



# Bases de Datos Masivas

Data Warehouse

**Bases de Datos Multidimensionales**

Banchero, Santiago

Septiembre 2015

# Introducción a Data Warehouse (DW)

Concepto de DW. Definición según W. H. Inmon:

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

**Características de un DW:**

- Orientado a un tema
- Integración
- Variante en el tiempo
- No volátil

# Introducción a Data Warehouse (DW)

## 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.

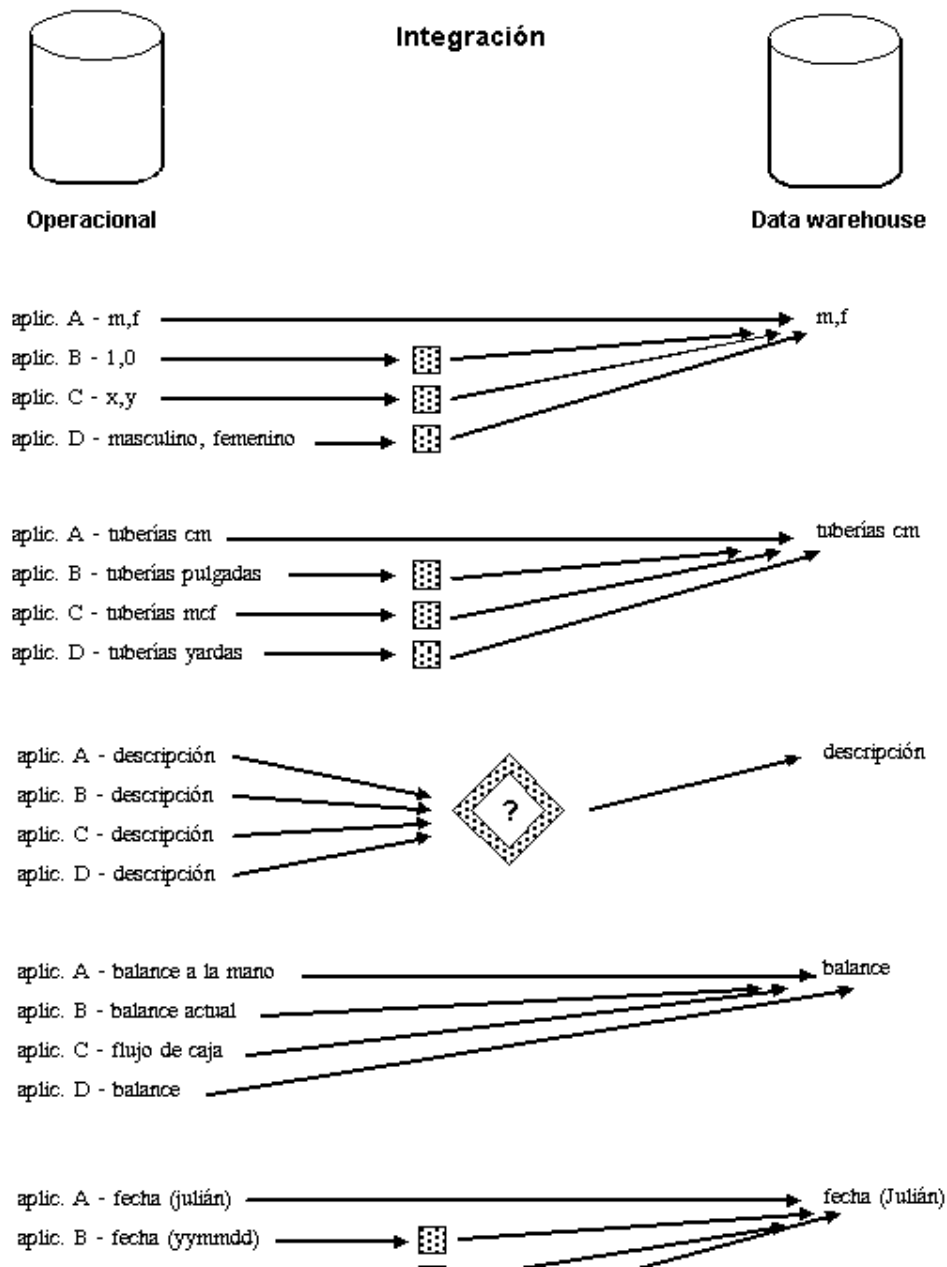
# Introducción a Data Warehouse (DW)

## Data Warehouse — Integrated

- Construido por la integración de múltiples y heterogeneas fuentes de datos
  - Bases de datos relacionales, archivos planos, XML, hojas de cálculo, etc.
- Técnicas de integración de datos y de limpieza de datos son aplicadas.
  - 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.

# Introducción a Data Warehouse (DW)

## Data Warehouse — Integrated



# Introducción a Data Warehouse (DW)

## Data Warehouse — Time Variant

El horizonte de tiempo en el DW es significativamente más largo que el de los sistemas de bases de datos operacionales.

- 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.

# Introducción a Data Warehouse (DW)

## 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



# Introducción a Data Warehouse (DW)

## OLTP y OLAP

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

Los sistemas tradicionales de gestión suelen realizar tareas repetitivas muy bien estructuradas e implican transacciones cortas y actualizaciones generalmente.

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*)

Los sistemas de soporte a la decisión requieren la realización de consultas complejas que involucran muchos datos e incluyen funciones de agregación.

De hecho, las actualizaciones son operaciones poco frecuentes en este tipo de aplicaciones, denominado genéricamente "procesamiento analítico"



# Introducción a Data Warehouse (DW)

## OLTP y OLAP

	<b>OLTP System Online Transaction Processing (Operational System)</b>	<b>OLAP System Online Analytical Processing (Data Warehouse)</b>
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

source: [www.rainmakerworks.com](http://www.rainmakerworks.com)

# Introducción a Data Warehouse (DW)

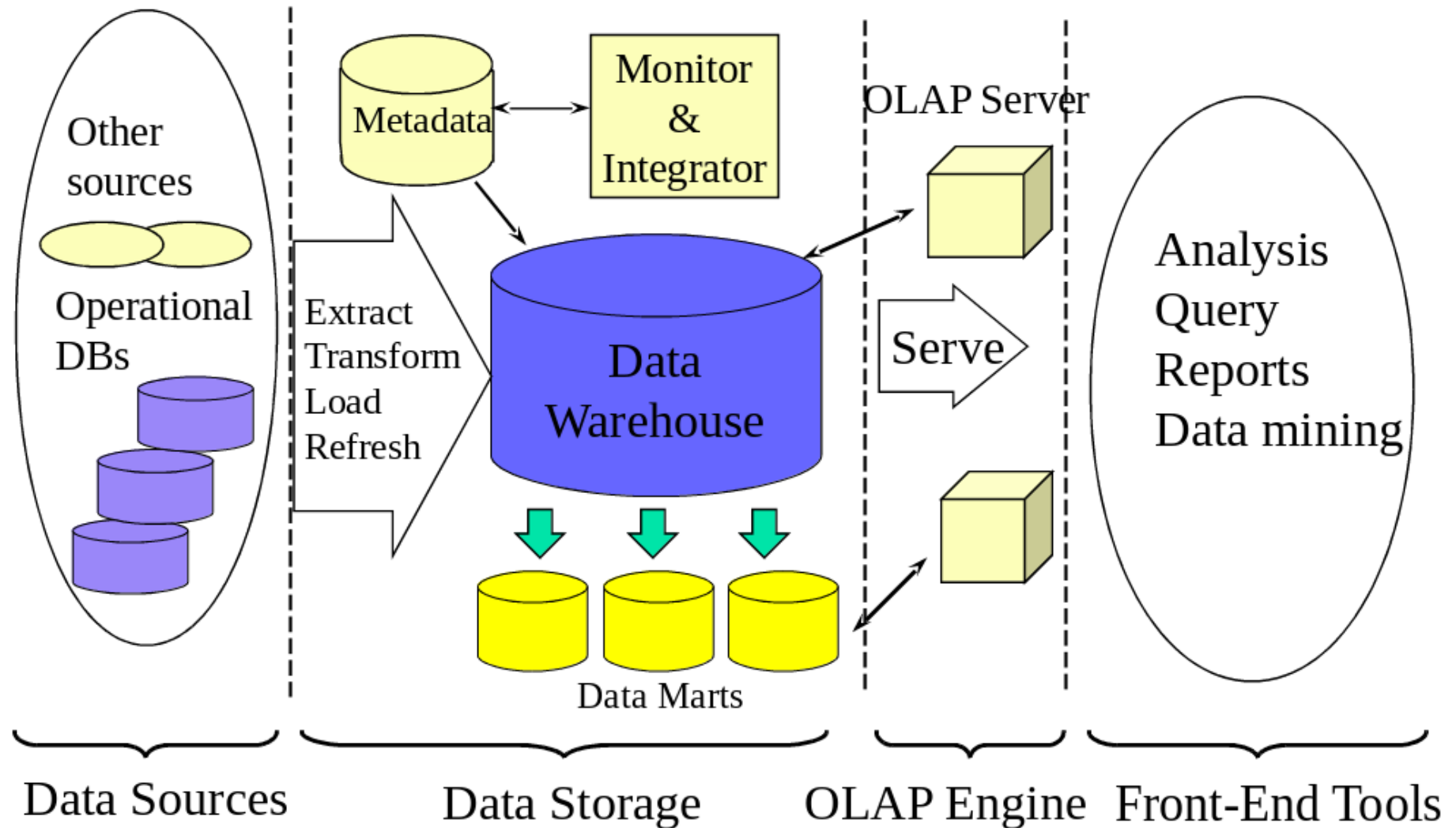
## ¿Por qué tener un DW separado?

- Mantener el **rendimiento** en ambos sistemas
  - DBMS están optimizados para OLTP. Métodos de acceso, indexación, control de concurrencia, mecanismos de recuperación.
  - DW está optimizado para OLAP. Resolver consultas complejas, vistas multidimensionales, consolidación, etc.
- Diferentes funciones y diferentes datos:
  - DSS requiere de **datos históricos**
  - Consolidación de datos: DSS<sup>1</sup> requieren consolidar (**agregación, sumariación**) datos heterogéneos.
  - Los OLTP se ocupan solo de las transacciones.

<sup>1</sup> *Decision Support System*

# Introducción a Data Warehouse (DW)

## Arquitectura de múltiples capas de un DW



# Introducción a Data Warehouse (DW)

## Tres modelos de DW

### DW Empresarial

recoge toda la información sobre temas que abarcan toda la organización

### Data Mart

un subconjunto de datos en toda la empresa que es de valor para un grupo específico de usuarios. Por ejemplo el ***data mart*** de marketing

### Virtual warehouse

Un conjunto de vistas sobre un sistema de **OLTP**

Solamente algunas de las posibles *sumarizaciones* pueden ser materializadas

# Introducción a Data Warehouse (DW)

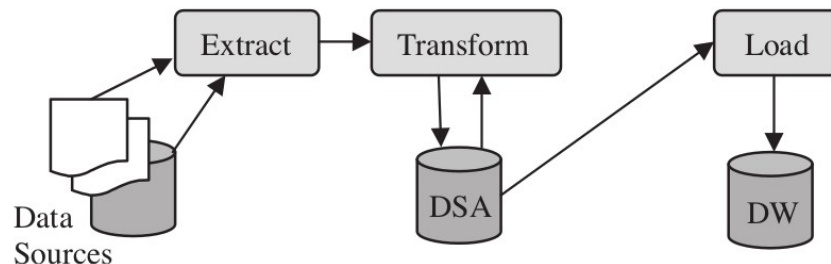
## Extraction, Transformation, and Loading (ETL)

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)



# Introducción a Data Warehouse (DW)

## Extraction, Transformation, and Loading (ETL)

### Data extraction

get data from **multiple, heterogeneous**, and external sources

### Data cleaning

**detect errors** in the data and rectify them when possible

### Data transformation

**convert data** from legacy or host format to warehouse format

### Load

sort, summarize, consolidate, compute views, check integrity, and build indices and partitions

### Refresh

**propagate the updates** from the data sources to the warehouse

# Introducción a Data Warehouse (DW)

## Metadata Repository

**Meta data** son los datos que definen a los objetos en el DW.

En él se almacenan:

- Descripciones de la estructura del DW: *schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents*
- **Operacional** meta-data: **el linaje de los datos** (historial sobre los datos migrados y las transformaciones), **datos en circulación** (active, archived, or purged), **información de monitoreo** (warehouse usage statistics, error reports, audit trails)
- Los **algoritmos** utilizados para la sumarización
- Cómo es el **mapeo** desde el OLTP al DW
- Datos relacionados con el rendimiento del sistema
  - warehouse schema, view and derived data definitions
- **Datos del negocio**
  - business terms and definitions, ownership of data, charging policies

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional

- Las herramientas de DW y OLAP se basan en un modelo de datos multidimensional
- Este modelo ve los datos como “**cubos**”
- Un **CUBO** permite que los datos sean modelados y visualizados en múltiples dimensiones.

Un cubo esta definido por 2 componentes:

- **Tablas de dimensiones**
  - **Tablas de Hechos**
- 
- **Dimension Tables:** tales como *items* (nombre, tipo, marca), o *tiempo* (días, semanas, meses, años)
  - **Fact Table:** Contiene las medidas (ej: ventas en pesos) y las claves para cada una de las tablas de dimensiones relacionadas.

En la literatura de almacenamiento de datos, un cubo de base de n-D se llama un **cuboide de base**. Más a la cima del esta el “cuboide” **0-D**, que tiene **el más alto nivel de resumen**, se llama el **cuboides ápice**.

El entramado de cuboides forma un cubo de datos.



# Introducción a Data Warehouse (DW)

## Modelo Multidimensional

### Tablas de dimensiones

- Representa lo que se quiere guardar en relación a un problema.
- Cada tabla a su vez puede tener asociadas otras tablas.
- Las Tablas de Dimensión pueden ser especificadas por usuarios o por expertos o generadas automáticamente y ajustadas a partir de la distribución de los datos.

### Claves Naturales vs Claves Subrogadas

Las claves existentes en los OLTP se denominan **claves naturales**;

Las **claves subrogadas** son aquellas que se definen artificialmente, son:

- de tipo numérico secuencial,
- no tienen relación directa con ningún dato
- y no poseen ningún significado en especial.

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional: ¿Por qué usar claves subrogadas?

**Fuentes heterogéneas.** El DW suele alimentarse de diferentes fuentes, cada una de ellas con sus propias claves, por lo que es arriesgado asumir un código de alguna aplicación en particular.

**Ejemplo:** Dos sistemas con claves su propia tabla de localidades.. ¿Qué ID le ponemos en el DW?

**Cambios en las aplicaciones origen.** Puede pasar que cambie la lógica operacional de alguna clave que hubiésemos supuesto única, o que ahora admite nulos.

**Ejemplo:** Algo raro... ¿Qué pasa si uno de los empleados no tiene nro de documento?

**Rendimiento.** Dado que un entero ocupa menos espacio que una cadena y además se lee mucho más rápido.

El problema en si no es el espacio, sino el **tiempo de lectura**.

Las claves subrogadas forman parte de la tabla de hechos, cada código se repite miles/millones de veces.

Será necesario optimizar todo lo posible.

**Lo mejor es crear nuestras propias claves subrogadas desde el inicio del proyecto.**

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional

### Tabla de Hechos

- El modelo multidimensional es organizado generalmente entorno a un tema.

**Ej: Ventas, Precipitaciones, etc.**

- Ese **tema tiene que estar representado** en la Tabla de Hechos.
- Los **hechos son medidas numéricas**, que se expresan generalmente en cantidades que van a permitir expresar las relaciones entre las dimensiones.
- La TH contiene los **nombres de los hechos o las medidas** y también las **claves para cada una de las Tablas de Dimensiones** que vamos a relacionar.

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional: Medidas

**Una medida consiste de dos componentes:**

- **propiedad numérica de un hecho**, como el **precio de venta o ganancia**
- **una fórmula**, por lo general una **función de agregación** simple, como suma, que pueden combinar varios valores de medida en una sola.

Las medidas pueden ser de tres clases:

**Aditivas:** Pueden ser combinadas a lo largo de una dimensión

Ventas totales del producto, localización, y el tiempo, porque esto no causa ningún solapamiento entre los fenómenos del mundo real que generaron los valores individuales.

**Semiaditivas:** No se las puede combinar a lo largo de una o más dimensiones

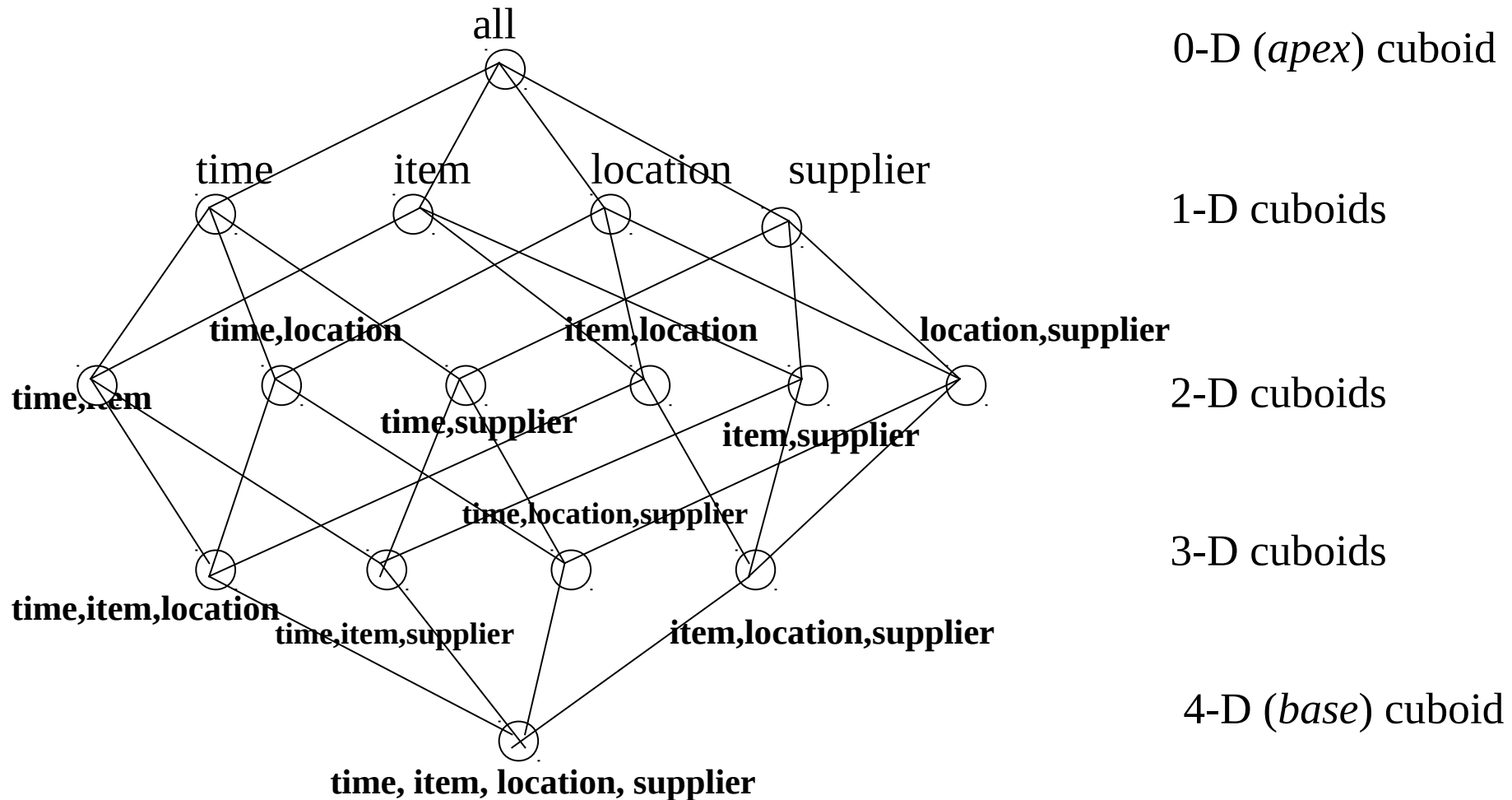
Resumir inventario a través de productos y almacenes es significativo, pero sumando los niveles de inventario a través del tiempo no tiene sentido

**No Aditivas:** No se puede combinar a lo largo de cualquier dimensión.

Por lo general debido a que la fórmula elegida impide que se combinen

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional



# Introducción a Data Warehouse (DW)

## Modelado conceptual del Data Warehouses

El **modelo de datos de ER** es utilizado en el diseño de bases de datos relacionales donde el esquema de la base consiste en un conjunto de **entidades y relaciones** entre ellas.

Este modelo es apropiado para OLTP

Un DW sin embargo, requiere un esquema conciso y orientado a un tema que facilite la tarea de OLAP

El abordaje más popular para diseño de DW es el **modelo multidimensional**

Este modelo, puede existir en forma de:

- Esquema de Estrella
- Esquema de copo de nieve
- Constelación de Hechos

# Introducción a Data Warehouse (DW)

## Esquema de Estrella

Es el esquema más utilizado, donde el DW contiene:

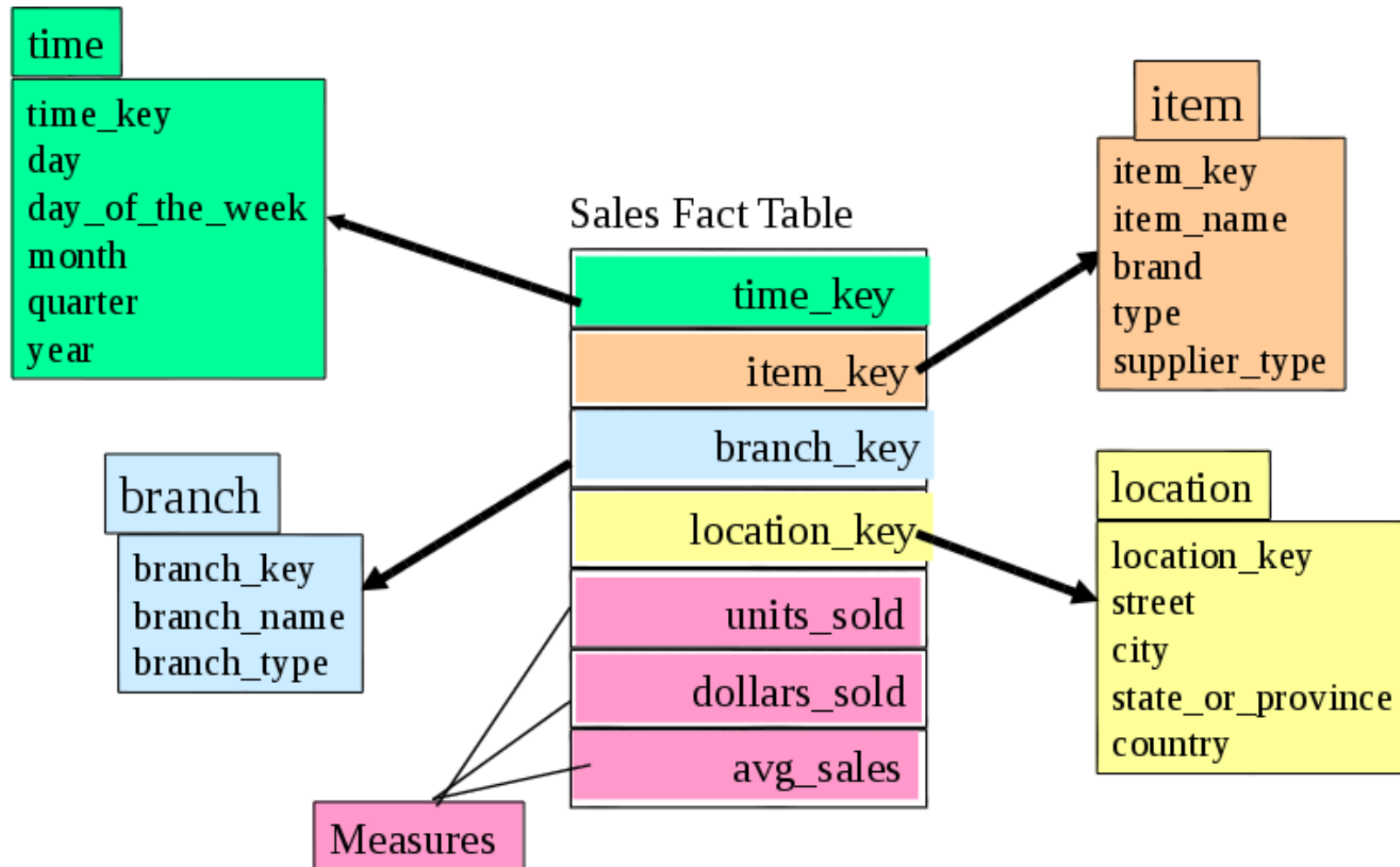
- 1) una gran tabla central (**Fact Table**) que contiene el volumen de datos sin redundancia
- 2) Un conjunto de tablas relacionadas (**Dimension Tables**) una por cada dimensión.

Cada dimensión es representada por una única tabla y cada tabla contiene un conjunto de atributos.

Los Atributos de una dimensión pueden formar una Jerarquía (Orden Total) o una grilla (lattice) (Orden Parcial)

# Introducción a Data Warehouse (DW)

## Esquema de Estrella





# Introducción a Data Warehouse (DW)

## Esquema de copo de nieve

Se trata de una variante del esquema Estrella donde algunas tablas de dimensiones son **Normalizadas**.

Con esta Normalización se generan tablas adicionales y el gráfico resultante forma una figura similar a un copo de nieve :D

El esquema snowflake **reduce la redundancia** generada en estrella a través de la normalización.

Las tablas son más fácil de mantener y **ahorra mas espacio de almacenamiento** (aunque es insignificante)

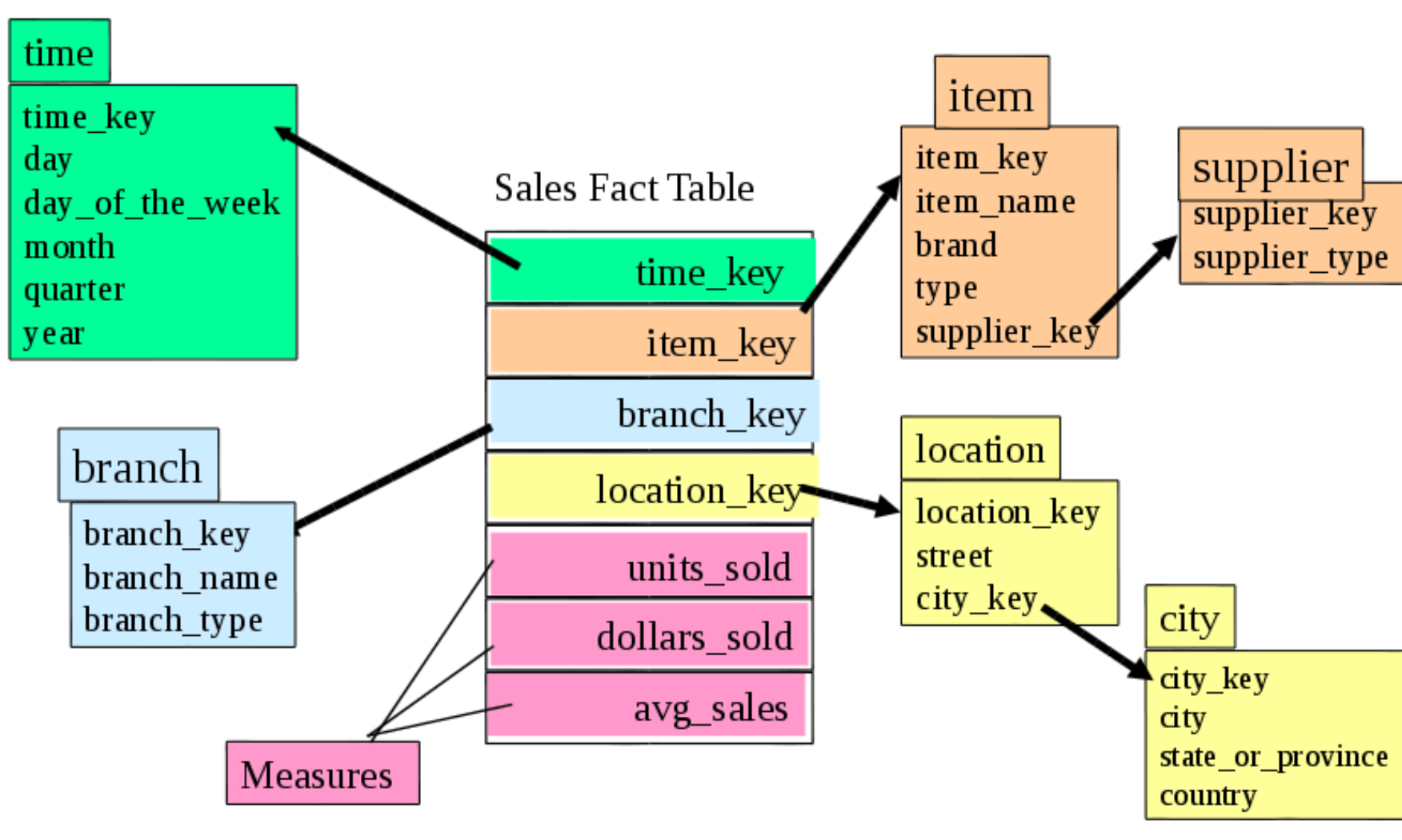
### Problema de snowflake:

La estructura puede reducir significativamente la efectividad de navegación debido a la cantidad de JOINS que son necesarios para correr una query.

Si bien reduce la redundancia **no es tan popular** como estrella en el diseño de DW

# Introducción a Data Warehouse (DW)

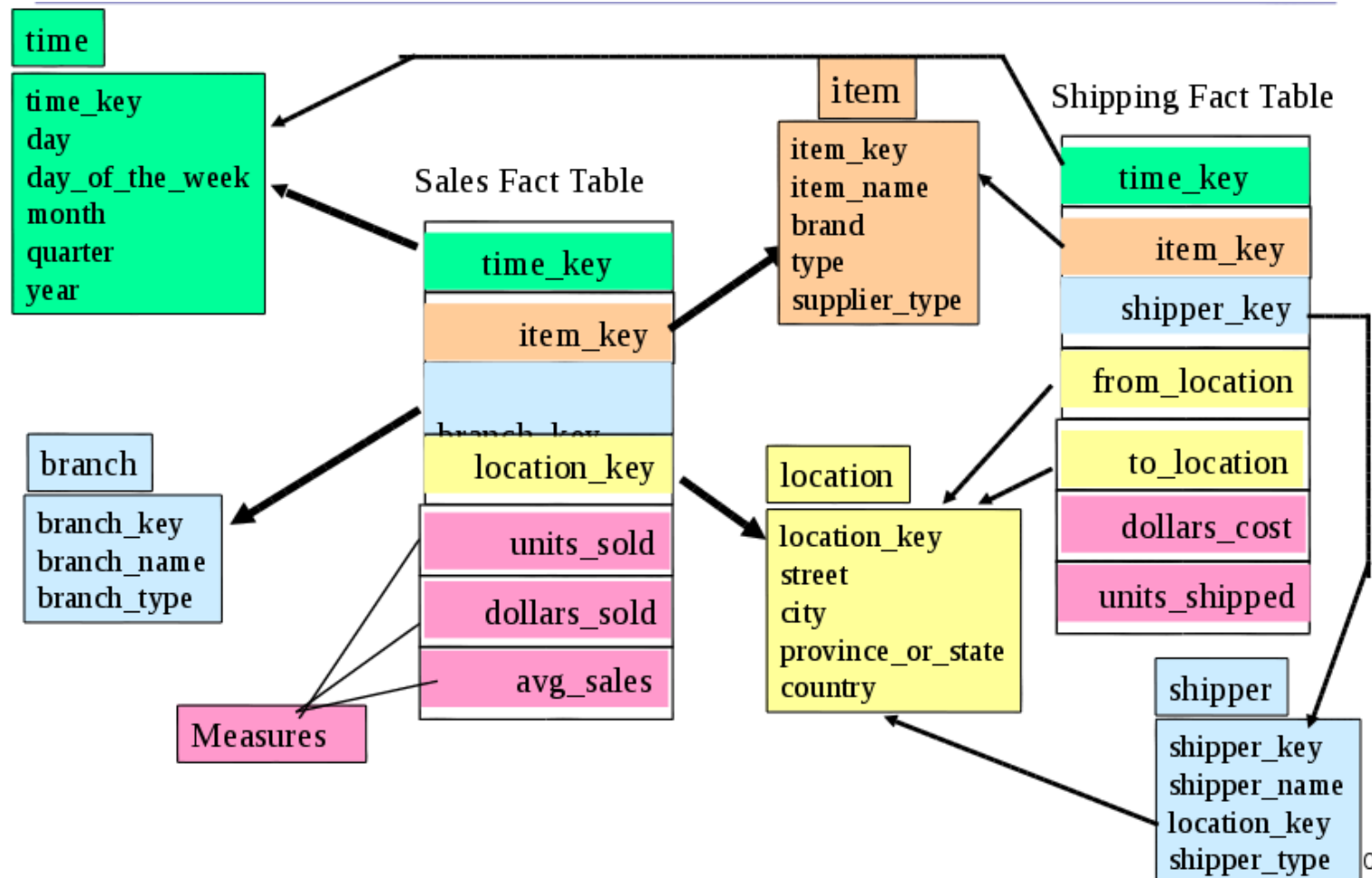
## Esquema de copo de nieve



# Introducción a Data Warehouse (DW)

## Esquema constelación de hechos

Son múltiples **tablas de hechos** que comparten **Tablas de Dimensiones** visto como una colección de esquemas de estrella, de ahí el nombre.



# Introducción a Data Warehouse (DW)

## Esquema Data Warehouse y Data Mart

En **data warehousing** Hay una distinción entre Data Warehouse y Data Mart:

**DW** recolecta información acerca de una temática que abarca a toda la organización (Clientes, personal, ventas)

En DW se utiliza habitualmente un esquema de constelación.

**Data Mart**, es un departamento/un subconjunto de los temas de la organización que se enfoca en un tema puntual, ej: ventas.

Para Data Mart, los esquemas de estrella y copo de nieve son los más utilizados.

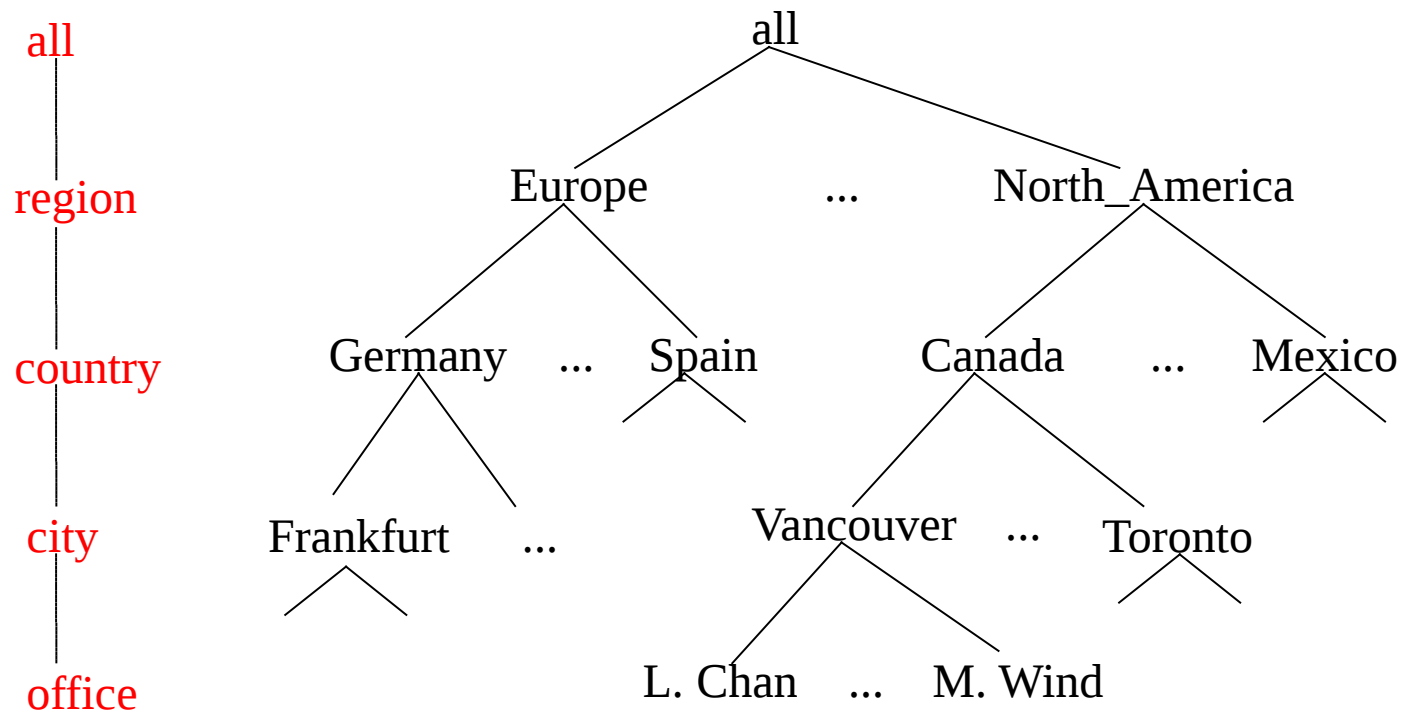
# Introducción a Data Warehouse (DW)

## Concepto de Jerarquía

El concepto de Jerarquía define una secuencia de mapeos de un conjunto de conceptos de bajo nivel a alto nivel, es decir, conceptos más generales.

Hay muchos conceptos de jerarquía que están implícitos en el DW, ejemplo de las ubicaciones.

El concepto de jerarquía permite que los datos se manejen en diferentes niveles de abstracción



# Introducción a Data Warehouse (DW)

## Modelo Multidimensional

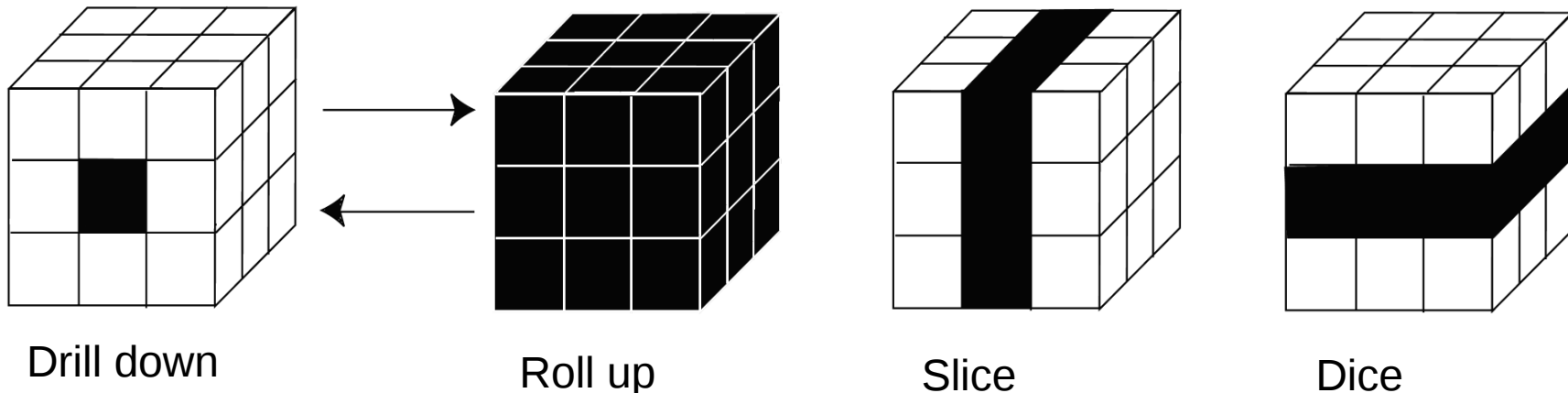
**Roll up (drill-up):** Datos Resumidos. Permite escalar la jerarquía o reducir dimensiones. Generalización y agregación

**Drill down (roll down):** Permite ir desde un alto nivel de resumen a un bajo nivel o datos detallados. Desagregación y especialización

**Slice:** Permite hacer un corte o proyección

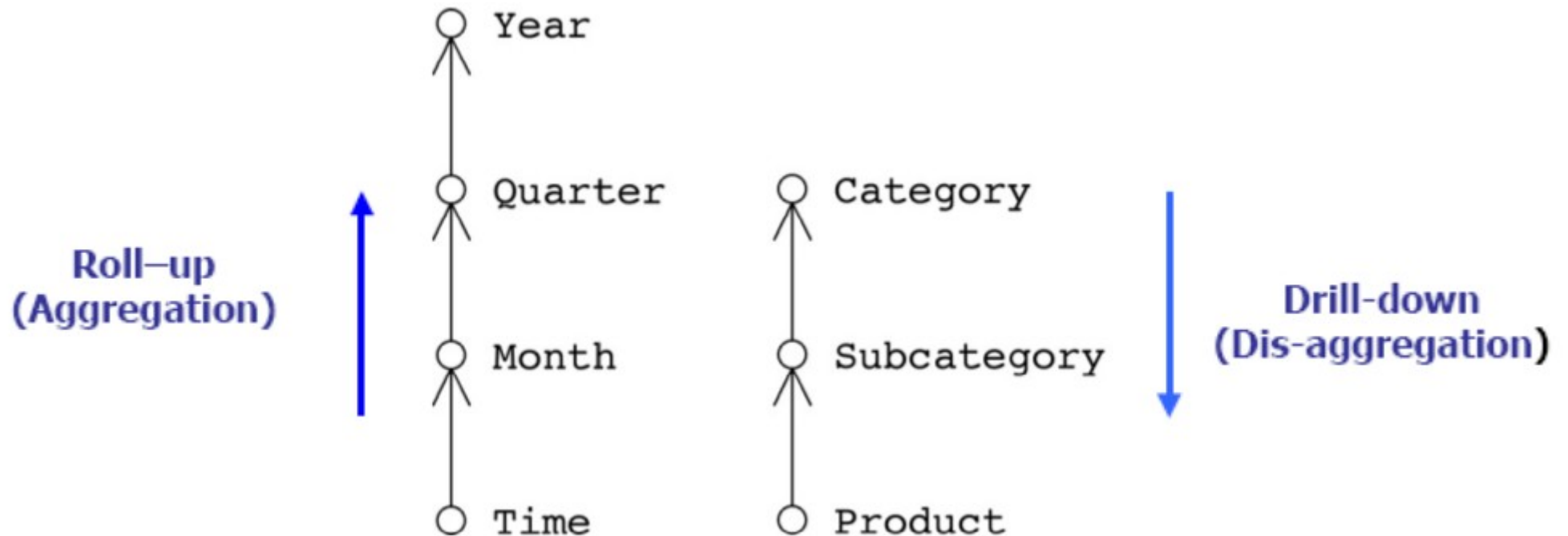
**Dice:** Permite seleccionar

**Pivot (Transpose):** Gira el cubo, lo rota en algún sentido



# Introducción a Data Warehouse (DW)

## Modelo Multidimensional



# Introducción a Data Warehouse (DW)


## Modelo Multidimensional

### Roll-Up

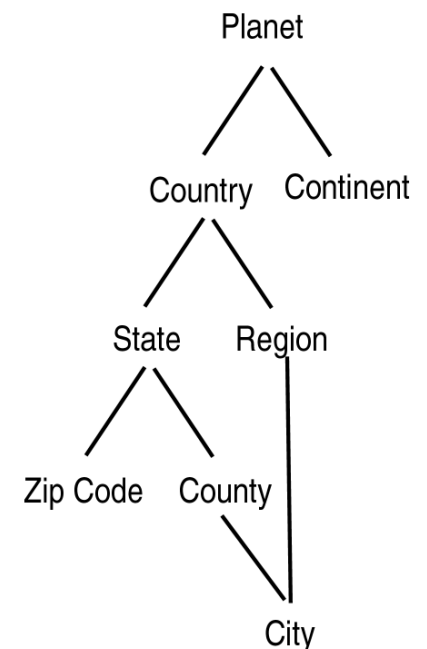
**Jerarquía: Ubicación**

Category	Year	Metrics Customer Region	Revenue						
			Northeast	Mid-Atlantic	Southeast	Central	South	Northwest	Southwest
Books	2005		\$416,183	\$316,104	\$36,517	\$207,850	\$137,502	\$19,062	\$187,368
	2006		\$534,932	\$401,908	\$42,027	\$239,806	\$138,683	\$22,655	\$183,275
Electronics	2005		\$1,860,172	\$6,517,723	\$1,226,825	\$3,719,752	\$915,633	\$1,434,575	\$3,625,191
	2006		\$2,403,311	\$8,253,620	\$1,451,397	\$4,631,259	\$999,611	\$1,615,848	\$4,298,985
Movies	2005		\$112,560	\$138,611	\$118,179	\$153,556	\$119,566	\$27,060	\$362,858
	2006		\$148,785	\$188,567	\$147,445	\$203,547	\$145,434	\$35,878	\$463,470
Music	2005		\$78,507	\$99,631	\$25,528	\$383,911	\$38,373	\$27,933	\$95,083
	2006		\$104,925	\$126,851	\$35,215	\$485,174	\$43,424	\$33,860	\$110,689

**Métrica: Ingresos**



Category	Year	Metrics	Revenue
Books	2005		\$1,320,585
	2006		\$1,563,287
Electronics	2005		\$19,299,870
	2006		\$23,654,030
Movies	2005		\$1,032,391
	2006		\$1,333,126
Music	2005		\$748,966
	2006		\$940,136



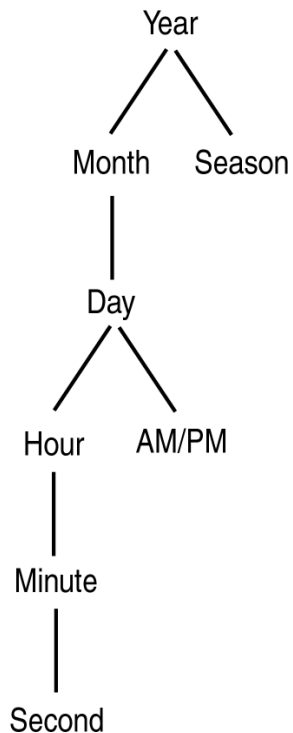


# Introducción a Data Warehouse (DW)

## Modelo Multidimensional

### Roll-Up

Jerarquía: Tiempo



Month	Metrics	Revenue						
	Customer Region	Northeast	Mid-Atlantic	Southeast	Central	South	Northwest	Southwest
Jan 2005		\$160,155	\$518,405	\$81,381	\$322,294	\$98,001	\$103,368	\$298,730
Feb 2005		\$170,777	\$491,628	\$80,399	\$314,466	\$91,222	\$114,341	\$373,645
Mar 2005		\$200,434	\$611,424	\$129,102	\$382,946	\$123,038	\$147,472	\$351,602
Apr 2005		\$194,811	\$502,241	\$93,654	\$357,188	\$90,663	\$124,824	\$304,416
May 2005		\$169,998	\$462,364	\$117,780	\$300,389	\$79,999	\$117,506	\$374,488
Jun 2005		\$202,477	\$559,466	\$109,979	\$309,683	\$115,318	\$117,008	\$373,526
Jul 2005		\$194,490	\$577,515	\$105,099	\$332,300	\$92,730	\$103,494	\$369,380
Aug 2005		\$203,085	\$599,761	\$118,805	\$410,885	\$119,178	\$131,148	\$384,555
Sep 2005		\$241,992	\$625,517	\$122,261	\$415,763	\$75,655	\$124,974	\$364,651
Oct 2005		\$217,477	\$641,340	\$137,925	\$382,321	\$89,679	\$124,276	\$337,489
Nov 2005		\$238,004	\$708,036	\$156,525	\$457,105	\$116,478	\$156,466	\$386,399
Dec 2005		\$273,721	\$774,372	\$154,139	\$479,729	\$119,113	\$143,753	\$414,983
Jan 2006		\$215,786	\$662,632	\$125,238	\$392,922	\$91,791	\$122,235	\$343,027
Feb 2006		\$253,128	\$711,937	\$123,725	\$415,742	\$97,309	\$137,589	\$391,277
Mar 2006		\$253,564	\$704,652	\$135,180	\$430,143	\$112,459	\$144,659	\$406,956
Apr 2006		\$255,352	\$710,402	\$126,717	\$426,423	\$113,233	\$140,976	\$395,924
May 2006		\$231,766	\$676,205	\$130,981	\$440,813	\$107,277	\$136,043	\$377,349
Jun 2006		\$290,534	\$769,788	\$123,743	\$507,166	\$125,631	\$131,549	\$439,321
Jul 2006		\$247,683	\$811,060	\$145,955	\$448,939	\$113,683	\$128,113	\$415,251
Aug 2006		\$252,313	\$719,509	\$125,944	\$427,188	\$108,987	\$153,966	\$421,310
Sep 2006		\$288,772	\$801,819	\$148,023	\$539,406	\$112,784	\$149,236	\$419,878
Oct 2006		\$307,610	\$710,458	\$163,254	\$450,006	\$105,218	\$144,906	\$440,856
Nov 2006		\$284,671	\$800,941	\$157,117	\$505,952	\$118,552	\$163,560	\$470,591
Dec 2006		\$310,775	\$891,543	\$170,207	\$575,086	\$120,228	\$155,409	\$534,680

Métrica: Ingresos

Quarter	Metrics	Revenue						
	Customer Region	Northeast	Mid-Atlantic	Southeast	Central	South	Northwest	Southwest
2005 Q1		\$531,366	\$1,621,457	\$290,882	\$1,019,706	\$312,261	\$365,181	\$1,023,977
2005 Q2		\$567,286	\$1,524,070	\$321,413	\$967,260	\$285,981	\$359,339	\$989,065
2005 Q3		\$639,567	\$1,802,793	\$346,165	\$1,158,948	\$287,563	\$359,616	\$1,118,587
2005 Q4		\$729,202	\$2,123,749	\$448,589	\$1,319,154	\$325,269	\$424,495	\$1,138,871
2006 Q1		\$722,478	\$2,079,221	\$384,143	\$1,238,807	\$301,559	\$404,483	\$1,141,260
2006 Q2		\$777,651	\$2,156,394	\$381,441	\$1,374,402	\$346,141	\$408,567	\$1,212,593
2006 Q3		\$788,768	\$2,332,388	\$419,923	\$1,415,533	\$335,455	\$431,315	\$1,256,439
2006 Q4		\$903,056	\$2,402,942	\$490,578	\$1,531,044	\$343,998	\$463,874	\$1,446,127

# Introducción a Data Warehouse (DW)

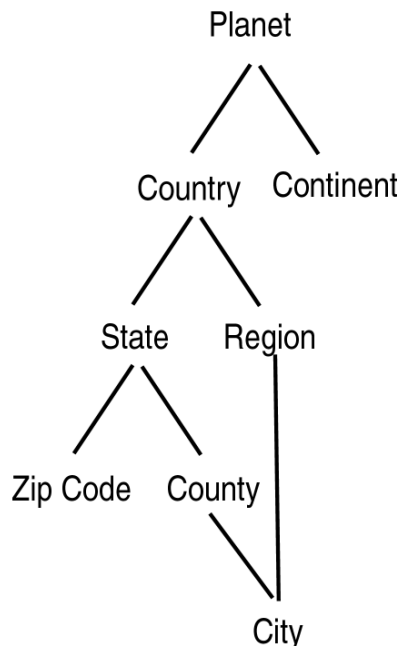
## Modelo Multidimensional

### Drill-Down

**Jerarquía: Ubicación**

**Métrica: Ingresos**

Quarter	Metrics	Revenue						
	Customer Region	Northeast	Mid-Atlantic	Southeast	Central	South	Northwest	Southwest
2005 Q1		\$531,366	\$1,621,457	\$290,882	\$1,019,706	\$312,261	\$365,181	\$1,023,977
2005 Q2		\$567,286	\$1,524,070	\$321,413	\$967,260	\$285,981	\$359,339	\$989,065
2005 Q3		\$639,567	\$1,802,793	\$346,165	\$1,158,948	\$287,563	\$359,616	\$1,118,587
2005 Q4		\$729,202	\$2,123,749	\$448,589	\$1,319,154	\$325,269	\$424,495	\$1,138,871
2006 Q1		\$722,478	\$2,079,221	\$384,143	\$1,238,807	\$301,559	\$404,483	\$1,141,260
2006 Q2		\$777,651	\$2,156,394	\$381,441	\$1,374,402	\$346,141	\$408,567	\$1,212,593
2006 Q3		\$788,768	\$2,332,388	\$419,923	\$1,415,533	\$335,455	\$431,315	\$1,256,439
2006 Q4		\$903,056	\$2,402,942	\$490,578	\$1,531,044	\$343,998	\$463,874	\$1,446,127



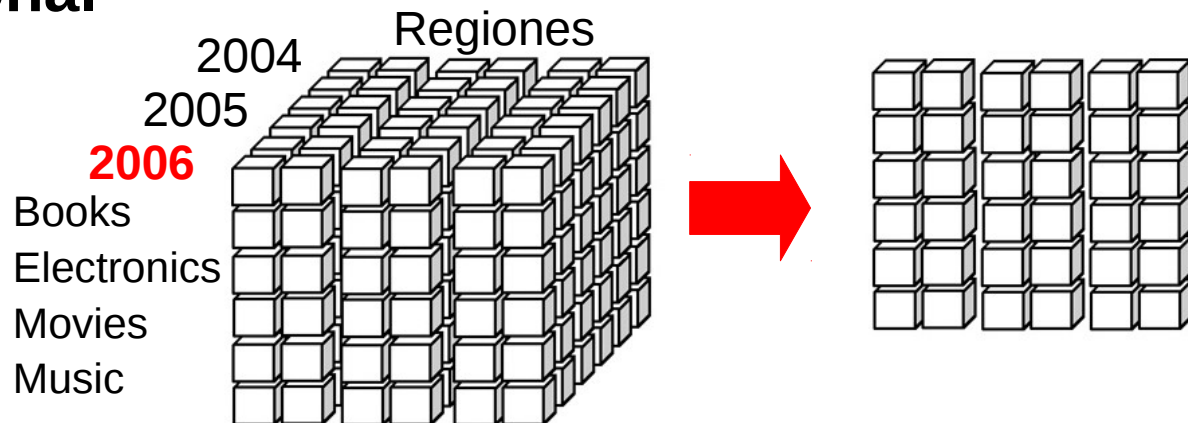
Quarter	Metrics	Revenue											
	Customer City	Addison	Akron	Albany	Albert City	Alexandria	Allentown	Anderson	Annapolis	Arden	Arlington Heights	Arlington	Arli
2005 Q1		\$7,713	\$30,140	\$4,626	\$6,686	\$29,042	\$4,579	\$1,948	\$40,066	\$23,341	\$10,481		
2005 Q2		\$15,903	\$48,029	\$8,959	\$2,088	\$17,590	\$7,268	\$2,416	\$42,764	\$21,026	\$7,514		\$
2005 Q3		\$10,091	\$30,510	\$5,763	\$11,380	\$26,389	\$10,195	\$568	\$51,650	\$25,132	\$10,784		\$
2005 Q4		\$12,425	\$48,588	\$10,939	\$10,463	\$28,016	\$10,426	\$3,412	\$67,515	\$28,398	\$14,692		
2006 Q1		\$7,256	\$26,183	\$7,998	\$5,603	\$35,959	\$10,273	\$2,732	\$50,121	\$26,351	\$7,276		\$
2006 Q2		\$10,411	\$49,540	\$6,065	\$5,670	\$25,166	\$6,992	\$1,377	\$68,198	\$27,556	\$16,755		\$
2006 Q3		\$10,325	\$38,414	\$9,108	\$7,760	\$33,170	\$16,978	\$821	\$69,858	\$33,579	\$10,235		\$
2006 Q4		\$16,613	\$49,133	\$13,137	\$10,637	\$30,344	\$13,875	\$747	\$48,305	\$32,842	\$13,311		\$

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional

### Slicing

Jerarquía: Tiempo



Category	Year	Metrics Customer Region	Revenue						
			Northeast	Mid-Atlantic	Southeast	Central	South	Northwest	Southwest
Books	2005		\$416,183	\$316,104	\$36,517	\$207,850	\$137,502	\$19,062	\$187,368
	2006		\$534,932	\$401,908	\$42,027	\$239,806	\$138,683	\$22,655	\$183,275
Electronics	2005		\$1,860,172	\$6,517,723	\$1,226,825	\$3,719,752	\$915,633	\$1,434,575	\$3,625,191
	2006		\$2,403,311	\$8,253,620	\$1,451,397	\$4,631,259	\$999,611	\$1,615,848	\$4,298,985
Movies	2005		\$112,560	\$138,611	\$118,179	\$153,556	\$119,566	\$27,060	\$362,858
	2006		\$148,785	\$188,567	\$147,445	\$203,547	\$145,434	\$35,878	\$463,470
Music	2005		\$78,507	\$99,631	\$25,528	\$383,911	\$38,373	\$27,933	\$95,083
	2006		\$104,925	\$126,851	\$35,215	\$485,174	\$43,424	\$33,860	\$110,689

**Slicing** en OLAP es una columna de datos correspondientes a un solo valor para uno o más elementos de la dimensión.

Ayuda a visualizar y recopilar información específica de una dimensión

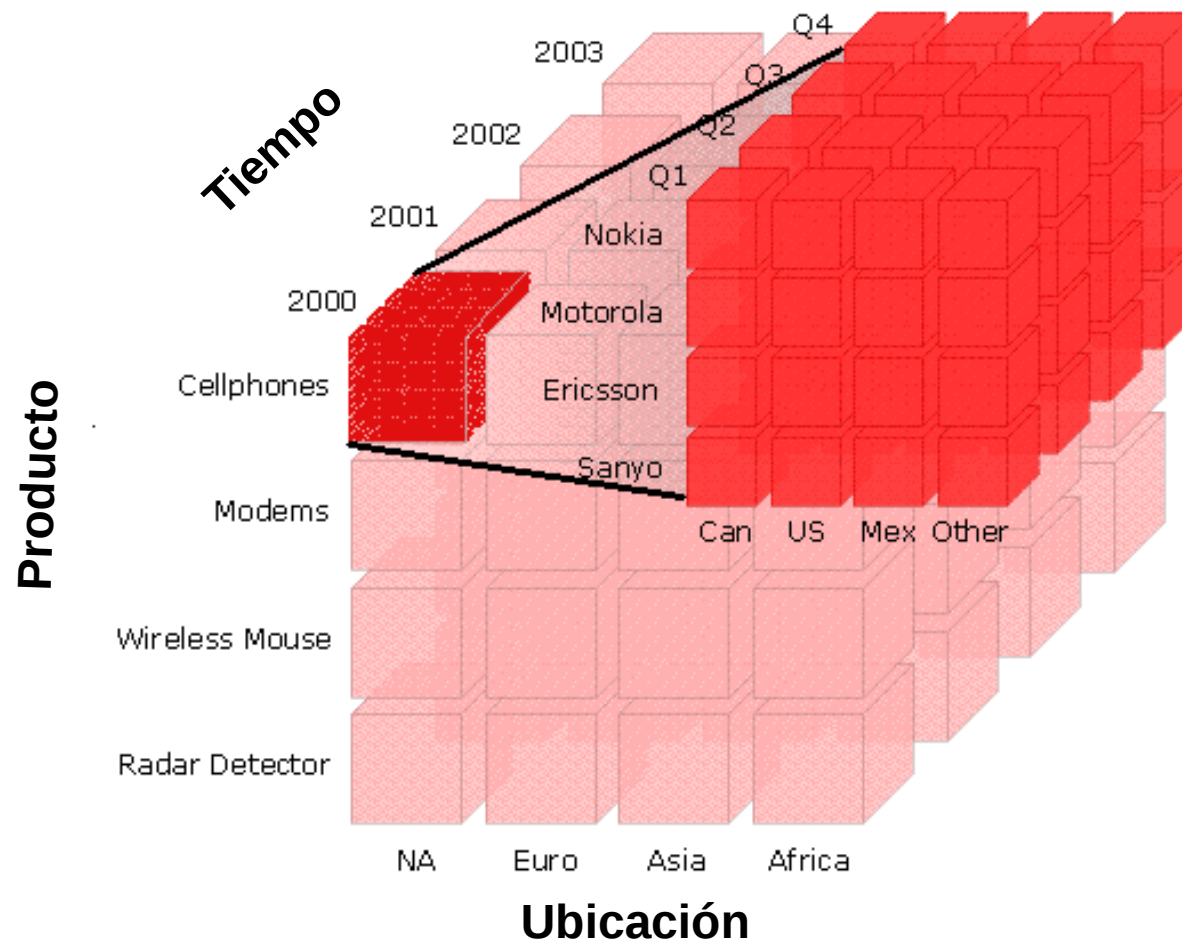
Report Filter (Local Filter):  
Year = 2006

Category	Metrics Customer Region	Revenue						
		Northeast	Mid-Atlantic	Southeast	Central	South	Northwest	Sout
Books		\$534,932	\$401,908	\$42,027	\$239,806	\$138,683	\$22,655	\$1
Electronics		\$2,403,311	\$8,253,620	\$1,451,397	\$4,631,259	\$999,611	\$1,615,848	\$4,2
Movies		\$148,785	\$188,567	\$147,445	\$203,547	\$145,434	\$35,878	\$4
Music		\$104,925	\$126,851	\$35,215	\$485,174	\$43,424	\$33,860	\$1

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional

### Dicing



La operación **Dice** es más una función de zoom que selecciona un subconjunto sobre todas las dimensiones, pero para valores específicos de la dimensión

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional: Servidores OLAP

### OLAP relacional (ROLAP)

Utilizar la **tecnología de base de datos relacional** para el almacenamiento, y también emplean estructuras de índices especializados, como *bit-map index*, para lograr un buen rendimiento de las consultas.

### OLAP multidimensional (MOLAP)

- Servidor especialmente desarrollado para almacenar y consultar datos multidimensionales
- Utiliza estructuras de datos basadas en arreglos

### OLAP híbrido (HOLAP)

- Los datos detallados se almacenan en una BD relacional
- Almacena datos agregados en forma multidimensional
- Se accede a los datos a través de herramientas MOLAP



# Introducción a Data Warehouse (DW)

## Modelo Multidimensional: Servidores OLAP

### OLAP multidimensional (MOLAP): Capacidad de análisis

- Ofrece vistas de objetos multidimensionales
- El **tiempo de respuesta cero**, ya que todo está previamente calculado.
- Si no se calcula previamente todo, **la capacidad de análisis se limita** a aquellas porciones del cubo que ya fueron previamente calculadas.

### Sistema de diseño suele ser propietario

- Generalmente el cubo se trata de una “caja negra” de datos encriptados que pueden residir de forma local o en un servidor MOLAP.
- Flexibilidad y escalabilidad limitados.
- Cambios en el modelo dimensional del negocio implican la generación de todos los cubos nuevamente.

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional: MOLAP Ventajas y Desventajas

### Ventajas

- **Consultas rápidas** debido a la optimización del rendimiento de almacenamiento, la indexación multidimensional y la memoria caché.
- **Ocupa menor tamaño en disco** en comparación con los datos almacenados en base de datos relacional debido a técnicas de compresión.
- **Automatización del procesamiento de los datos agregados** de mayor nivel.
- Muy **compacto** para conjuntos de datos de pocas dimensiones.
- El **modelo de almacenamiento en vectores/matrices** proporciona una indexación natural.
- **Eficaz extracción de datos** lograda gracias a la pre-estructuración de los datos agregados.

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional: MOLAP Ventajas y Desventajas

### Desventajas

- La **etapa de procesamiento y carga de datos**, puede ser bastante larga, sobre todo para grandes volúmenes de datos. (Puede evitarse haciendo un procesamiento incremental)
- Las herramientas MOLAP tradicionalmente tienen **dificultades para consultar** con modelos con **dimensiones muy altas (del orden de millones de miembros)**.
- Algunas herramientas MOLAP (por ejemplo, Essbase) tienen **dificultades** para actualizar y consultar los **modelos con más de diez dimensiones**.
  - Este límite varía en función de la **complejidad y la cardinalidad** de las dimensiones de que se trate.
  - También **depende de la cantidad de hechos** o medidas almacenados. Otras herramientas MOLAP puede manejar cientos de dimensiones.
- El **enfoque MOLAP introduce redundancia** en los datos.



# Introducción a Data Warehouse (DW)

## Modelo Multidimensional: ROLAP - Capacidad de análisis

- Ofrece vistas de objetos multidimensionales.
- Tiempos de respuestas que rondan entre los **segundos y los minutos**.
- Existen técnicas de **tuning, caching, materialización de vistas, indexación y esquema de diseño** que mejoran la performance de respuesta de los ROLAP.
- Los datos se almacenan en tablas relacionales de DB Relacionales.
- Uso de esquemas:
  - Esquema estrella (star)
  - Esquema copo de nieve (snowflakes)
- **Es el enfoque más común en la práctica**

# Introducción a Data Warehouse (DW)

## Modelo Multidimensional: ROLAP - Capacidad de análisis

### Sistema de diseño abierto

- El cliente interactúa directamente contra el RDBMS vía SQL en distintos motores.
- Provee **flexibilidad y escalabilidad**.
- Los **cambios en el modelo dimensional** del negocio son trasladados al DW e **inmediatamente** se encuentra disponible para consultar.
- La **ventana de carga** del data warehouse es menor pues no existe el tiempo de generación de los multi-cubos.

Los ambientes adecuados para ROLAP son:

- Modelos dimensionales grandes y dinámicos.
- Grandes volúmenes de datos.
- Necesidad de **análisis a nivel transaccional**.

# Referencias

- Pedersen, T. B., & Jensen, C. S. (2001). Multidimensional database technology. *Computer*, 34(12), 40-46.
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques: concepts and techniques*. Elsevier.
- Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.