



UNIVERSIDAD
DE LOS ANDES
MÉRIDA VENEZUELA



Datawarehousing

Jose Aguilar
CEMISID, Escuela de Sistemas
Facultad de Ingeniería
Universidad de Los Andes
Mérida, Venezuela

Introducción a Data Warehouse

Analizando la información de una empresa

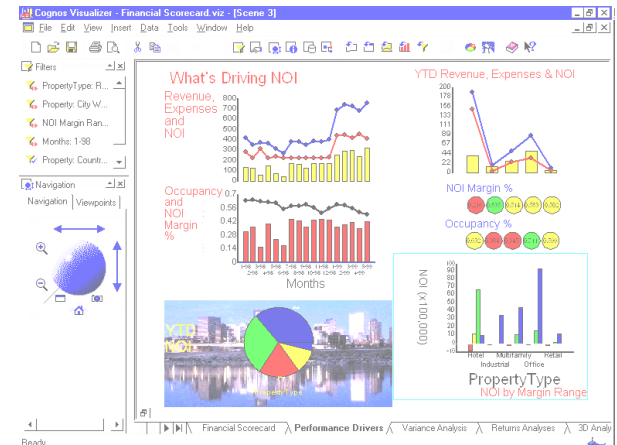
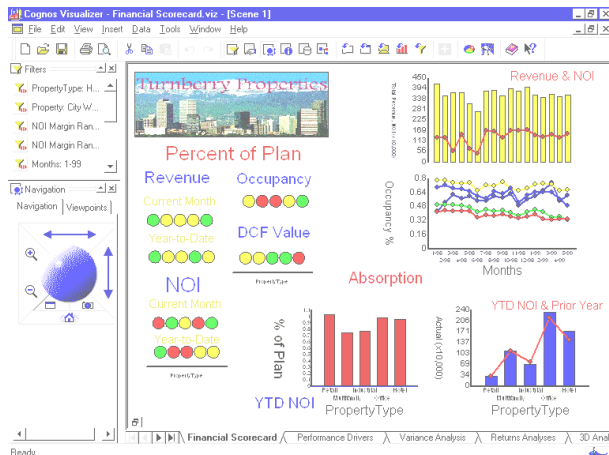
- ✓ Información periódica de las ventas
- ✓ Información del esfuerzo comercial
- ✓ Información sobre los pedidos a los proveedores

Por qué no integrarla y cruzarla para obtener:

- ✓ En qué zonas se está vendiendo más de cada línea de productos?
- ✓ Quienes son los clientes más rentables?
- ✓Cuál es la relación entre el esfuerzo comercial y las operaciones cerradas?
- ✓ De qué proveedores se está comprando la mayor parte de los productos vendidos ?

Introducción a Data Warehouse

- ✓ Se necesita entender no solo **QUE** está pasando, sino **CUANDO, DONDE, QUIEN, COMO Y POR QUE**.
- ✓ Requerimientos de información con **OPORTUNIDAD**.
- ✓ **ESCALAR, ENRIQUECER Y COMPARTIR** a todos los usuarios en la organización



Introducción a Data Warehouse

Ventas

- Número de pedidos
- Productos pedidos
- Clientes que realizaron los pedidos

- Clientes más rentables
- Pedidos más frecuentes
- Productos más rentables
- % de nuevos clientes

Servicio al Cliente

- Datos de llamadas de nuestros clientes
- Información sobre el log de nuestra página web

- Qué clientes visitan nuestra página web ?
- % pedidos realizados por nuestros canales de ventas
- Qué consulta es más frecuente?

Marketing

- Número de campañas realizadas y características de cada una

- % éxito de las campañas
- Qué tipo de clientes han respondido mejor a cada una de las campañas realizadas ?

Distribución

- Productos que salen diariamente del almacén
- Tiempo teórico de entrega

- Número de pedidos retrasados
- Distribuidor que tiene el mayor número de retrasos
- Tiempo medio de entrega

DATO

INFORMACION

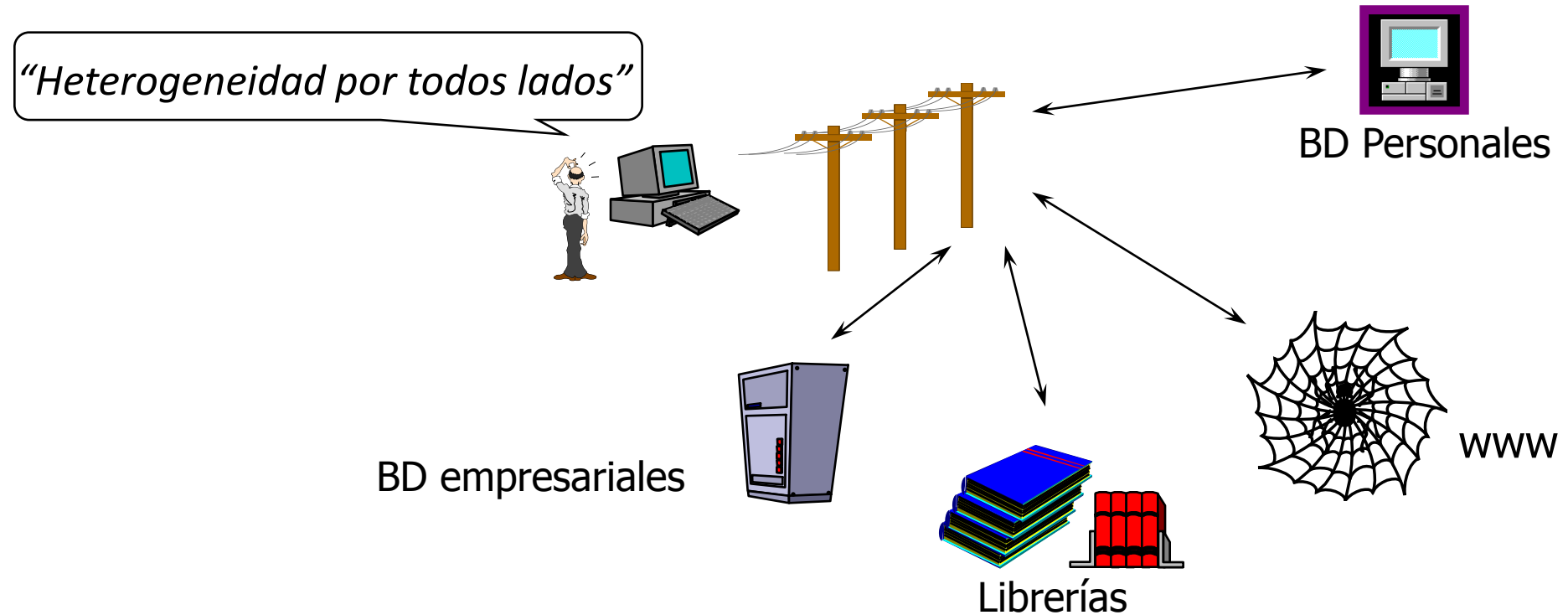
Introducción a Data Warehouse

El Poder de la Información

"Hace diez años les pude decir cuántos productos vendimos al Oeste del Mississipi. Hoy no sólo les puedo decir eso mismo, sino cuántos vendimos y de que tipo en California, en el Condado de Orange, en la ciudad de Irvine, en el Supermercado local Von's, en una promoción especial, al final del pasillo 4, los Jueves".

D. Wayne Calloway Director Ejecutivo de Operaciones de Pepsico
en una asamblea general de accionistas

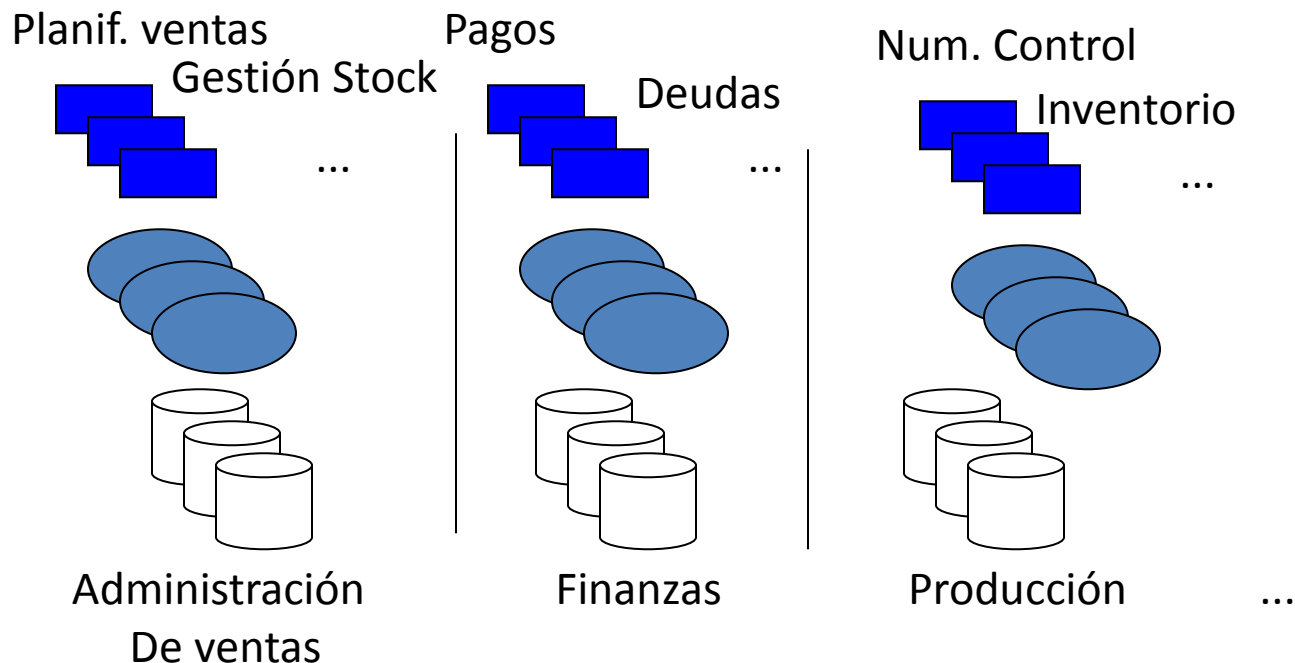
Problemas: Heterogeneidad de las fuentes de Información



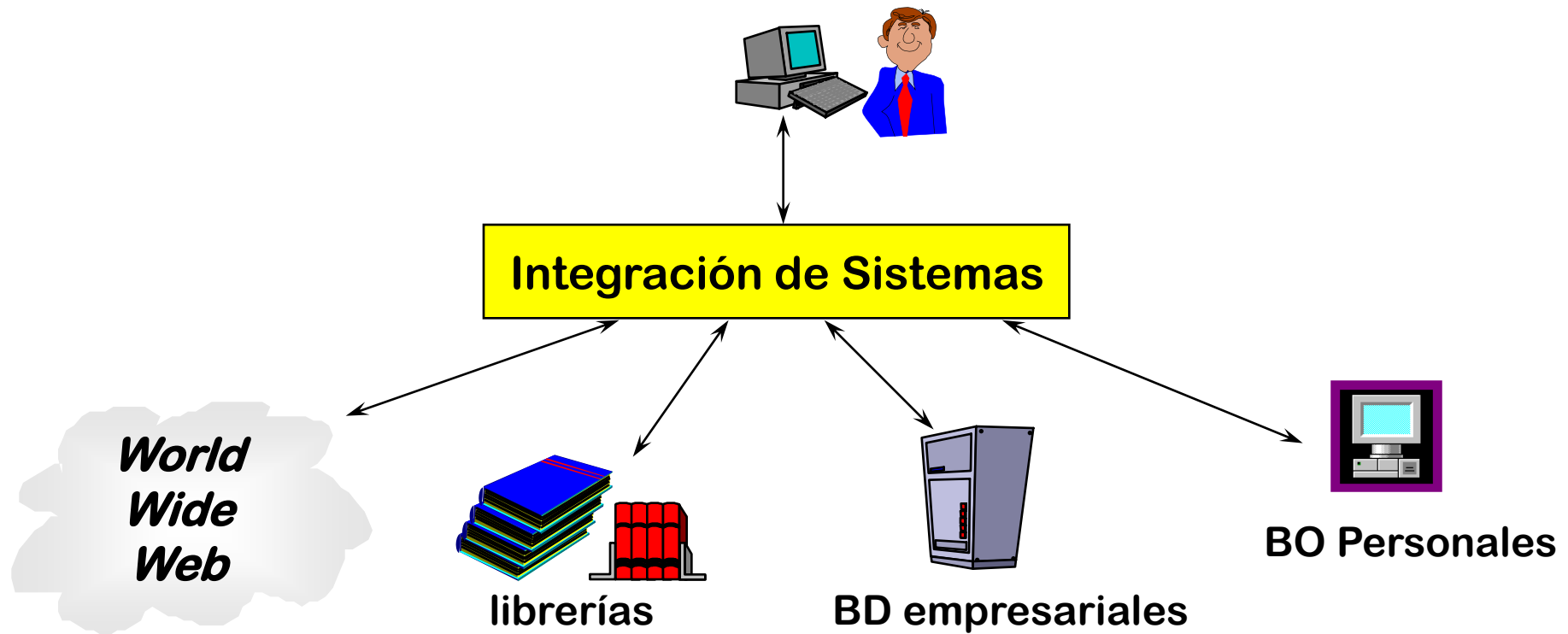
- Diferentes interfaces
- Diferentes representaciones de datos
- Información duplicada e inconsistente

Problema: Gestión de datos en grandes empresas

- Fragmentación vertical de los sistemas de información
- Desarrollo de las aplicaciones guiadas por los sistemas operativos



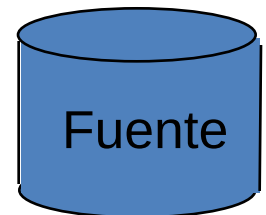
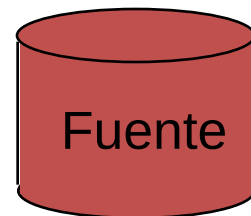
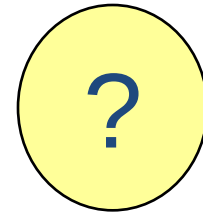
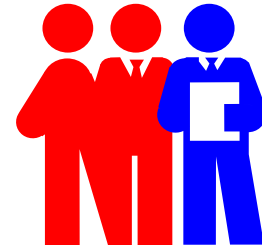
Objetivo: Unificar Acceso a los Datos



- Recopilar y combinar la información
- Proporcionar visión integrada, en una interfaz de usuario uniforme
- Soportar el intercambio

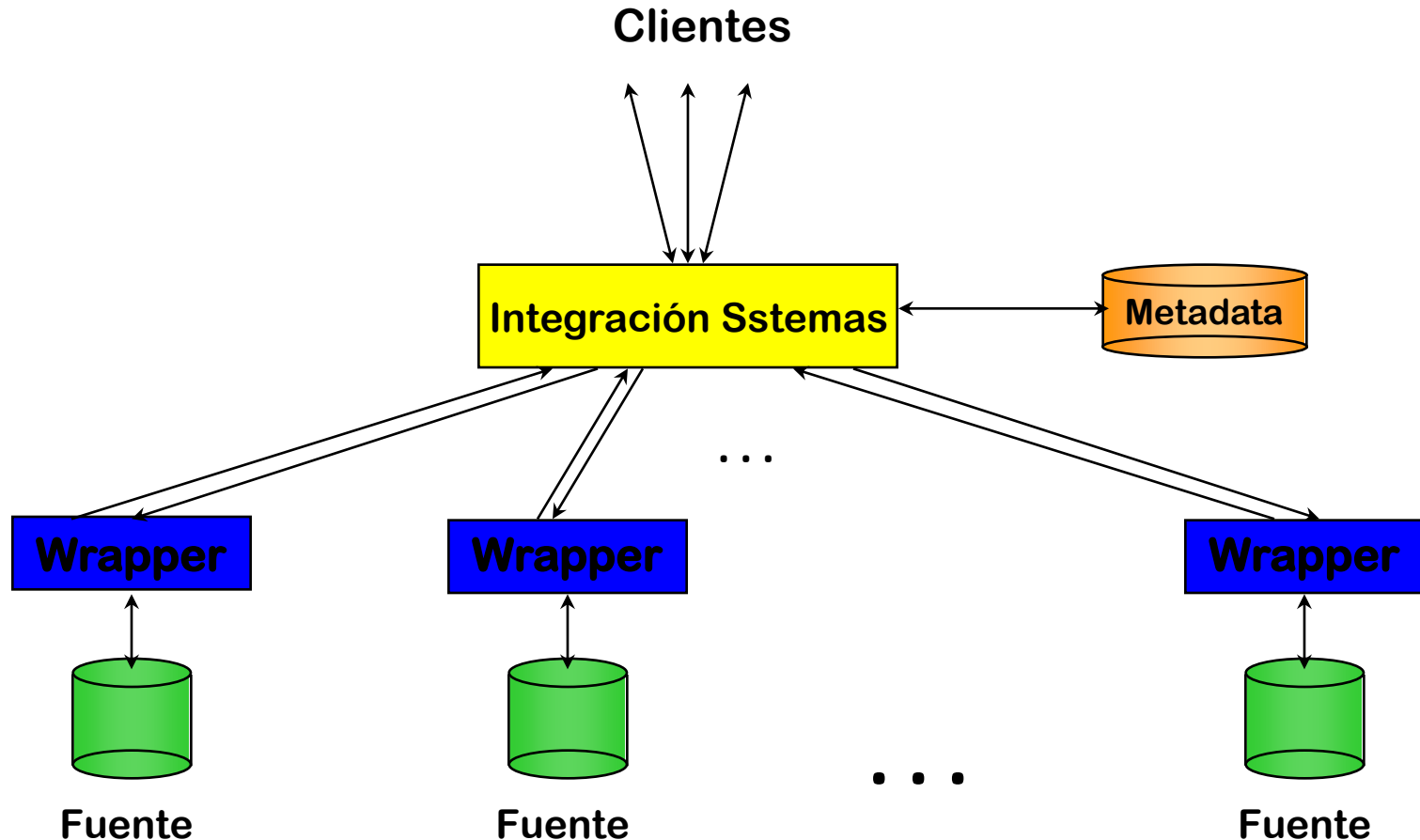
Objetivo: Unificar Acceso a los Datos

- Dos enfoques:
 - Guiado por la consulta (perezoso)
 - Warehouse (ansioso)



Enfoque tradicional

- Guiado por la consulta (perezoso, on-demand)

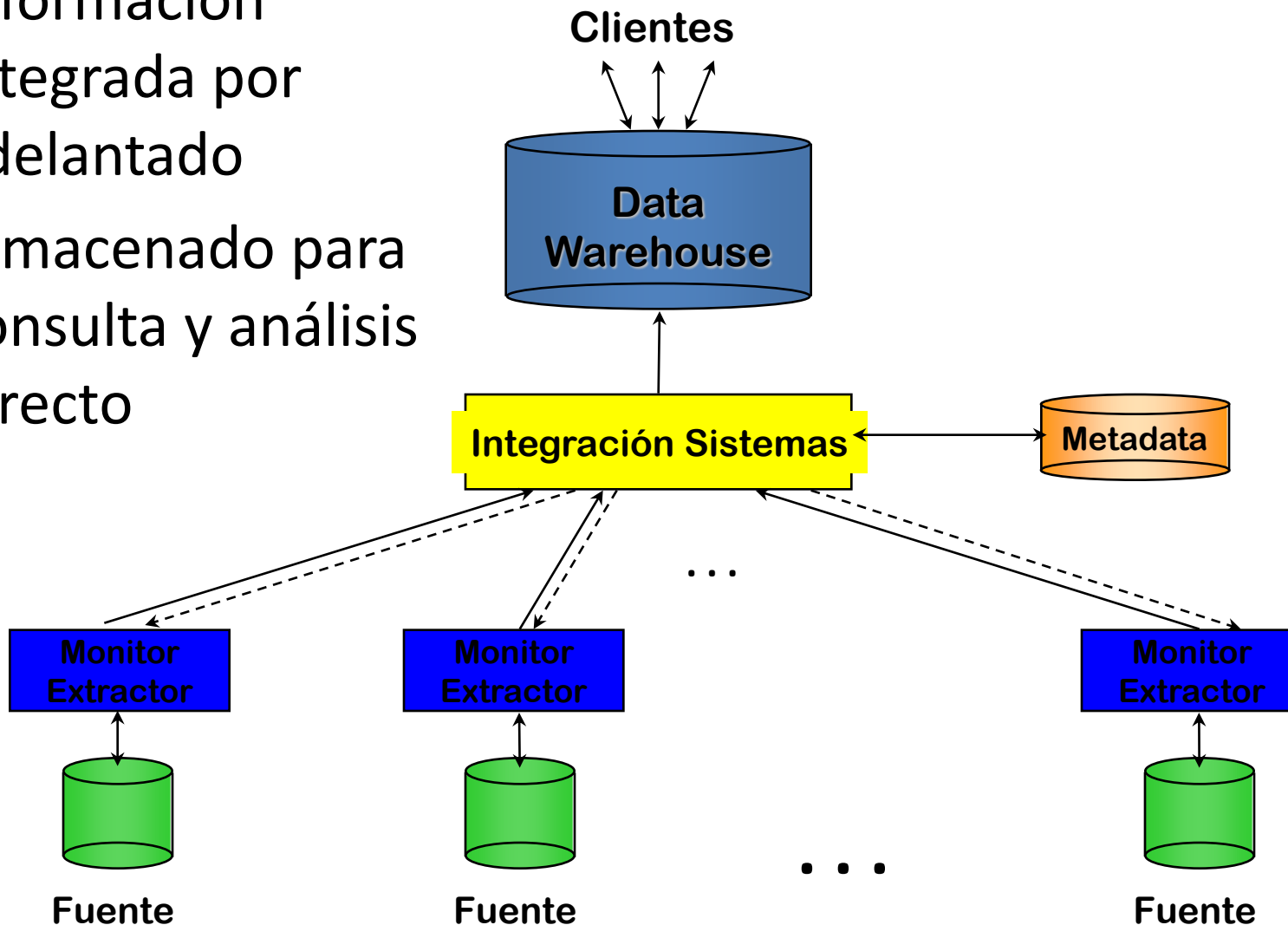


Problema del Enfoque tradicional

- ♦ El retraso en el procesamiento de consultas
- ♦ Fuentes de información lentas o no disponibles
 - ♦ Filtrado e integración son complejas
 - ♦ Ineficiente y potencialmente costoso para las consultas frecuentes
- ♦ Compite con el procesamiento local en el sitio fuente
- ♦ No ha calado en la industria

Enfoque Warehousing

- Información integrada por adelantado
- Almacenado para consulta y análisis directo



Ventaja Enfoque Warehousing

- **Alto rendimiento de la consulta**
 - Pero no necesariamente la información más actualizada
- **No interfiere con el procesamiento local en el sitio origen**
 - Las consultas complejas en warehouse
 - OLTP en las fuentes de información
- **Información copiada en el almacén**
 - Se puede modificar, anotar, resumir, reestructurar, etc.
 - Puede almacenar información histórica
 - Seguridad, sin auditoría
- **Usada en la industria**

Concepto Data Warehousing

- *Es un **gran almacén de datos** para consultar
- *Es un **repositorio de datos** de muy fácil acceso, alimentado de numerosas fuentes, transformadas en grupos de información sobre temas específicos de negocios, para permitir nuevas consultas, análisis, toma de decisiones.

- *Tiene **gran capacidad de almacenamiento**, pues los datos pueden ser de grandes periodos de tiempo.



Concepto Data Warehousing (Bodegas de Datos)

Se trata, de **un expediente completo de una organización**, más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos.

Concepto Data Warehousing (Bodegas de Datos)

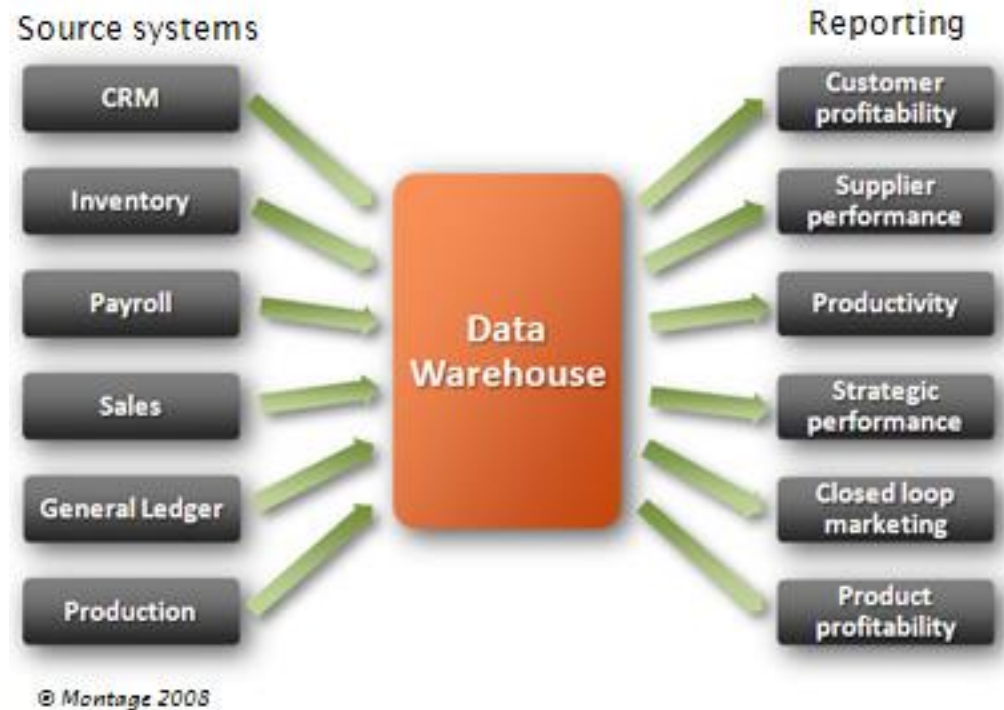
Conjunto de **datos integrados y orientados a una materia, varían con el tiempo, soportan el proceso de toma de decisiones de la administración** y esta orientada al manejo de grandes volúmenes de datos provenientes de diversas fuentes o diversos tipos.

Estos datos **cubren largos períodos de tiempo** lo que trae consigo que se tengan diferentes esquemas de los datos. **Previo a su utilización,** se debe aplicar procesos de análisis, selección y transferencia de datos seleccionados desde diversas fuentes.

Concepto Data Warehousing

Basado en el **concepto de Metadatos** (datos que describen otros datos). Por ejemplo, en una biblioteca se usan fichas que especifican autores, títulos, casas editoriales y lugares para buscar libros. Así, los metadatos ayudan a ubicar datos.

Cumple el principio de arquitectura fundamental que es **separar los sistemas transaccionales de los informacionales** en dos entornos de manera que el **análisis de los datos existentes no interfiera** con el procesamiento y registro de nuevos datos.



Concepto Data Warehousing (Bodegas de Datos)

- Datawarehouse (Almacén de datos) se define como una colección de datos que verifican las siguientes propiedades:
 - Está orientado a un tema
 - Datos integrados
 - No volátiles
 - Variante en el tiempo

que surgieron como una herramienta de soporte para la toma de decisiones a nivel gerencial

Orientado hacia temas



Los datos se almacenan y agrupan por temas de interés

Se agrupa por temas orientados a la organización, tales como :

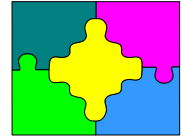
- Clientes
- Productos
- venta

en lugar de las transacciones individuales.

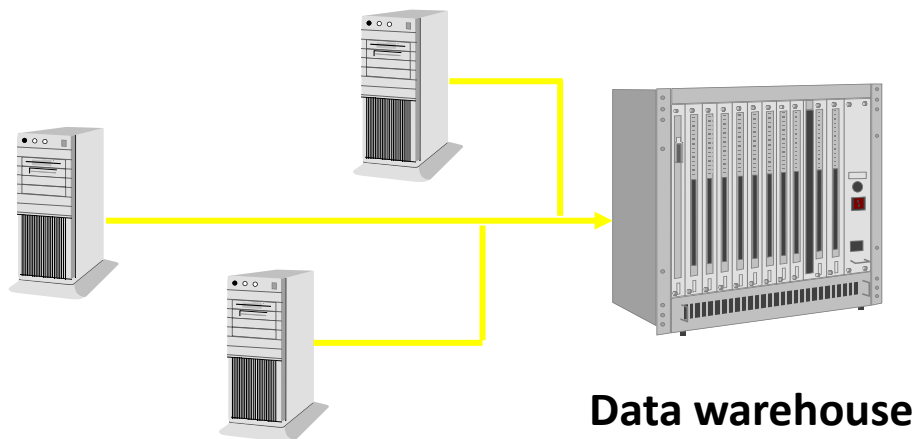
La normalización no es relevante

- **Centrándose en el modelado y análisis de datos** para la toma de decisiones, no en las operaciones diarias o el procesamiento de transacciones.
- Proporcionar una **visión sencilla y concisa en torno a cuestiones temáticas particulares** mediante la exclusión de los datos que no son útiles en el proceso de apoyo a la decisión.

Integración de Datos



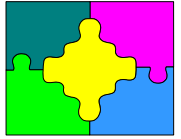
- El almacén de datos **integra datos** que provienen de varias fuentes. Parte de una base de datos operacional y mediante un proceso de **carga de datos** hace el Datawarehouse.
- El proceso de carga es lo más complicado por problemas de codificación, preparación de los datos, ...



Sistemas Operacionales

Los datos en el almacén deben ser:

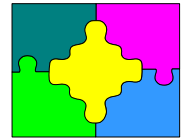
- Limpios
- Validados
- Adecuadamente integrados



Integración de Datos

- Integración de **múltiples fuentes de datos heterogéneas**,
 - bases de datos relacionales, archivos planos, registros de transacciones en línea
- Se aplican **técnicas de limpieza y de integración de datos**.
- Garantiza la **coherencia** en los convenios de denominación, estructuras de codificación, medidas, etc., entre las diferentes fuentes de datos
 - Por ejemplo, el precio del hotel: moneda, impuestos, desayuno, etc.
- Cuando los datos se mueven a la bodega, **se convierten** al estandar usado.

Ejemplo Integración de Datos



Sistema de Cuenta de cheques

Jane Doe (name)
Female (gender)
Bounced check #145 on 1/5/95
Opened account 1994

← *Datos operacionales*

Sistema de cuentas de ahorro

Jane Doe
F (gender)
Opened account 1992

cliente

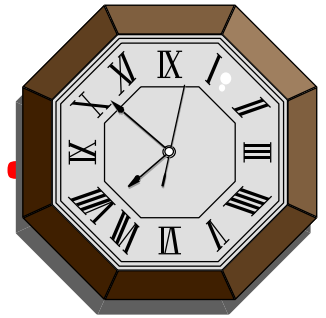
Jane Doe
Female
Bounced check #145
Married
Owns 25 Shares Exxon
Customer since 1992

Sistema de inversiones de clientes

Jane Doe
Owns 25 Shares Exxon
Opened account 1995

↑ *data warehouse*

... variante en el tiempo ...



- Todos los datos en el almacén de datos tienen una **marca de tiempo** en el momento de entrada en el almacén o cuando se usan en el almacén.
- Esta grabación cronológica de datos ofrece posibilidades **históricas y análisis de tendencias**.
- Normalmente, en las BD operativas **los datos se sobrescribe**, ya que los valores anteriores no son de interés.

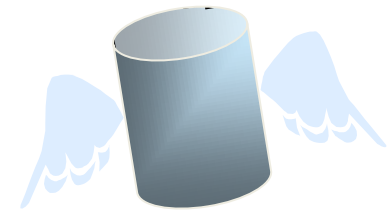
TIEMPO

id_tiempo

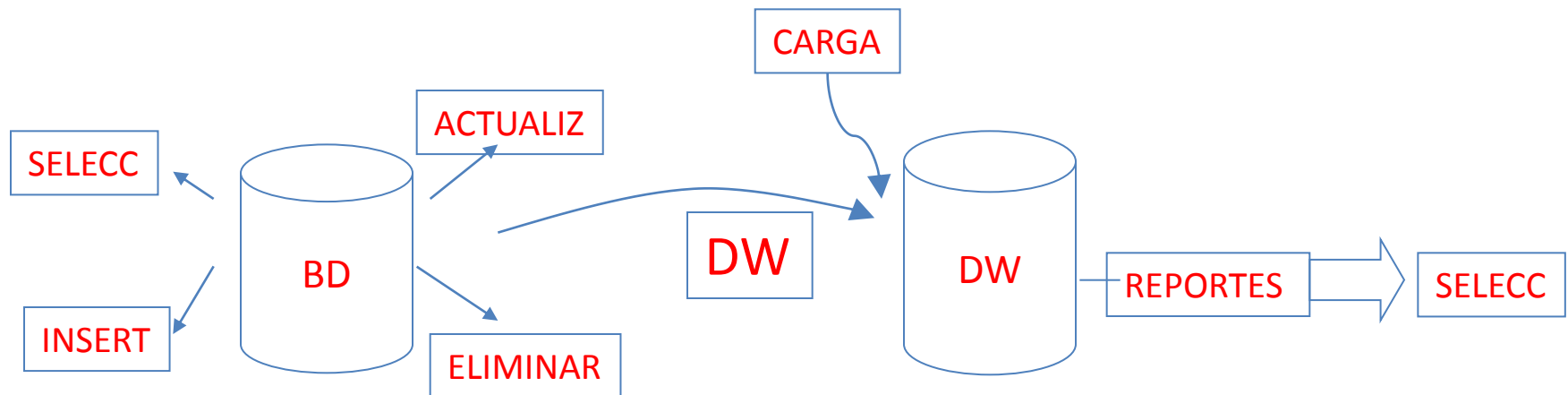
* periodo

porque se hacen cargas de datos continuamente, cuando los datos van cambiando, se actualizan los históricos y se guardan en ficheros temporales.

No volátiles



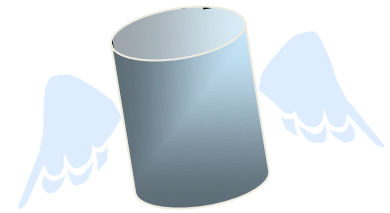
Son estables, una vez almacenados los datos **no se modifican**.



Datos actúan como un recurso estable para informes coherentes y análisis comparativo.

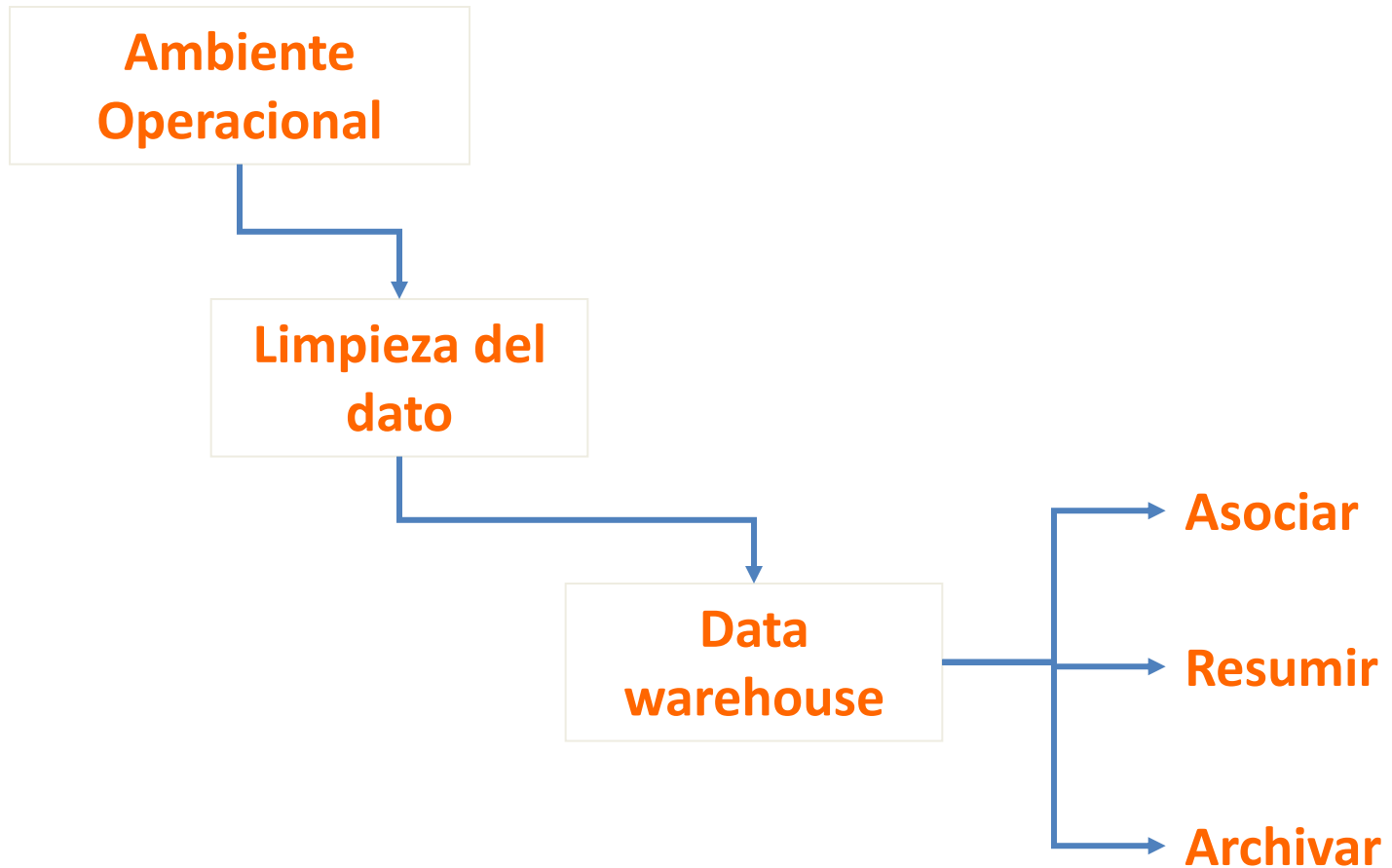
Por el contrario, los datos operativos se actualizan (insertar, eliminar, modificar, etc.)

No volátiles



- **Físicamente separado** de los datos del entorno operacional.
- **Actualización de los datos no se produce** en el entorno de almacén de datos.
- **No requiere el procesamiento de transacciones**, ni mecanismos de recuperación y control de concurrencia
 - Requiere sólo dos operaciones en el Acceso a los datos:
carga inicial de datos y el acceso a los datos.

Flujo del Dato



Diferencia Data Warehousing y BD

BD

- Una **colección estructurada** de registros o datos

Data Warehouse

- Una **colección lógica** de información, recogida de muchos diferentes bases de datos operacionales, que apoya las actividades de análisis organizacional y las tareas de toma de decisiones

Formato

Modelos BD

relacional
jerárquica
red
orientada a objetos
Etc.

Data Warehouse

- Tradicionalmente heterogénea
=> integración
- Se construyen envolturas/
mediadores en la parte superior
- Cuando una consulta se plantea
a un sitio, un **meta-diccionario**
se usa para traducir la consulta
en consultas apropiadas para el
sitio
- Los resultados se integran en un
conjunto de respuestas globales
Filtradas

La información de fuentes heterogéneas se
integra con antelación y se almacena en
bodegas para consultas directas y análisis

BD

Diseñado para
tiempos
operaciones en
tiempo real

En su mayoría
permanente
actualizaciones

Muchas pequeñas
transacciones

Mb - Gb de datos

Data Warehouse

Diseñado para el análisis de
la organización por
categorías y atributos

Sobre todo se leen

Las consultas son largas y
complejas

Gb - Tb de datos

BD

- Instantánea
- Los datos en bruto
- Cientos de usuarios (por ejemplo, los usuarios de una organización)

Data Warehouse

- Histórica
- Resumidos y conciliados los datos
- Decenas de usuarios (por ejemplo, los gerentes, los analistas)

Data Warehouse vs. SMBD

- **OLTP (on-line transaction processing)**
 - Major tarea de SMBD relacionales tradicionales
 - Las operaciones del día a día: compras, inventario, financiera, nómina, contabilidad, etc.
- **OLAP (on-line analytical processing)**
 - Mayores tareas de data warehouse
 - Análisis de datos y toma de decisiones

Características diferentes (OLTP vs. OLAP):

- **Orientación de la BD:** usuario vs organización
- **Contenido de los datos:** actual, detallada vs histórica, consolidada
- **Diseño de base de datos:** aplicación + ER vs estrella + por tema
- **Vista:** actual, local vs integrada, , evolutiva
- **Patrones de acceso:** actualizar vs consultas complejas

Las diferencias de diseño

BDs

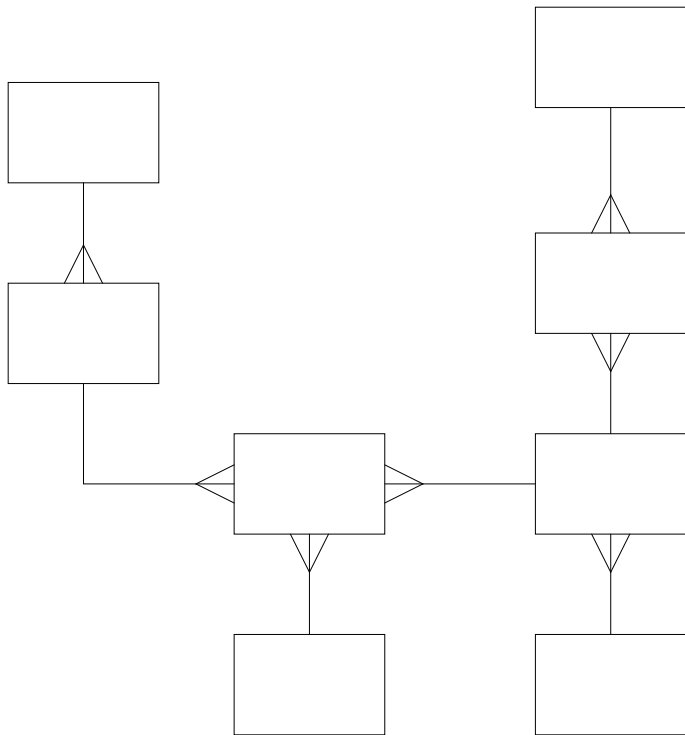
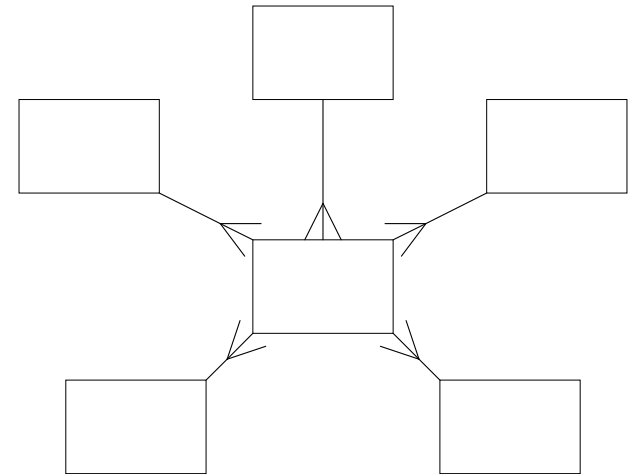


Diagrama ER

Data Warehouse



Estrella

RESUMEN DIFERENCIAS

BD OPERACIONAL

- Datos operacionales
- Orientado a aplicaciones
- Datos Actuales
- Datos Detallados
- Datos en continuo cambio

DATAWAREHOUSE

- Datos de negocio
- Orientado al sujeto
- Actuales + Histórico
- Datos Resumidos
- Datos Estables

Por qué Data Warehouse?

Diferentes funciones y datos:

- **Datos que faltan:** apoyo a las decisiones requiere datos históricos que BDs operacionales no suelen tener
- **Consolidación de datos:** Se requiere de la consolidación (agregación, resumen) de los datos de fuentes heterogéneas
- **Calidad de los datos:** las diferentes fuentes suelen utilizar representaciones de datos inconsistentes, códigos y formatos que deben ser conciliados, etc.

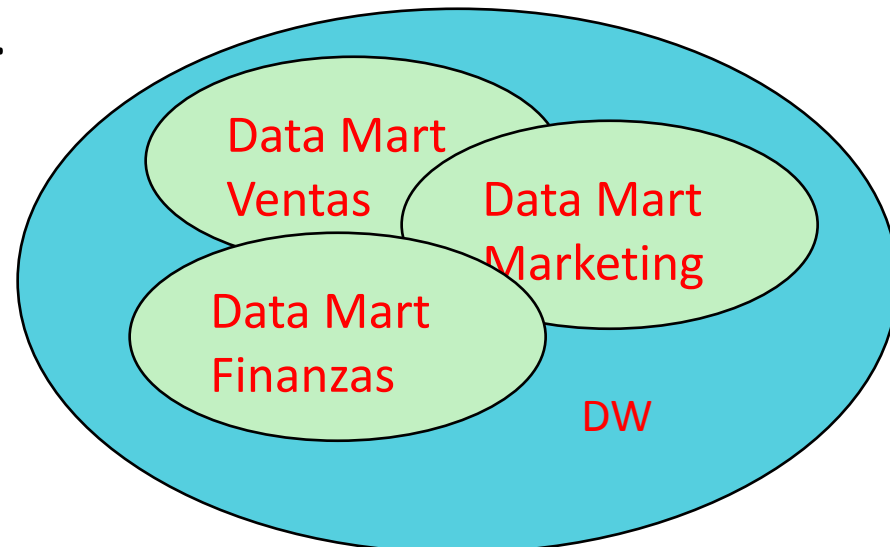
Data Mart

Las grandes cantidades de datos en el Data Warehouse a veces se subdividen en unidades lógicas más pequeñas (**data marts**)

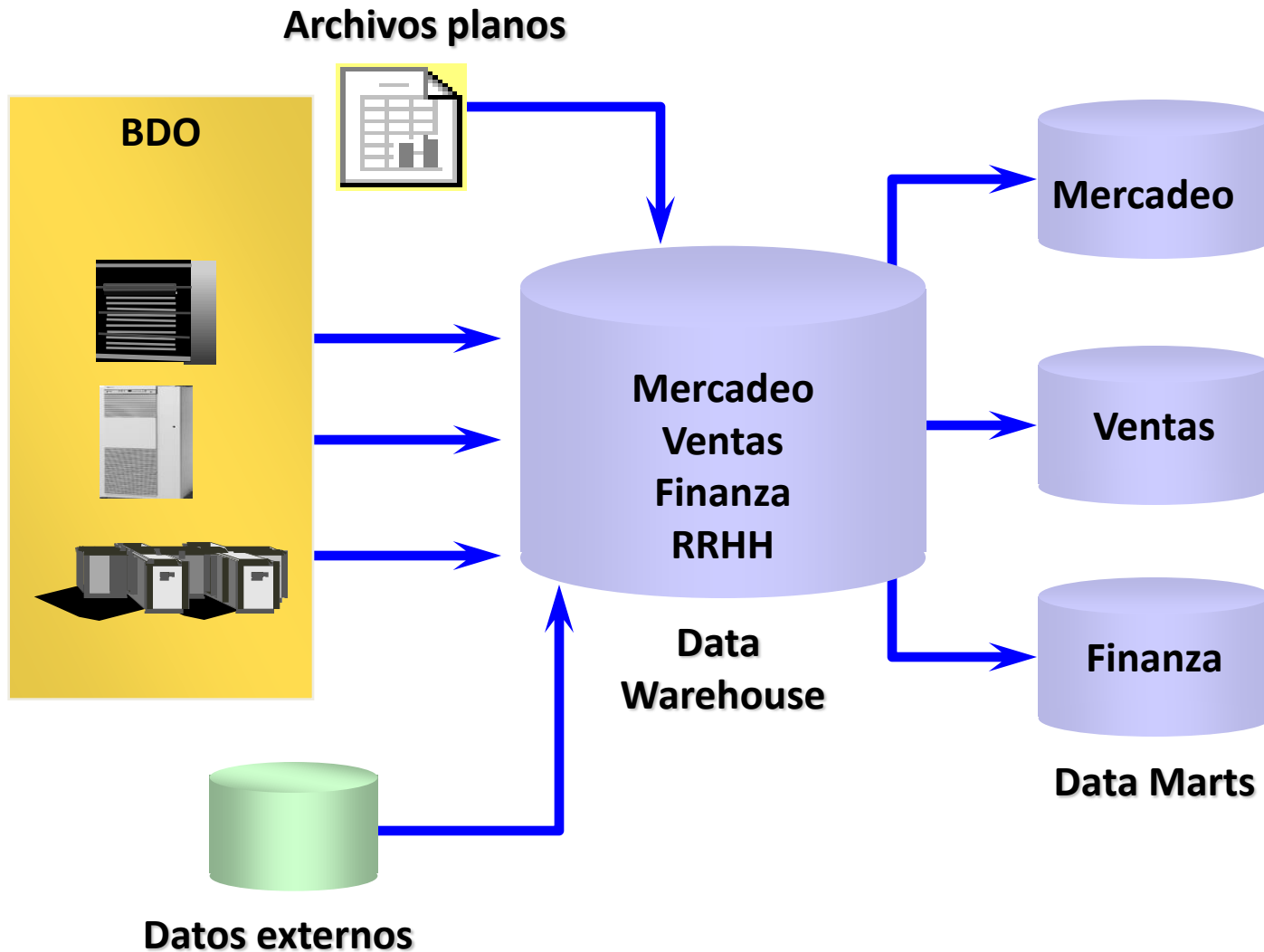


Data Mart

- Una versión reducida de menor costo que el data warehouse, diseñado para una **unidad de negocios estratégica o departamento**.
- Entre las **características** de un data mart destacan:
 - Usuarios limitados.
 - Área específica.
 - Tiene un propósito específico.
 - Tiene una función de apoyo.

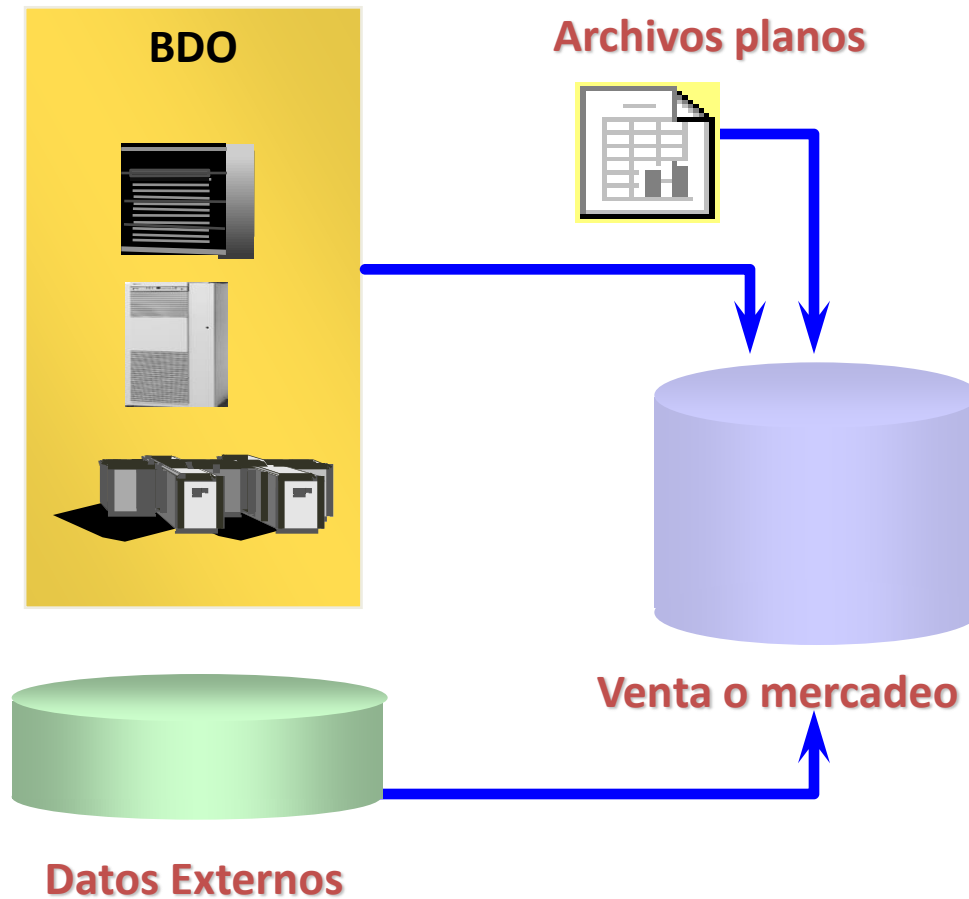


Tipo Data Mart



Dependiente de DW

Tipo Data Mart



Independiente de DW

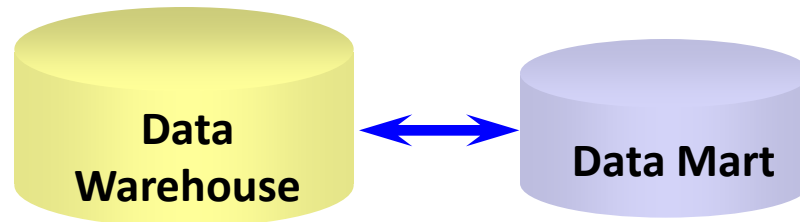
Razones para la creación de un Data Mart

- Para dar a los usuarios un **acceso más flexible** a los datos que necesitan para analizar con más frecuencia.
- Para proporcionar datos en una forma que coincide con el **punto de vista** colectivo de un grupo de usuarios
- Para mejorar el **tiempo de respuesta** del usuario final.
- Los usuarios potenciales de un data mart están **claramente definidos**

Razones para la creación de un Data Mart

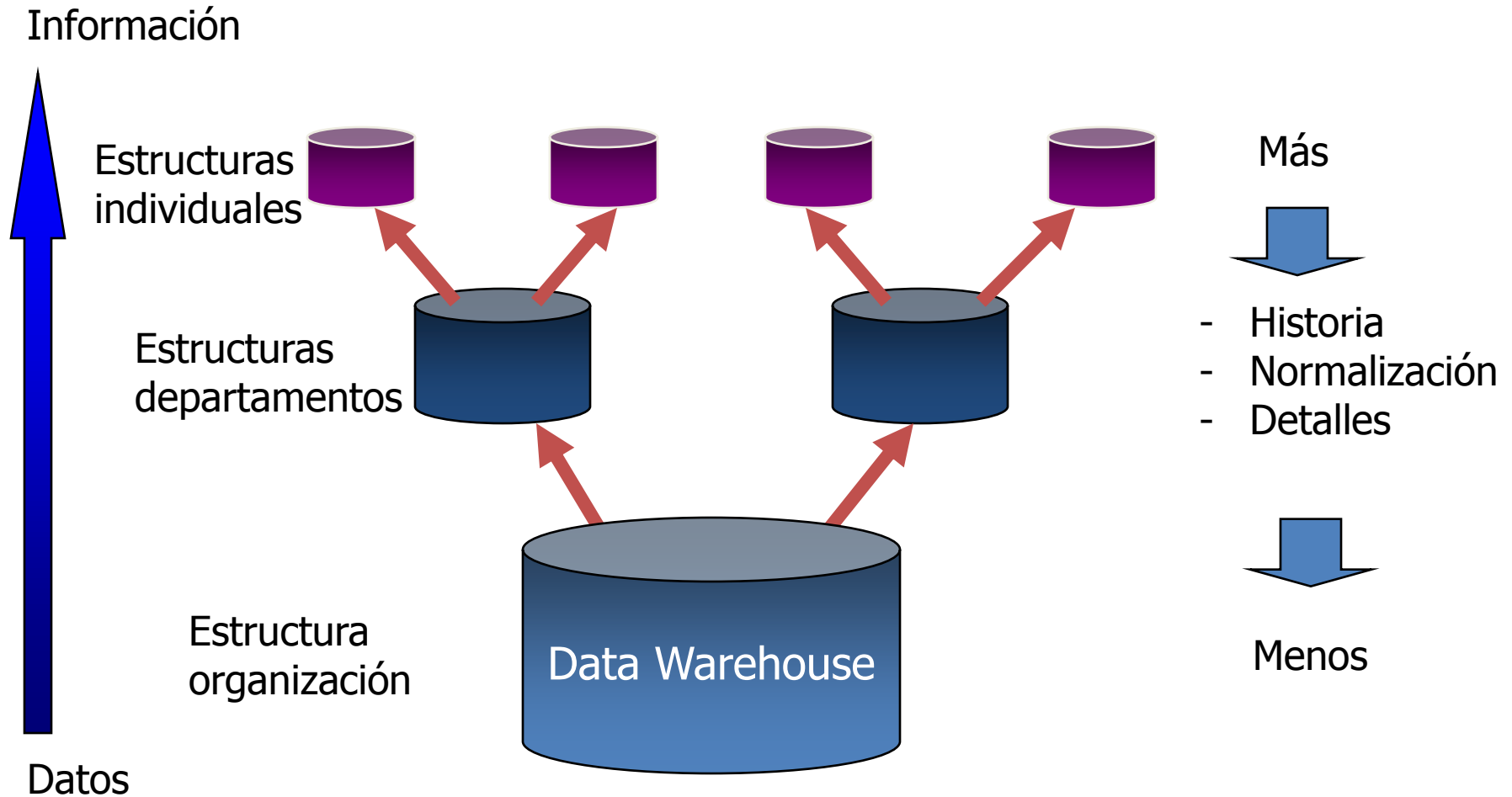
- Proporcionar datos estructurados según las **exigencias de las herramientas** de acceso de los usuarios finales.
- La construcción de un data mart **es más sencilla** en comparación con el establecimiento de un data warehouse corporativo.
- El **costo de la implementación** de data mart es **mucho menor** que el requerido para un data warehouse

Data Warehouses Vs Data Marts



	Data Warehouse	Data Mart
Alcance	Empresa	Departamento
Sujetos	Múltiple	Grupo
Fuente de datos	Muchas	Pocas
Tamaño (típico)	100 GB a > 1 TB	<100 GB
Tiempo de implementación	Meses o años	Meses

Desde Data Warehouse a Data Marts



Resumen conceptual

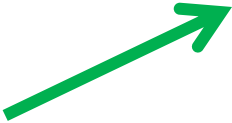
- **Data Warehouse** - Almacén de datos consultable en la empresa. Está formada por la unión de todos los data mart que lo constituyen.
- **Data Mart** - Un subconjunto lógico Data Warehouse. A menudo visto como una mirada del Data Warehouse para un único proceso de la organización o para un grupo de procesos relacionados.
- **Datos Operacionales** - BDs operativas que se desarrollaron independientes una de otra. Apoyan operaciones del día a día de la organización, por lo que debe actualizarse continuamente.

Funcionamiento de Data warehouse

Tres funciones esenciales:

1. Recopilación de los datos desde Bases de Datos
2. Gestión de los datos en el almacenamiento
3. Análisis de datos para toma de decisiones

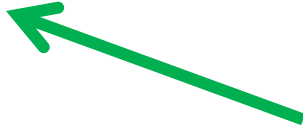
Sistemas de
Soporte a
Decisiones
(DSS)



herramientas para
hacer consultas o
informes



Sistemas de
información
ejecutiva (EIS)



Elementos que integran un almacén de datos

Metadatos

"datos acerca de los datos".

Su función es recoger la descripción de la estructura del almacén de datos:

Tablas

Columnas en tablas

Jerarquías y Dimensiones de datos

Relaciones entre tablas

Entidades y Relaciones

Elementos que integran un almacén de datos

Metadato también cuenta con

- **Meta-datos de los datos Operacionales**
 - Historia de los datos (migraciones, transformaciones),
 - Estado de los datos (activo, archivados, o limpiado),
 - Información de seguimiento (estadísticas de uso de almacén, informes de errores, trazas de auditoría)
- **Los algoritmos utilizados para resumir**
- **El mapeo del entorno operativo al almacén de datos**
- **Los datos relacionados con el rendimiento del sistema**
- ...

Elementos que integran un almacén de datos

Middleware

- Ofrece servicios a aplicaciones distribuidas que requieren usar el DW.
- Su función es asegurar la conectividad entre todos los componentes de la arquitectura de un almacén de datos (datawarehouse) para brindar servicios.

Funciones ETL, OLAP

12 reglas de un Data Warehouse

1. Data Warehouse y los entornos operativos estan separados
2. Los datos se integran
3. Contiene datos históricos durante un largo período de tiempo
4. Datos son capturados en un punto dado en el tiempo
5. Datos son orientados a temas

12 reglas de un Data Warehouse

6. Principalmente es usado sólo de lectura, con actualizaciones periódicas por lotes
7. Ciclo de Vida de Desarrollo sigue enfoque dirigido por los datos frente al enfoque tradicional basado en procesos
8. Datos contiene varios niveles de detalle: actual, viejo, ligeramente resumido, muy resumido

12 reglas de un Data Warehouse

9. Medio Ambiente se caracteriza por transacciones de sólo lectura a conjuntos de datos muy grandes
10. Sistema rastrea las fuentes de los datos, las transformaciones y el almacenamiento
11. Los metadatos son un componente crítico
12. Contiene mecanismos para el uso óptimo de los recursos

Aplicaciones de Data warehouse

- **Procesamiento analítico**

análisis multidimensional de datos en el almacén de datos
apoya las operaciones básicas de OLAP,

- **Minería de datos**

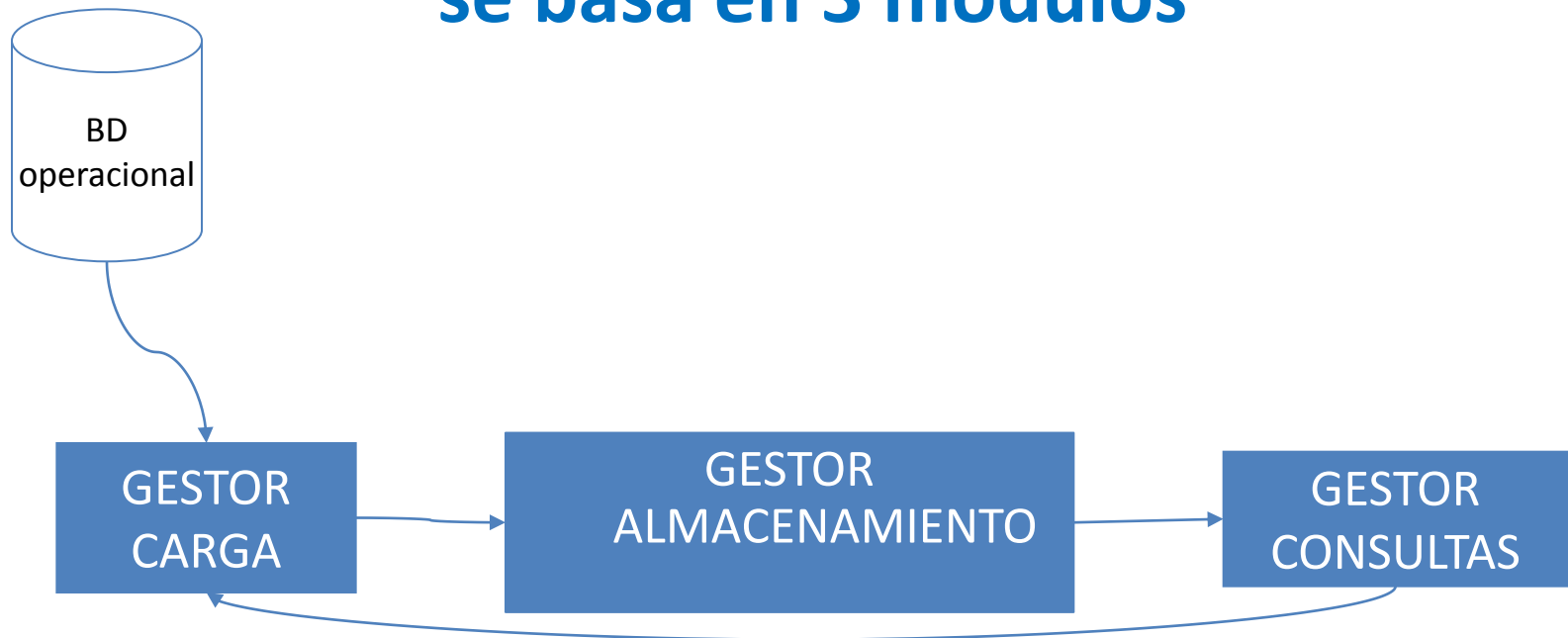
descubrimiento de conocimiento a partir de patrones ocultos:
asociaciones, clasificación, predicción.

- **Visualización de la información**

apoya la consulta utilizando tablas cruzadas, tablas, cuadros,
gráficos, etc.

ARQUITECTURA PLANA DW

se basa en 3 módulos



GESTOR DE CARGA

- Permite hacer la carga.
- Dificultades:
 - La integración de los datos
 - La elección del momento de la carga
 - Minimizar el tiempo de carga
 - Buen diccionario de datos o METADATA (para evitar cometer errores en la carga)
 - Diseño de procedimientos SQL

GESTOR DE ALMACENAMIENTO

- Se encarga del almacenamiento, de la estructura,....
- Existe una tabla llamada FACT (Hechos) y unas tablas llamadas dimensiones o tablas dimensionales.
 - Entre la tabla FACT y las tablas dimensionales suele haber relaciones 1:N
 - Este modelo tiene forma de estrella, por eso se denomina MODELO ESTRELLA

Ejemplo 1:

CATEGORIAS_PROF

c_categoria
* descripción
* salario_min
* salario_max

PERSONAS

c_persona
* nombre
* apellido1
* apellido2
* c_categoria
* direccion

dimensiones

PROF_OFERTAS

c_persona
c_categoria
c_oferta
c_trabajo
* salario_deseado

Hechos

TRABAJO

c_trabajo
* descripcion

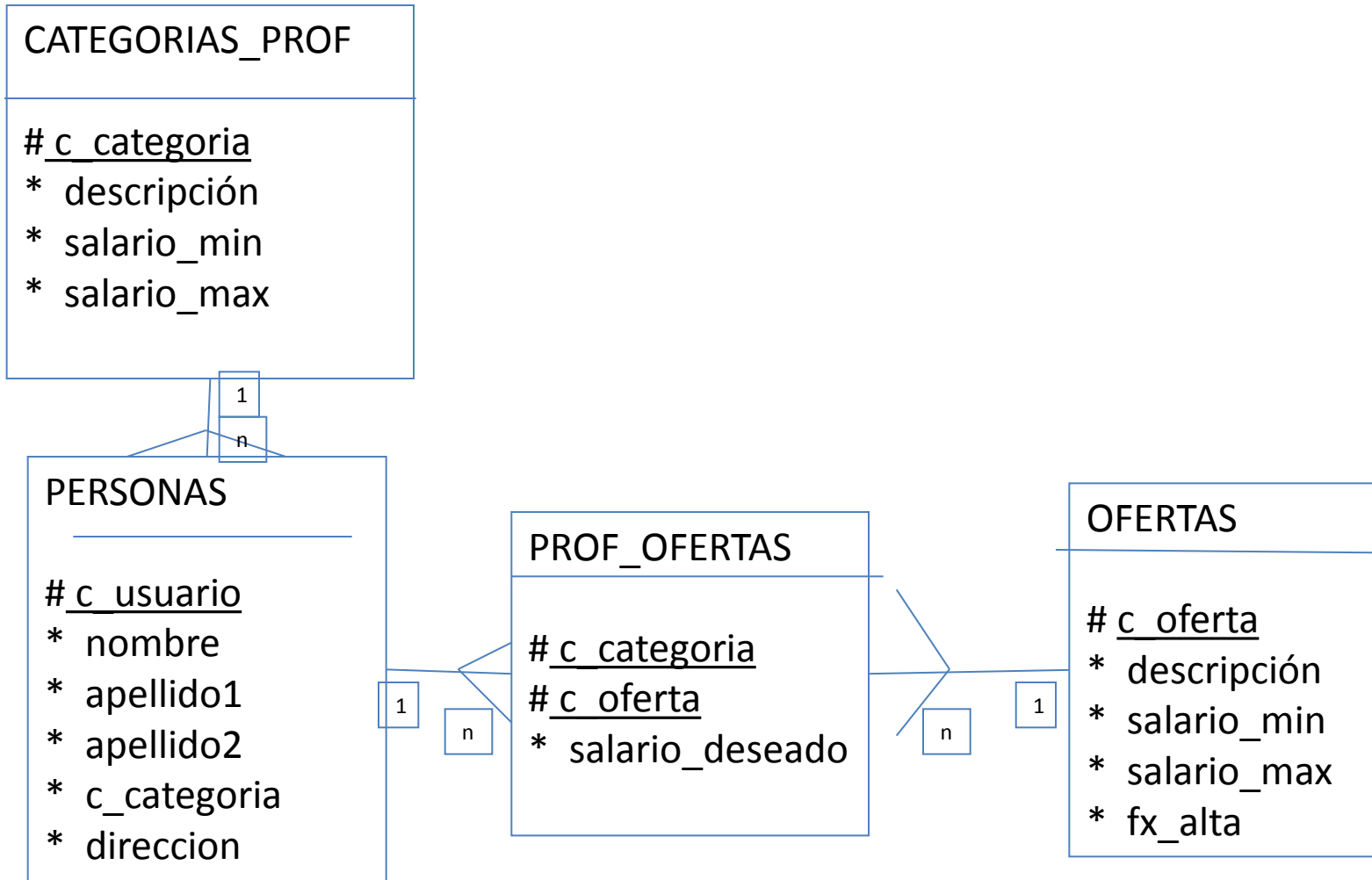
OFERTAS

c_oferta
* descripción
* salario

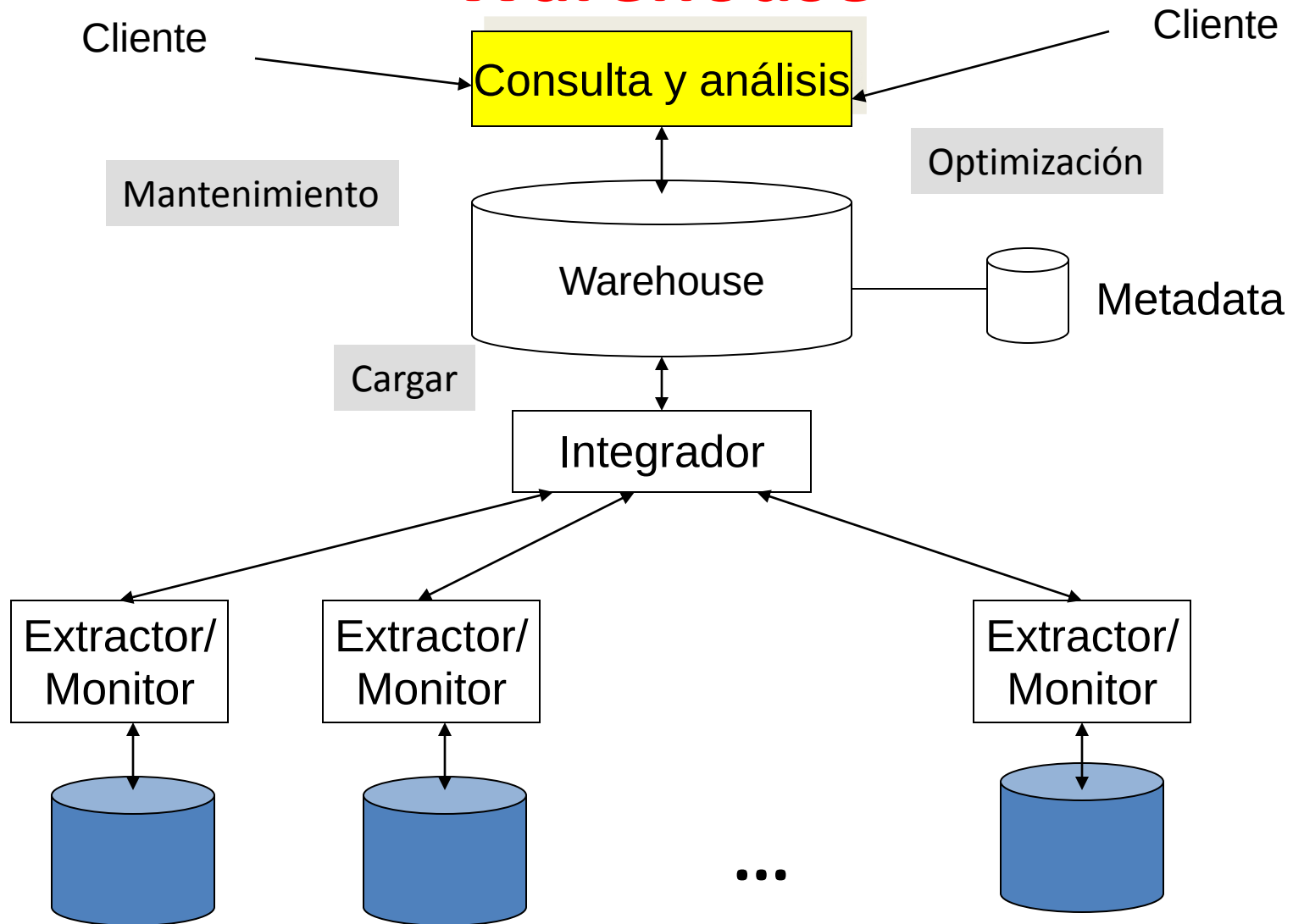
GESTOR DE CONSULTAS

- Las consultas se hacen sobre la tabla FACT.
- También define perfiles de consultas, pues ellas son diferentes dependiendo del usuario y sus necesidades.

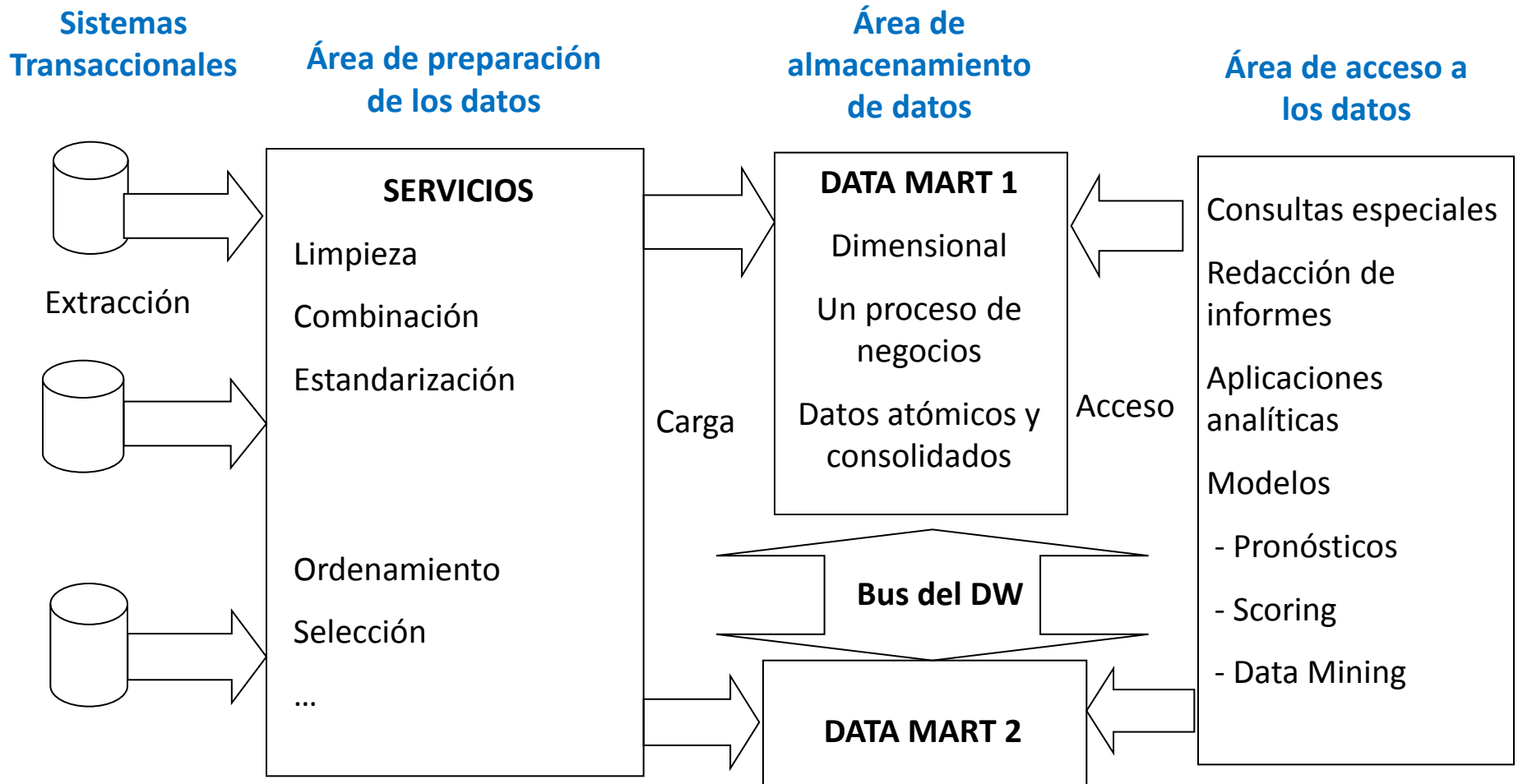
Ejemplo 1:



Arquitectura genérica Data Warehouse



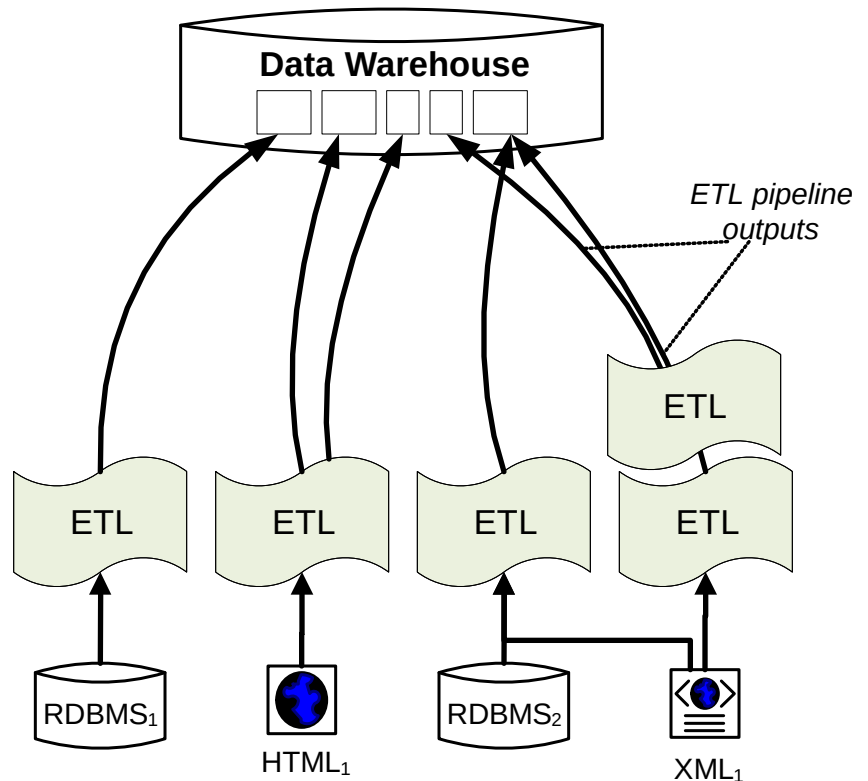
Arquitectura genérica Data Warehouse con Data Mart



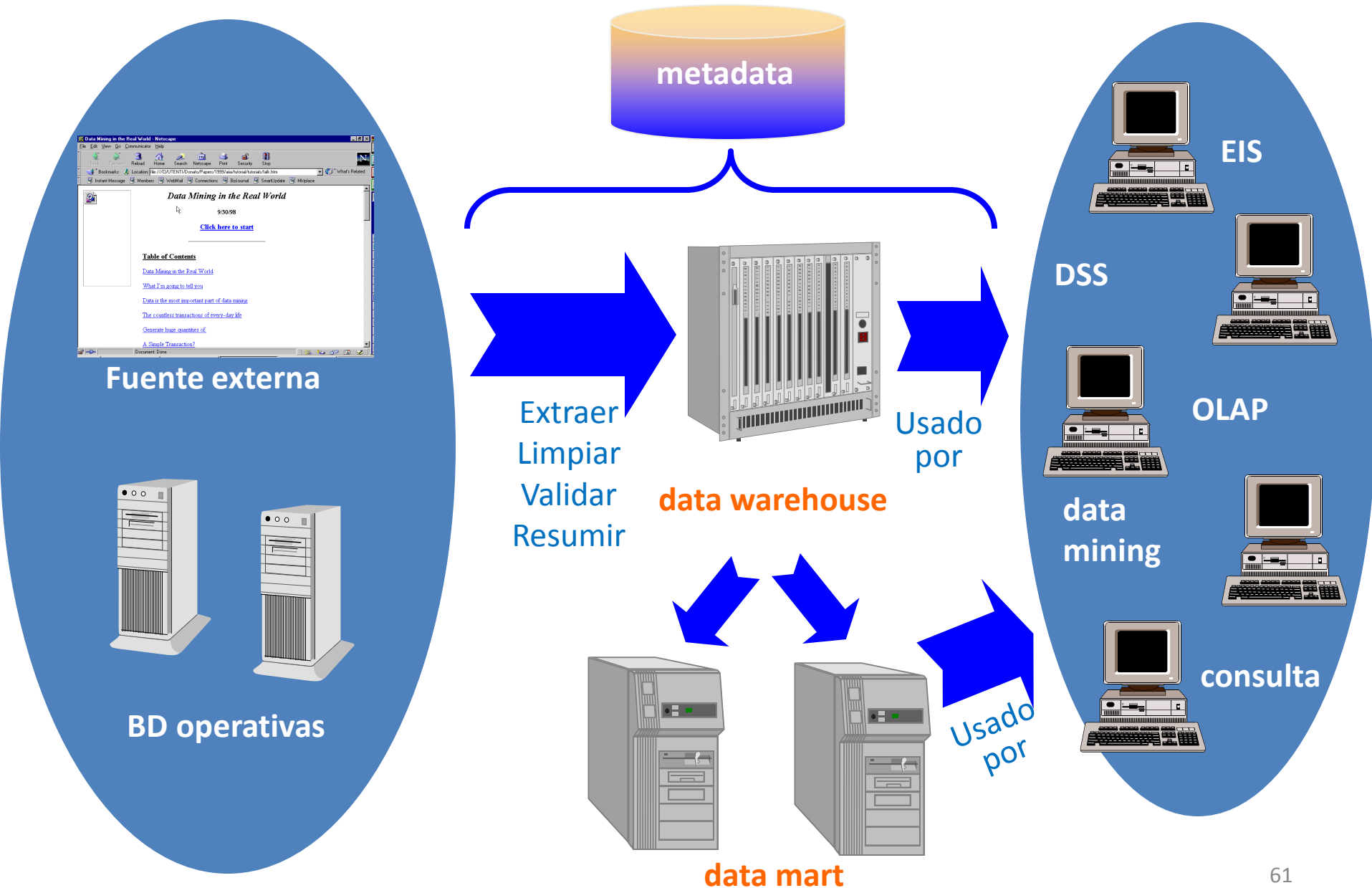
Arquitectura Data Warehouse

primer modelo

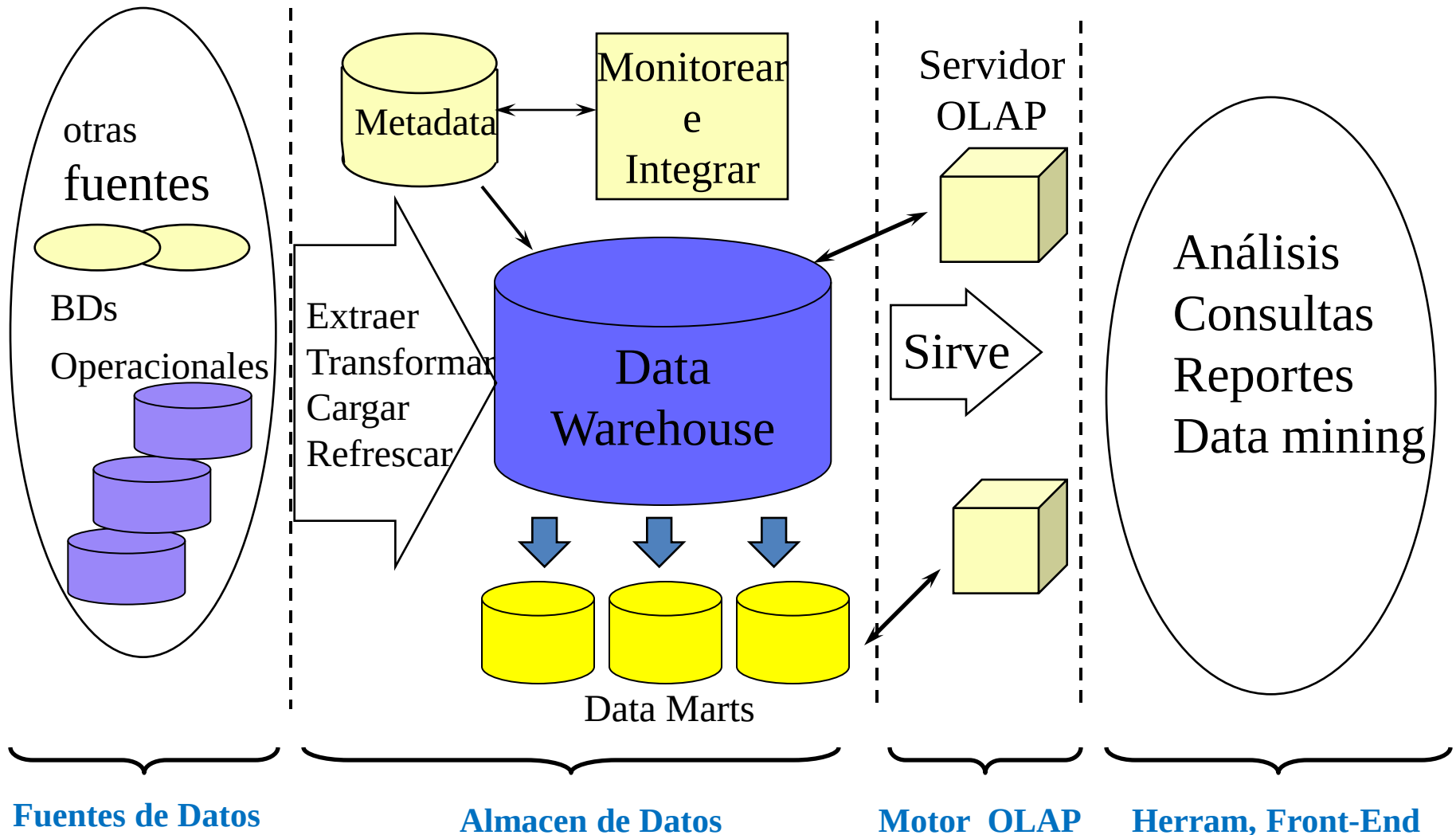
- En la parte superior una base de datos centralizada
 - Generalmente configurado para consultas y agregar – no transacciones
 - Muchos índices, vistas , etc.
- Los datos se cargan y actualizan periódicamente a través de ETL (Extract / Transform / Load)



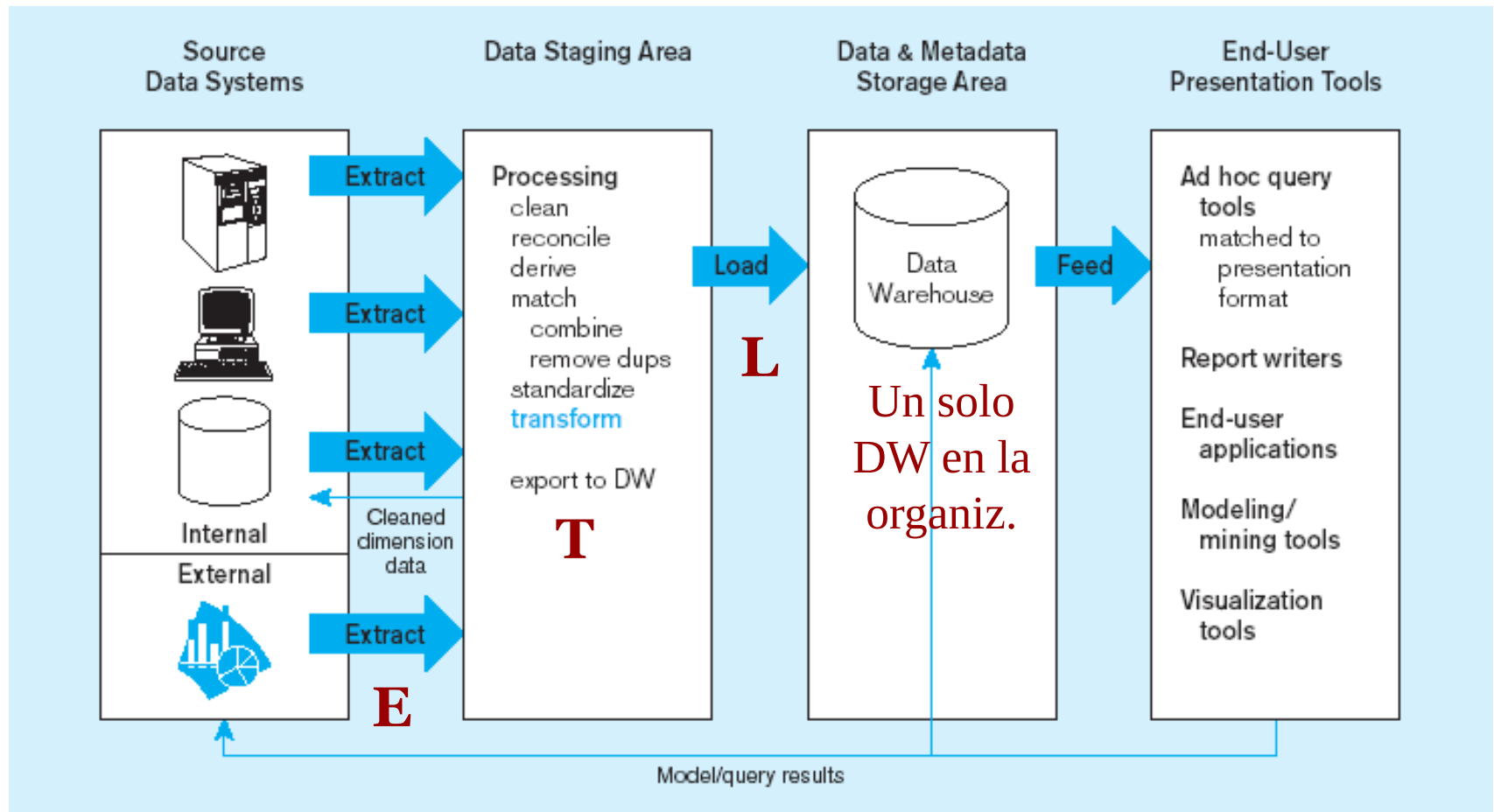
Arquitectura varios niveles



Arquitectura varios niveles



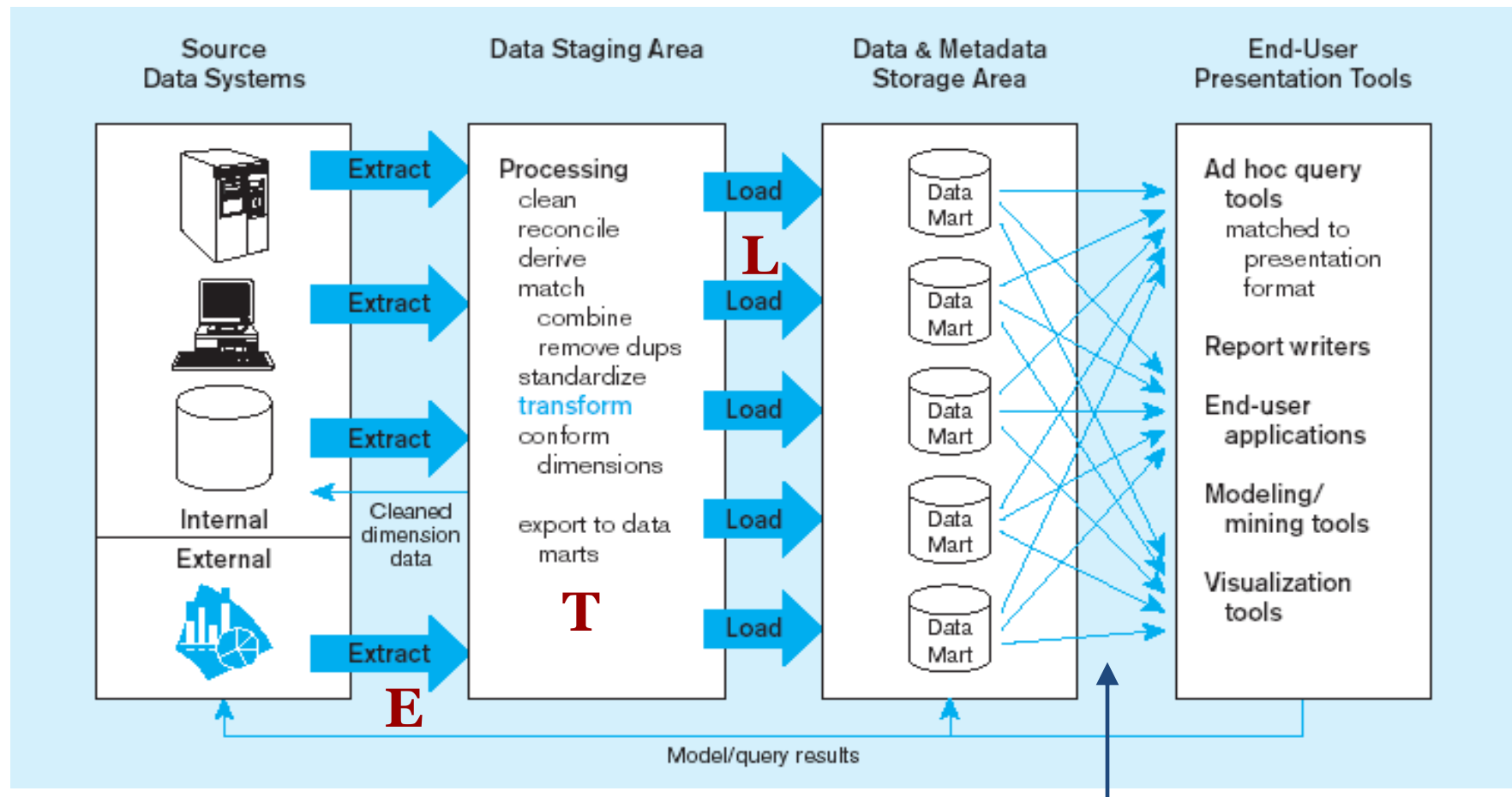
Genérica arquitectura de niveles de DW



Centralizado

Extracción de datos periódica
Datos no están completamente al día

Arquitectura independientes Data mart

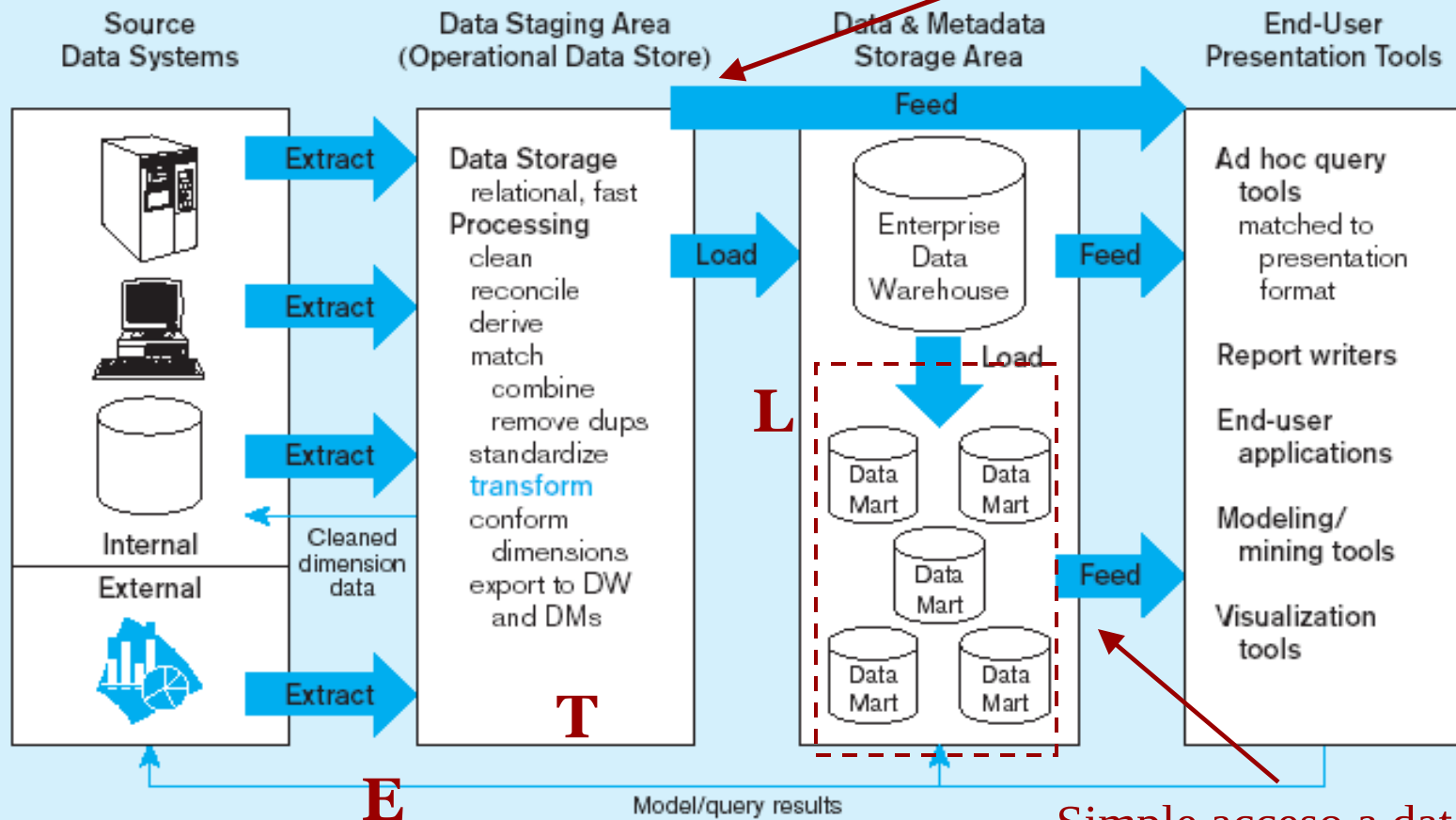


Separado ETL para c/data mart

Acceso datos por
*multiple*s data marts

Arquitectura Híbrida

Se pueden obtener
actuales datos



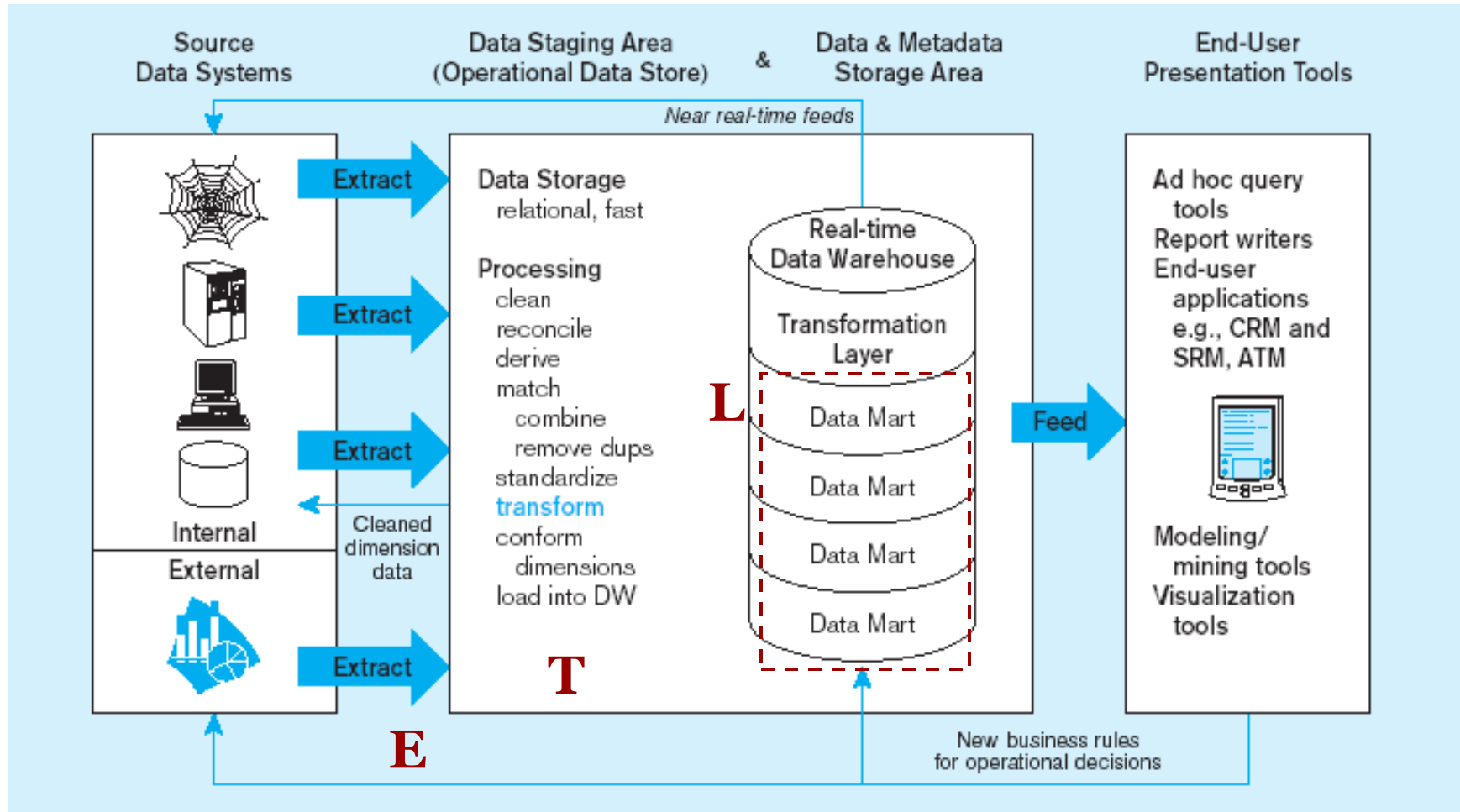
Simple acceso a datos

Un solo ETL para
enterprise data warehouse
(EDW)

Data marts *dependientes*
del EDW

Federado

Data mart Lógico y DW-RT



ETL RT para
Data Warehouse

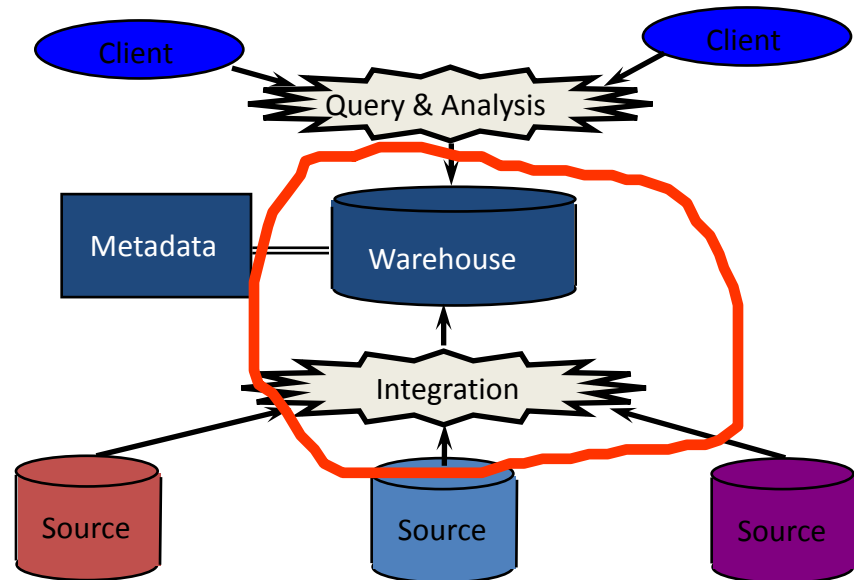
Data marts no separado de BDs, vistas lógicas de los datos de data warehouse
➔ Fácil crear nuevos data marts

Datawarehousing: ETL

Jose Aguilar
CEMISID, Escuela de Sistemas
Facultad de Ingeniería
Universidad de Los Andes
Mérida, Venezuela

Integración

- Selección de los datos
- Transformación de datos
- Carga de datos



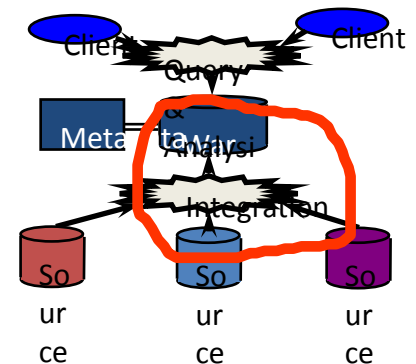
Proceso ETL

ETL (Extracción, Transformación y Carga)

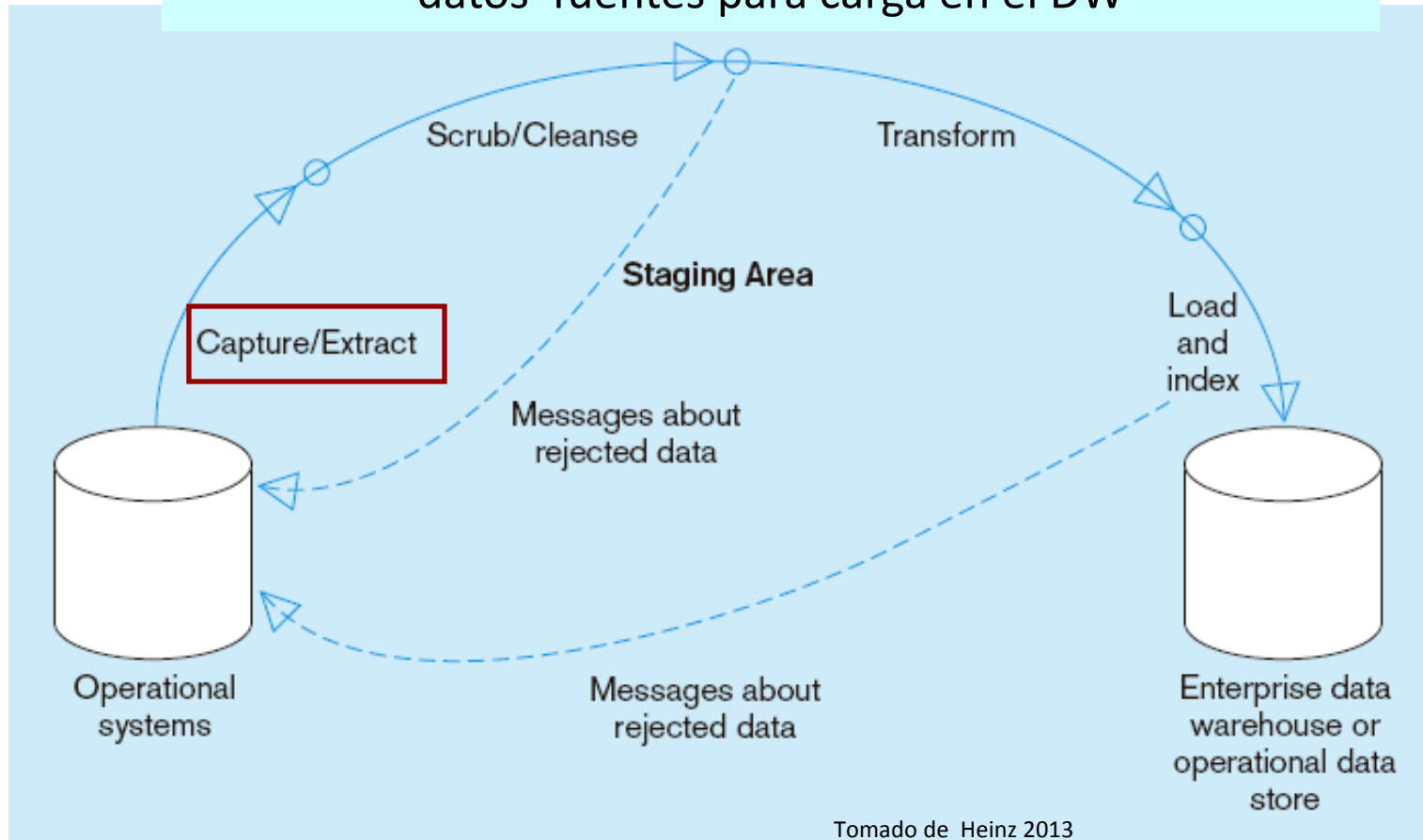
Extracción: Obtención de información de las distintas fuentes, tanto internas como externas.

Transformación: Filtrado, limpieza, depuración, homogeneización y agrupación de la información.

Carga: Organización y actualización de los datos y los metadatos en el DW.



Captura / Extrae...... obtiene un subconjunto de los datos fuentes para carga en el DW



Extracción estática: captura los datos en un momento puntual

Extracción Incremental: captura cambios que se van produciendo

Extracción

- Obtener datos de múltiples, heterogéneas fuentes externas
- Periódica
- Claves:
 - Manipular los datos sin interrumpir ni paralizar los OLTP, ni tampoco el DW.
 - No depender de la disponibilidad de los OLTP.
 - Almacenar y gestionar los metadatos que se generarán en los procesos ETL.
 - Facilitar la integración de las diversas fuentes, internas y externas.

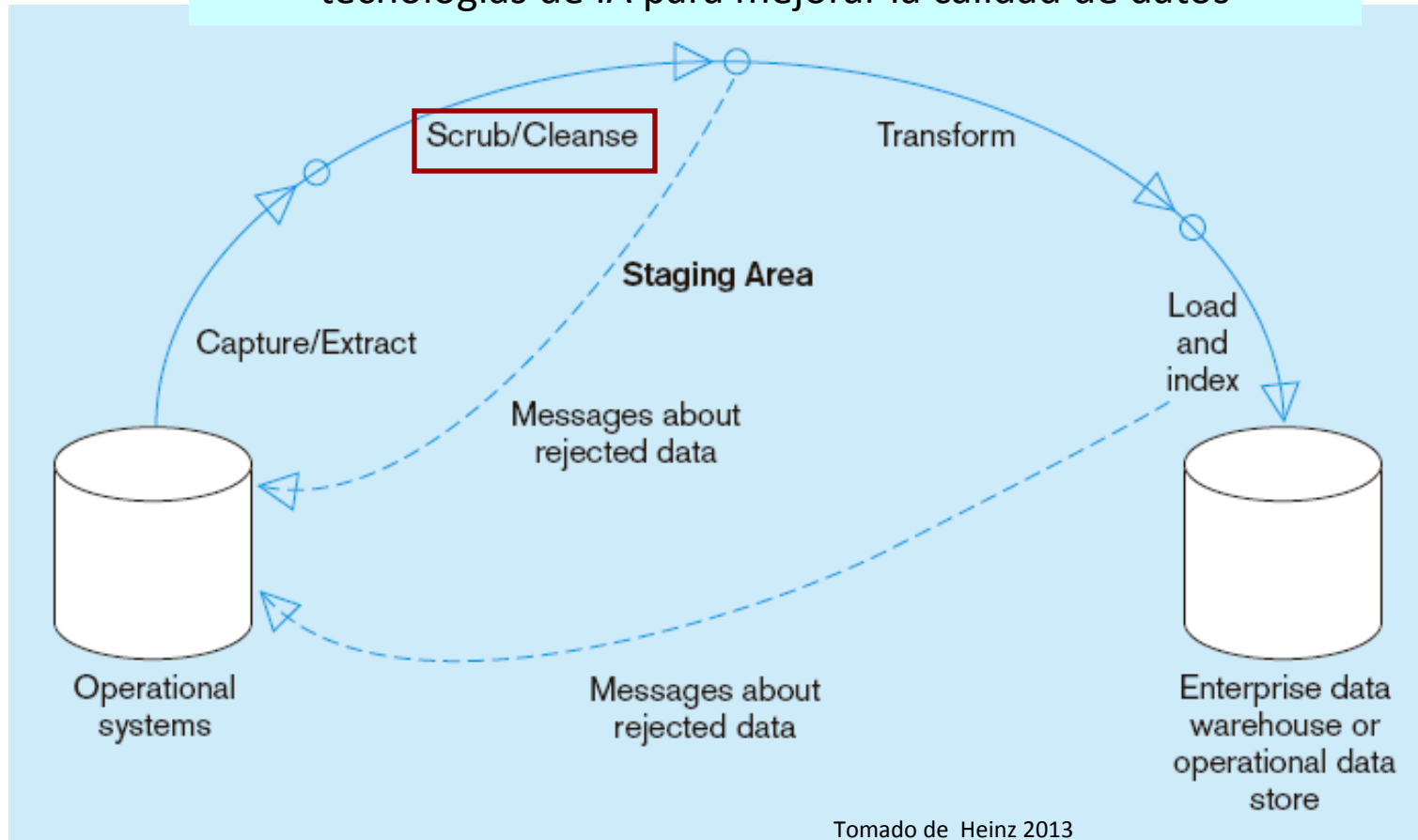
ETL: Técnicas de Monitoreo

- Instantáneas cada cierto tiempo
- Disparadores de base de datos (reglas (triggers))
- Logs de registros
- Envío de datos
- Envío de transacciones
- Consultas a las BDs

Métodos de Extracción

- **Extracciones a granel**
 - Todo el DW se actualiza periódicamente
 - Pesado para las conexiones de red entre el origen y destino
 - Más fácil de configurar y mantener
- **Extracciones basadas en intercambio**
 - Sólo los datos que han sido recién insertadas o actualizadas en los sistemas de origen se cargan en el DW
 - Menos carga en la red pero requiere una programación más compleja para determinar cuando un nuevo registro de DW debe insertarse o cuando un registro DW tiene que actualizarse

Limpieza... utiliza reconocimiento de patrones y tecnologías de IA para mejorar la calidad de datos



Solución de errores: faltas de ortografía, fechas erróneas, uso de campo incorrecto direcciones no coinciden, datos faltantes, datos duplicados, inconsistencias

También decodifica, reformatea, convierte, genera claves, fusiona, detecta errores registro, localiza datos faltantes

Limpieza de datos

Sistemas de origen tiene "datos sucios" que deben ser limpiados

Razones de datos "sucios"

- valores ficticios
- La falta de datos
- datos encriptados
- datos contradiciéndose
- El uso inapropiado de datos
- Violación de Reglas de Negocio
- Claves principales reutilizados
- Identificadores no únicos

detectar errores
en los datos y
rectificarlos

Pasos en la Limpieza de datos

1. Análisis (parsing)
2. Corrección
3. Estandarización
4. Mapeo (matching)
5. Consolidación

Análisis (Parsing)

Localiza e identifica **elementos individuales** en los archivos de origen y luego los aísla.

- **Ejemplos**

- Análisis del primer nombre, segundo nombre y apellido;
- Analizar número y nombre de la calle;
- Analizar la ciudad y el estado.

Corregir

Corrige los componentes de **datos individuales** utilizando algoritmos sofisticados y fuentes de datos secundarias.

Acciones típicas con Datos Anómalos (Outliers):

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.

Acciones contra Datos Faltantes (Missing Values):

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.
- Esperar hasta que los datos faltantes estén disponibles.

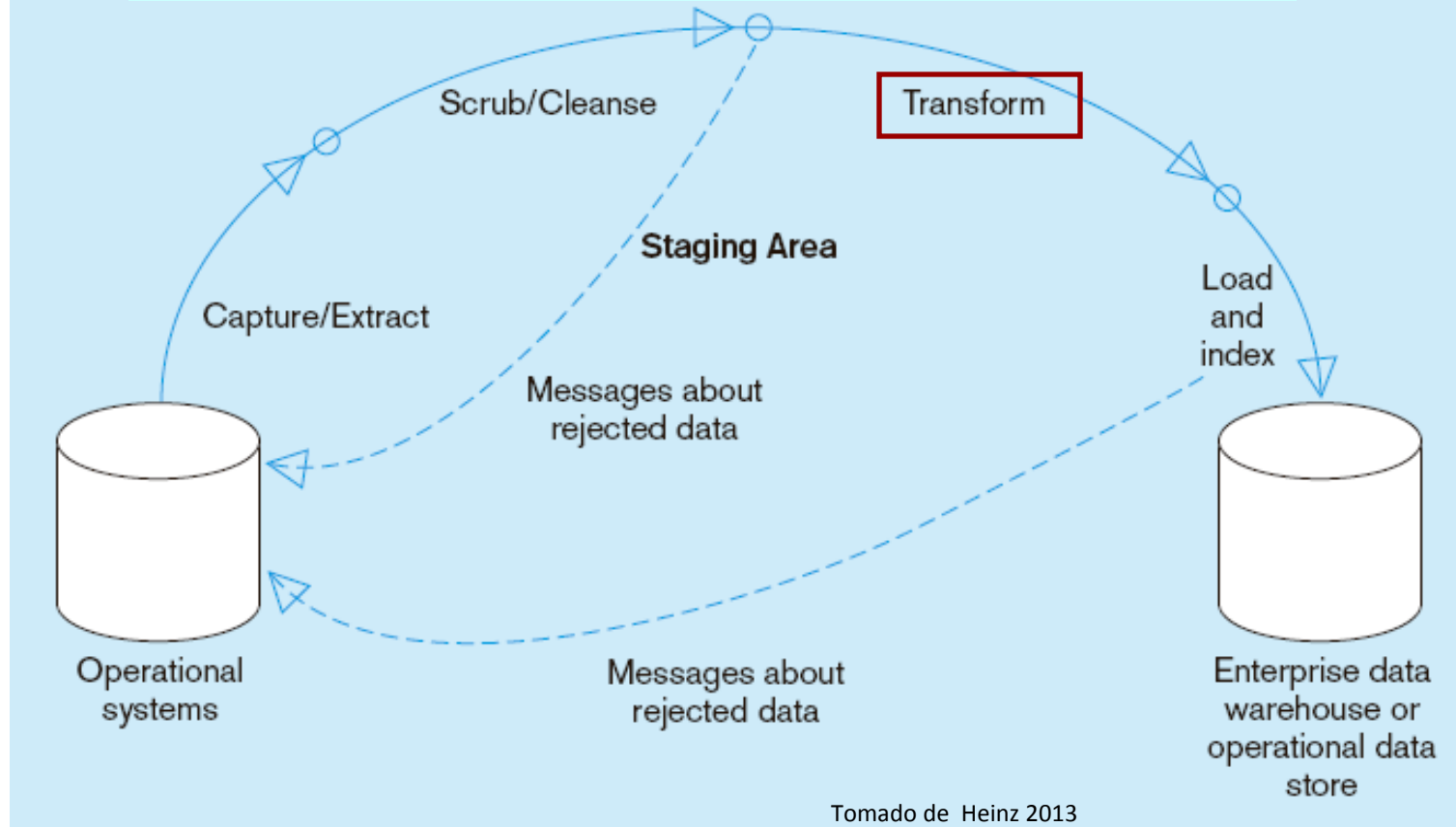
Estandarización

- Aplica **rutinas de conversión** para transformar los datos en formatos preferidos (y coherentes), utilizando reglas de estandarización.
- **Ejemplos:** la sustitución de un apodo, uso de un nombre de calle preferido, etc.

Mapeo (matching)

- búsqueda de **coincidentes en registros** en los datos analizados, corregidos y estandarizados, basados en reglas predefinidas para eliminar la duplicación.
- **Ejemplos:** identificación de nombres y direcciones similares.

Transforma... convierten los datos desde las BD operacionales a formato del **DW**



A nivel de registro:

- Partición de Datos (selección)
- Juntar datos (combinación)
- Resumir datos (agregación)

A nivel de campo:

- de un solo campo: de un campo a un campo
- multi-campo: de muchos campos a uno, o uno a muchos campos

Transformación de datos

- Transforma los datos de acuerdo con reglas y normas establecidas
- Clásicos problemas:
 - Codificación.
 - Medida de atributos.
 - Convenciones de nombramiento.
 - Fuentes múltiples,

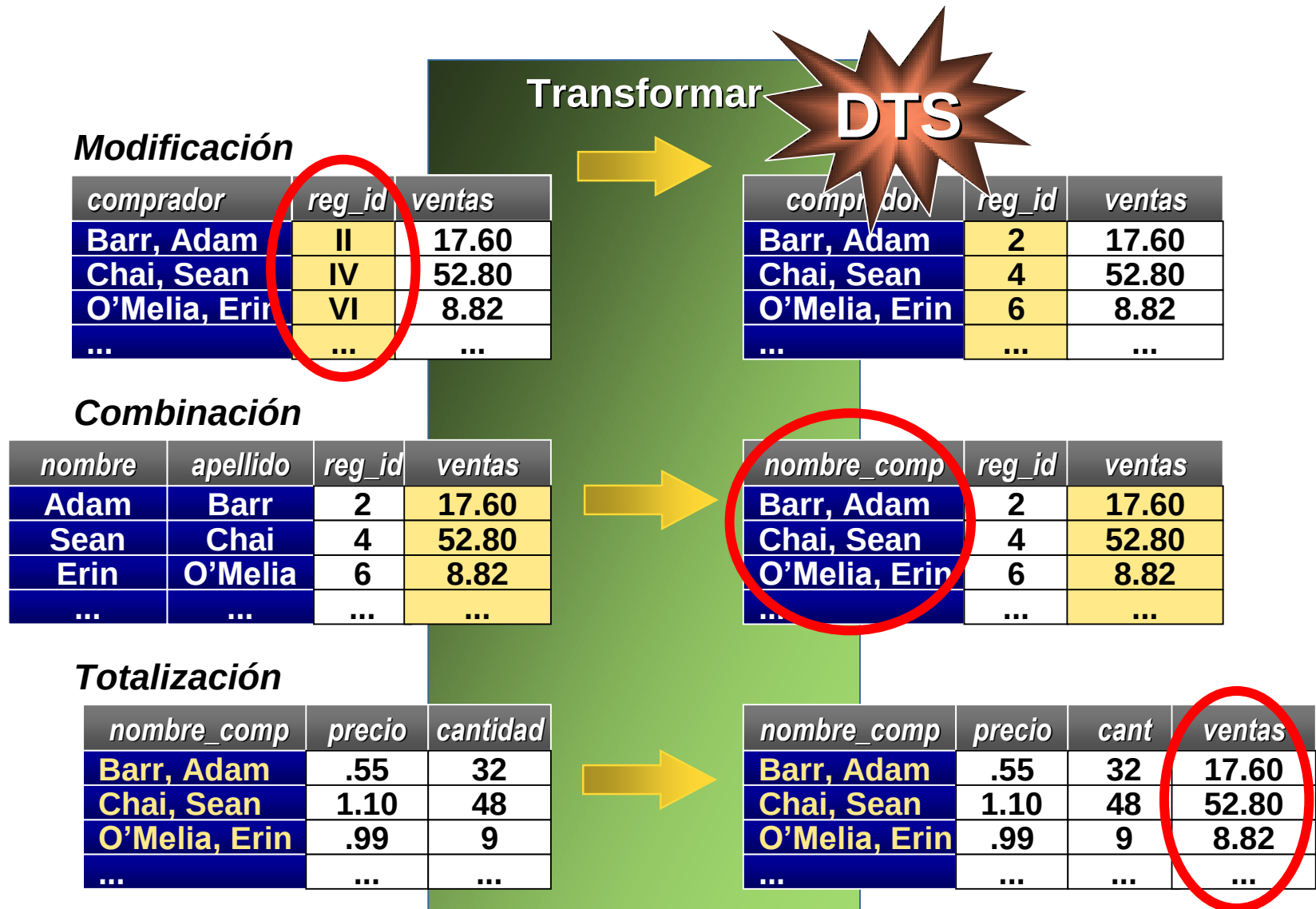
ordena

calcula

resume

consolida

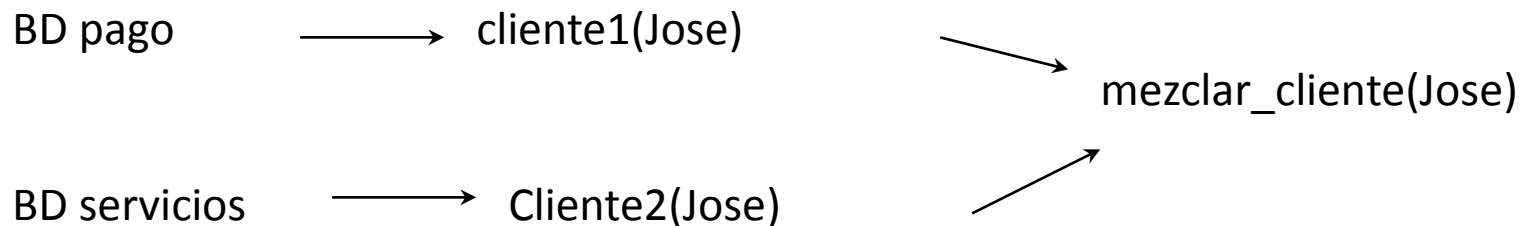
Proceso ETL; Consideraciones



ETL: Transformación de datos

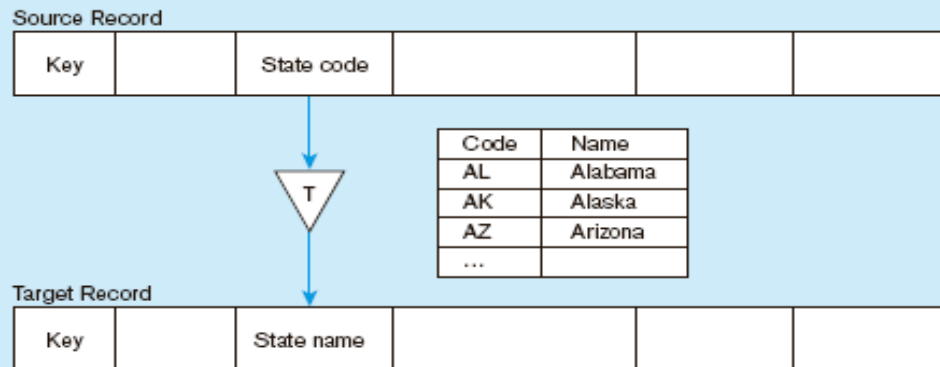
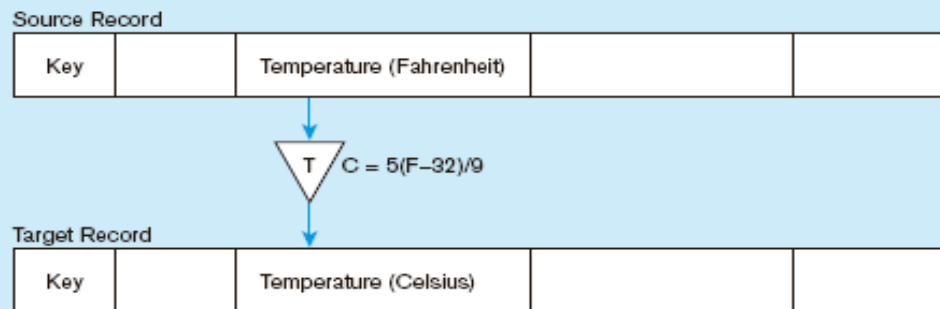
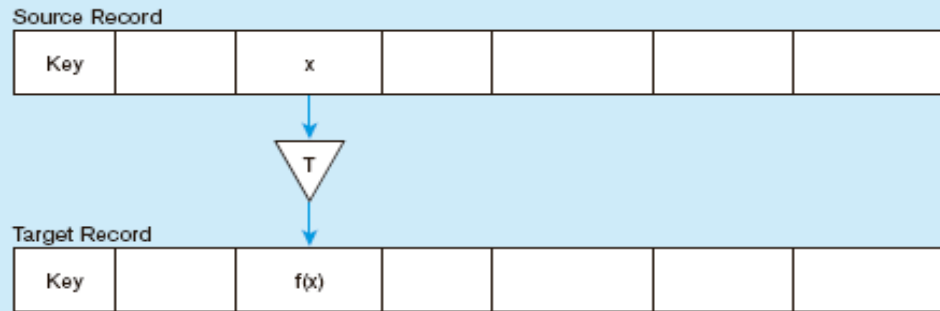
Posibles tareas

- **Migrar** (por ejemplo, yen a dólares)
- **Refinar**: utilizar el conocimiento específico de dominio (por ejemplo, números de seguro social)
- **Fusionar** (por ejemplo, lista de correo con la de clientes)



Transformación de un solo campo

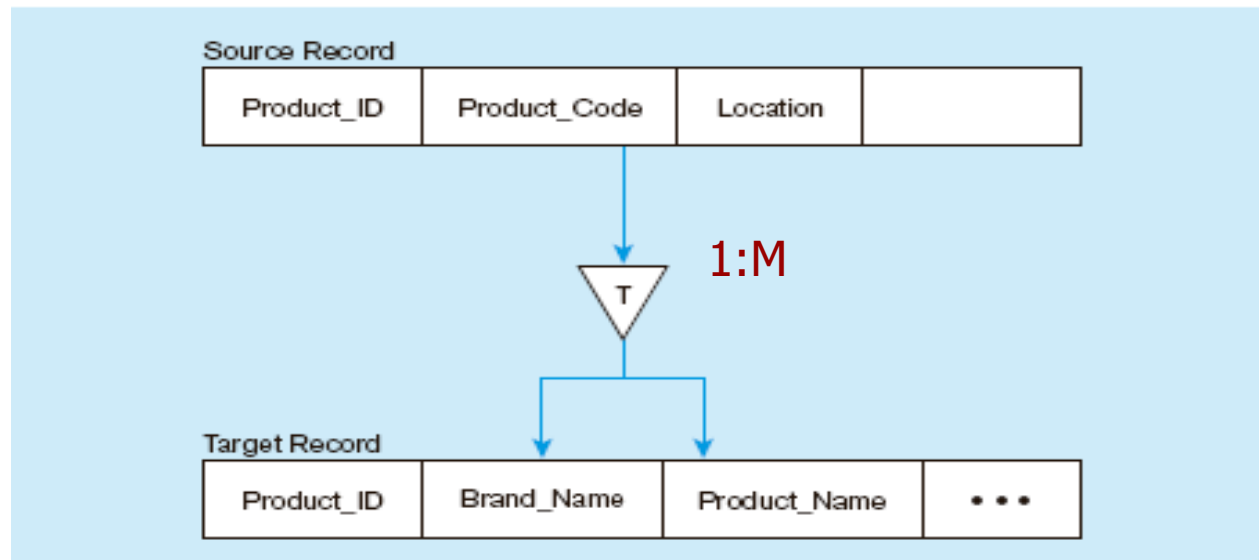
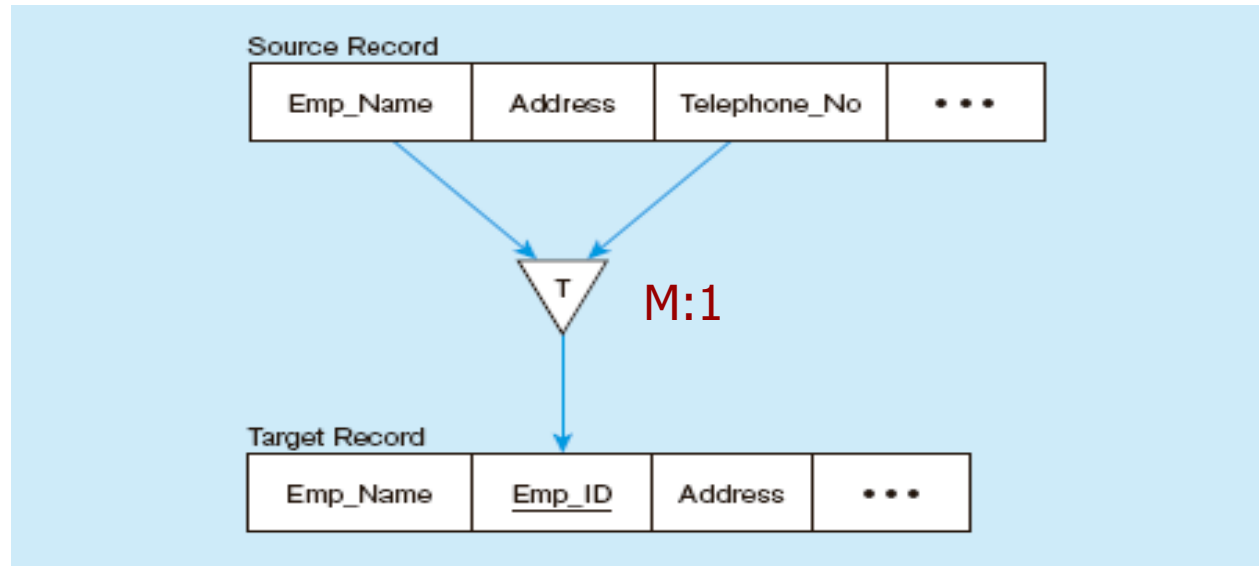
Una función de transformación traduce los datos de una forma antigua a una nueva



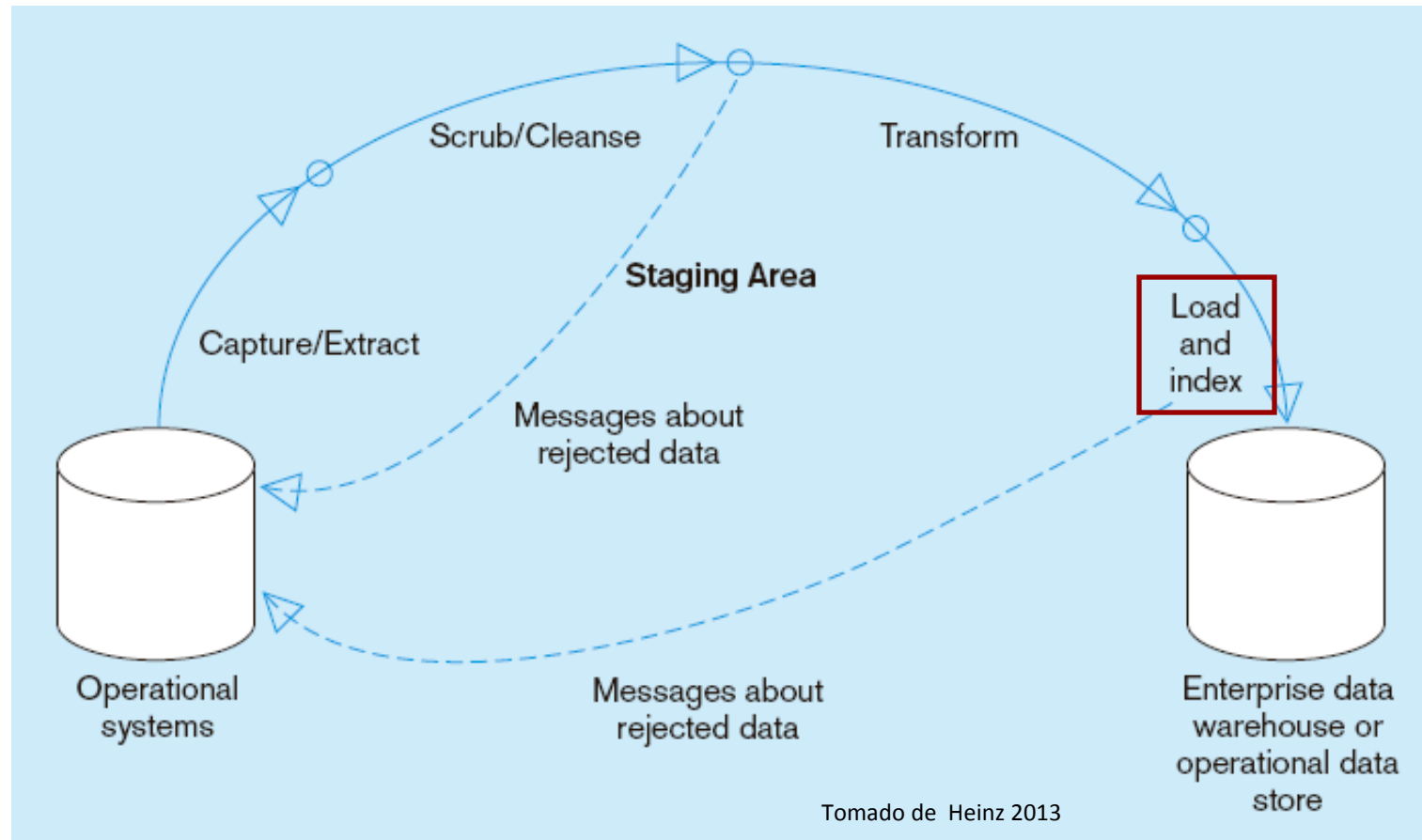
Transformación algorítmica: utiliza una fórmula o expresión lógica

Table lookup: utiliza una tabla separada basada en códigos

Transformación multicampos



Cargar/Indizar... Transforma datos y crea índices



Modo de Actualización 1:
reescritura masiva de datos en
destino a intervalos periódicos

Modo de actualización 2:
solamente cambios en los
datos de origen se escriben en
el warehouse

Carga de Datos

- Los datos físicamente se almacena en el almacén de datos
- La carga ocurre en una "ventana de carga"
- La tendencia cada vez mayor es actualizaciones en tiempo real