renovación y las necesidades del negocio**. Este además, captura sólo datos modificados desde la última extracción utilizando muchas técnicas como **columnas de auditoría, logs de la base de datos, fecha del sistema o archivos delta**.

---

<sup>2</sup>Libro de Kimball

La etapa de extracción es conceptualmente la **tarea más simple de todas**, con el objetivo de identificar el subconjunto correcto de datos de origen que tiene que ser presentado al workflow de ETL para su posterior procesamiento.

Las dificultades de la tarea se deben a dos restricciones:

- La fuente debe sufrir una **sobrecarga mínima durante la extracción**, ya que otras actividades administrativas también tienen lugar durante ese período.
- Tanto por **razones técnicas como políticas**, los administradores son bastante reacios a aceptar intervenciones importantes en la configuración de su sistema; Por lo tanto, tiene que haber un mínimo de interferencia con la configuración del software en el lado de la fuente.



El segundo paso en cualquier escenario ETL es la transformación de datos.

Se ocupa del **mejorar la calidad de los datos** para que estos sean precisos, correctos, completos, consistentes y sin ambigüedad.

Este proceso incluye la limpieza, transformación e integración de datos.

Define la granularidad de las tablas de hechos, las tablas de dimensiones, el esquema DW (Estrella o Copo de nieve), los hechos derivados, las dimensiones que cambian lentamente, las tablas de hechos sin hechos.

Todas las reglas de transformación y los esquemas resultantes se describen en el **repositorio de metadatos**.

Dependiendo de la aplicación y la herramienta utilizada, los procesos ETL pueden contener una gran cantidad de transformaciones.

Las **tareas de transformación y de limpieza** se ocupan de las clases de conflictos y problemas que se pueden distinguir en tres niveles:

- **Schema-level:** (a) los conflictos de nombres, cuando se utilice el mismo nombre para diferentes objetos (homónimos) o se utilizan nombres diferentes para el mismo objeto (sinónimos) y (b) los conflictos estructurales.
- **Record-level:** registros duplicados o contradictorios.
- **Value-level:** numerosos problemas técnicos de bajo nivel pueden ser atendidas en diferentes escenarios de ETL. Representación de fechas, sexo, etc.

Cargar datos a la estructura multidimensional resultante es el paso final de ETL.

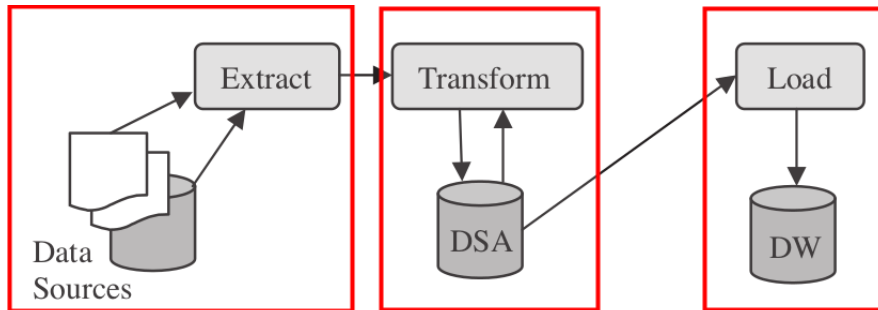
En este paso, los datos extraídos y transformados se **escriben en las estructuras multidimensionales** a las que acceden realmente los usuarios finales y los sistemas de aplicación.

El paso de carga incluye tanto la carga a las **tablas de dimensiones** como la carga a las **tablas de hechos**.


# Framework de ETL

Datos extraídos de fuentes de almacenamiento OLTP o formatos semiestructurados.

Carga al DW central y a todas sus distintas partes (datamarts, vistas).



Los datos son propagados a un área del DW llamada **Data Staging Area** (DSA), donde su transformación homogeneización y limpieza tiene lugar

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