

ALMACENES DE DATOS Y MINERÍA DE DATOS

Dra. Amparo López Gaona
Fac. Ciencias, UNAM

Marzo 2018

INTRODUCCIÓN

- Cada día crece, en forma espectacular, la cantidad de datos generados y registrados.
- Principales fuentes de datos:



INTRODUCCIÓN


- Cada día crece, en forma espectacular, la cantidad de datos generados y registrados.
- Principales fuentes de datos:
 - Comercio: electrónico, transacciones, almacenes, etc.
 - Ciencia: simulación científica, bioinformático, procesamiento de imágenes, etc.
 - Diario: noticias, cámaras digitales, YouTube,...
- Estamos ahogados en datos, pero sedientos de conocimiento.




INTRODUCCIÓN

- Cada día crece, en forma espectacular, la cantidad de datos generados y registrados.
- Principales fuentes de datos:
 - Comercio: electrónico, transacciones, almacenes, etc.
 - Ciencia: simulación científica, bioinformático, procesamiento de imágenes, etc.
 - Diario: noticias, cámaras digitales, YouTube,...
- Estamos ahogados en datos, pero sedientos de conocimiento.



- En la actualidad, las aplicaciones sobre **bases de datos** son muy importantes para la vida de una organización.
 - Reúnen, almacenan y procesan todos los datos necesarios para la ejecución exitosa de las operaciones diarias de las organizaciones.
 - Proporcionan información en línea y producen reportes para monitorear las organizaciones.
 - Se han desarrollado métodos eficientes para el procesamiento de transacciones en línea OLTP, donde una consulta se ve como una transacción.
 - Se han optimizado los algoritmos para edición e inserción de datos.
- Aproximadamente el 90 % de los SABD son relacionales.

- En la actualidad, las aplicaciones sobre **bases de datos** son muy importantes para la vida de una organización.
 - Reúnen, almacenan y procesan todos los datos necesarios para la ejecución exitosa de las operaciones diarias de las organizaciones.
 - Proporcionan información en línea y producen reportes para monitorear las organizaciones.
 - Se han desarrollado métodos eficientes para el procesamiento de transacciones en línea OLTP, donde una consulta se ve como una transacción.
 - Se han optimizado los algoritmos para edición e inserción de datos.
- Aproximadamente el 90 % de los SABD son relacionales.
- **Datos históricos → almacenamientos externos**

... INTRODUCCIÓN

- Al expandirse los negocios, su complejidad crece; los ejecutivos y administradores requieren información para ser competitivos.
 - Necesitan información para formular las estrategias del negocio, establecer metas, alcanzar objetivos y monitorear resultados.
- Ejemplos de objetivos de negocios:



- Al expandirse los negocios, su complejidad crece; los ejecutivos y administradores requieren información para ser competitivos.
 - Necesitan información para formular las estrategias del negocio, establecer metas, alcanzar objetivos y monitorear resultados.
- Ejemplos de objetivos de negocios:
 - Conservar su clientela base.
 - Aumentar su clientela un $x\%$ en los n años siguientes.
 - Mejorar los niveles de calidad de sus principales productos.
 - Incrementar sus ventas un $x\%$ en cierta región, etc.
 - Mejorar el servicio al cliente en ...
 - etc.



... INTRODUCCIÓN

Para lograr estos objetivos, los ejecutivos necesitan información para



- Conocer a profundidad las operaciones de la compañía.
- Revisar y monitorear los indicadores de rendimiento, notar cómo afectan unos a otros.
- Llevar registro de cómo cambian los factores de negocios en el tiempo y comparar el rendimiento de su compañía en relación a la competencia e industria.
- Enfocar su atención en las necesidades y preferencias de los clientes.
- Conocer resultados de mercadotecnia y ventas.
- Conocer niveles de calidad, de productos y servicios.

Este tipo de información esencial se llama **información estratégica**.

- La información estratégica no pretende ayudar en la producción de facturas, hacer envíos, etc. es más importante para la salud y supervivencia de la corporación.
- Las decisiones críticas dependen de la información estratégica apropiada de una empresa.
- Ejemplos de información estratégica:



- La información estratégica no pretende ayudar en la producción de facturas, hacer envíos, etc. es más importante para la salud y supervivencia de la corporación.
- Las decisiones críticas dependen de la información estratégica apropiada de una empresa.
- Ejemplos de información estratégica:
 - Características de clientes que han comprado x producto en los últimos 10 años.
 - ¿Han cambiado los gustos de los compradores del producto x? (por ejemplo después de casarse)
 - ¿Qué descuentos ofrecer para incrementar significativamente las ventas?



... INFORMACIÓN ESTRATÉGICA

- Características deseadas de información estratégica.
 - **Integrada.**



- Características deseadas de información estratégica.
 - **Integrada.** Debe tener una vista completa de la empresa.
 - **Íntegra.**



- Características deseadas de información estratégica.

- **Integrada.** Debe tener una vista completa de la empresa.
- **Íntegra.** La información debe ser precisa y conforme a las reglas del negocio.
- **Accesible.**



- Características deseadas de información estratégica.

- **Integrada.** Debe tener una vista completa de la empresa.
- **Íntegra.** La información debe ser precisa y conforme a las reglas del negocio.
- **Accesible.** Fácilmente accesible con trayectorias intuitivas de acceso y listas para analizar.
- **Creíble.**



- Características deseadas de información estratégica.

- **Integrada.** Debe tener una vista completa de la empresa.
- **Íntegra.** La información debe ser precisa y conforme a las reglas del negocio.
- **Accesible.** Fácilmente accesible con trayectorias intuitivas de acceso y listas para analizar.
- **Creíble.** Cada factor del negocio debe tener uno y sólo un valor.
- **A tiempo.**



- Características deseadas de información estratégica.



- **Integrada.** Debe tener una vista completa de la empresa.
- **Íntegra.** La información debe ser precisa y conforme a las reglas del negocio.
- **Accesible.** Fácilmente accesible con trayectorias intuitivas de acceso y listas para analizar.
- **Creíble.** Cada factor del negocio debe tener uno y sólo un valor.
- **A tiempo.** La información debe estar disponible en el tiempo estipulado.

¿SON APROPIADOS LOS OLTP PARA TALES TAREAS?

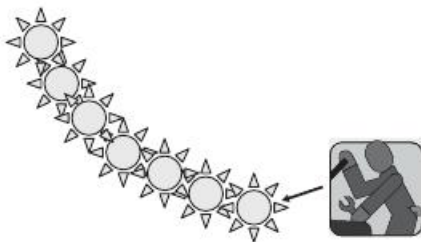
Los sistemas operacionales son sistemas para procesamiento de transacciones en línea usados para ayudar en los procesos diarios de las organizaciones/negocios.



Get the data in

Making the wheels of business turn

- ◆ Take an order
- ◆ Process a claim
- ◆ Make a shipment
- ◆ Generate an invoice
- ◆ Receive cash
- ◆ Reserve an airline seat



EJEMPLO: COMPAÑÍA DE VENTAS

Una compañía de venta de electrónicos tiene una base de datos como siguiente:



customer

<u>cust_ID</u>	name	address	age	income	credit_info	category	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	\$78000	1	3	...
...

item

<u>item_ID</u>	name	brand	category	type	price	place_made	supplier	cost
I3	hi-res-TV	Toshiba	high resolution	TV	\$988.00	Japan	NikoX	\$600.00
I8	Laptop	Dell	laptop	computer	\$1369.00	USA	Dell	\$983.00
...

employee

<u>empl_ID</u>	name	category	group	salary	commission
E55	Jones, Jane	home entertainment	manager	\$118,000	2%
...

branch

<u>branch_ID</u>	name	address
B1	City Square	396 Michigan Ave., Chicago, IL
...

purchases

<u>trans_ID</u>	cust_ID	empl_ID	date	time	method_paid	amount
T100	C1	E55	03/21/2005	15:45	Visa	\$1357.00
...

items_sold

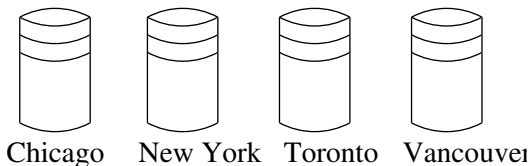
<u>trans_ID</u>	<u>item_ID</u>	qty
T100	I3	1
T100	I8	2
...

works_at

<u>empl_ID</u>	<u>branch_ID</u>
E55	B1

... EJEMPLO: COMPAÑÍA DE VENTAS

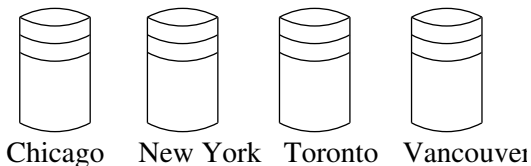
- Compañía con sucursales en todo el mundo y cada una con su propia fuente de datos.



- El presidente desea hacer un análisis de las ventas de la compañía, en el último semestre por tipo de artículo y por sucursal.???

... EJEMPLO: COMPAÑÍA DE VENTAS

- Compañía con sucursales en todo el mundo y cada una con su propia fuente de datos.



- El presidente desea hacer un análisis de las ventas de la compañía, en el último semestre por tipo de artículo y por sucursal.???
- Difícil por la dispersión de los datos en distintas bases de datos y en diferentes lugares.
- ¿Problemas de usar estos datos para el análisis del negocio?

... EJEMPLO COMPAÑÍA DE VENTAS

Problemas con los datos:



... EJEMPLO COMPAÑÍA DE VENTAS

Problemas con los datos:



- El mismo dato se encuentra en diferentes sistemas.
 - Ejemplo: los datos de los clientes en diferentes tiendas y departamentos.
 - El mismo concepto definido en forma diferente
- Fuentes heterogéneas
 - BDR, OLTP,
 - Hojas de cálculo, ...
- Los datos son adecuados para los sistemas operacionales.
 - Contabilidad, ventas, etc.
 - No soportan análisis de las funciones del negocio. (distintas vistas)
- La calidad de los datos es mala.
 - Los datos pueden ser imprecisos, estar perdidos, etc.
- Los datos son “volátiles”
 - Los datos pueden ser borrados (6meses)
 - Los datos pueden cambiar con el tiempo – no hay información histórica.

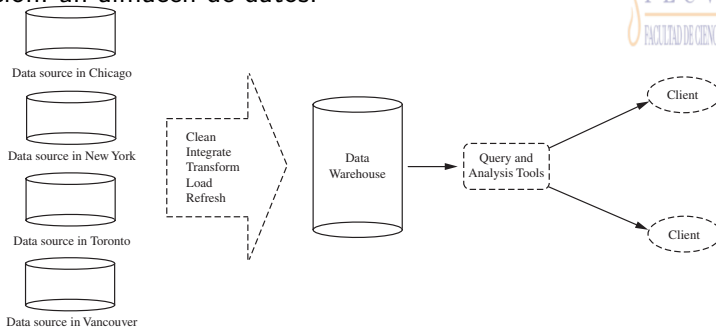
... EJEMPLO: COMPAÑÍA DE VENTAS

- Solución:



... EJEMPLO: COMPAÑÍA DE VENTAS

- Solución: un almacén de datos.



- Un DWH es un repositorio de datos recolectados de varias fuentes, almacenados bajo un esquema unificado.
- Son sistemas, principalmente para la extracción de información a partir de datos históricos.
- El análisis de datos se conoce como OLAP (*Online analytical processing*)

tes datos
facilita a los


-
- Diagrama de arquitectura de un Data Warehouse:
- Fuentes de Datos:** Representadas por iconos de discos duros y documentos.
 - Proceso ETL:** Representado por un icono de flecha verde con un símbolo de ETL.
 - Almacén de Datos (Data Warehouse):** Representado por un icono de cilindro de base de datos.
 - Herramientas de Consulta y Análisis:** Representadas por un icono de gráficos y documentos.
 - Usuarios:** Representados por un icono de una persona.
- El flujo de datos es el siguiente:
- ```

 Fuentes de Datos → ETL → Almacén de Datos ↔ Herramientas de Consulta y Análisis ↔ Usuarios

```



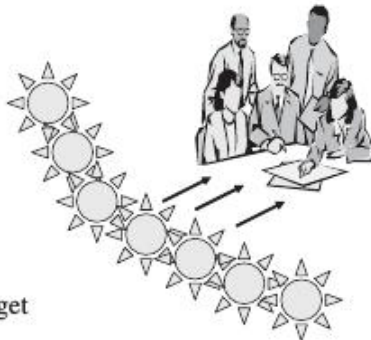
# SISTEMAS PARA APOYO A TOMA DE DECISIONES

Los almacenes de datos son sistemas especialmente diseñados y  construidos para la toma de decisiones, se usan para observar cómo trabaja el negocio y luego tomar decisiones estratégicas que lo lleven a mejorar.

## *Get the information out*

### *Watching the wheels of business turn*

- ◆ Show me the top-selling products
- ◆ Show me the problem regions
- ◆ Tell me why (drill down)
- ◆ Let me see other data (drill across)
- ◆ Show the highest margins
- ◆ Alert me when a district sells below target



# INFORMACIÓN ESTRATÉGICA DE UN DWH

## Información operacional



Sistemas operacionales  
(Procesos básicos del negocio)



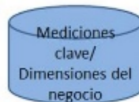
Extraer, limpiar y agregar



Transformación de la  
información



## Información estratégica



Data warehouse

# INFORMACIÓN ESTRATÉGICA DE UN DWH

## Información operacional



Sistemas operacionales  
(Procesos básicos del negocio)



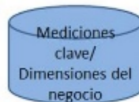
Extraer, limpiar y agregar



Transformación de la  
información



## Información estratégica



Data warehouse

Un dwh es un concepto sencillo: Toma todos los datos que hay en una organización, los limpia, transforma y luego proporciona información estratégica útil. :)

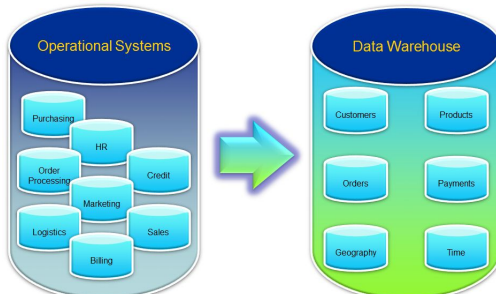
# DEFINICIÓN DE ALMACÉN DE DATOS DWH

- “Un almacén de datos es una colección de datos orientados a un tema, integrados, históricos y no volátiles para apoyar el proceso de toma de decisiones de los ejecutivos” - W. H. Inmon.



# DEFINICIÓN DE ALMACÉN DE DATOS DWH

- “Un almacén de datos es una colección de datos orientados a un tema, integrados, históricos y no volátiles para apoyar el proceso de toma de decisiones de los ejecutivos” - W. H. Inmon.
- Orientada a un tema/proceso, en lugar de procesamiento de transacciones de una organización.
  - Está enfocado a responder eficientemente consultas estratégicas:
    - Actividades/temas de interés: compras, ventas, alquileres, ...
    - Contexto de análisis: clientes, vendedores, productos, etc...

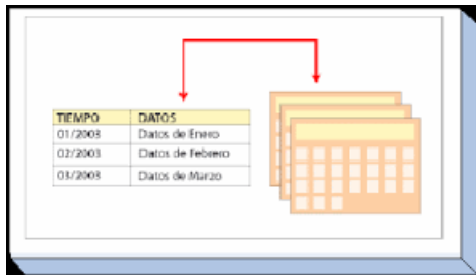






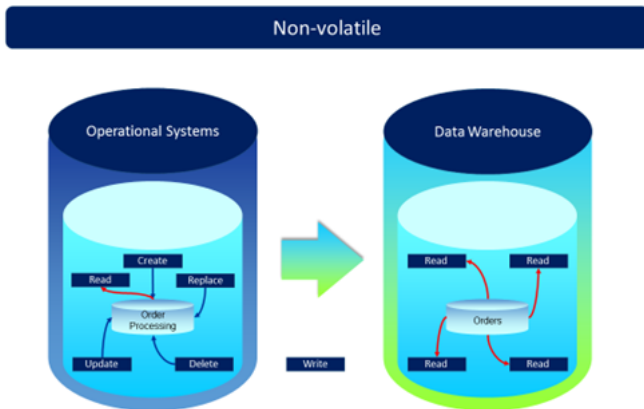
## ... DEFINICIÓN DE ALMACÉN DE DATOS DWH

- Históricos. Los datos se almacenan para proporcionar información desde una perspectiva histórica. Cada elemento clave contiene explícita o implícitamente un elemento de tiempo.



## ... DEFINICIÓN DE ALMACÉN DE DATOS DWH

- No volátil. Los datos nuevos se agregan (refreshing) al almacén pero no reemplazan nada, por lo tanto se va conservando la historia. No hay supresión ni actualización.



- La razón de ser de un DWH es responder preguntas acerca de una organización: desempeño de varias operaciones, tendencias de la negociación, qué puede hacerse para mejorar la organización.
- Proporcionar una vista integrada y total de la organización.
- Presentar la información de la organización de manera consistente.
- Facilitar el acceso a datos actuales e históricos de la organización para la toma de decisiones estratégicas.
  - El análisis de datos se conoce como OLAP (Online analytical processing)
  - Utilizan un modelo multidimensional (cubos, hipercubos, etc.).
- Permitir realizar operaciones de apoyo a decisiones sin obstruir a los sistemas operacionales.
- Ser adaptable y resistente a los cambios.
- Ser un bastión seguro que protege la información.



- Beneficios esperables:
  - Acceso interactivo e inmediato a información estratégica de un área de negocios.
  - Toma de decisiones basadas en datos objetivos.
- Beneficios tangibles:
  - Aumento de ganancias.
  - Reducción de gastos.

## Ejemplo:

- Reducir pérdidas debido a la mejora en la detección de fraudes.
  - Incrementar la fidelidad de clientes mediante campañas de marketing dirigidas a ciertos sectores.
  - Reducir costos de inventarios mediante mejoras en previsión de demanda de productos.
- Los beneficios aumentan:
  - Cuanto más importantes son las decisiones.
  - Cuanto más crítico es el factor tiempo.

Un proyecto de DWH se considera exitoso si:

- Integra información heterogénea.
- Hace visible y manejable la información útil.
- Incluye datos de calidad validada.
- Ofrece acceso directo a usuarios.
- El sistema se populariza.



Un proyecto de DWH se considera exitoso si:

- Integra información heterogénea.
- Hace visible y manejable la información útil.
- Incluye datos de calidad validada.
- Ofrece acceso directo a usuarios.
- El sistema se populariza.

Esto tiene como consecuencia:

- Incremento en la cantidad y complejidad de consultas y reportes solicitados al DWH, por los usuarios.
- Incremento del número de usuarios activos.
- Decremento notorio del tiempo requerido para obtener información estratégica.



Un proyecto de DWH se considera exitoso si:

- Integra información heterogénea.
- Hace visible y manejable la información útil.
- Incluye datos de calidad validada.
- Ofrece acceso directo a usuarios.
- El sistema se populariza.



Esto tiene como consecuencia:

- Incremento en la cantidad y complejidad de consultas y reportes solicitados al DWH, por los usuarios.
- Incremento del número de usuarios activos.
- Decremento notorio del tiempo requerido para obtener información estratégica.

**El DWH cumple los objetivos y produce los resultados deseados.**



Se debe evitar:

- Establecer expectativas demasiado altas.
- Cargar el DWH con todo lo disponible.
- Diseñar el DWH igual que un sistema de producción.
- Ignorar fuentes de datos externas.
- Ignorar que los sistemas evolucionan.



# BASES DE DATOS VS ALMACENES DE DATOS



# BASES DE DATOS VS ALMACENES DE DATOS

## Comparación de usuarios:

- OLTP: Profesional de TI.
- OLAP: Analista de información.



## Comparación de objetivos:

- Operaciones diarias.
- Apoyo a la toma de decisiones.

## Comparación de entornos de procesamiento

- Procesamiento de transacciones.
  - Datos primarios de transacciones.
  - Operaciones diarias y decisiones a corto plazo.
- Procesamiento de consultas.
  - Datos secundarios transformados.
  - Decisiones a mediano y largo plazo.

# ... BASES DE DATOS VS ALMACENES DE DATOS

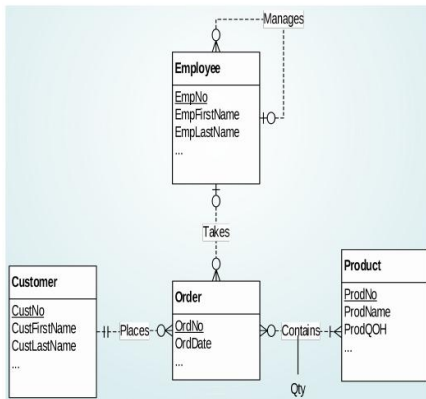
Comparación de datos:



| <b>Característica</b>  | <b>BD Operacional</b>      | <b>Data Warehouse</b>                            |
|------------------------|----------------------------|--------------------------------------------------|
| Actualidad             | Actual                     | Histórico                                        |
| Nivel de detalle       | Individual                 | Individual y resumido                            |
| Orientación            | Proceso                    | Tema                                             |
| Registros por consulta | Decenas                    | Millones                                         |
| Nivel de normalización | Principalmente normalizado | Normalización relajada                           |
| Nivel de actualización | Alta volatilidad           | No volátil<br>Principalmente refrescado          |
| Modelo de datos        | Relacional                 | Relacional (estrella)<br>multidimensional (cubo) |

# ... BASES DE DATOS VS ALMACENES DE DATOS

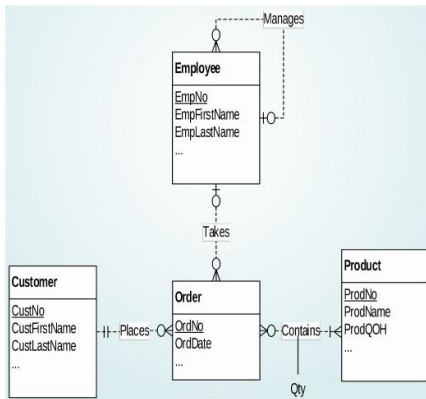
## Comparación de esquemas: Operational database



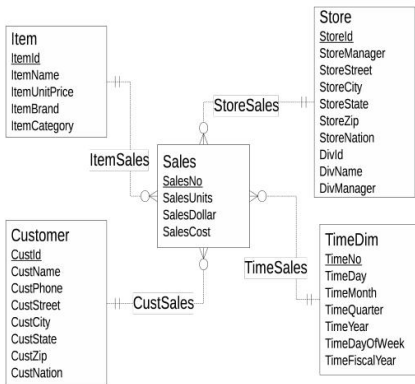
# ... BASES DE DATOS VS ALMACENES DE DATOS

Comparación de esquemas:

## Operational database



## Data warehouse



# ¿PORQUÉ SEPARARLO?



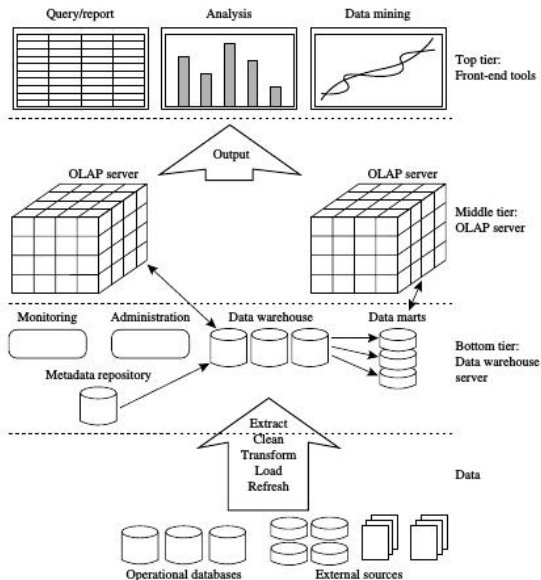
# ¿PORQUÉ SEPARARLO?

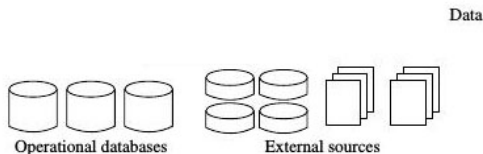
- La BD está diseñada y afinada para ciertas tareas como búsqueda de un registro particular.
- Las consultas del DWH son complejas involucran el cálculo de grandes grupos de datos a niveles de resúmenes y pueden requerir el uso de una organización de datos especial, acceso y consultas.
- Las BD soportan procesamiento concurrente de múltiples transacciones. Por lo que se requieren mecanismos para asegurar la consistencia y robustez de las transacciones.
- Una consulta a un OLAP necesita sólo leer los datos para aplicar operaciones de agrupación y agregación. Por lo tanto no requieren los mecanismos antes mencionados.





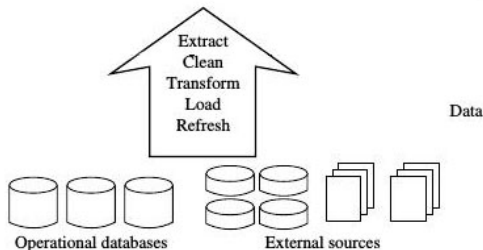
# ARQUITECTURA DE UN DWH



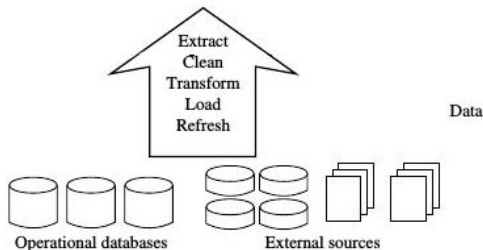


Se pueden agrupar en 4 categorías:

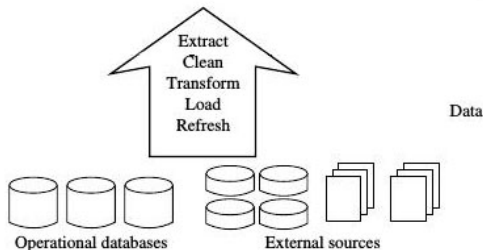
- Datos de producción.
- Datos internos.
- Datos archivados.
- Datos externos.



- La capa de preparación de datos se encarga del proceso que extrae, integra y limpia los datos de las fuentes y alimenta a la capa del DWH.

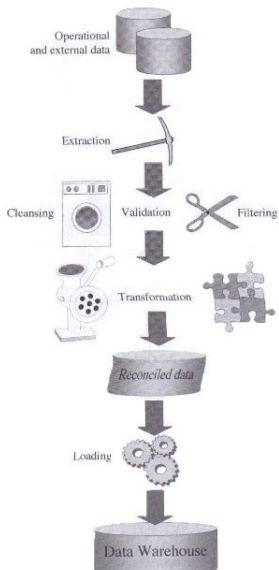


- La capa de preparación de datos se encarga del proceso que extrae, integra y limpia los datos de las fuentes y alimenta a la capa del DWH.
- El proceso ETL, también se conoce como **reconciliación de datos**.



- La capa de preparación de datos se encarga del proceso que extrae, integra y limpia los datos de las fuentes y alimenta a la capa del DWH.
- El proceso ETL, también se conoce como **reconciliación de datos**.
- Tiene 4 etapas: extracción, limpieza, transformación y carga.
- Limpieza = rectificar valores
- Transformación = formatos de datos.

# ... PREPARACIÓN DE DATOS Y ETL



Extracción. Esta etapa se trata con numerosas fuentes de datos.

Hay dos tipos de extracciones posibles:



- Estática. Utilizada al poblar el almacén por primera vez.
- Incremental. Utilizada para actualizar regularmente el almacén. Basada en la bitácora de la BD, si se utilizan marcas de tiempo (timestamp) asociadas con los datos operacionales, éstas se utilizan para la extracción.  
¿Donde colocar los datos extraídos para su preparación?

Es la fase crucial en la creación de un DWH, pues de ella depende la calidad de los datos.

Errores o inconsistencias que hacen que los datos se consideren “sucios”.

- Datos duplicados.
- Valores inconsistentes que son lógicamente asociados.
- Datos perdidos.
- Uso inesperado de campos.
- Valores imposibles o erróneos.
- Valores inconsistentes para una entidad debido a diferentes prácticas.
- Valores inconsistentes para una entidad debido a errores de dedo.

Objetivo de la limpieza:





Es la fase crucial en la creación de un DWH, pues de ella depende la calidad de los datos.

Errores o inconsistencias que hacen que los datos se consideren “sucios”.

- Datos duplicados.
- Valores inconsistentes que son lógicamente asociados.
- Datos perdidos.
- Uso inesperado de campos.
- Valores imposibles o erróneos.
- Valores inconsistentes para una entidad debido a diferentes prácticas.
- Valores inconsistentes para una entidad debido a errores de dedo.

Objetivo de la limpieza: **rectificación y homogeneización.**



- Es el corazón de la reconciliación. Convierte los datos de su formato operacional en un formato específico para el DWH.
- Difícil en presencia de múltiples fuentes heterogéneas.
- Ejemplos de principales transformaciones:
  - Conversión y normalización, que opera tanto sobre formatos de almacenamiento como sobre unidades de medida para uniformar los datos.
  - Matching. que asocia campos equivalentes en fuentes diferentes.
  - Selección, que reduce la cantidad de campos y registros fuentes.



# EJEMPLO DE DATOS SUCIOS

## FORMULARIO DE MATRICULA

No. Estudiante:

1200

1- Pueden existir registros de base de datos múltiples para un solo estudiante, debido un cambio de nombre o un movimiento de grupo.

Apellido:

Fernández

Nombre:

Sabina

2- Diferentes lugares de la Universidad registran la misma persona (en la facultad, en la biblioteca, etc.), por lo que el DW cuenta con el mismo dato múltiples veces.

Edad:

20

Sexo:

Femenino

3- Diferentes registros pueden proporcionar la misma información en el mismo campo, pero en formatos diferentes (por ejemplo, 'Femenino' y 'Masculino' versus 'F' y 'M')

Dirección:

Av. Central, # 23, Distrito Oeste

4- No hay cuidado a la hora de la entrada de datos y se introducen con errores o incompletos.

# EJEMPLO DE LIMPIEZA Y TRANSFORMACIÓN

Juan Pérez Cruz

Tlalcoligia 98 Edif. A-201

Tlallpan 41430, México D.F.



# EJEMPLO DE LIMPIEZA Y TRANSFORMACIÓN



Juan Pérez Cruz  
Tlalcoligia 98 Edif. A-201  
Tlallpan 41430, México D.F.

## Normalización:

|                |                |
|----------------|----------------|
| Nombre:        | Juan           |
| Apellido P:    | Pérez          |
| Apellido M:    | Cruz           |
| Calle Num-ext: | Tlalcoligia 98 |
| Interior:      | Edif. A-201    |
| Delegación:    | Tlallpan       |
| CP:            | 41430          |
| Ciudad:        | México D.F.    |

# ... EJEMPLO DE LIMPIEZA Y TRANSFORMACIÓN

## Estandarización

Nombre: Juan  
Apellido P: Pérez  
Apellido M: Cruz  
Calle Num-ext: Tlalcoligia 98  
Interior: A-201 <--  
Delegación: Tlallpan  
CP: 41430  
Ciudad: Ciudad de México <---



# ... EJEMPLO DE LIMPIEZA Y TRANSFORMACIÓN

## Estandarización

Nombre: Juan  
Apellido P: Pérez  
Apellido M: Cruz  
Calle Num-ext: Tlalcoligia 98  
Interior: A-201 <--  
Delegación: Tlalpan  
CP: 41430  
Ciudad: Ciudad de México <---

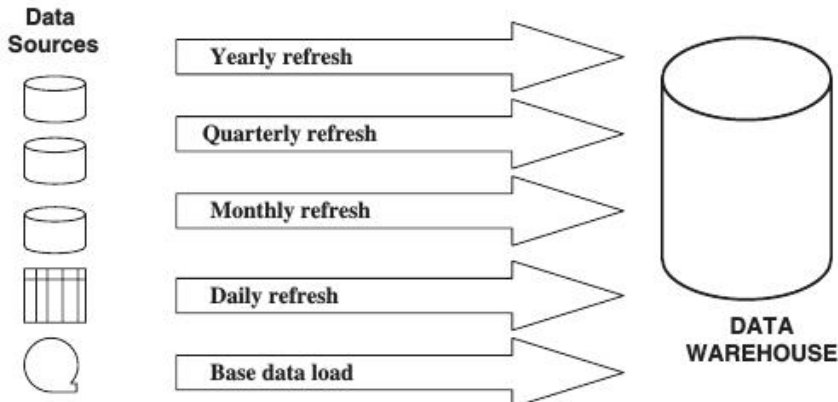
## Corrección:

Nombre: Juan  
Apellido P: Pérez  
Apellido M: Cruz  
Calle Num-ext: Tlalcoligia 98  
Interior: A-201  
Delegación: Tlalpan <--  
CP: 14430 <---  
Ciudad: Ciudad de México

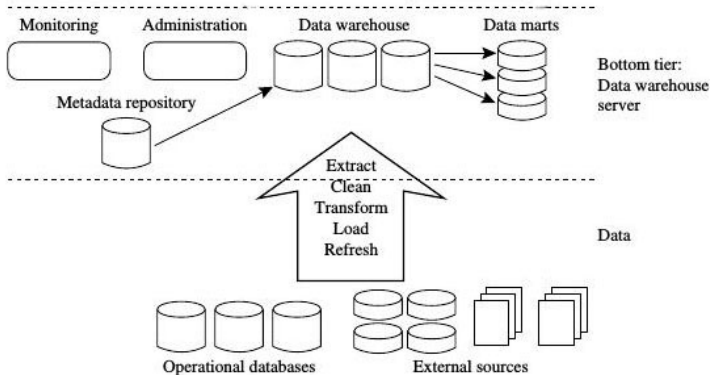


## ... ETL (CARGA)

- Esta función consume mucho tiempo, especialmente la carga inicial.
- La carga inicial transfiere grandes volúmenes de datos.
- Las condiciones del negocio determinan los ciclos de refresco.







- El **DWH** **primario, principal o corporativo** actúa como un sistema de almacenamiento centralizado para que los datos puedan ser resumidos.
- Un **datamart** es un DWH que contiene la información necesaria para un dominio de aplicación (área) específico.
- Aunque no son estrictamente necesarios, son útiles para DWH de tamaño medio a grande porque:
  - Se usan como bloques de construcción en tanto el DWH se desarrolla incrementalmente.
  - Delimitan la información requerida por un grupo específico de usuarios para resolver consultas,
  - Puede conseguirse mejor rendimiento en ellos debido a que su tamaño es menor que el DWH primario.



- Metadatos =



- Metadatos = datos que definen otros datos.



- Metadatos = datos que definen otros datos.
- En DWH especifican la fuente, valores, uso y características de los datos en él y definen cómo pueden cambiar y procesarse los datos en cada capa de la arquitectura.
- Los metadatos en un DWH caen en tres categorías principales:
  - Metadatos operacionales.
  - Metadatos de extracción y transformación.
  - Metadatos de/para el usuario final.

- Metadatos = datos que definen otros datos.
- En DWH especifican la fuente, valores, uso y características de los datos en él y definen cómo pueden cambiar y procesarse los datos en cada capa de la arquitectura.
- Los metadatos en un DWH caen en tres categorías principales:
  - Metadatos operacionales.
  - Metadatos de extracción y transformación.
  - Metadatos de/para el usuario final.
- El repositorio de metadatos es consultado por todos los componentes de la arquitectura del DWH.

- Metadatos = datos que definen otros datos.
- En DWH especifican la fuente, valores, uso y características de los datos en él y definen cómo pueden cambiar y procesarse los datos en cada capa de la arquitectura.
- Los metadatos en un DWH caen en tres categorías principales:
  - Metadatos operacionales.
  - Metadatos de extracción y transformación.
  - Metadatos de/para el usuario final.
- El repositorio de metadatos es consultado por todos los componentes de la arquitectura del DWH.
- Importancia:

- Metadatos = datos que definen otros datos.
- En DWH especifican la fuente, valores, uso y características de los datos en él y definen cómo pueden cambiar y procesarse los datos en cada capa de la arquitectura.
- Los metadatos en un DWH caen en tres categorías principales:
  - Metadatos operacionales.
  - Metadatos de extracción y transformación.
  - Metadatos de/para el usuario final.
- El repositorio de metadatos es consultado por todos los componentes de la arquitectura del DWH.
- Importancia:
  - Actúan como pegamento que conecta todas las partes del DWH.
  - Proporcionan a los desarrolladores información acerca del contenido y estructura.
  - Abren la puerta para que los usuarios finales reconozcan el contenido del DWH en sus propios términos.



## ... METADATOS

Al hacer una consulta al almacén de datos, ¿qué metadatos pueden necesitarse?



Al hacer una consulta al almacén de datos, ¿qué metadatos pueden necesitarse?



- ¿Hay consultas predefinidas que puedan usarse?
- ¿Cuales son los diferentes tipos de datos en el dwh?
- ¿Cómo y dónde buscar para saber lo que hay disponible?
- ¿De qué sistemas fuente se obtuvieron los datos para el dwh?
- ¿Cómo se integraron los datos de los sistemas transaccionales?
- ¿Qué grado de detalle puedo tener?
- ¿Qué jerarquías hay en las dimensiones?
- etcétera.

## ... METADATOS (EJEMPLO)

|                           |                       |
|---------------------------|-----------------------|
| <b>Nombre de entidad:</b> | Cliente               |
| <b>Alias:</b>             | Consumidor, comprador |



**DEFINICIÓN:** Una persona u organización que compra bienes o servicios de la compañía.

**OBSERVACIONES:** La entidad cliente incluye clientes regulares, actuales y pasados.

**SISTEMAS FUENTE:** Ordenes de compra terminadas, mantenimiento de contactos, ventas en-línea.

**FECHA DE CREACIÓN:** 15 de Enero de 2006

**FECHA DE ÚLTIMA ACT.:** 28 de Febrero de 2017

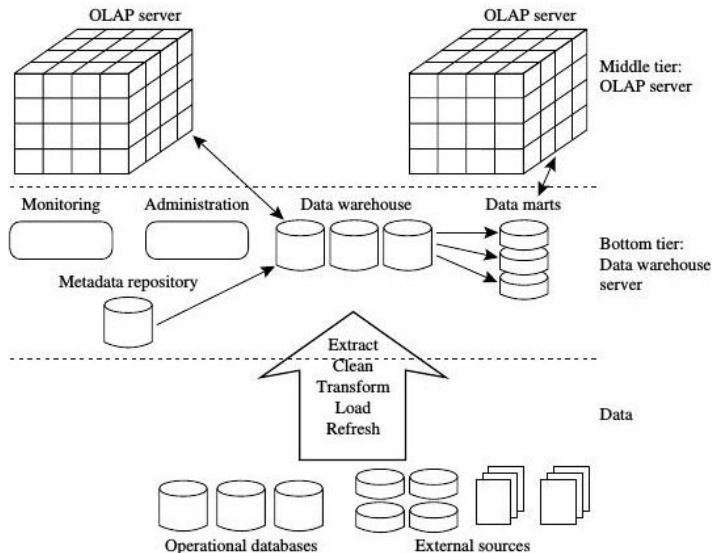
**CICLO DE ACTUALIZACIÓN:** semanal

**ÚLTIMO REFRESCO COMPLETO:** 29 de Diciembre de 2016

**CALIDAD DE DATOS REVISADA:** 27 de Febrero de 2017.

**PROYECTOS DE ARCHIVO:** Cada seis meses

**USUARIO RESPONSABLE:** Juan Pérez



- Es la forma de modelar los datos en un DWH.
- Es mejor para el análisis de datos que el modelo E/R:
  - Es menos flexible.
  - No es adecuado para sistemas OLTP
- Está orientado hacia:
  - ¿Qué es importante?
  - ¿Qué describe lo importante?
  - ¿Qué se desea optimizar?
- Los datos se dividen en hechos (con medidas) y dimensiones.



### Ejemplo:



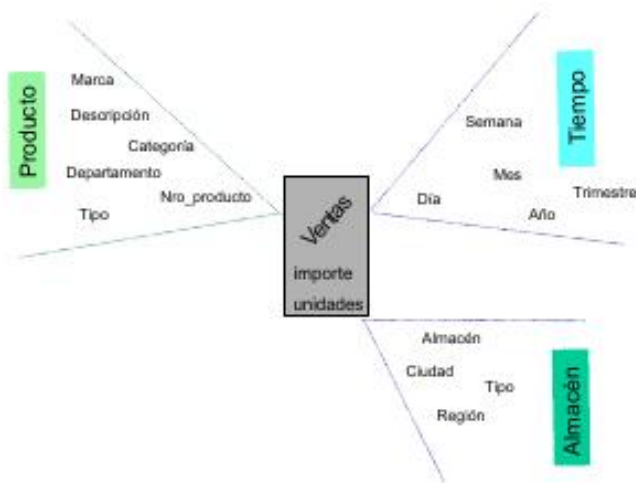
- Organización: cadena de tiendas departamentales.
- Actividad que se va a analizar: ventas de productos.
- Ejemplo de información registrada:
  - 5 refrigeradores LG vendidos en Liverpool de Perisur el día 24 de diciembre de 2016, por un importe de \$45,000.00
- Para hacer el análisis no se toma en cuenta cada venta individual, lo importante son las ventas diarias de productos en las distintas tiendas.

## ... MODELO MULTIDIMENSIONAL

- Los factores que afectan la toma de decisiones son **hechos** específicos de la organización.
  - Ejemplo: ventas.
- Las instancias de los hechos corresponden a **eventos** que ocurren.
  - Ejemplos: una venta, un envío.
- Cada hecho se describe por los valores de un conjunto de **medidas** relevantes que proporciona una descripción cuantitativa de eventos.
  - Ejemplos: Cantidades vendidas, importe de ventas.
- Las **dimensiones** describen los hechos:
  - Ejemplo: ventas con respecto a las dimensiones *producto*, *fecha* y *sucursal*.
- Objetivo del modelado dimensional:
  - Rodear los hechos del mayor contexto (dimensiones) posible.

Los hechos (datos) “viven” en un **cubo** multidimensional

# ... MODELO MULTIDIMENSIONAL





- Las dimensiones son el corazón del modelo multidimensional.
- Las dimensiones se usan para:
  - Seleccionar datos.
  - Agrupar datos en diferentes niveles de detalle.
- Las dimensiones tienen atributos con valores:
  - La dimensión producto tiene “tipo” con valores: “refrigerador”, “lavadora”, ...
  - La dimensión Fecha tiene valores: “1/1/2017”, “29/2/2016”, ...
- Los atributos de las dimensiones pueden tener un orden:
  - Usado para comparar datos del cubo a lo largo de ciertos valores.
  - Especialmente usados para la dimensión Fecha:
    - Porcentaje de ventas comparado con el último mes, para ver si hubo un incremento.

- Las dimensiones se organizan sus atributos en jerarquías con niveles de detalle.



- Producto: Producto -- > tipo -- > categoría
- Tienda: Tienda -- > área -- > ciudad -- > país  
Tienda -- > tipo -- > departamento
- Fecha:

- Las dimensiones se organizan sus atributos en jerarquías con niveles de detalle.
  - Producto: Producto -- > tipo -- > categoría
  - Tienda: Tienda -- > área -- > ciudad -- > país  
Tienda -- > tipo -- > departamento
  - Fecha: Día -- > mes -- > trimestre -- > año
- Las dimensiones tiene un nivel superior denominado ALL.
- Los elementos en un nivel de la jerarquía no están en ningún orden.
- Una dimensión puede tener más de una jerarquía.



- Las dimensiones se organizan sus atributos en jerarquías con niveles de detalle.
  - Producto: Producto -- > tipo -- > categoría
  - Tienda: Tienda -- > área -- > ciudad -- > país  
Tienda -- > tipo -- > departamento
  - Fecha: Día -- > mes -- > trimestre -- > año
- Las dimensiones tiene un nivel superior denominado ALL.
- Los elementos en un nivel de la jerarquía no están en ningún orden.
- Una dimensión puede tener más de una jerarquía.
  - Fecha: día -- > mes -- > semestre -- > año
  - Fecha: día -- > semana -- > año
  - 
  - Tienda: Tienda -- > área -- > ciudad -- > país  
Tienda -- > tipo -- > departamento



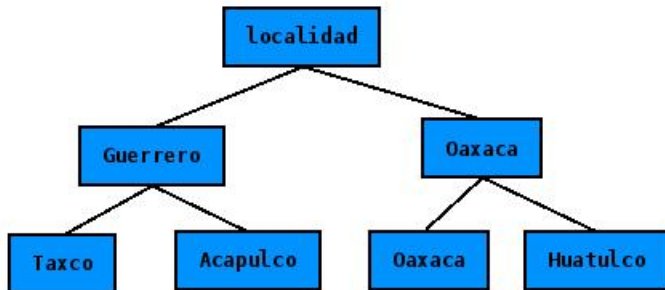
- Las dimensiones se organizan sus atributos en jerarquías con niveles de detalle.
  - Producto: Producto -- > tipo -- > categoría
  - Tienda: Tienda -- > área -- > ciudad -- > país  
Tienda -- > tipo -- > departamento
  - Fecha: Día -- > mes -- > trimestre -- > año
- Las dimensiones tiene un nivel superior denominado ALL.
- Los elementos en un nivel de la jerarquía no están en ningún orden.
- Una dimensión puede tener más de una jerarquía.
  - Fecha: día -- > mes -- > semestre -- > año
  - Fecha: día -- > semana -- > año
  - 
  - Tienda: Tienda -- > área -- > ciudad -- > país  
Tienda -- > tipo -- > departamento
- Regla general: La dimensiones deberían contener mucha información.

# ... JERARQUÍA DE DIMENSIONES

Esquema

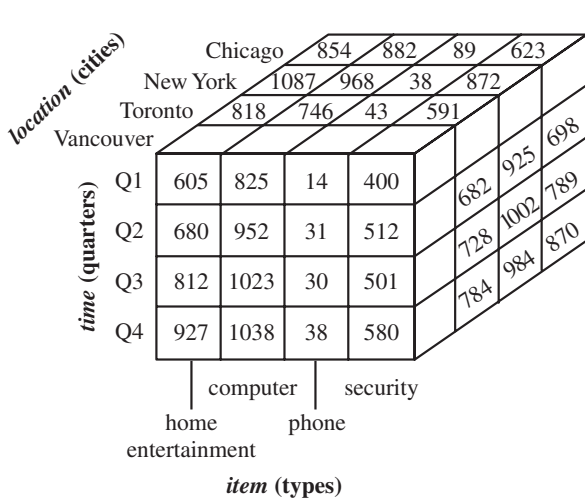


Instancias



- Un cubo puede tener muchas dimensiones:
  - Si tiene más de tres se llama hipercubo.
  - Teóricamente no hay límite para la cantidad de dimensiones.
  - Típicamente tienen entre 4 y 12 dimensiones.
  - Pero, sólo se pueden ver 2 o 3 a la vez.
- Un cubo consta de celdas:
  - Una combinación de valores de las dimensiones.
  - Una celda puede estar vacía.
  - Un cubo poco denso tiene muchas celdas vacías.
  - Un cubo denso tiene pocas celdas vacías.
  - Los cubos son poco densos cuando se tiene gran cantidad de dimensiones.

# ... EL MODELO MULTIDIMENSIONAL: CUBO





# IMPLEMENTACIÓN DEL MODELO DIMENSIONAL

Se tienen tres enfoques para implementar el DWH y se relacionan con el modelo lógico usado para representar datos:



- **ROLAP = *Relational OLAP***

- Ventaja: Se implementa sobre el modelo más conocido para manejo de BD, aunque requiere de una capa entre el cliente OLAP y el servidor relacional.
- Desventaja: Rendimiento.

- **MOLAP = *Multidimensional OLAP***

- Ventaja: Rapidez de repuesta. Representación de los datos de manera natural y “sencilla”.
- Desventaja: No hay bases de datos multidimensionales por lo tanto no hay estándares, complejo el almacenamiento.

- **HOLAP = *Hybrid OLAP***

# ESQUEMAS PARA BD MULTIDIMENSIONAL (ROLAP)

El modelo más popular es de **estrella**, el cual contiene:

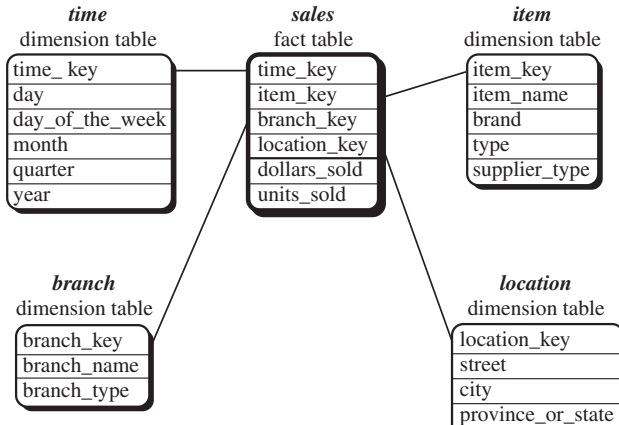
- Una tabla central (tabla de hechos) que contiene los datos **sin** redundancia.
- Un conjunto de tablas asistentes para cada dimensión.



# ESQUEMAS PARA BD MULTIDIMENSIONAL (ROLAP)

El modelo más popular es de **estrella**, el cual contiene:

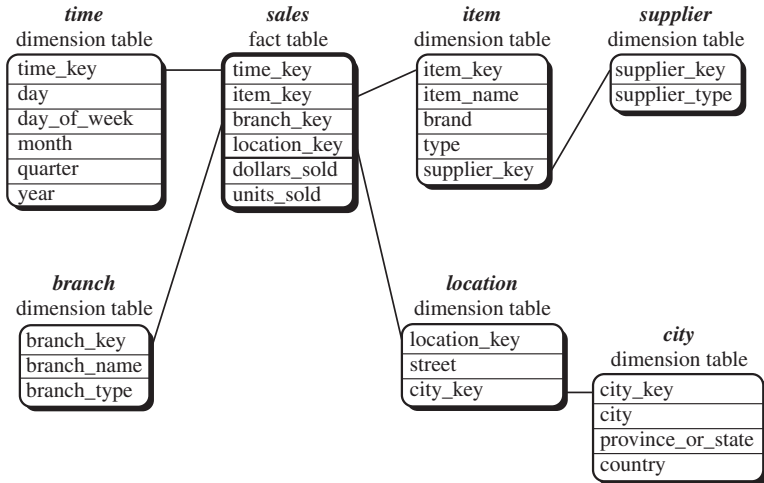
- Una tabla central (tabla de hechos) que contiene los datos **sin** redundancia.
- Un conjunto de tablas asistentes para cada dimensión.



- Ventajas:
  - Fácil de usar.
  - Relativamente flexible.
  - Tabla de hechos normalizada.
  - Las tablas de dimensiones son relativamente pequeñas.
- Desventajas:
  - Las jerarquías de atributos se “ocultan” en las columnas.
  - Las tablas de dimensiones pueden no estar normalizadas.

# MODELO COPO DE NIEVE

Es una variante del esquema de estrella donde cada tabla de dimensión es normalizada, por lo tanto se divide en varias tablas.



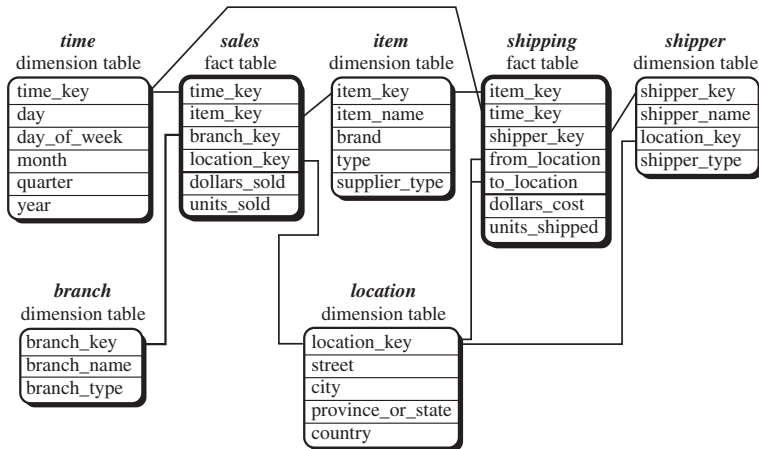
## ... MODELO COPO DE NIEVE

- Ventajas:
  - Las tablas de las dimensiones son más pequeñas.
  - Mínima redundancia.
- Desventajas:
  - Necesidad de hacer joins al momento de realizar consultas.

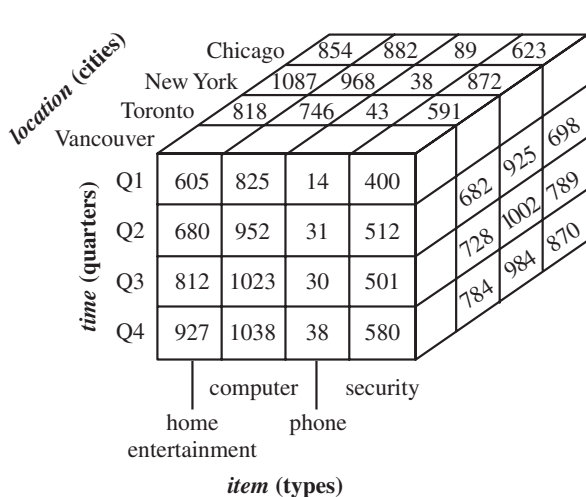


# MODELO CONSTELACIONES DE HECHOS

Cuando se requiere que múltiples tablas de hechos compartan tablas de dimensión. Este tipo de esquema puede verse como una colección de estrellas y por eso se llama galaxia o bien constelación de hechos.



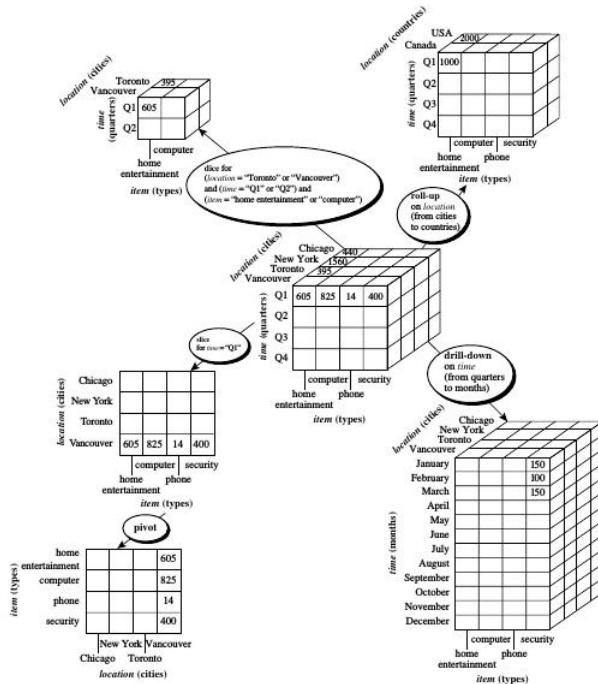
# OPERACIONES OLAP



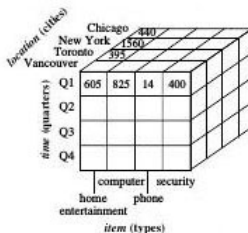


- Roll up: realiza agregación sobre un cubo, subiendo en la jerarquía de conceptos de una dimensión.
- Drill-down: Inversa de roll-up. Va de menos detalle a más detalle. De un nivel mayor a un nivel menor o datos detallados introduciendo nuevas dimensiones.
- Slice: cubo con la selección de una dimensión dada.
- Dice: cubo con la selección de más de una dimensión.
- Pivot (rotate): cambia la orientación del cubo.

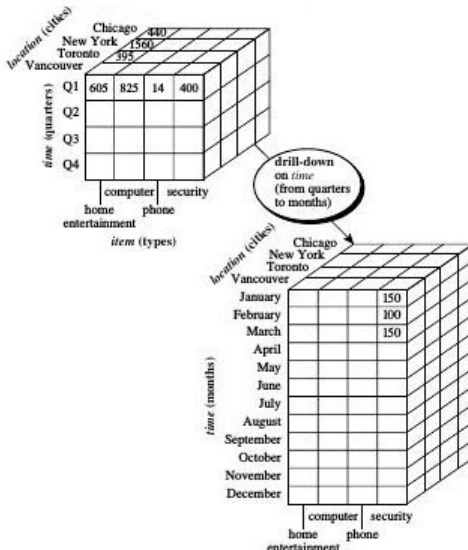




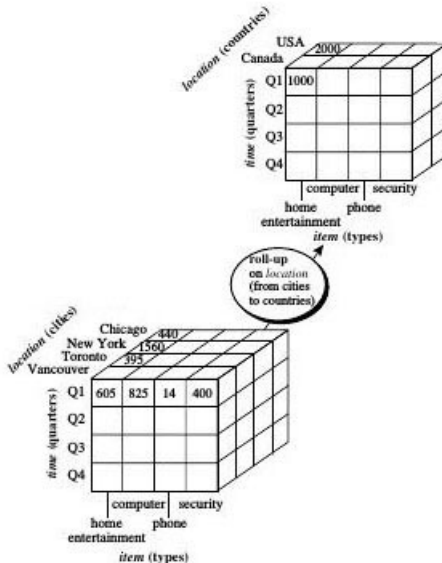
# ... OPERACIONES OLAP (DRILL-DOWN)



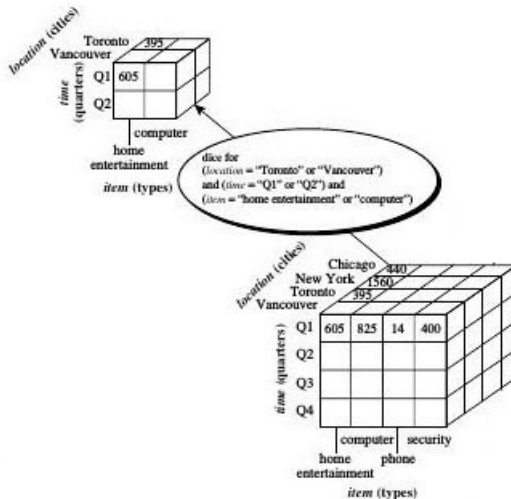
# ... OPERACIONES OLAP (DRILL-DOWN)



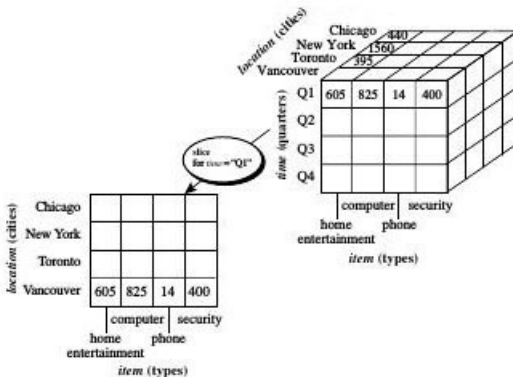
# ... OPERACIONES OLAP (ROLL-UP)



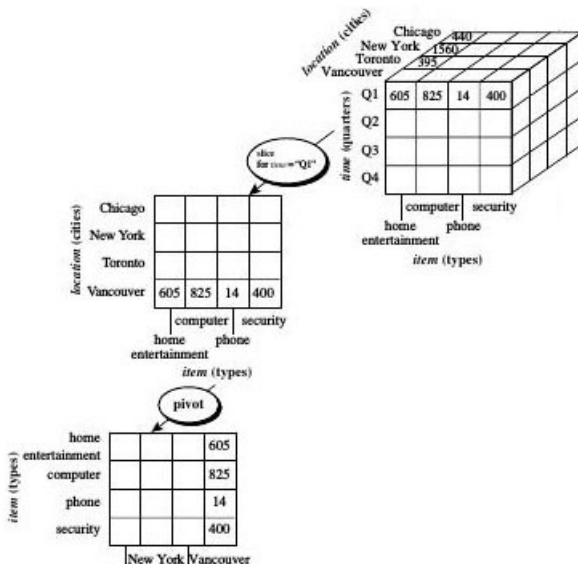
# ... OPERACIONES OLAP (DICE)



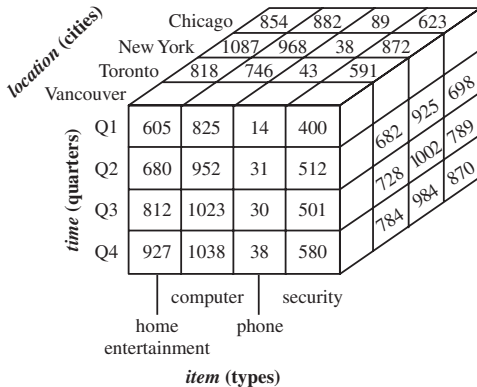
# ... OPERACIONES OLAP (SLICE)



# ... OPERACIONES OLAP (PIVOT)







The diagram illustrates a 3D data cube representing an OLAP fact table. The dimensions are:

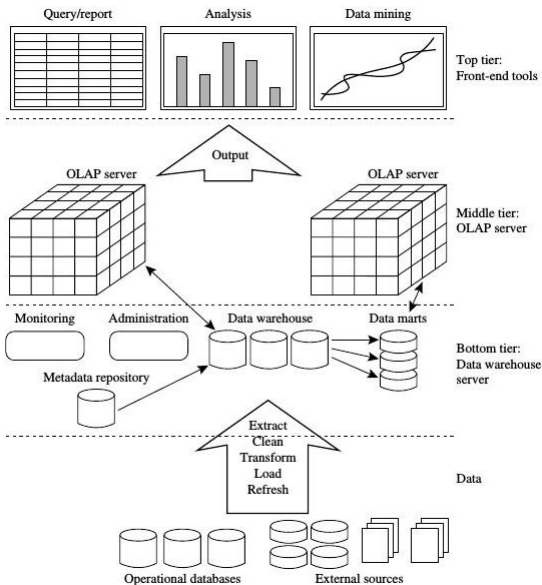
- location (cities)**: Chicago, New York, Toronto, Vancouver
- time (quarters)**: Q1, Q2, Q3, Q4
- item (types)**: computer, security, home entertainment, phone

The data values for the 'computer' and 'security' item types are shown in the table below:

| location (cities) | time (quarters) | computer | security |
|-------------------|-----------------|----------|----------|
| Chicago           | Q1              | 605      | 825      |
| Chicago           | Q2              | 680      | 952      |
| Chicago           | Q3              | 812      | 1023     |
| Chicago           | Q4              | 927      | 1038     |
| New York          | Q1              | 14       | 400      |
| New York          | Q2              | 31       | 512      |
| New York          | Q3              | 30       | 501      |
| New York          | Q4              | 38       | 580      |
| Toronto           | Q1              | 682      | 925      |
| Toronto           | Q2              | 728      | 1002     |
| Toronto           | Q3              | 784      | 984      |
| Toronto           | Q4              | 870      | 870      |
| Vancouver         | Q1              | 698      | 698      |
| Vancouver         | Q2              | 698      | 698      |
| Vancouver         | Q3              | 698      | 698      |
| Vancouver         | Q4              | 698      | 698      |

Ejemplos:

- Obtener el total de ventas en computadoras durante el segundo semestre en New York.
- Obtener el total de ventas en computadoras en diciembre en Canadá.



- El acceso a reportes está orientado y limitado a aquellos usuarios que requieren tener acceso regular a la información en una forma casi estática.
  - Ejem. Una autoridad local en salud debe enviar a sus oficinas estatales reportes mensuales con la información de los costos de admisión de pacientes.
- Un reporte es definido por una consulta y una disposición o formato.
- Una consulta generalmente implica una restricción y una agregación en datos multidimensionales.
  - Buscar los recibos del último trimestre para cada categoría de productos. La presentación puede ser como una tabla o una gráfica (diagrama, histograma, pay, etc.)
- Una buena herramienta de reportes debe permitir flexibilidad en la generación de éstos. Un reporte puede explícitamente ejercitarse por usuarios o automática y regularmente enviarlos a usuarios registrados.

# ACCESO AL DWH (OLAP)

- Principal forma de explotar la información en un DWH.
- Dirigida a los usuarios finales cuyas necesidades de análisis no son definidas de antemano y le permite analizar y explorar datos de manera interactiva en base al modelo multidimensional.
- Una sesión OLAP consta de una trayectoria de navegación que corresponde a un proceso de análisis para hechos de acuerdo a diferentes puntos de vista y a diferentes niveles de detalle.
- Cada paso de una sesión de análisis se caracteriza por un operador OLAP que convierte la última consulta en una nueva.



# ACCESO AL DWH (DASHBOARD)

Un panel de indicadores, tablero o dashboard es una GUI que muestra una cantidad limitada de datos relevantes en un formato breve y fácil de leer.



# ACCESO AL DWH (DASHBOARD)

Un panel de indicadores, tablero o dashboard es una GUI que muestra una cantidad limitada de datos relevantes en un formato breve y fácil de leer.

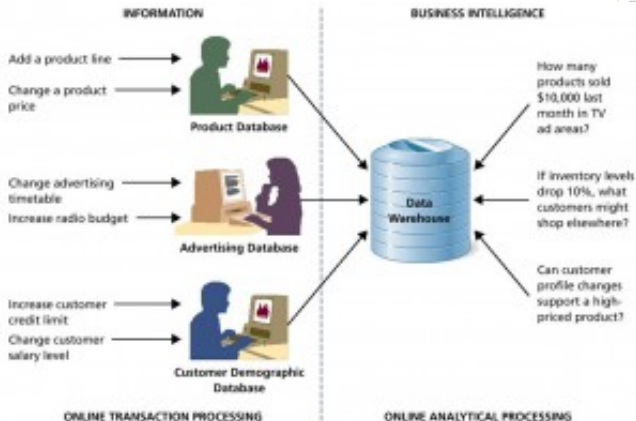


# ACCESO AL DWH (DASHBOARD)

Un panel de indicadores, tablero o dashboard es una GUI que muestra una cantidad limitada de datos relevantes en un formato breve y fácil de leer.



Pueden proporcionar un panorama en tiempo real de las tendencias para un fenómeno específico o para varios si es que están interrelacionados.





- 1 P. Ponniah, Data Warehousing Fundamentals. John Wiley & Sons, 2010.

Presenta un panorama muy completo del tema. Es sólo teórico pero muy claro.

- 2 R. Kimball, The Data Warehouse Toolkit: the definitive guide to dimensional modeling, John Wiley & sons, 3rd edition. 2013.

Libro sobre diseño. Cada capítulo presenta el diseño de un almacén e introduce conceptos.

- 3 A. Tennick, Practical MDX Queries. McGraw Hill 2010.  
Ejemplos de consultas en MDX.

