Almacenes y minería de datos

Dra. Amparo López Gaona
Fac. Ciencias, UNAM

Introducción

- Cada día crece, en forma espectacular, la cantidad de datos generados y registrados.
  kilobytes, megabytes, gigabytes, terabytes, petabytes, exabytes, zettabytes, yottabytes, etc.
- Principales fuentes de datos:
  - Comercio: e-commerce, transacciones, almacenes, etc.
  - Ciencia: simulación científica, bioinformático, procesamiento de imágenes, etc.
  - Día a día: noticias, cámaras digitales, YouTube,...
- Estamos ahogados en datos, pero sedientos de conocimiento.

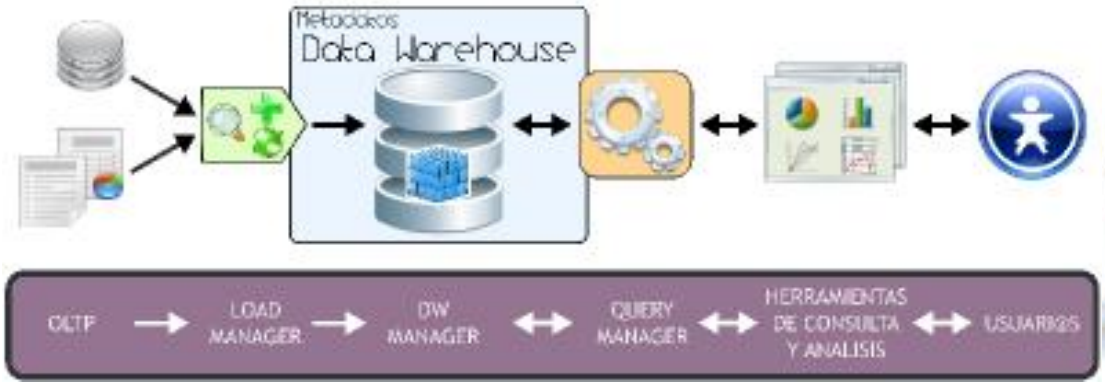
¿cómo puedo analizar estos datos?

... Introducción

- Las aplicaciones sobre **bases de datos** son muy importantes para la vida de una organización.
  - Soportan las operaciones día a día de los negocios.
  - Sin estos sistemas de cómputo, los negocios no pueden sobrevivir.
  - Reúnen, almacenan y procesan todos los datos necesarios para la ejecución exitosa de las operaciones diarias rutinarias. Proporcionan información en línea y producen reportes para monitorear y realizar los negocios.
- Las BDs han evolucionado desde el procesamiento de archivos hasta los SABD como los que conocemos. Pasando por diversos modelos de datos.
- Se han desarrollado métodos eficientes para el procesamiento de transacciones en línea OLTP, donde una consulta se ve como una transacción. Esto significó el uso masivo de BDR.

... Introducción

- Al expandirse los negocios, la complejidad de estos crece; los ejecutivos requieren información para ser competitivos y mejorar su línea de producción.
- En los 90 se empieza a tomar ventaja competitiva con la construcción de almacenes de datos (dwh).
- El **almacén de datos** visto como un repositorio de fuentes datos heterogéneas bajo un esquema uniforme en un solo sitio facilita a los ejecutivos la toma de decisiones.

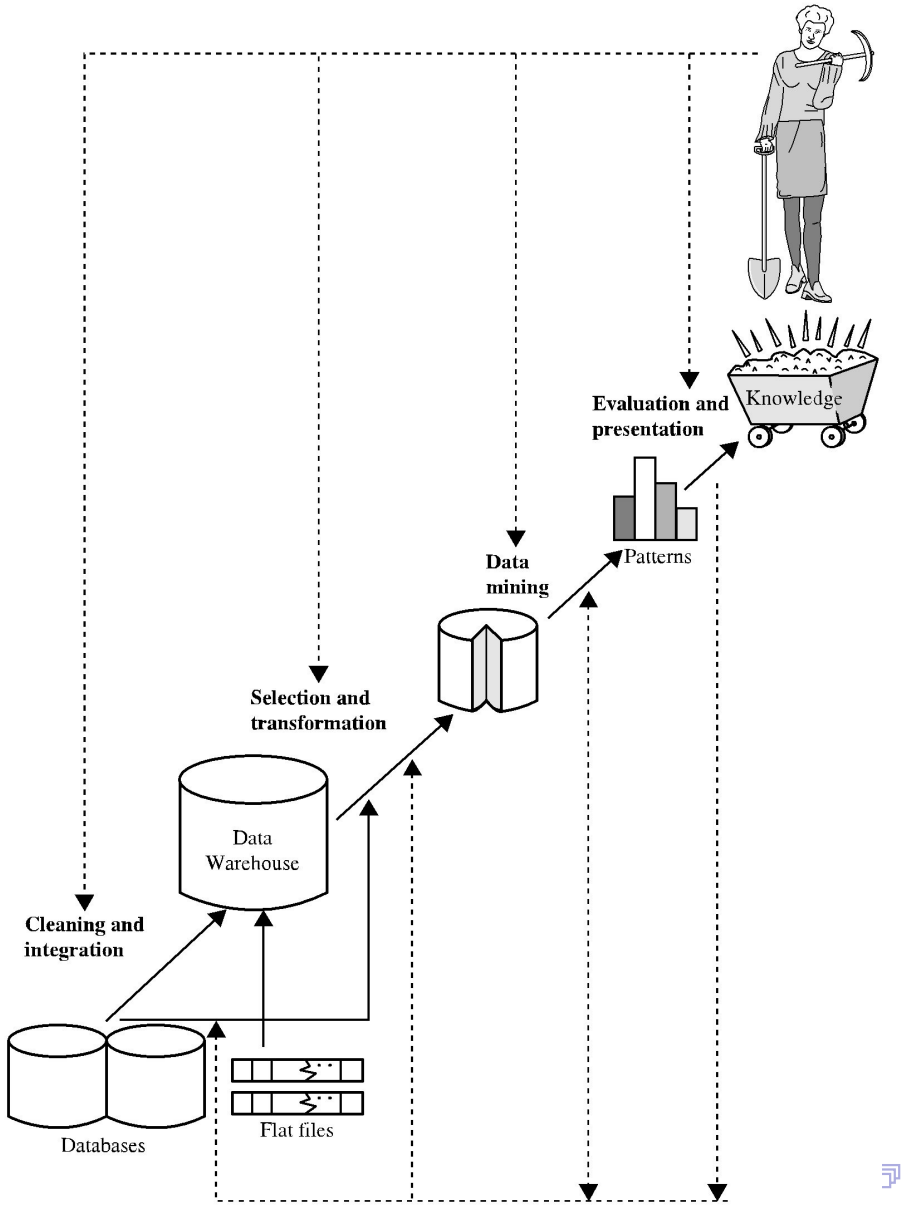


... Introducción

- La **minería de datos** es la extracción o "minado" de conocimiento de grandes volúmenes de datos.
  - es un proceso que intenta descubrir patrones (útiles, inesperados) en grandes volúmenes de datos.

- Ejemplo

Proceso de descubrimiento de conocimiento



Inteligencia de negocios (BI)

Inteligencia de negocios

- Extrae conocimiento de grandes cantidades de datos almacenados en alguna organización "moderna".
- Almacenes de datos, minería de datos.

Información estratégica

¿Quién necesita información estratégica? ¿Qué es información estratégica?
R. Los responsables de mantener la competitividad de una empresa.
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.
- Mejor el servicio al cliente en ...
- etc.

... Información estratégica

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 tecnologías emergentes.
- Conocer resultados de mercadotecnia y ventas.
- Conocer niveles de calidad, de productos y servicios.

Estos tipos de información esencial se llaman **información estratégica**.

... Información estratégica

La información estratégica no pretende producir una factura, hacer un envío, etc. es más importante para la salud y supervivencia de la corporación.
Las decisiones críticas dependen de la información estratégica apropiada de una empresa.
Características deseadas de información estratégica.

- **Integrada**.
- **Integra**.
- **Accesible**.
- **Creíble**.
- **A tiempo**.

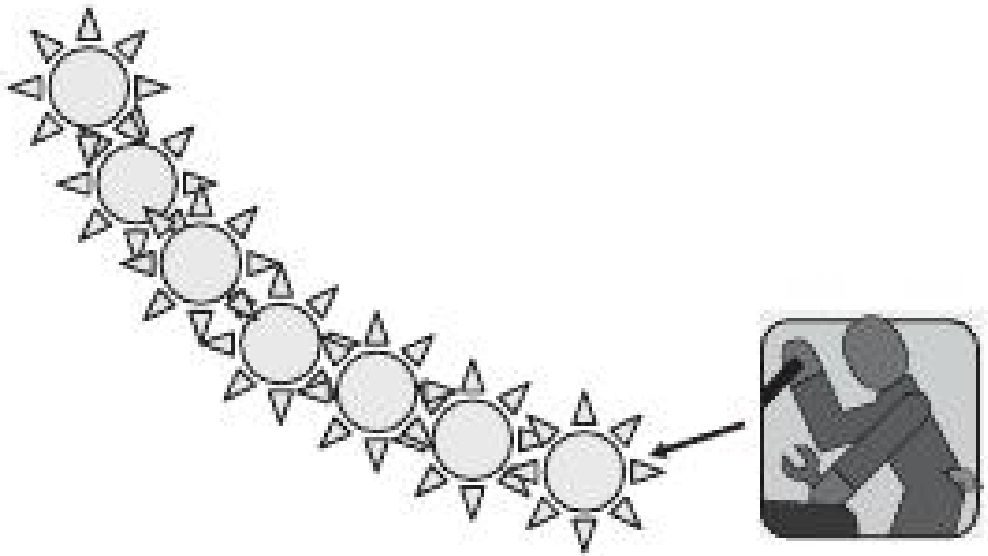
Sistemas operacionales vs. Apoyo a toma de decisiones

Los sistemas operacionales son sistemas para procesamiento de transacciones en línea (OLTP). Son sistemas que se usan para ejecutar el día a día de los 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



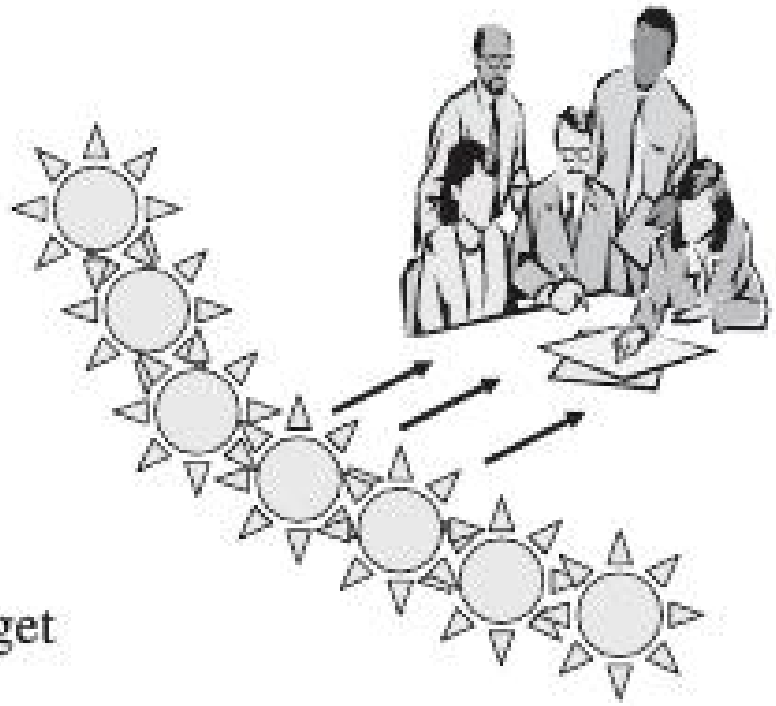
... Sistemas operacionales vs. Apoyo a toma de decisiones

Los 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



OLTP vs. OLAP

- Procesamiento de transacciones en línea (OLTP)
  - Muchas consultas "pequeñas" sobre una cantidad pequeña de tuplas de varias tablas que requieren unirse.
  - Actualizaciones frecuentes. El sistema siempre está disponible para actualizaciones y consultas.
  - Volumen pequeño de datos (unos cuantos históricos).
  - Modelo de datos complejo (normalizado).
- Procesamiento analítico en línea (OLAP)
  - Menos consultas, pero más grandes, generalmente requieren rastrear una gran cantidad de registros y hacer agregaciones.
  - Lecturas frecuentes, actualizaciones frecuentes (diariamente, semanalmente).
  - Operaciones en dos fases: lectura o actualización.
  - Grandes volúmenes de datos (colección de datos históricos).
  - Modelo de datos sencillo (multidimensional/de-normalizado).

Características de estos sistemas de apoyo

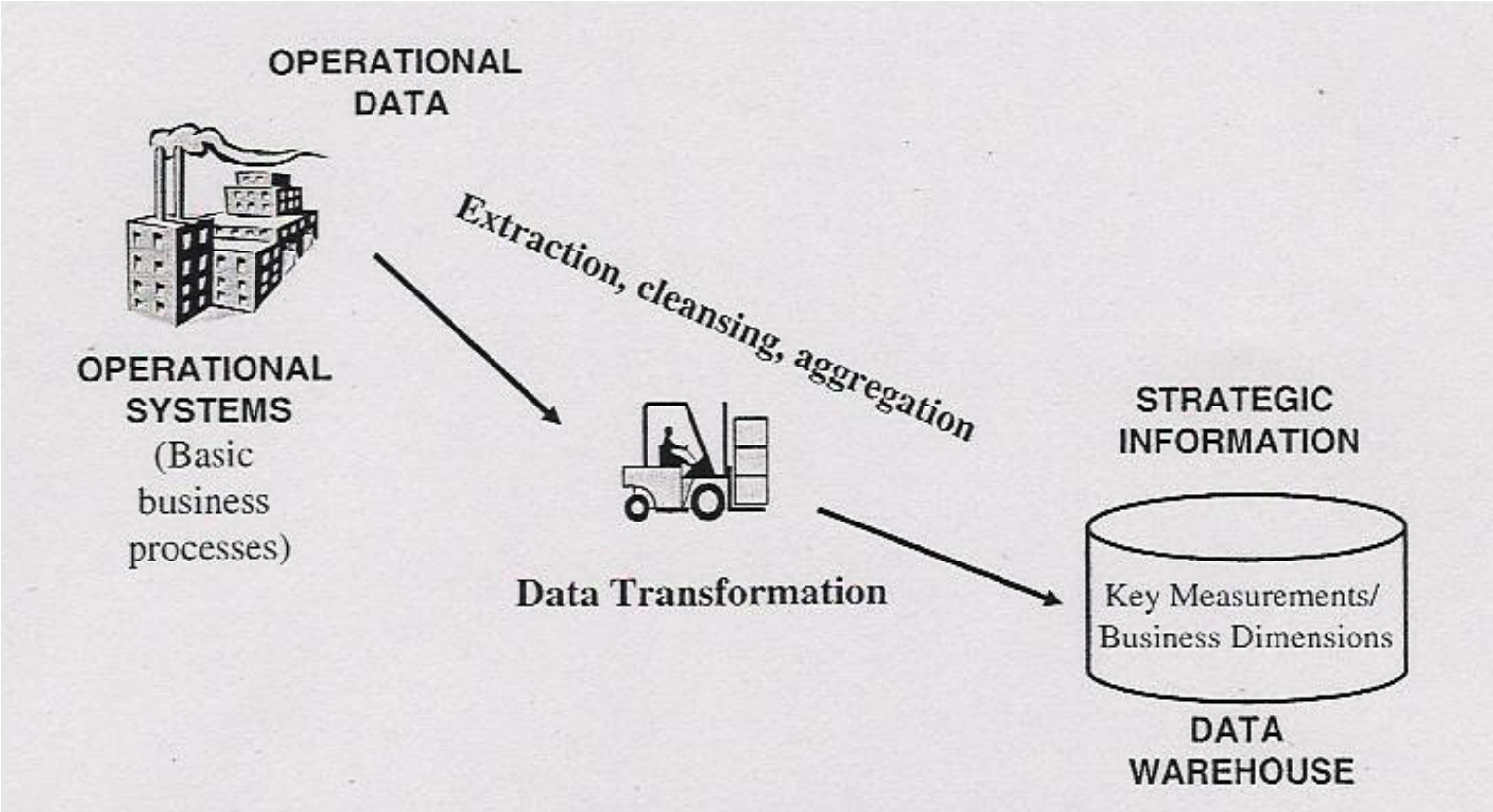
- Bases de datos diseñadas para tareas analíticas.
- Datos de múltiples aplicaciones
- Facilidad de uso y propicio para largas sesiones interactivas de los usuarios.
- Uso intenso de lectura de datos.
- Interacción directa con el sistema por los usuarios sin ayuda.
- Contenido estable y actualizado periódicamente
- Contenido con datos actuales e históricos
- Habilidad de los usuarios de ejecutar consultas y conseguir resultados en línea.
- Habilidad de los usuarios para iniciar reportes.

Requerimientos de procesamiento en el nuevo entorno

Hay al menos cuatro niveles de requerimientos de procesamiento analítico:

- Ejecutar consultas sencillas y obtener reportes de datos actuales e históricos.
- Realizar análisis "what if" en diferentes formas.
- Consultar, regresar, analizar y continuar el proceso tantas veces como se quiera.
- Detectar/marcar tendencias históricas y aplicarlas en procesos interactivos futuros.

Información estratégica de un DWH



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

Ejemplo: Compañía de ventas

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

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	...
...	...	...	...	...	...	...	...

item

item_ID	*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

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
...	...	...	...	...	...

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
...	...	...

purchases

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

items_sold

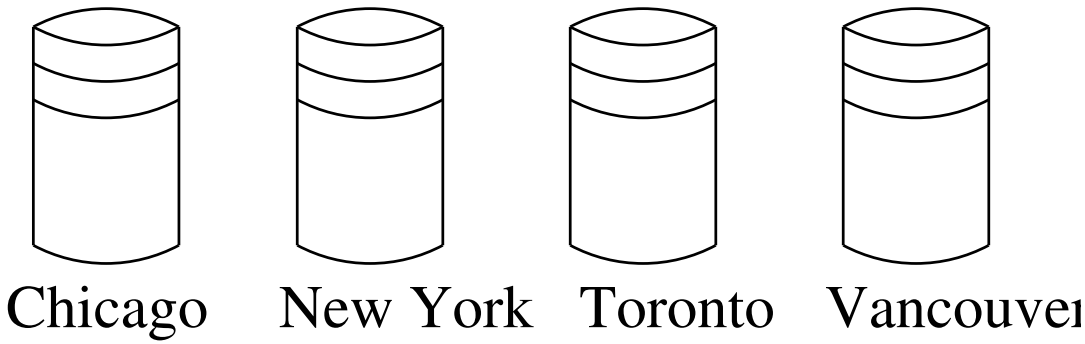
trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
...	...	...

works_at

empl_ID	*branch_ID*

... Ejemplo: Compañía de ventas

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



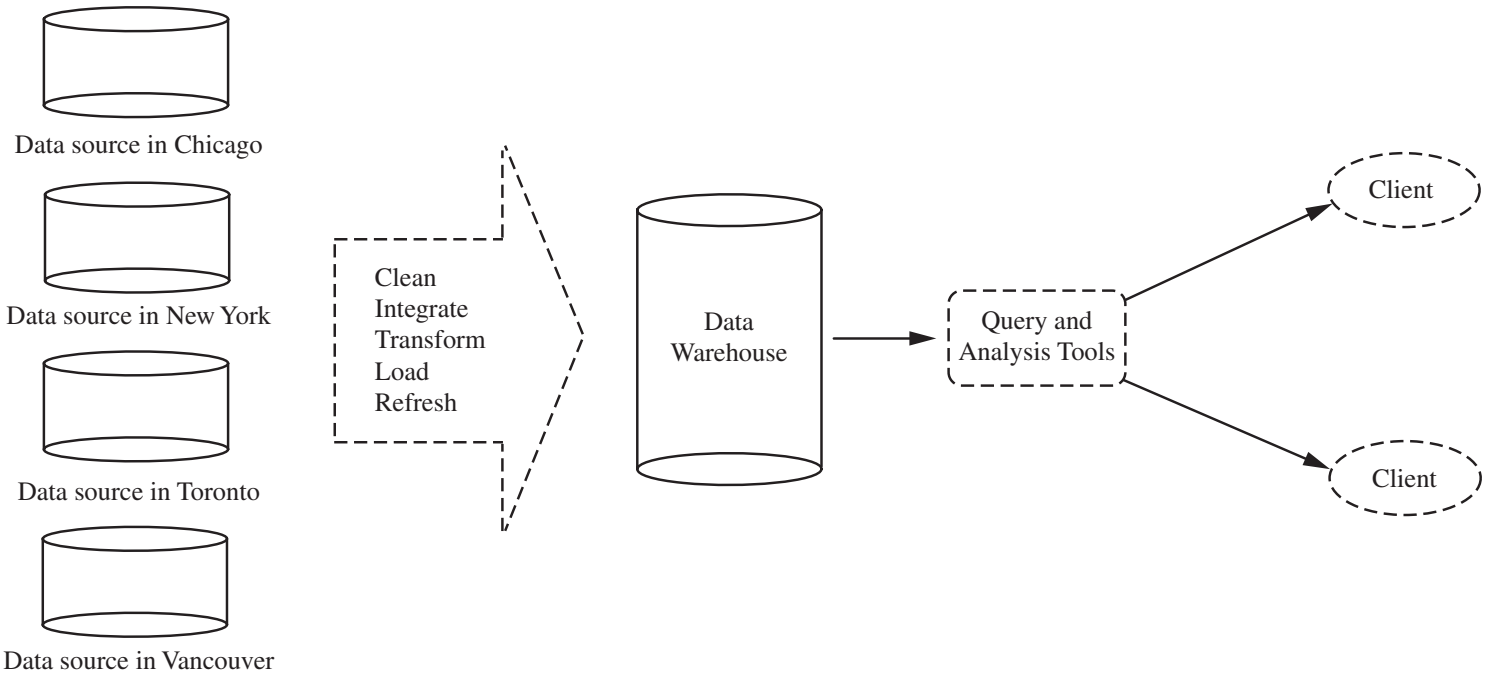
- Cada tienda tiene sus propios registros de clientes y de ventas.
- El mismo cliente puede ser visto como un cliente distinto para diferentes tiendas; difícil detectar información duplicada de los clientes.
- Datos imprecisos o perdidos en la dirección de algunos clientes.
- Los registros de compra se mantienen en el sistema operaciones por un tiempo corto (ej., 6 meses) y luego se borrar o archivan.
- El mismo producto puede tener diferente precio, o diferente descuento en las difretents tiendas

... 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
- Fuetnes heterogeneas
  - BDR, OLTP,
  - Hojas de cálculo, ...
- Los datos son adecuados para los sistemas operacionales.
  - Contabilidad, ventas, etc.
  - No soportan análisis de las funciones del negocio.
- La calidad de los datos es mala.
  - Los datos pueden ser imprecisos, estar perdidos, etc.
- Los datos son "volatiles"
  - Los datos pueden ser borrados (6meses)
  - Los datos pueden cambiar con el tiempo – no hay información histórica.

... Ejemplo: Compañía de ventas

- 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.
- Solución: un almacén de datos.



- Un DWH es un repositorio de información recolectada de varias fuentes, almacenada bajo un esquema unificado y que usualmente

... ¿ Qué es un 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. Tema como cliente, proveedor, producto, ventas. En lugar de procesamiento de transacciones de una organización.
- Integrada. Usualmente se construye integrando múltiples fuentes heterogéneas. Se requieren técnicas de limpieza e integración de datos para asegurar la consistencia entre los datos.
- Históricos. Los datos se almacenan para proporcionar información desde una perspectiva histórica. Cada elemento clave contiene explícita o implícitamente un elemento de tiempo.
- No volátil. No requiere mecanismos para procesamiento de transacciones, recuperación y control de concurrencia. Sólo requiere dos operaciones para accesar los datos: carga inicial y acceso de datos.

Datawarehousing

- Data warehousing es el proceso de construir y usar almacenes de datos.
- Los datos de los sistemas operacionales se:
  - Extraen.
  - Limpian.
  - Transforman.
  - Agregan/Resumen (?).
  - Cargan en el DWH.
- Un buen DWH es un pre-requisito para una BI exitosa.

¿Porqué separarlo?

La separación se basa en las distintas estructuras, contenido y uso de los datos en los dos sistemas.

- Alto rendimiento en ambos sistemas:
  - DBMS - afinado para OLTP: métodos de acceso, índices, control de concurrencia, recuperación de fallas.
  - DWH - afinado para OLAP: consultas complejas, vistas multidimensionales, consolidación.
- Diferentes funciones sobre diferentes datos:
  - Datos "perdidos": El soporte a decisiones requiere datos históricos que las BDs normalmente no mantienen.
  - Datos consolidados: El soporte a decisiones requiere consolidación (agregación, resúmenes) de datos de fuentes heterogéneas.
  - Calidad de los datos: diferentes fuentes típicamente utilizan diferentes representación, códigos y formatos de datos que deben ser unificados.

Factores de éxito

Un proyecto de DW 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.

Errores a evitar

Se debe evitar:

- Establecer expectativas demasiado altas.
- Cargar el DW con todo lo disponible.
- Elegir un administrador del DW sin orientación al negocio.
- Diseñar el DW igual que un sistema de producción.
- Ignorar fuentes de datos externas.
- Ignorar que los sistemas evolucionan.

Beneficios esperables

Se obtiene:

- Acceso interactivo e inmediato a información estratégica de un área de negocios.
- Permite toma de decisiones basadas en datos objetivos.
- Los beneficios aumentan:
  - cuanto más importantes son las decisiones.
  - cuanto más crítico es el factor tiempo.
- Capitalización de datos en bases heterogéneas.