

Aux 2. Introducción a la Minería de Datos

Gastón L'Huillier^{1,2}, Richard Weber²
glhuilli@dcc.uchile.cl

¹Departamento de Ciencias de la Computación
Universidad de Chile

²Departamento de Ingeniería Industrial
Universidad de Chile

2010



Escenario Actual...

- Mucha información en la empresa
- Se pierde conocimiento sobre que información se podría utilizar para la toma de decisiones
- Información no consolidada entre distintas unidades de negocio
- Lento acceso y muy costoso acceder a la información relevante para tomar decisiones
- Etc...

Definición - Data Warehousing

Un Data Warehouse es una arquitectura

“Colección de datos utilizada para la toma de decisiones la cual es orientada al negocio, integrada, que varía en el tiempo y no volátil.”

Características

- **Orientada al negocio:** Definida para temas importantes del negocio
- **Integrada:** Usualmente construida integrando múltiples fuentes de datos heterogéneas
- **Varía en el tiempo:** Datos almacenados proveen información desde una perspectiva histórica
- **No volátil:** No ocurren muchos cambios a los datos disponibles en el data warehouse.



Data Warehouse vs BBDD Operacionales [1]

BBDD. Operacionales	Data Warehouse
Datos básicos con gran nivel de detalle .	Datos derivados de alguna agregación .
Reflejan el “ahora” .	Reflejan la “historia” .
Constantemente actualizados.	Rara vez modificados o eliminados.
Mínima redundancia.	Alta redundancia.
Estructura estática .	Estructura dinámica .
Soportan funciones del día a día .	Soportan requerimientos de información .

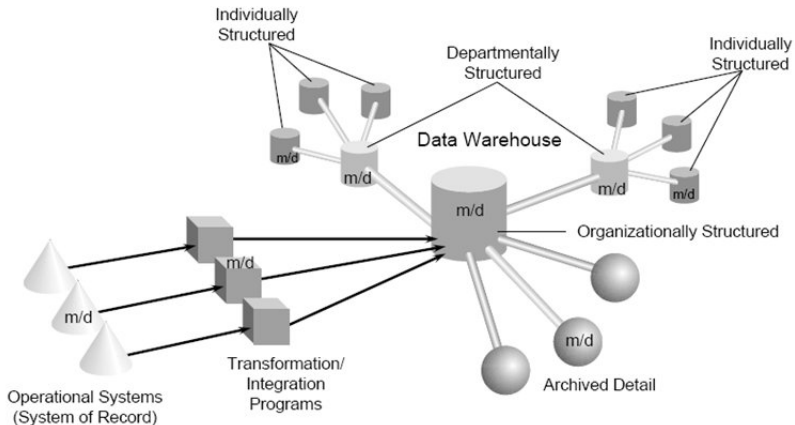


Data Warehouse vs BBDD Operacionales [2]

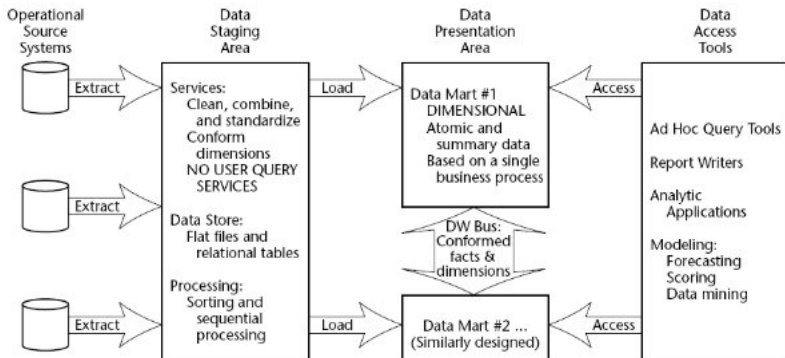
BBDD. Operacionales	Data Warehouse
Transaccionales (OLTP).	Analíticos (OLAP).
Actualizaciones no programadas y frecuentes.	Actualizaciones periódicas y programadas.
Unidades de tiempos de respuesta en segundos.	Unidades de tiempos de respuesta en minutos.
Muchos usuarios concurrentes.	Muy pocos usuarios concurrentes.
Principales prioridades: Alta disponibilidad y Rápido.	Principales prioridades: Altamente flexible y orientado al usuario final.



Arquitectura Data Warehousing [1]



Arquitectura Data Warehousing [2]



Arquitectura Data Warehousing [3]

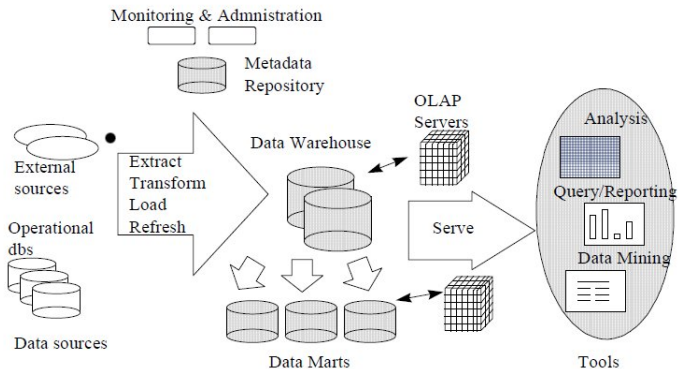


Figura: [Chaudhuri and Dayal, 1997]

Top Down approach ([Inmon, 2005])

- Diseñar completamente el Data Warehouse y luego ir desplegando los Data Marts.
- Esquema rígido y de largo desarrollo, pero fácil de mantener y muy integrado.

Bottom Up approach ([Kimball and Ross, 2002])

- Desde la perspectiva del usuario: “me interesa sólo mi área de trabajo”
- Desarrollo de proyectos departamentales (Data Marts) para una posterior construcción del Data Warehouse.
- Alto ROI, rápido de construir, pero difícil de mantener.

Data Warehouse vs Data Mart

Data Warehouse	Data Mart
A nivel corporativo .	A nivel departamental .
Modelo más complejo (modelo constelación).	Puede ser suficiente un simple modelo estrella .
Se alimenta de muchas fuentes operacionales .	Se alimenta de pocas fuentes operaciones .
Desarrollo complejo al largo plazo (años).	Desarrollo simple y al corto plazo (meses).
Resultados en el largo plazo .	Resultados en el corto plazo .



Diseño de un Data Warehouse

Tabla de dimensiones

Tablas con los descriptores textuales del negocio sobre los cuales se quiere agregar o desagregar la información.

Tablas de Hechos (Fact Table)

Tabla central con las medidas e indicadores proyectados sobre las dimensiones.

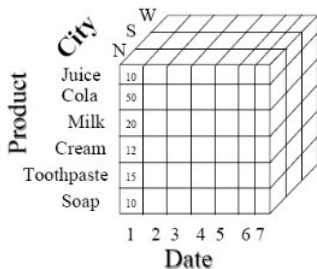
Granularidad

Nivel mínimo sobre el cual está diseñada una determinada dimensión.

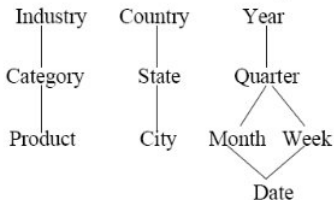


Diseño de un Data Warehouse [2]

Dimensiones y Granularidad



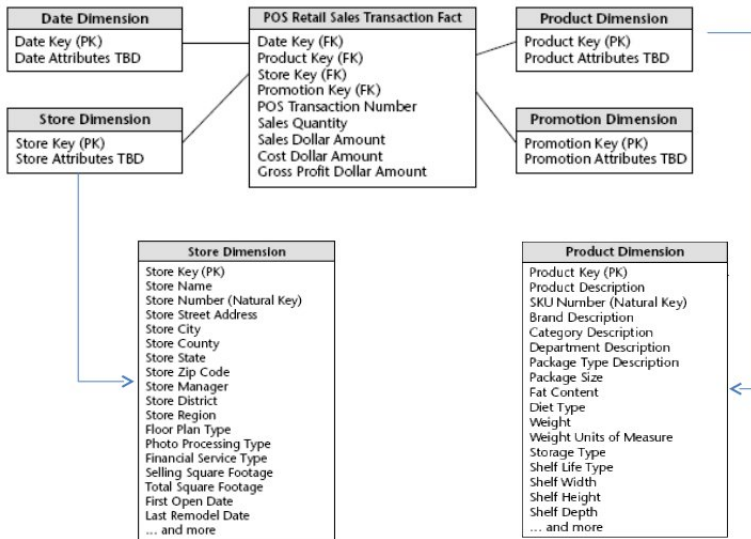
Dimensions: Product, City, Date
Hierarchical summarization paths



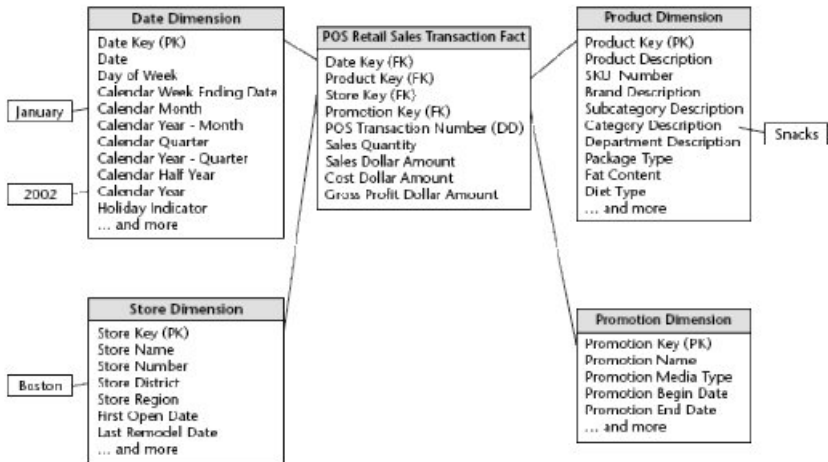
Modelo Estrella (Retail Case Study)



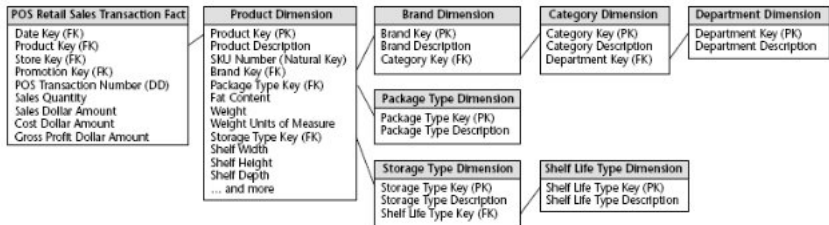
Modelo Estrella (Retail Case Study) [2]



Modelo Estrella (Retail Case Study) [3]



Modelo Snowflake (Retail Case Study) [2]



Operaciones sobre un cubo OLAP

Un usuario final debe poder realizar fácilmente las siguientes operaciones:

- **Slice and Dice:** “Cortes” o filtros de la información en para poder examinar mejor.
- **Roll up:** Agregar información sobre una determinada dimensión.
- **Drill Down:** desagregar información sobre una determinada dimensión.
- **Pivot:** Rotar el cubo sobre alguno de las dimensiones de interés.



Operaciones sobre un cubo OLAP [2]

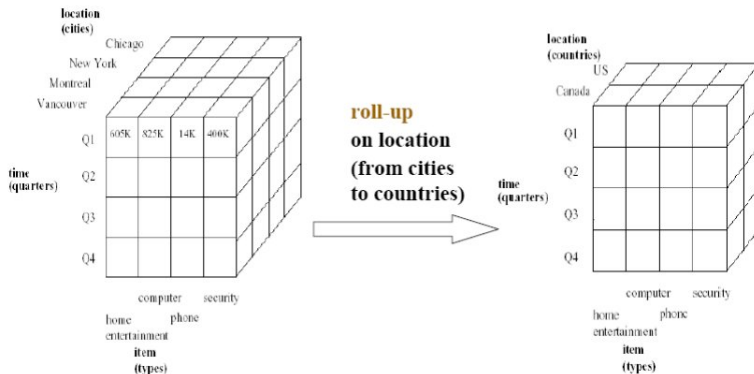


Figura: Operación **roll-up**

Operaciones sobre un cubo OLAP [3]

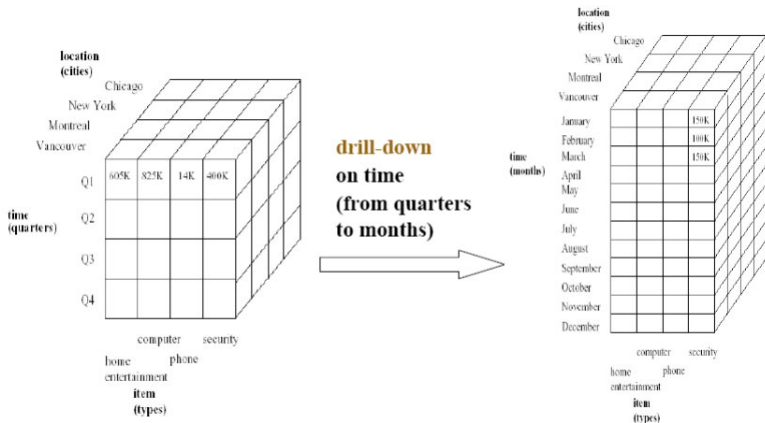


Figura: Operación **drill-down**

Operaciones sobre un cubo OLAP [4]

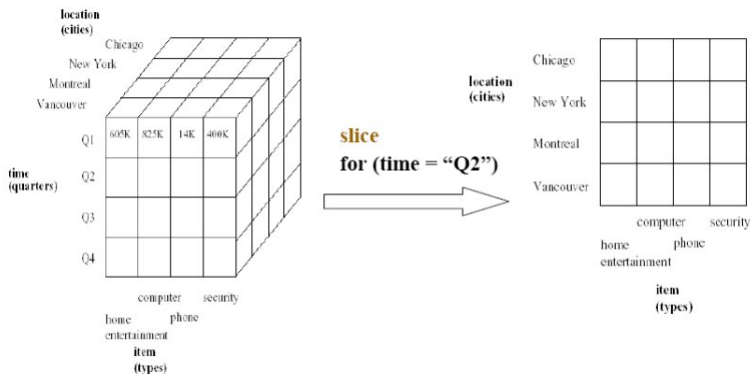


Figura: Operación **slice**

Operaciones sobre un cubo OLAP [5]

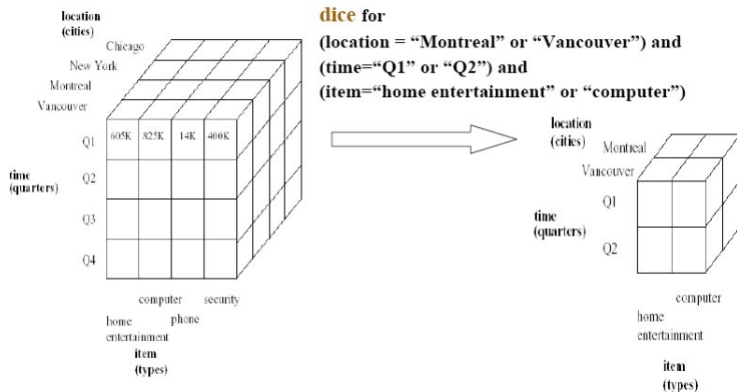


Figura: Operación **dice**

Operaciones sobre un cubo OLAP [6]

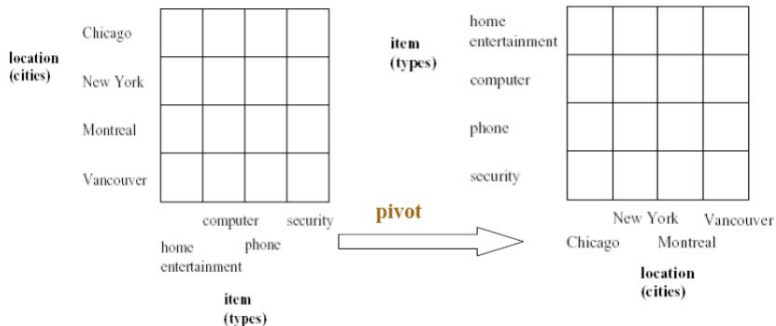


Figura: Operación **pivot**

[Kimball and Ross, 2002]'s top ten pitfalls

- Pitfall¹⁰: Enamorarse de la tecnología y los datos, en vez de concentrarse en los requerimientos del negocio y objetivos.
- Pitfall⁹: Fallar en adoptar una administración visionaria, con influencia, accesible y razonable cómo sponsor del data warehouse.
- Pitfall⁸: Atacar un proyecto galáctico de muchos años, en vez de perseguir un proyecto más manejable.
- Pitfall⁷: Utilizar mucha energía en construir una estructura normalizada, quedándote sin presupuesto antes de construir un prototipo viable y operativo



[Kimball and Ross, 2002]'s top ten pitfalls [2]

- Pitfall⁶: Poner más atención al rendimiento operacional y para un mejor desarrollo que a una interfaz rápida y fácil de utilizar.
- Pitfall⁵: “Keep it simple”: no desarrollar modelos muy complejos que usuarios finales no podrán utilizar, o que necesitarán una fuerte capacitación.
- Pitfall⁴: Alimentar modelos dimensionales con una visión standalone, sin considerar una arquitectura tecnológica adecuada.
- Pitfall³: Cargar solamente datos agregados.



- Pitfall²: Presumir que el negocio, sus requerimientos, el análisis, los datos, la tecnología que lo soporta y la gente es estática y no cambia en el tiempo.

Pitfall¹

Olvidar que el éxito del Data Warehouse está directamente relacionado a la aceptación del usuario. Si un usuario no lo acepta cómo un sistema para la toma de decisiones, el proyecto fracasó.

Data Warehousing en el SII

- **Objetivo:** “Reducir tasa de evasión de impuestos”
- **Problema:** Información con problemas de calidad, redundancia de datos, no consolidada, sin políticas de seguridad, etc.
- **Situación preliminar:** base de datos FoxPro, aplicaciones independientes no centralizadas en Clipper.
- **Propuesta:**
 - 1er paso: Data Mart en el depto. estudios económicos y tributarios como primera experiencia. (tecnología: MySQL y PHP)
 - 2do paso: Data Warehouse corporativo (tecnología Oracle)

Data Warehousing en Salcobrand

- **Objetivo:** “Maximizar fidelización de mercado”
- **Problema:** Estrategia de fidelización mediante tarjetas personales, altos volúmenes de datos en tiempo record.
- **Situación preliminar:** Sistemas operacionales relacionales.
- **Propuesta:** Desarrollo sistema CRM (Customer Relationship Management) en utilizando una arquitectura Data Warehouse.
- **Tecnología:** Sybase para almacenamiento, Microstrategy para reportes y análisis.

Data Warehousing en CORFO-Innova

- **Objetivo:** “Desarrollar Sistema de Indicadores de Gestión en la Industria del Salmón”
- **Problema:** Necesario para el país tener un sistema de indicadores que permitan dar a cuenta de las actividades económicas, sociales y ambientales relacionadas con la industria del salmón.
- **Situación preliminar:** Inexistente.
- **Propuesta:** Desarrollo de un Data Warehouse que permita consolidar la información necesaria para el calculo de los indicadores.
- **Tecnología:** MySQL, Jasper Reports.

Aplicaciones en Chile [3] (cont.)

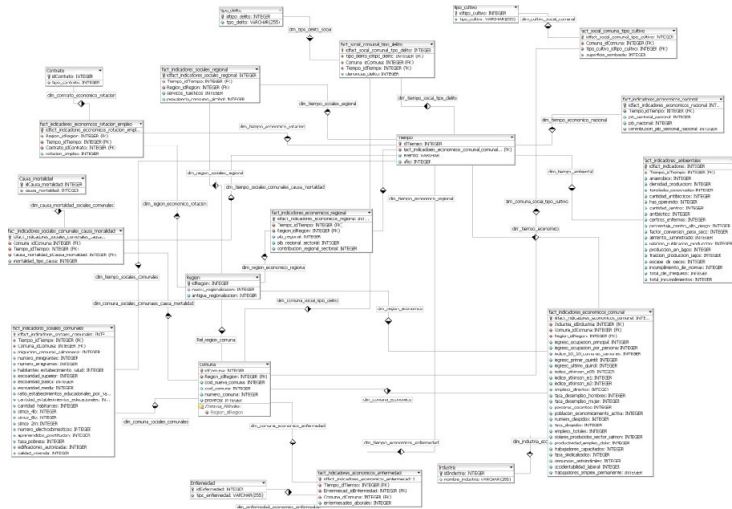


Figura: Modelo “constelación” proyecto ISIS

References I



Chaudhuri, S. and Dayal, U. (1997).
An overview of data warehousing and olap technology.
SIGMOD Record, vol. 26(1):65–74.



Inmon, W. H. (2005).
Building the Data Warehouse, 4rd Edition.
Wiley Publishing.



Kimball, R. and Ross, M. (2002).
The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling
(Second Edition).
Wiley.

