Bases de Datos Masivas

Data Warehouse

Bases de Datos Multidimensionales

Agosto 2016

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

	OLTP System Online Transaction Processing (Operational System)	OLAP System Online Analytical Processing (Data Warehouse)
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

Introducción a Data Warehouse (DW)

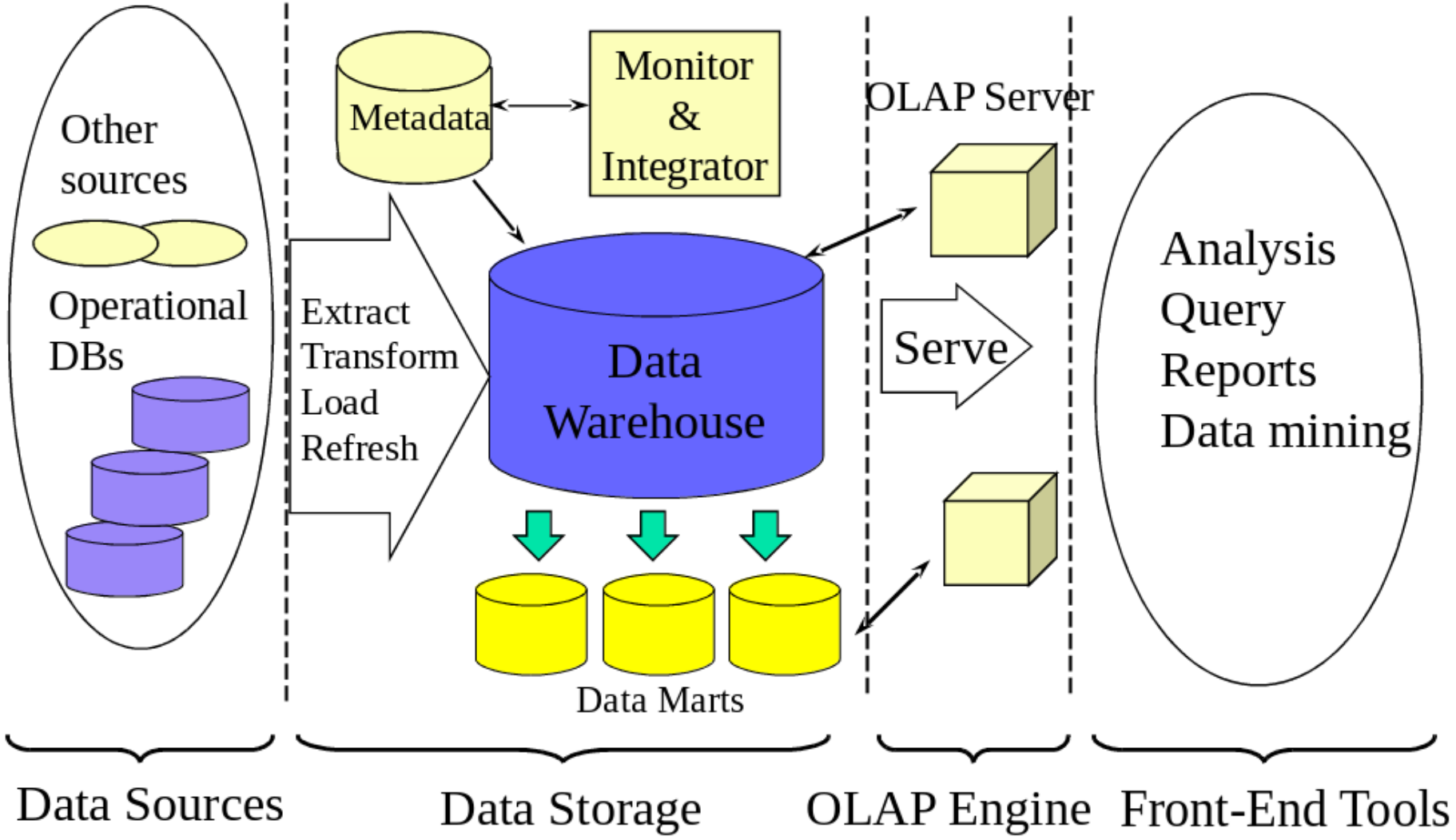
¿Por qué tener un DW separado?

- 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[1] requieren consolidar (**agregación**, **sumarización**) datos heterogéneos.
  - Los OLTP se ocupan solo de las transacciones.

[1] ***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 a **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

Son solo un **conjunto de vistas** sobre un sistema de **OLTP.**

Es fácil de construir aunque solamente soporta algunas operaciones de resumen y agregación.

La construcción de un DW virtual requiere un exceso de capacidad en los servidores de las DB operacionales.

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: esquema*, vistas, dimensiones, jerarquías, definiciones de datos derivados, ubicaciones de los data mart y contenidos.*
- 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
- Datos del negocio
  - Definiciones y términos del negocio, propietario de los datos, políticas de facturación

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**

- **Tablas de dimensiones**: tales como ***items*** (nombre, tipo, marca), o ***tiempo*** (días, semanas, meses, años)

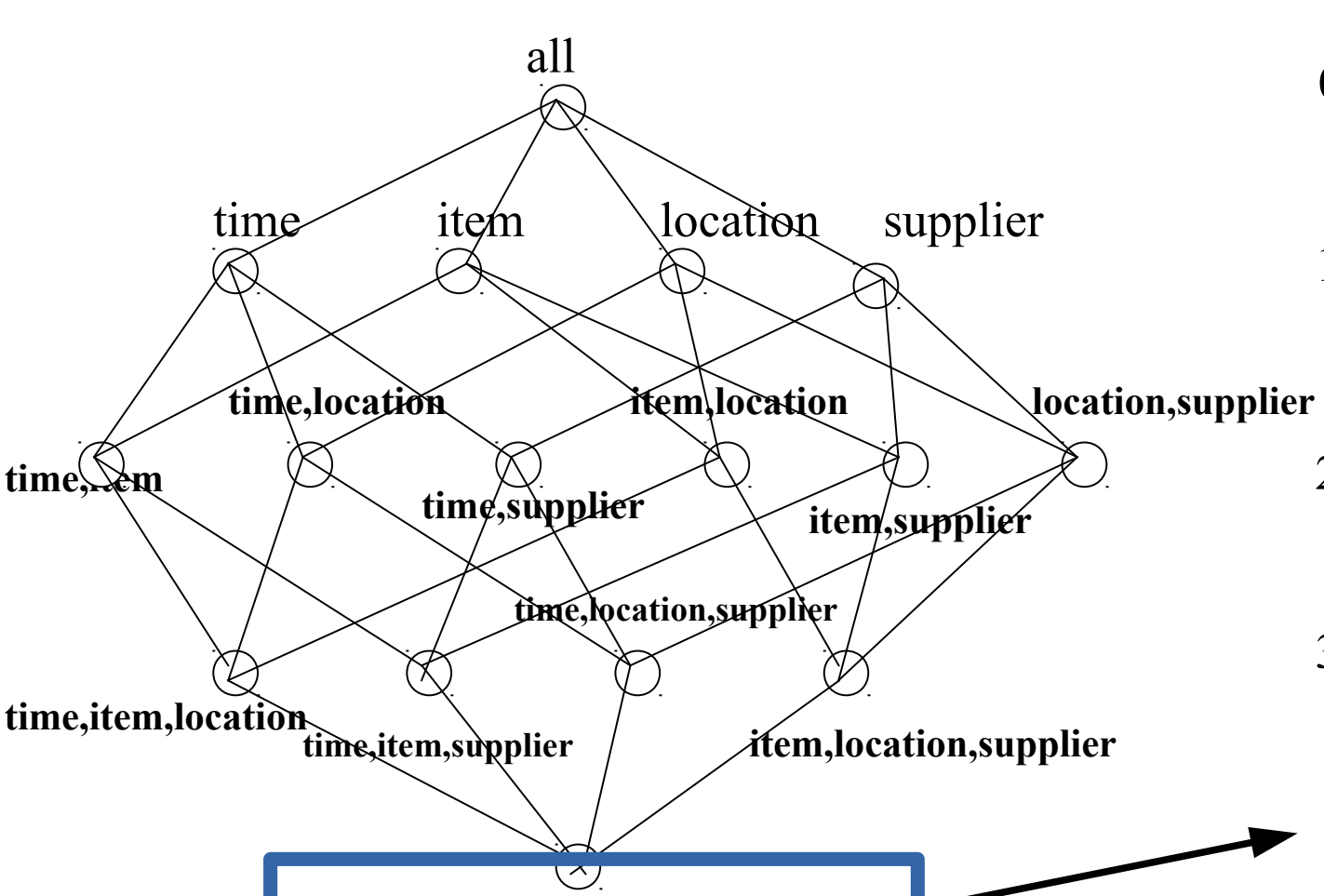
- **Tablas de Hechos**: Contiene las medidas (ej: ventas en $$$) y las claves para cada una de las tablas de dimensiones relacionadas.

En la literatura de almacenamiento de datos, un cubo de base de n-D se llama un **cuboide de base**. Más a la cima 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)

Cubos como Grafos (GDA)



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (*base*) cuboid

Conforme se unen los cuboides hay mayor nivel de agregación

Zhao, 2011

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 múltidimensional 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 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)

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

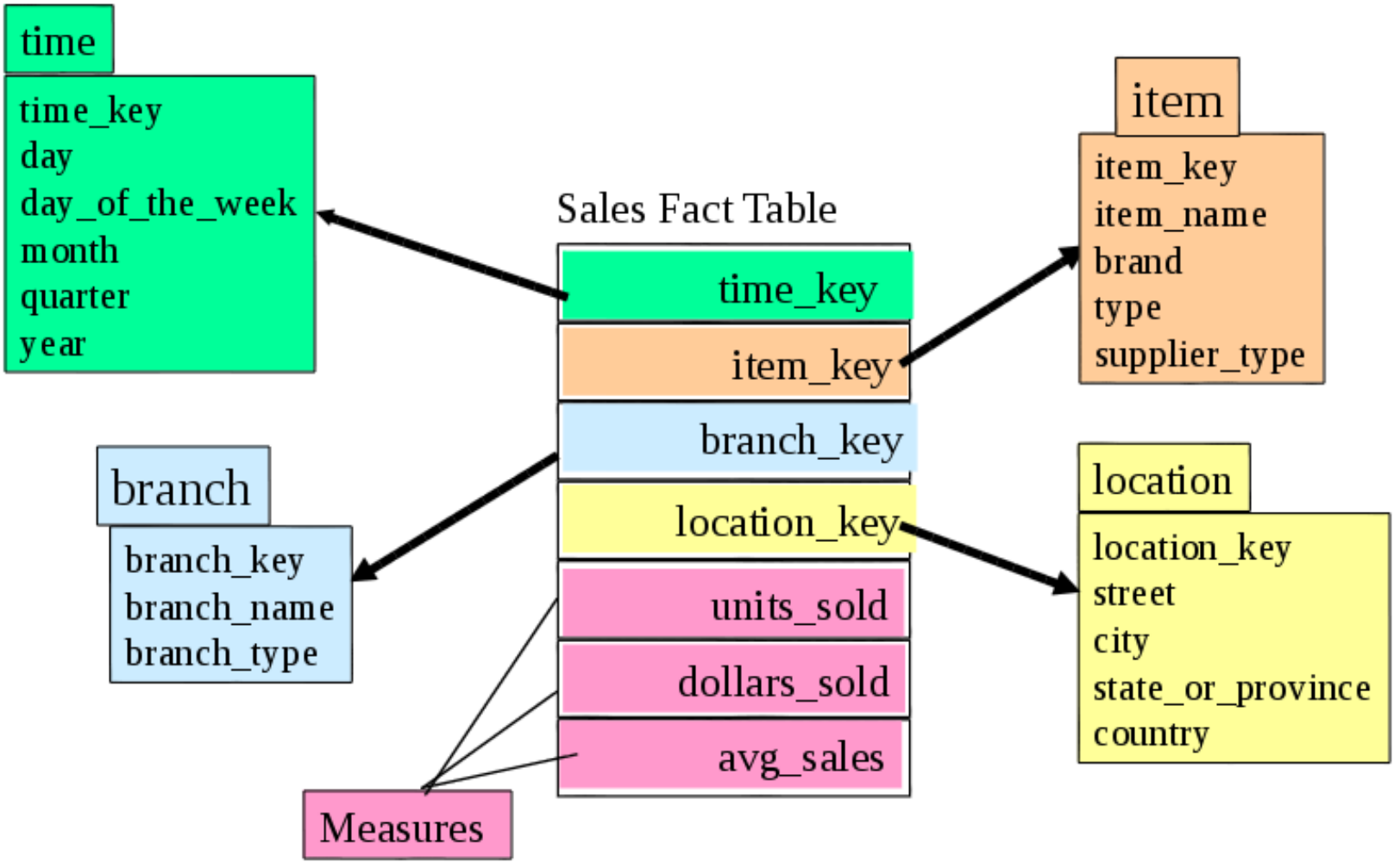
2)Un conjunto de tablas relacionadas (**Dimension Tables**) una por cada dimensión.

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)

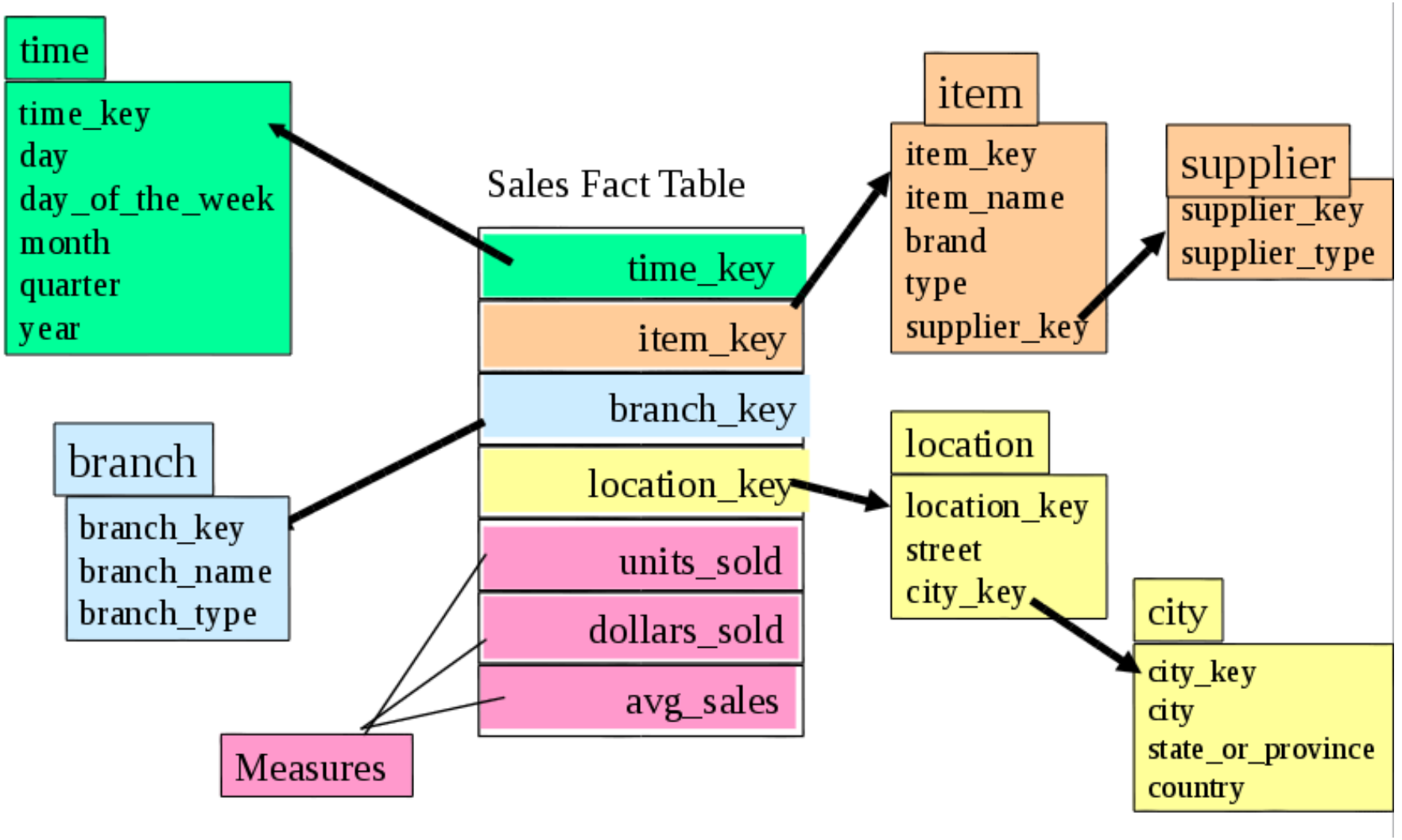
Problema de snowflake:

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)

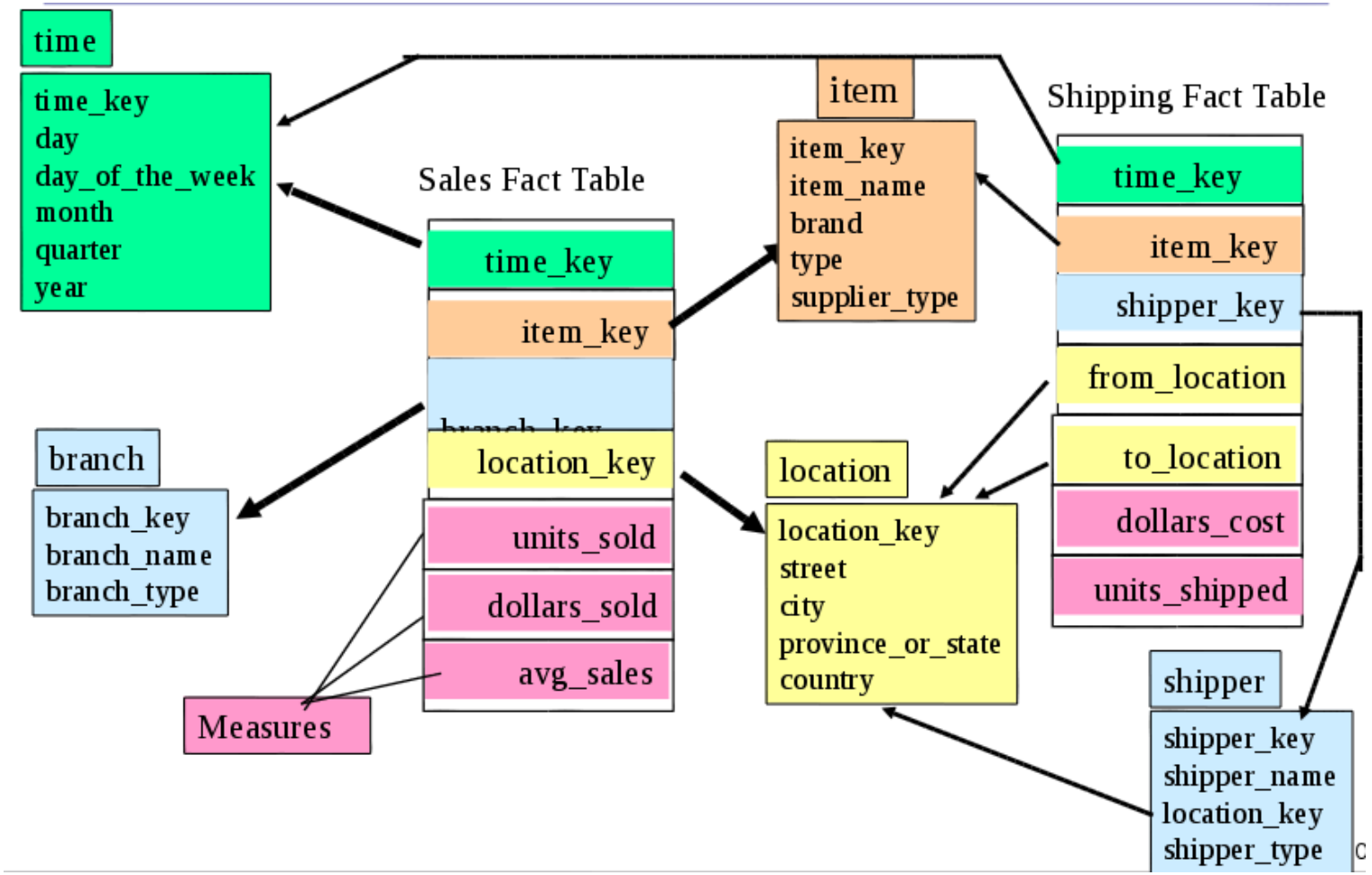
Esquema de copo de nieve



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 wharehousing** 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.

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.
- Zhao (2011) Graph Cube: On Warehousing and OLAP Multidimensional Networks.