



Definición

Data Warehousing: almacenamiento, transformación y distribución de datos útiles para los responsables de tomar decisiones



Definición (cont.)

“Un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia.”

W.H. Inmon



Características

- **Orientado al Negocio** - organiza y presenta los datos desde la perspectiva del usuario.
- **Maneja gran volumen de datos** - contiene datos históricos.
- **Almacena información sobre diversos medios** - a causa del gran volumen que debe manejar.



Características(cont.)

- **Abarca múltiples versiones de un esquema de base de datos** - debido a la información histórica que contiene.
- **Sumariza y agrega información** - para presentarla de una manera comprensible para los usuarios.
- **Integra y asocia información proveniente de diversas fuentes** - datos recolectados durante años por diversas aplicaciones.



Motivación

- Mercados altamente dinámicos y competitivos.
- Necesidad de tomar decisiones rápidamente.
- Aumento de la capacidad de almacenamiento.
- Crecientes volúmenes de información disponible.
- Baja de costos del Hardware.



OLTP - On Line Transaction Processing

- Procesamiento de los datos operacionales.
- Gran nivel de detalle.
- Sistemas diseñados para soportar actualizaciones consistentes (normalización).
- Ineficiente para toma de decisiones.
- Consultas orientadas a obtener como respuesta unos pocos registros.



OLAP - On Line Analytical Processing

- Sistemas que permiten recolectar y organizar la información analítica realmente necesaria y disponer inmediatamente de ella en diversos formatos (tablas, gráficos, reportes, etc.).
- Analizan los datos desde diferentes perspectivas (dimensiones) del negocio.
- Soportan análisis complejos de grandes volúmenes de datos.
- En consecuencia:
 - Distintas técnicas de diseño requeridas (p.ej. desnormalización)
 - Distintos mecanismos de procesamiento de consultas (orientados a consultas de agregación)



OLTP vs. OLAP

	OLTP	OLAP
Usuario Tipico	empleado	profesional
Uso del sistema	operacional	análisis
Interaccion usuarios	predeterminada	ad-hoc
Unidad de trabajo	transaccion	consulta
Caracteristicas	lectura/escritura	lectura
Registros accedidos	decenas	millones
Cant. de usuarios	miles	cientos
Focalizacion	ABM de datos	extraer información



Componentes

- **Fuentes de datos.** Sistemas operacionales, información externa, etc.
- **Meta Datos.** Estructura, definición y origen de los Datos.
- **Data Warehouse.** Datos organizados y herramientas para su análisis.
- **Usuarios .** Responsables de tomar decisiones.

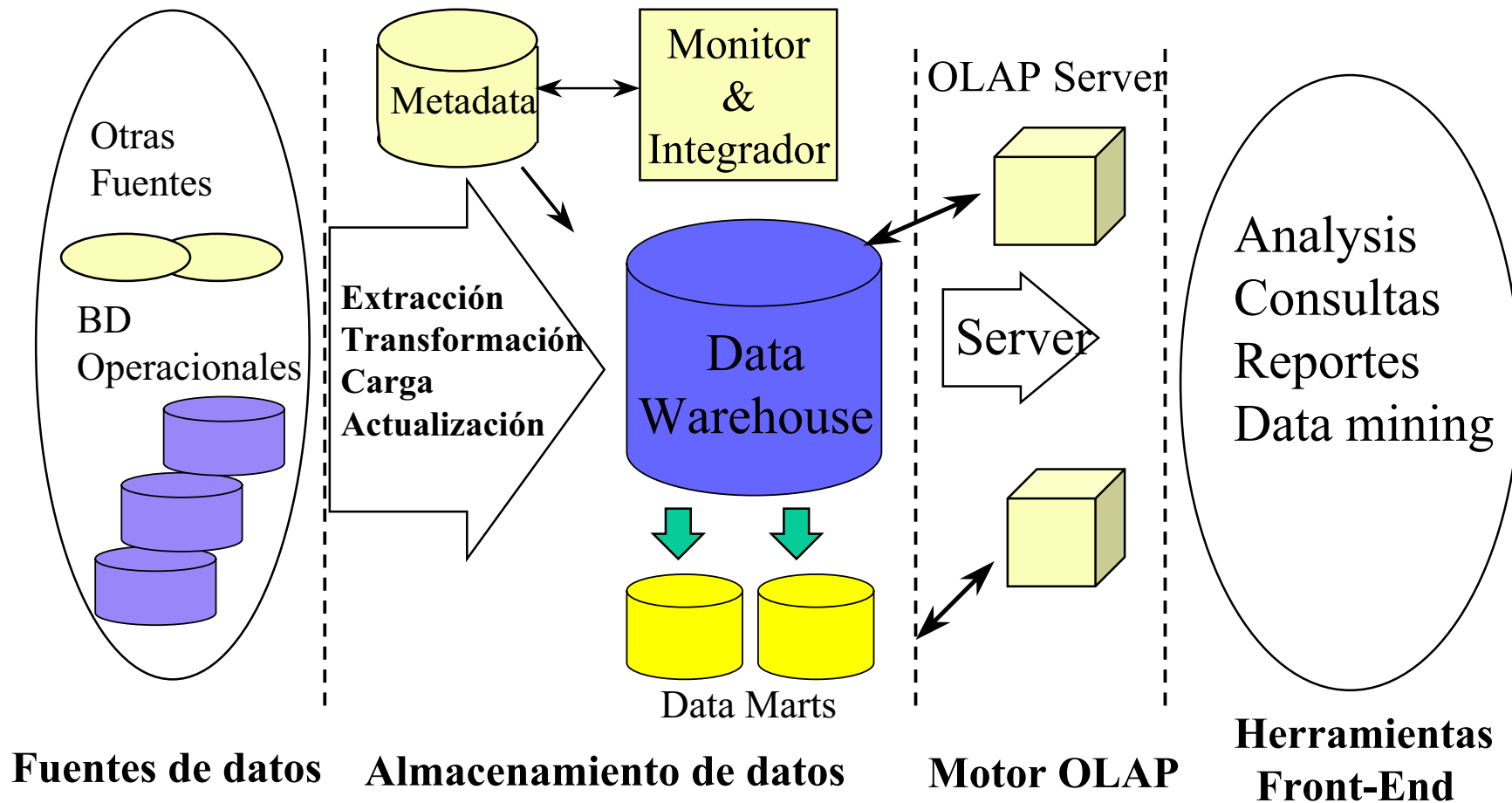


Tres Clases de Data Warehouse

- Enterprise Warehouse
 - Representa la información de toda la organización
- Data Mart
 - Un subconjunto de la información de la organización, que es de valor para grupos específicos de usuarios.
- Virtual Warehouse
 - Un conjunto de vistas sobre los datos operacionales
 - Solo unas pocas se materializan



Arquitectura Típica



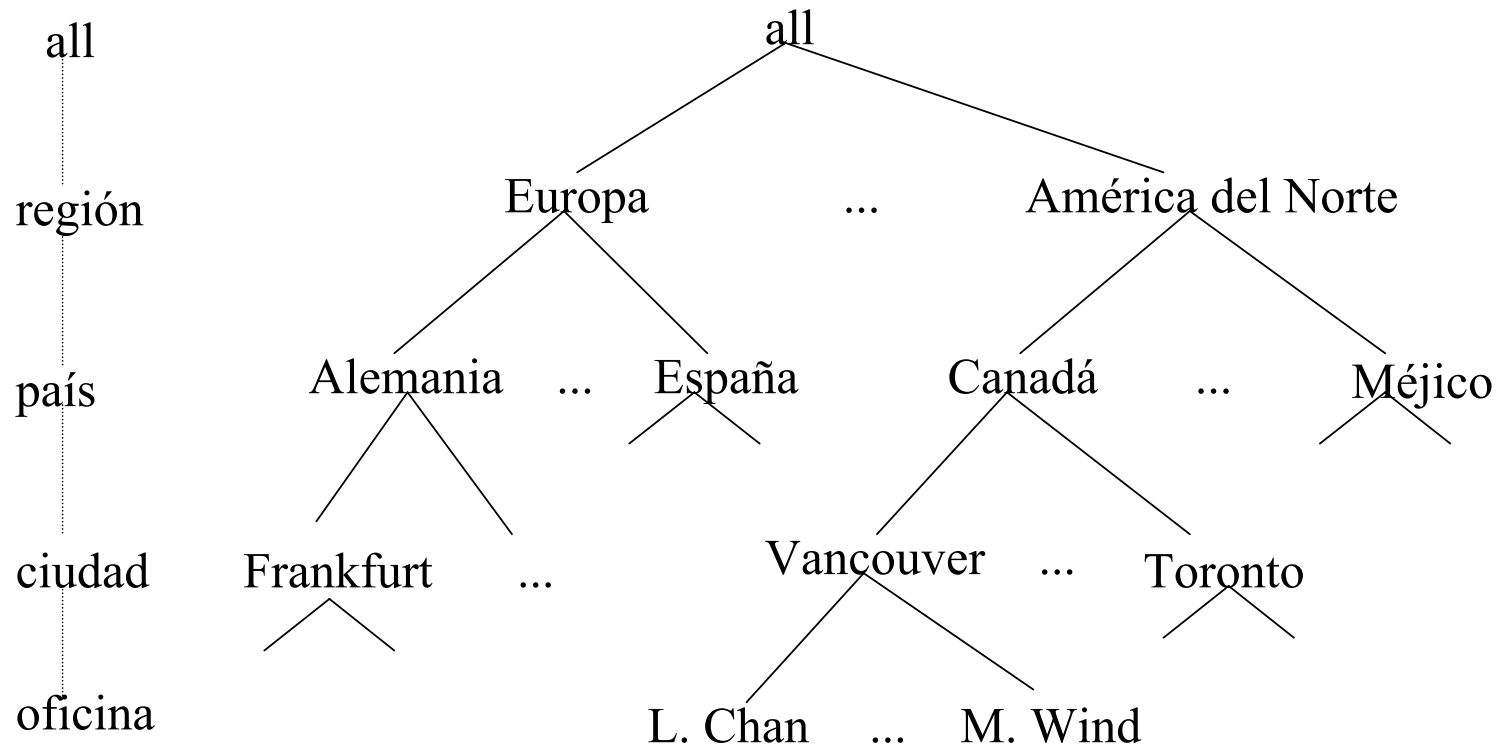


El Modelo Multidimensional

- Vista multidimensional del data warehouse => influencia el diseño de la base de datos, las herramientas front-end, y los motores OLAP.
- Modelo multidimensional de datos: un conjunto de medidas numéricas son los objetos de análisis.
 - Ej: ventas, beneficios, duración de llamadas, etc.
- Adicionalmente existen, asociadas a las medidas, las **dimensiones** de análisis, que proveen el contexto a las medidas, y se describen mediante atributos.
 - El modelo define una medida como un valor en un espacio multidimensional. Estas medidas pueden también representar datos agregados.
- Las dimensiones se pueden organizar en jerarquías de agregación.



Jerarquías Dimensionales



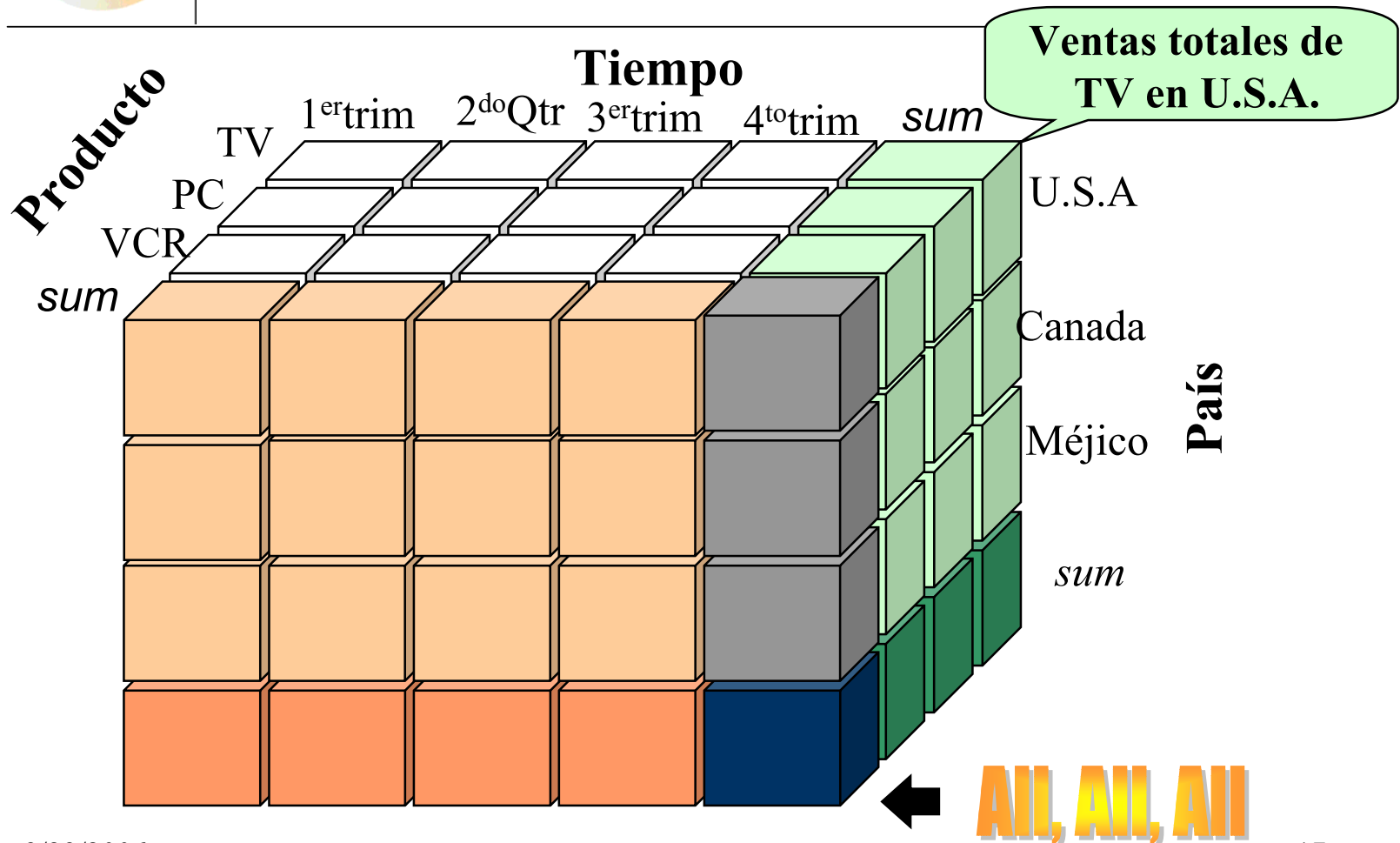
Esquema

9/29/2006

Instancia



Ejemplo de Data Cube





Diseño

- El esquema **estrella** (Kimball, 1995) describe el modelo multidimensional de datos mediante tablas de hechos y tablas de dimensión.
- Ejemplo: queremos modelar y analizar las ventas a través de múltiples dimensiones.
 - Tablas de Dimensión: **Productos (item_id, marca, tipo), o Tiempo (día, semana, mes, trimestre, año), Geografía (sucursal, ciudad, region)**
 - Tablas de Hechos: contienen medidas (como **ventas_totales**) y las claves de las tablas de dimensión; ej: **Ventas (item_id,día,sucursal,ventas_totales)**.
- Variante normalizada: el esquema **snowflake**.
- No provee soporte directo a las jerarquias dimensionales



Diseño Físico: ROLAP vs. MOLAP

- El modelo multidimensional es implementado directamente por los llamados *servidores MOLAP* (**M**ultidimensional **OLAP**).
 - Soportan la visión multidimensional de datos mediante un motor de almacenamiento multidimensional, conformado por arrays propietarios.
 - No requieren un mapping entre modelos.
 - Excelente performance; problema: dimensiones esparzas.
- Si se utilizan BD relacionales como servidores, el modelo y sus operaciones deben ser mapeados a relaciones y consultas SQL => implementación *ROLAP* (**R**elational **OLAP**)
 - Extienden el modelo relacional los servidores relacionales con middleware que soporta consultas multidimensionales.
 - Utilizan diversas técnicas de materialización de vistas.