Datawarehousing

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

Introducción a Data Warehouse

Ejemplos de preguntas de los gerentes ....

- ✓ Quién compró? Dónde y cómo compraron?
- ✓ Cuáles son las características de mis clientes ?
- ✓ Qué está requiriendo mi cliente ?
- ✓ Que región nos provee mayores ventas ?
- ✓ Qué segmento de clientes respondió mejor a la reciente promoción?
- ✓Qué productos llegaron con retraso y por qué?
- ✓ En qué periodo se vendió más ? Como vendimos ?
- ✓ Cuáles han sido los valores históricos de mis activos, de mi patrimonio, de mis ventas?

Introducción a Data Warehouse

Frustración de los gerentes debido a ...

- Desfase entre disponibilidad de información y oportunidad comercial
- Usuarios no tienen acceso autónomo a la información
- Muchos datos pero no suficiente información
- Los reportes no coinciden
- Poco tiempo para análisis de la información
- No se soporta una visión global de la empresa
- No se dispone de análisis histórico de información
- No pueden hacerse preguntas complejas del negocio

Introducción a Data Warehouse

Analizando la información del negocio

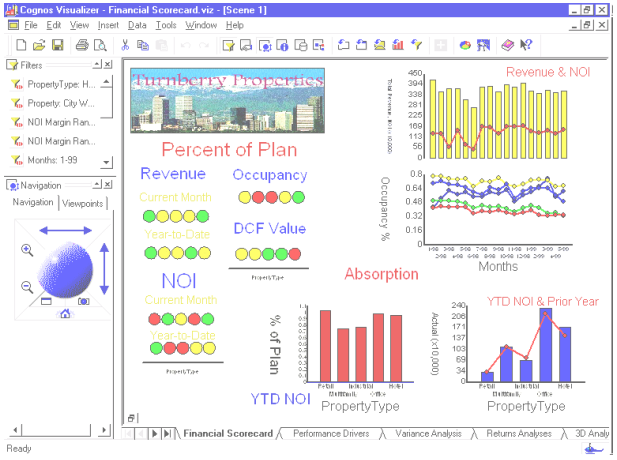
- ✓ Información periódica de las ventas
- ✓ Información del esfuerzo comercial
- ✓ Información sobre los pedidos a los proveedores

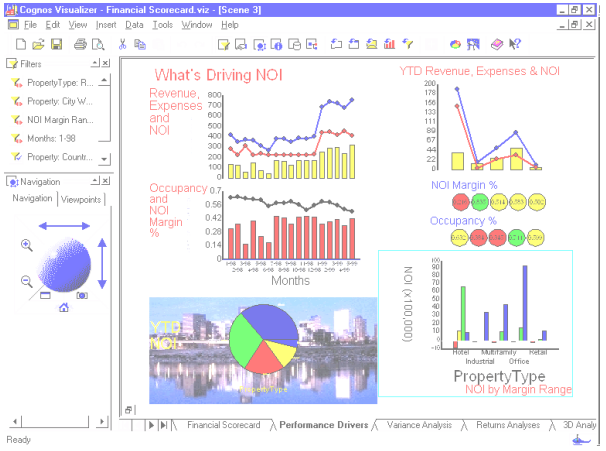
Por qué no integrarla y cruzarla para obtener:

- ✓ 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 ?

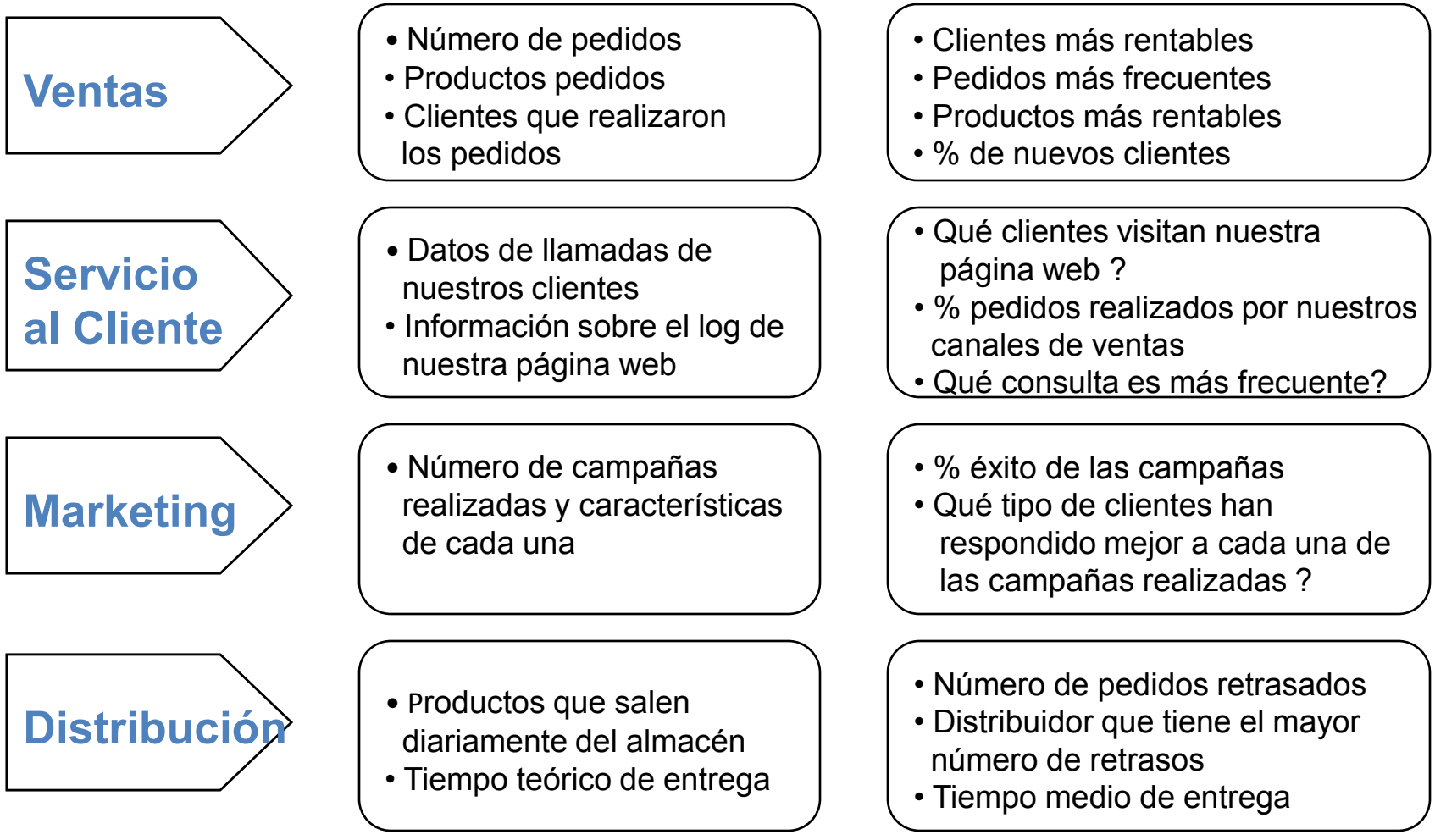
Introducción a Data Warehouse

- ✓ 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



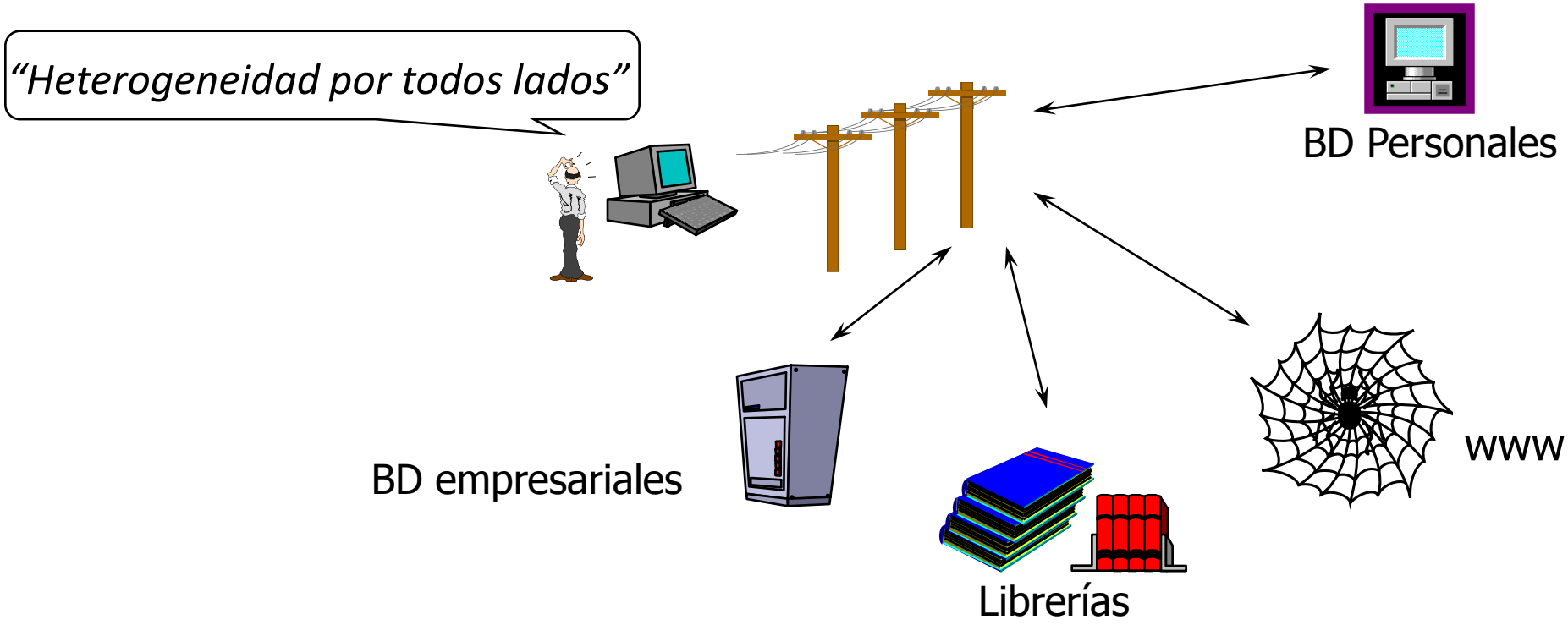
Introducción a Data Warehouse

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

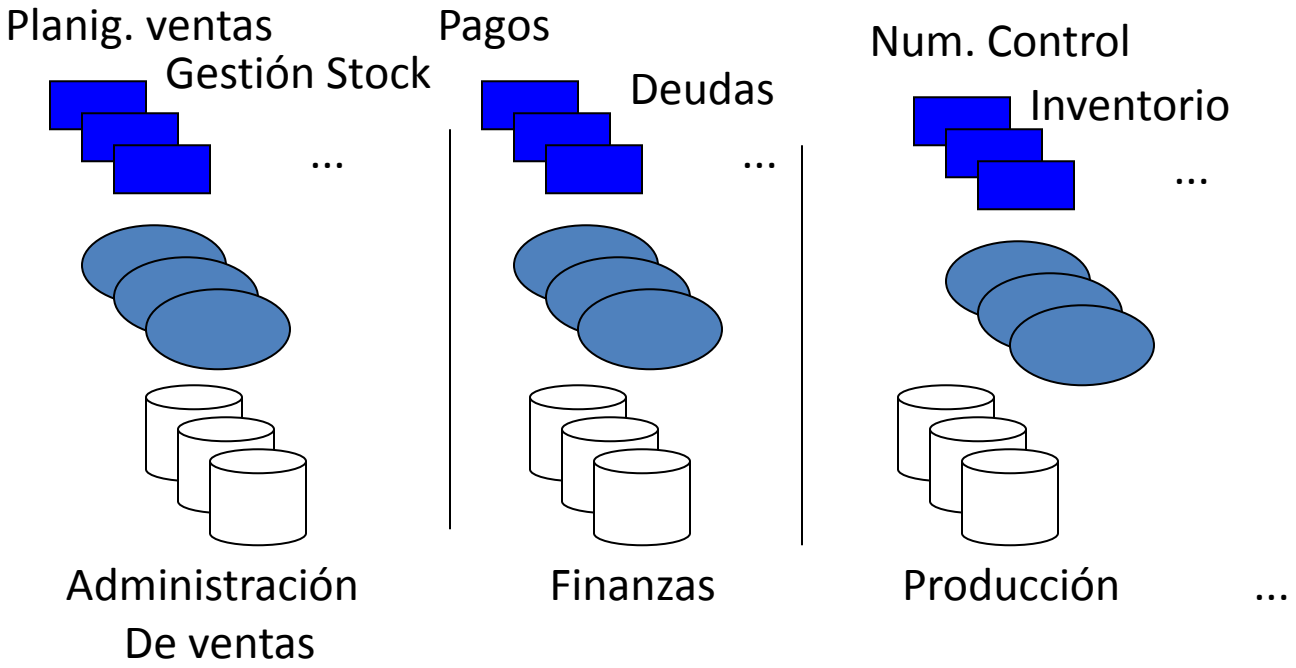
Problemas: Heterogeneidad de las fuentes de Información



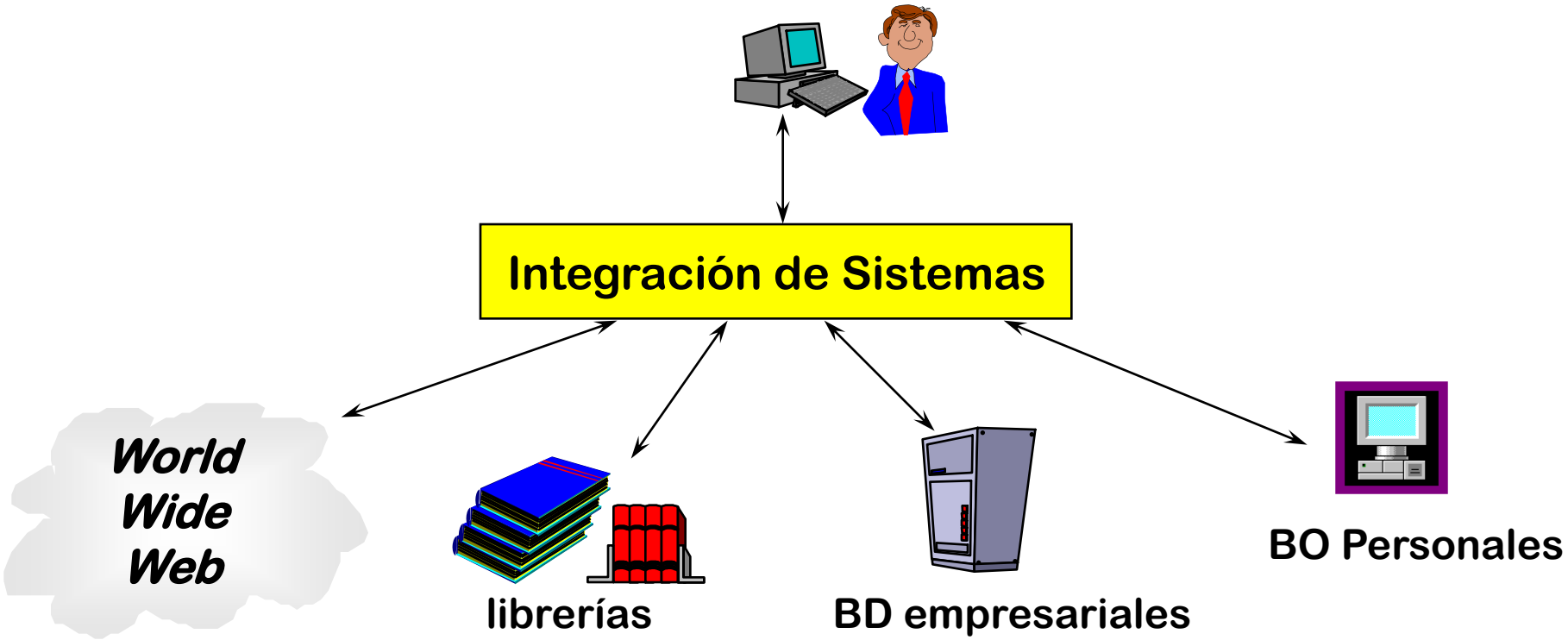
- Diferentes interfaces
- Diferentes representaciones de datos
- Información duplicar e inconsistente

Problema: Gestión de datos en grandes empresas

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



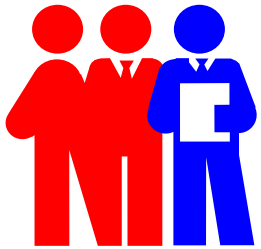
Objetivo: Unificar Acceso a los Datos

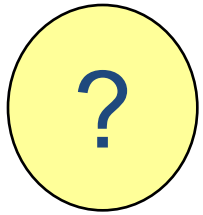


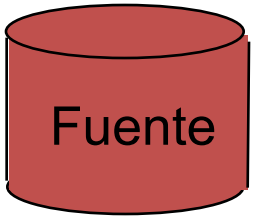
- Recopilar y combinar la información
- Proporcionar visión integrada, una interfaz de usuario uniforme
- Soportar el intercambio

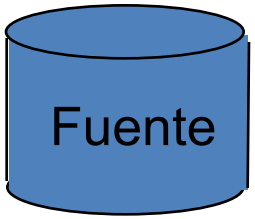
Objetivo: Unificar Acceso a los Datos

- Dos enfoques:
  - Guiador la consulta (perezoso)
  - Warehouse (ansioso)



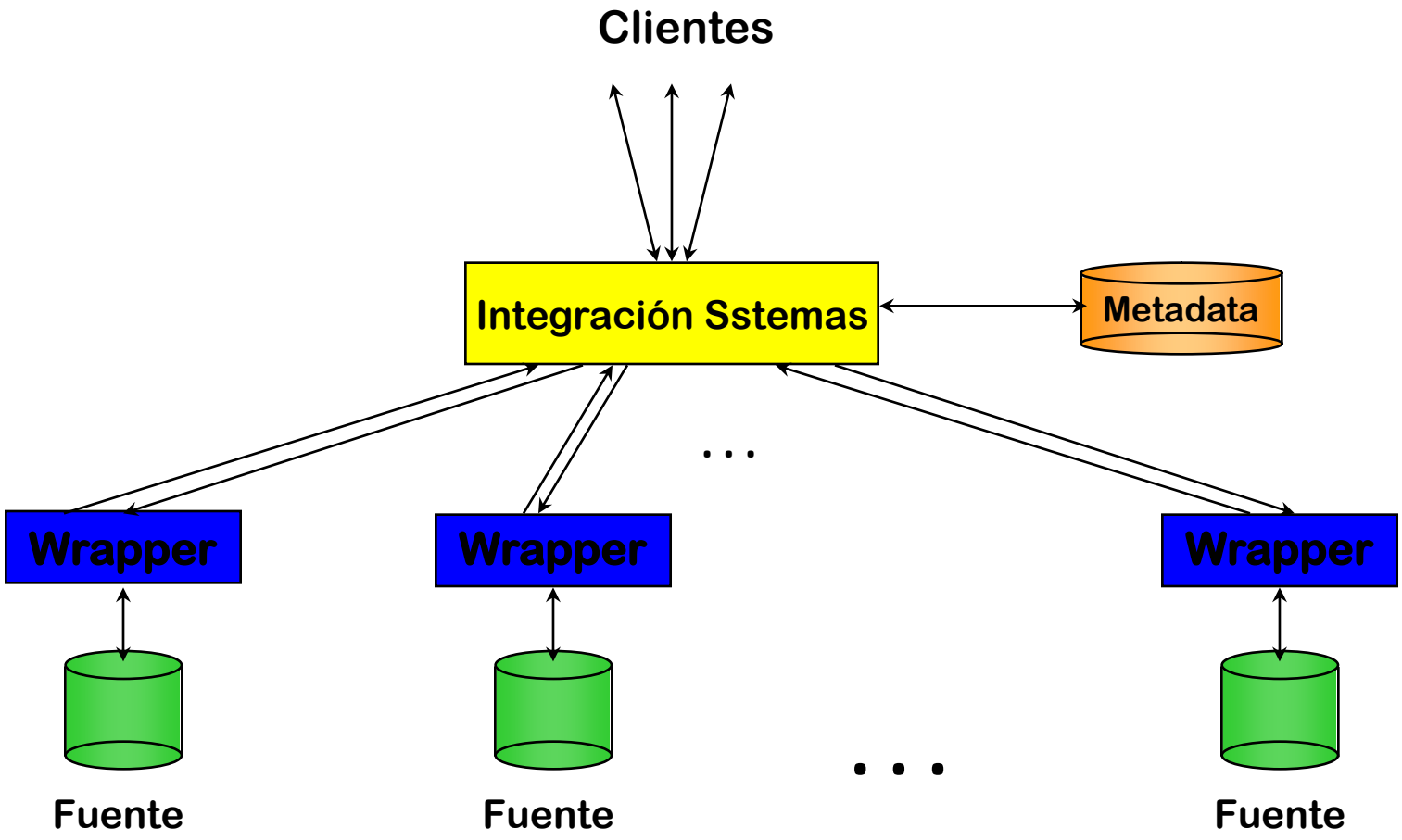






Enfoque tradicional

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

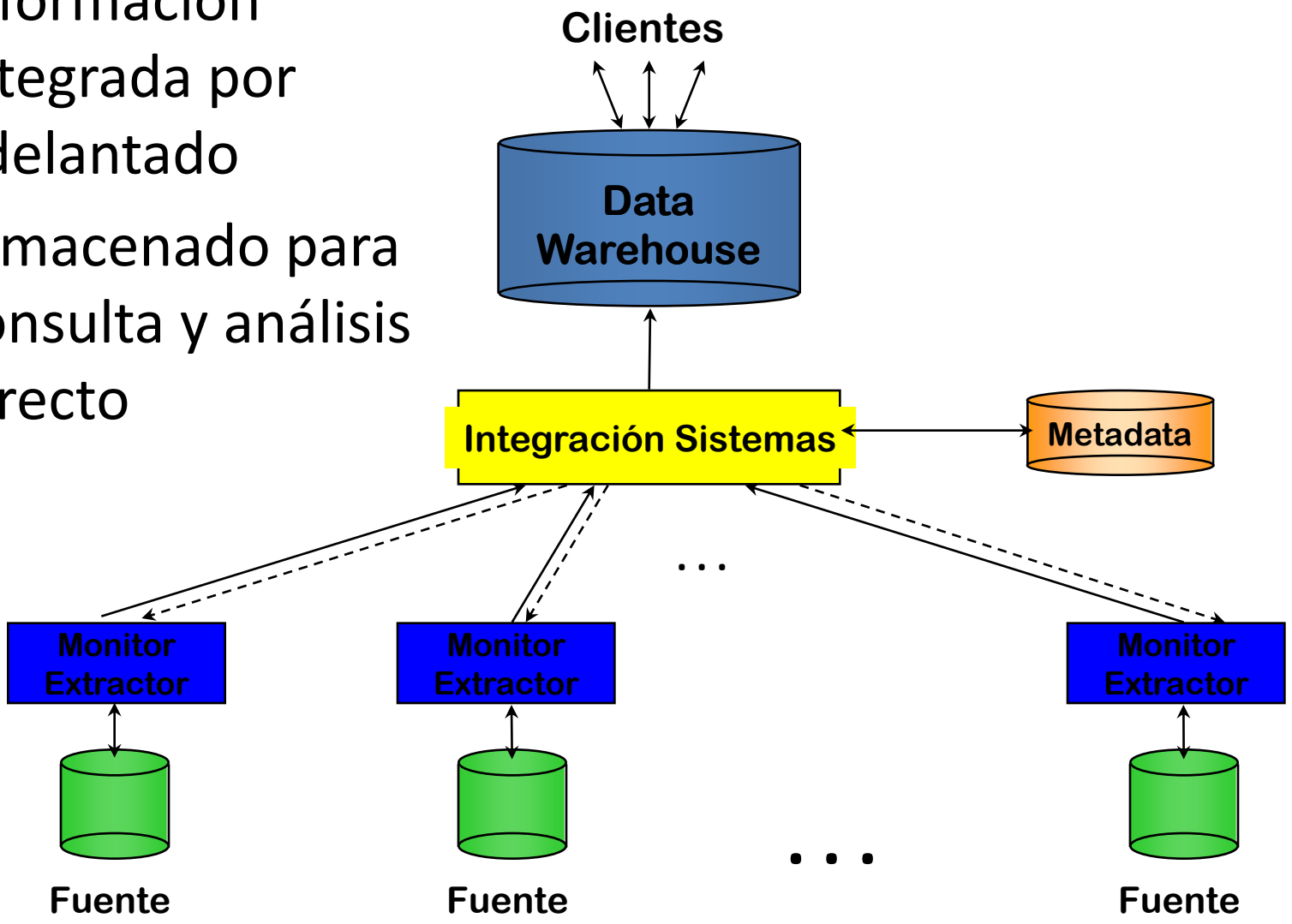


Problema Enfoque tradicional

- El retraso en el procesamiento de consultas
- Fuentes de información lentas o no disponibles
  - El filtrado y la integración es Compleja
  - 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

-Emplea 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 intefiera con el procesamiento y registro de nuevos 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 las fuentes.

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)

- Datawarehouse (Almacen de datos) se define como una colección de datos que verifican las siguientes propiedades:

  - Está orientado (a un tema) a objetos
  - 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 empresa, tales como

- Clientes
- Productos
- venta

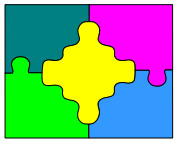
en lugar de las transacciones individuales.

La normalización no es relevante

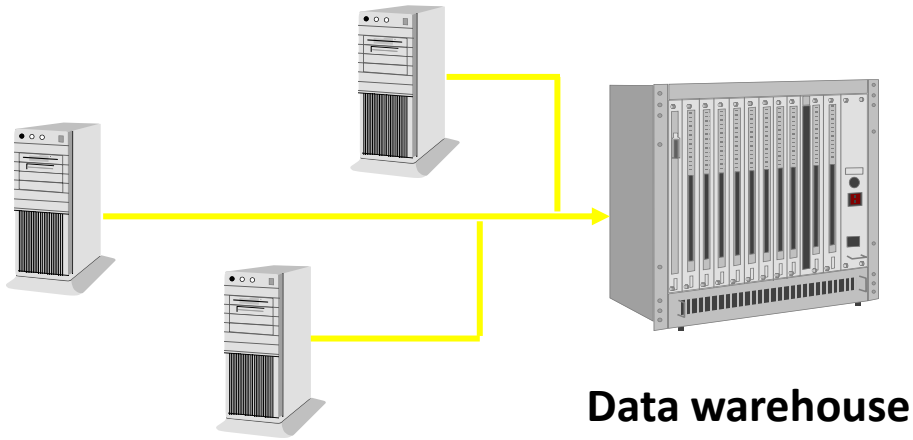
- Centrándose en el modelado y análisis de datos para la toma de decisiones, no en las operaciones diarias o el procesamiento de transacciones.

- 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.

Datos integrados



- el almacén de datos integra datos que provienen de varias fuentes. Partimos de una base de datos (operacional) y mediante un proceso de carga de datos hacemos el Datawarehouse.
- El proceso de carga es lo más complicado por problemas de codificación, medidas de los atributos… de las bases de datos.



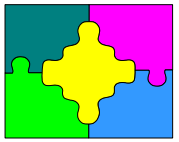
Sistemas Operacionales

Data warehouse

Los datos en el almacén deben ser:

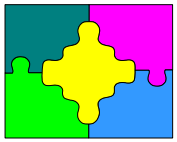
- Limpios
- Validados
- Adecuadamente integrados

Datos integrados



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

Ejemplo Datos integrados



← *Dato operational*

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

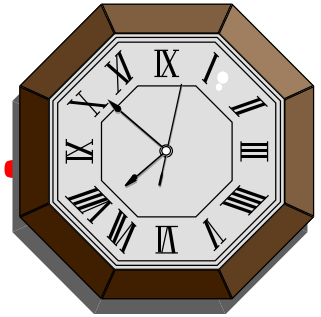
Sistema de cuentas de ahorro
Jane Doe
F (gender)
Opened account 1992

Sistema de inversiones de clientes
Jane Doe
Owns 25 Shares Exxon
Opened account 1995

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

↑ *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 al almacén o cuando se resume en el almacén.
- Esta grabación cronológica de datos ofrece posibilidades históricas y análisis de tendencias.
- Por el contrario, en las BD operativas los datos se sobrescribe, ya que los valores anteriores no son de interés.

TIEMPO

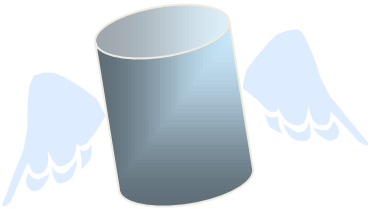
id_tiempo

* periodo

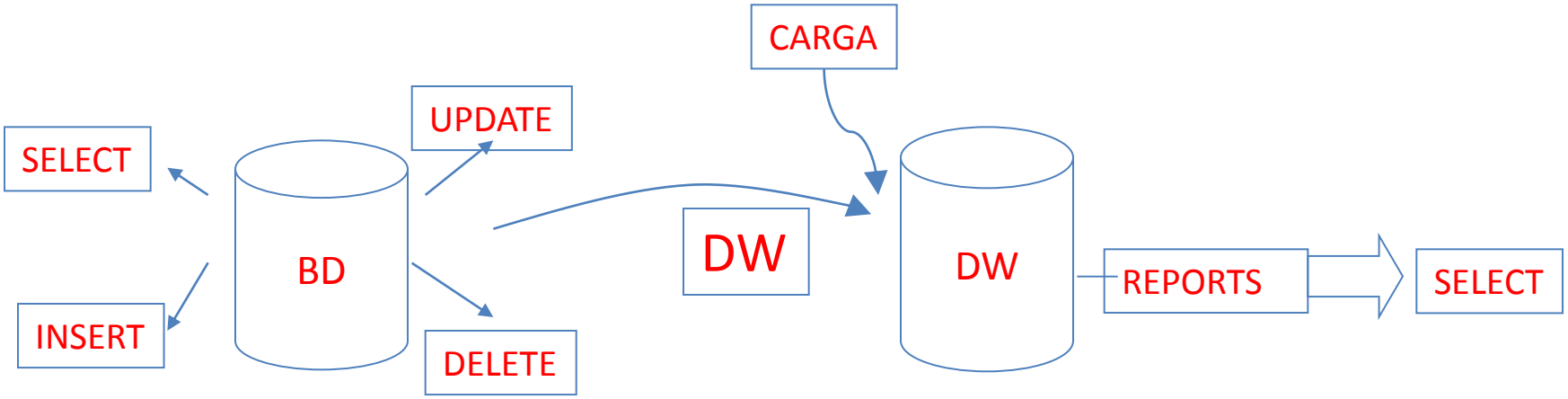
porque se hacen cargas de datos continuamente, cuando los datos van cambiando, se actualizan los históricos y se guardan en ficheros temporales.

la clave de los datos operativos puede o no puede contener "elemento de tiempo".

No volátiles



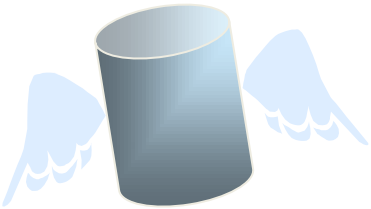
Son estables, una vez almacenados los datos no se modifican.



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

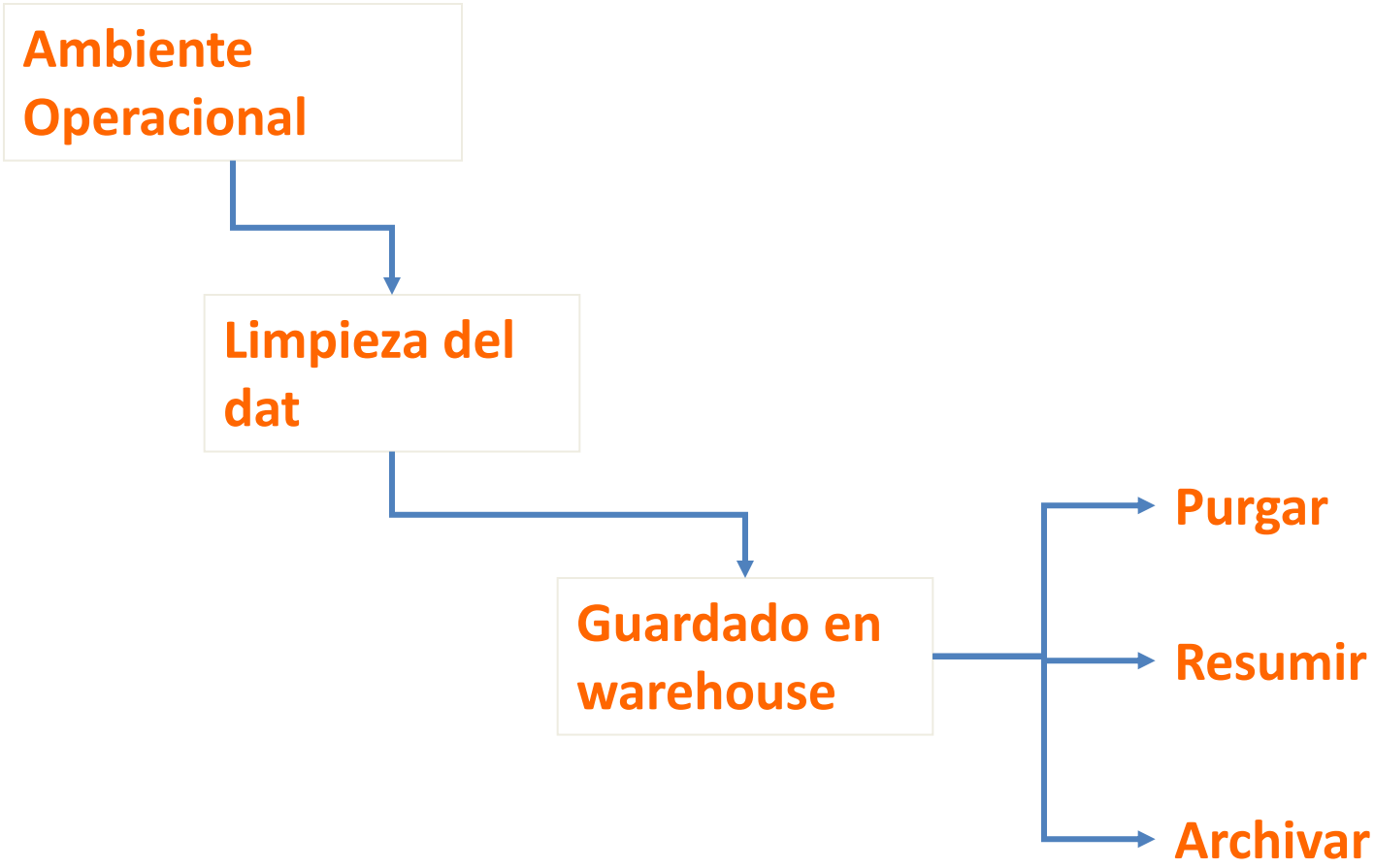
Por el contrario, los datos operativos se actualizan (insertar, eliminar, modificar, etc.)

No volátiles

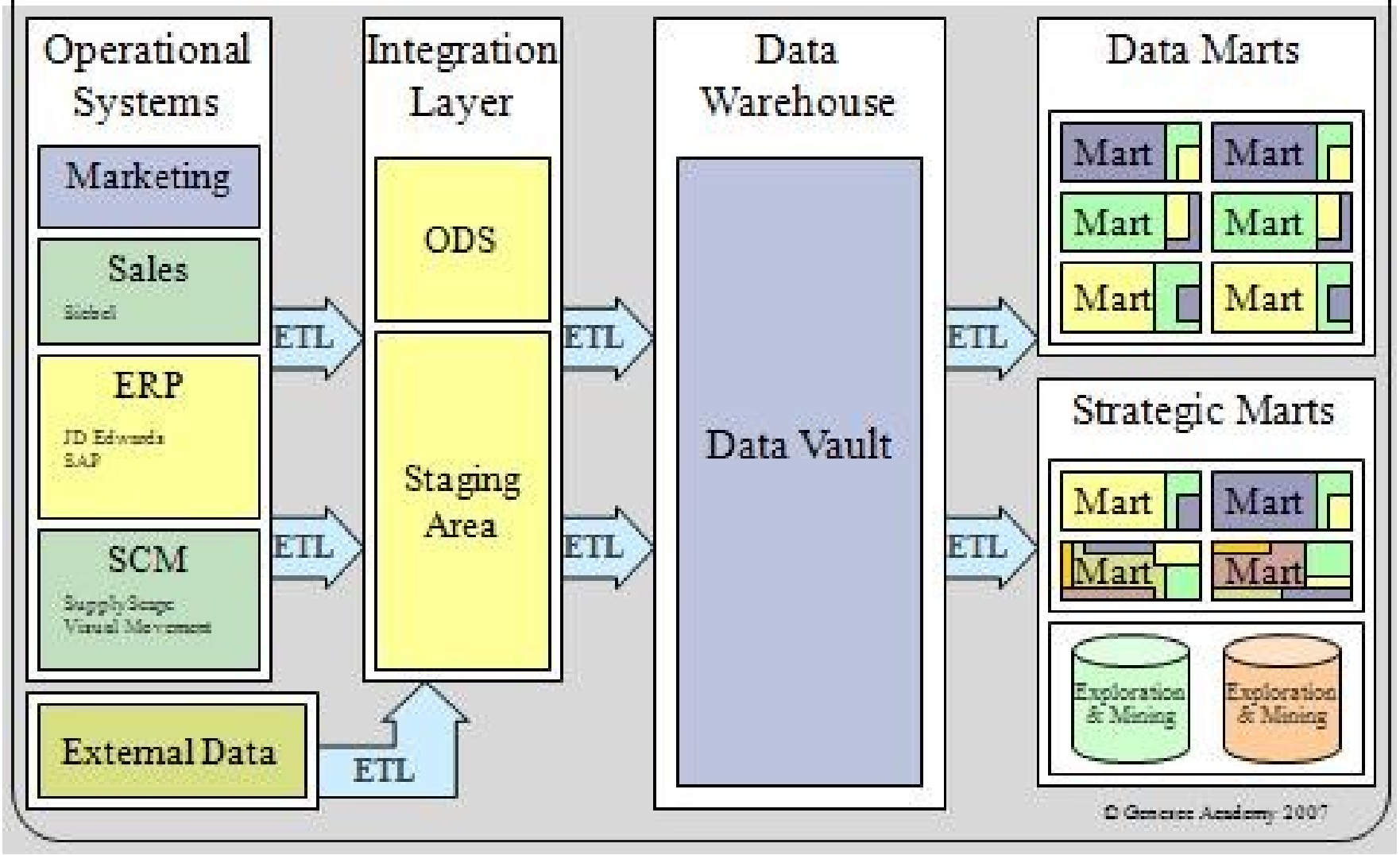


- Físicamente separado de los datos del entorno operacional.
- Actualización de los datos no se produce en el entorno de almacén de datos.
- No requiere el procesamiento de transacciones, ni mecanismos de recuperación y control de concurrencia
  - Requiere sólo dos operaciones en el Acceso a los datos: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

Flujo del Dato



Ubicación Data Warehousing (Bodega de Datos)



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 de negocio y las tareas de toma de decisiones

Estructure o formato

Modelos BD

relacional
plans
jerárquica
red
dimensional
orientada a objetos

Data Warehouse

- Tradicionalmente heterogénea => integración
- Construir 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 Filtrado de información

La información de fuentes heterogéneas se integra con antelación y se almacena en bodegas para consulta directa y análisis

BD	Data Warehouse
Diseñado para las tiempo operaciones de negocio tiempo real	Diseñado para el análisis de las medidas del negocio por categorías y atributos
En su mayoría actualizaciones	Sobre todo se lee
Muchas pequeñas transacciones	Las consultas son largas y complejoa
Mb - Gb de datos	Gb - Tb de datos

BD

- Instantánea
- Los datos en bruto
- Miles de usuarios (por ejemplo, los usuarios de oficina)

Data Warehouse

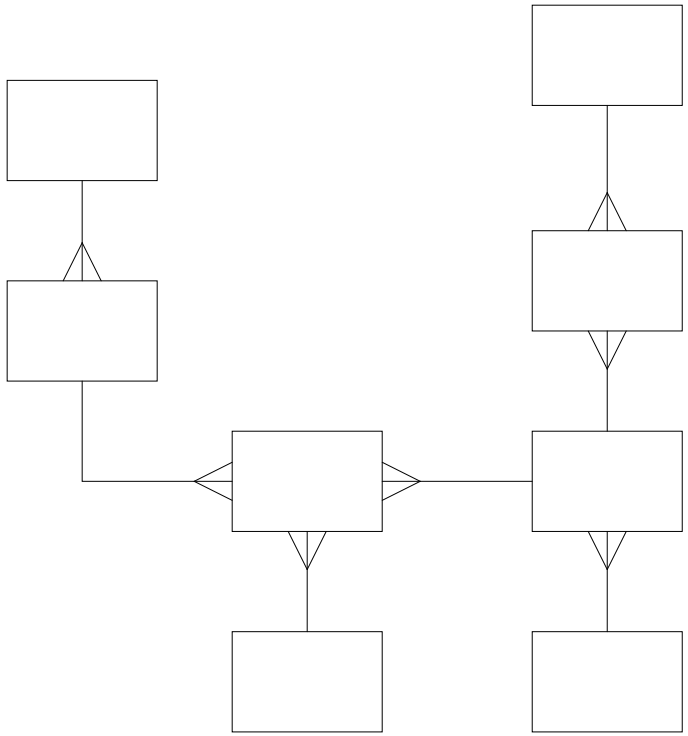
- Historica
- Resumidos y reconciliados los datos
- Cientos de usuarios (por ejemplo, los gerentes, los analistas)

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)
  - Majores tareas de data warehouse
  - Análisis de datos y toma de decisiones
- Características diferentes (OLTP vs. OLAP):
  - Orientación del usuario y del sistema: usuario vs cliente
  - Contenido de los datos: actual, detallada vs histórica, consolidada
  - Diseño de base de datos: aplicación + ER vs estrella + sujeto
  - Vista: actual: local vs integrada, , evolutiva
  - Patrones de acceso: actualizar vs consultas complejas

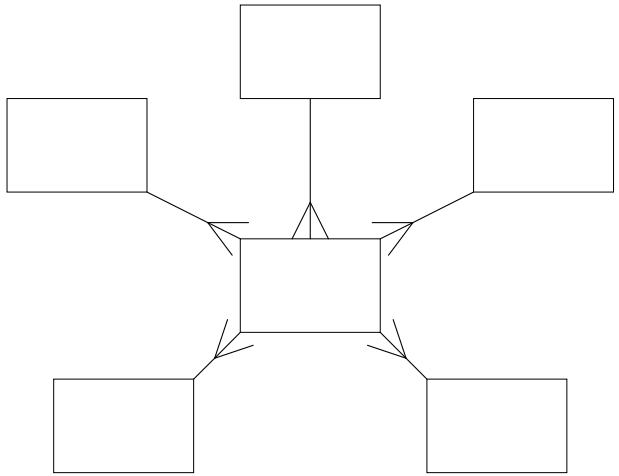
Las diferencias de diseño

BDs



Diagráma ER

Data Warehouse



Estrella

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 diferentes:

- **Datos que faltan**: apoyo a las decisiones requiere datos históricos que BDs operativos no suelen tener
- **Consolidación de datos**: SSD requieren 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 reconciliados

Resumen conceptual

- **Data Warehouse** - La fuente 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 restricción del Data Warehouse para un único proceso de negocio o para un grupo de procesos de negocio relacionados dirigida hacia un grupo empresarial en particular.

- **Datos Operacionales** - sistemas BDs operativos que se desarrollaron independientes uno de otro. Apoyan operaciones del día a día de la organización, por lo que debe actualizarse continuamente.

Data Warehouses

Grandes cantidades de datos - a veces subdivididas en unidades lógicas más pequeñas (data marts dependientes)

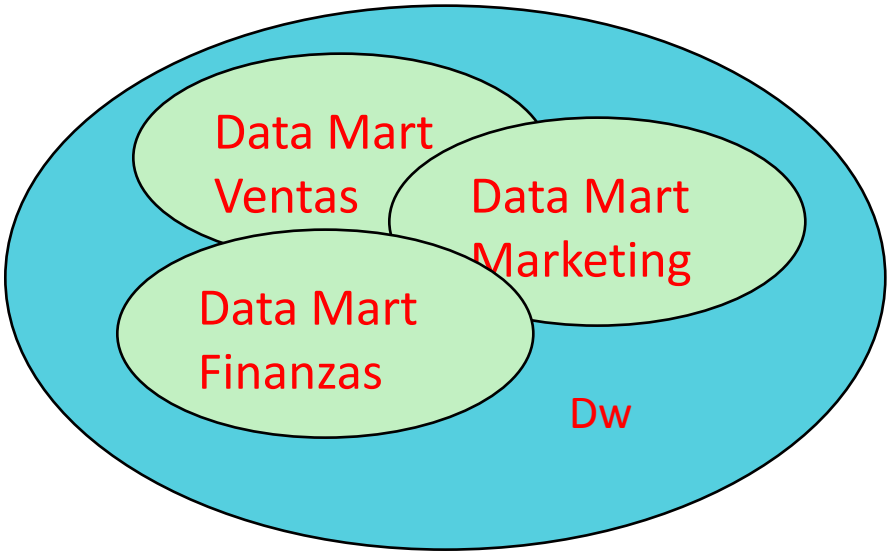
Datos en data warehouses son temáticamente coherente y para problemas y/o instituciones específicas

Data Mart

- Una versión reducida de menor costo del data warehouse diseñado para una unidad de negocios estratégica o departamento.

- Principal problema: data marts a menudo difieren de un departamento a otro.

- Dos enfoques:
  - data marts $\Rightarrow$ sistema en toda la empresa
  - data warehouse $\Rightarrow$ data marts

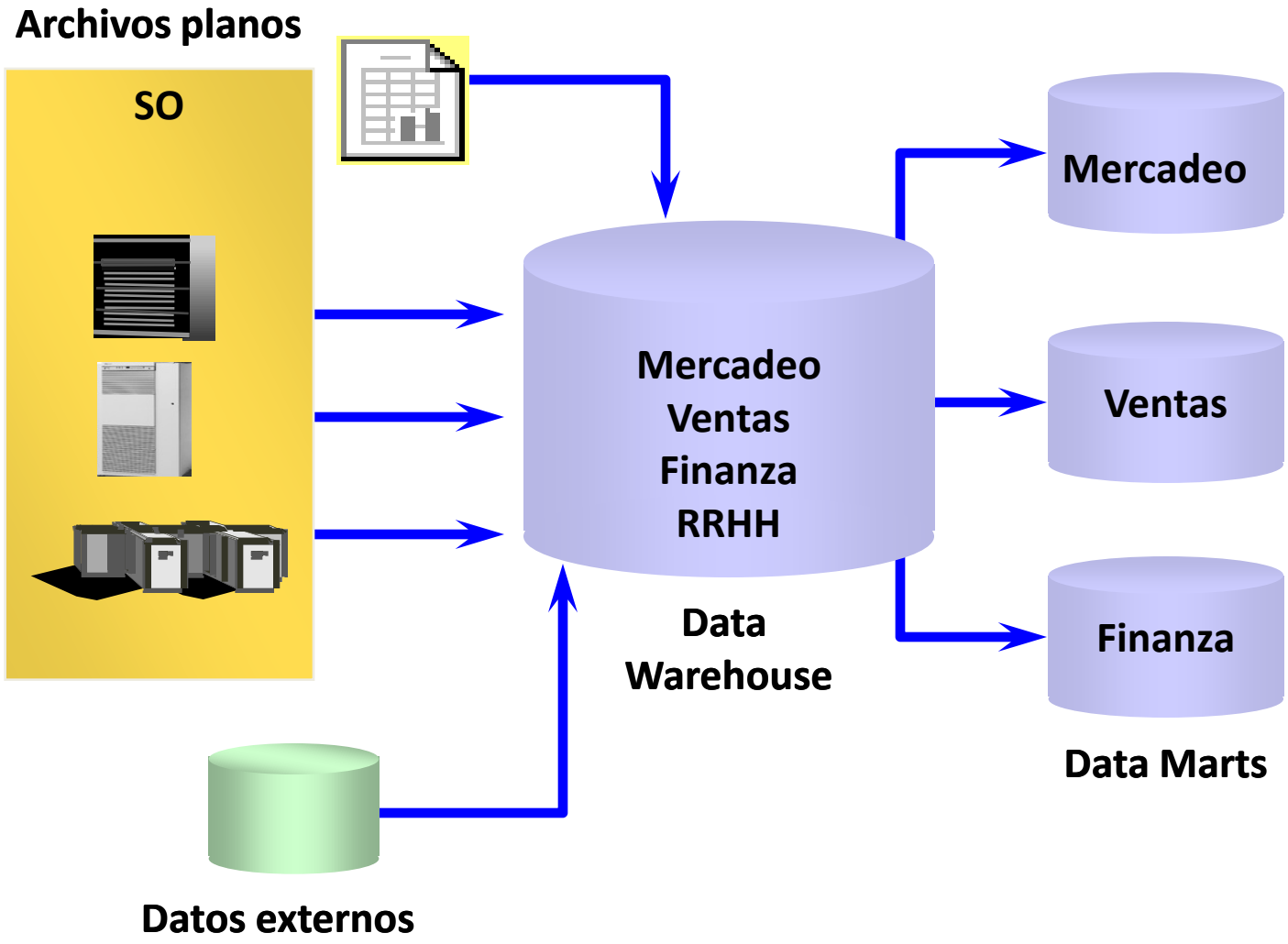


Objetivo de Data Mart

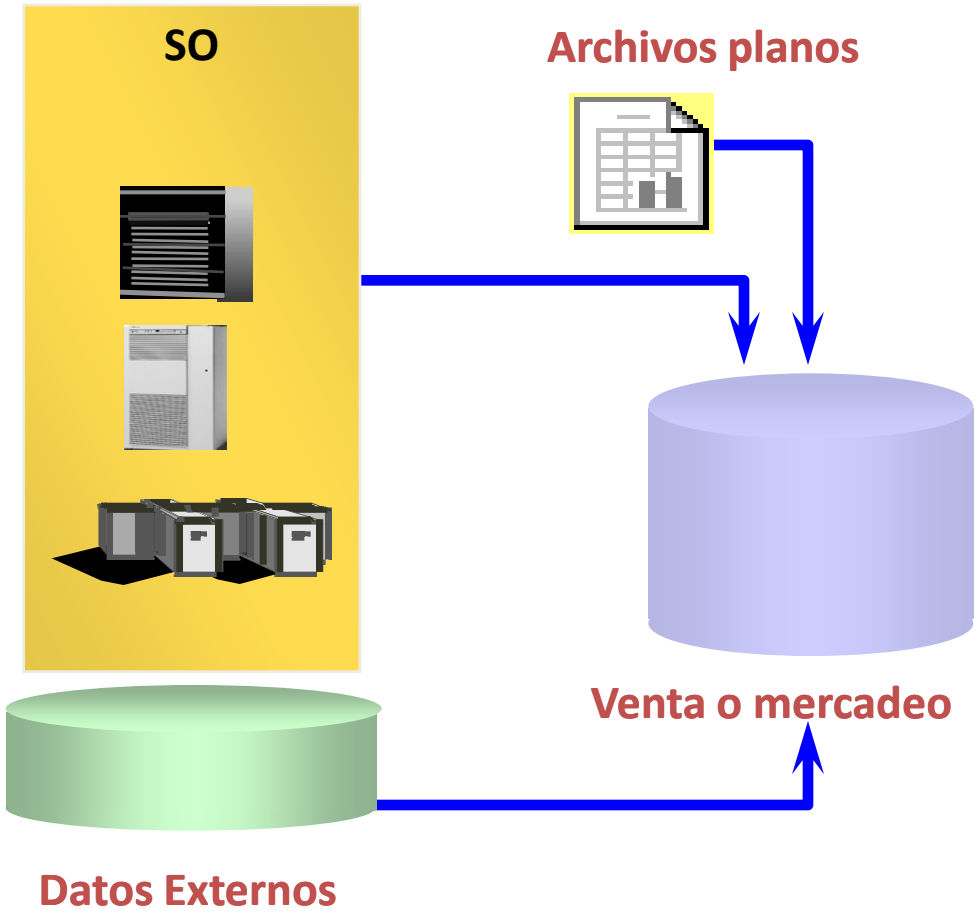
- Su propósito es soportar la toma de decisiones en un negocio (no las operaciones del negocio).

- Entre las características de un data mart destacan:
  - Usuarios limitados.
  - Área especifica.
  - Tiene un propósito especifico.
  - Tiene una función de apoyo.

Dependiente Data Mart



Independiente Data Mart



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 y pueden ser objeto de apoyo

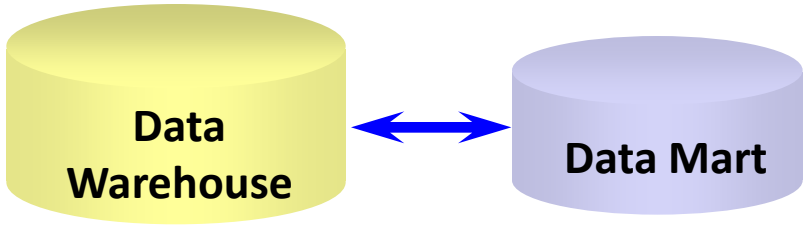
Razones para la creación de un Data Mart

- Proporcionar datos apropiadamente 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 la requerida para un data warehouse

Data Warehouses Vs Data Marts



propiedad	Data Warehouse	Data Mart
alcance	Empresa	departamento
sujetos	Múltiple	Sujeto/grupo
Fuente de datos	Muchos	pocos
Tamaño (típico)	100 GB a> 1 TB	<100 GB
Tiempo de implementación	Meses o años	Meses

Nueva Definición de Data Warehouses de Ralph Kimball

- Unión de todos los Data marts de una entidad.

- Por tanto una metodología ascendente (Modelo bottom-up (las partes individuales se diseñan con detalle y luego se enlazan para formar componentes más grandes, que a su vez se enlazan hasta que se forma el sistema completo))

Por qué datos en Warehouse?

Un almacén de datos contiene cinco tipos de datos:

- Datos actuales de detalle
- Viejo datos de detalle
- Datos ligeramente resumidos
- Datos altamente resumidos
- Metadatos

Granularidad de los datos: una cuestión clave de diseño

Funcionamiento de Data warehouse

- Almacén o contenedor de datos que son necesarios o útiles para una organización.

- Debe entregar la información correcta a la gente indicada en el momento óptimo y en el formato adecuado.

- El almacén de datos da respuesta a las necesidades de usuarios expertos, utilizando Sistemas de Soporte a Decisiones (DSS), Sistemas de información ejecutiva (EIS) o herramientas para hacer consultas o informes.

Funcionamiento de Data warehouse

Tres funciones esenciales:

Recopilación de datos de bases de datos y cargados

Gestión de los datos en el almacenamiento

Análisis de datos para la toma de decisiones

Funcionamiento de Data warehouse

- **Integración de los datos** provenientes de bases de datos distribuidas por las diferentes unidades de la organización, que con frecuencia tendrán diferentes estructuras (fuentes heterogéneas).

- **Separación de los datos** usados en operaciones diarias en el almacén de datos para los propósitos de divulgación, ayuda en la toma de decisiones, para el análisis y para operaciones de control. Ambos tipos de datos no deben coincidir en la misma base de datos, obedecen a objetivos muy distintos y podrían entorpecerse entre sí.

Elementos que integran un almacén de datos

- **Metadatos**
  - "datos acerca de los datos", documenta, entre otras cosas, qué tablas existen, qué columnas posee cada una de las tablas y qué tipo de datos se pueden almacenar.

  - Su función es recoger todas las definiciones de la organización y el concepto de los datos en el almacén de datos concerniente a:

*Tablas , *Columnas de tablas , *Relaciones entre tablas , *Jerarquías y Dimensiones de datos y *Entidades y Relaciones

Elementos que integran un almacén de datos

Metadato cuenta con

- Descripción de la estructura del almacén
  - esquema, dimensiones, jerarquías, ubicaciones data mart, etc .
- Meta-datos de los datos Operacionales
  - Historia de los datos (migraciones, transformaciones), estado de los datos (activo, archivados, o limpiado), la 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
- Los datos empresariales
  - propiedad de los datos, políticas de uso, etc.

Elementos que integran un almacén de datos

- **Middleware** ofrece servicios u operaciones que hacen posible el funcionamiento de aplicaciones distribuidas sobre plataformas heterogéneas que se sitúa entre las capas de aplicaciones y las capas inferiores (sistema operativo y red). Su función es asegurar la conectividad entre todos los componentes de la arquitectura de un almacén de datos (datawarehouse).

- **Funciones ETL, OLAP.**

Aplicaciones de Data warehouse

- **Tratamiento de la información**
  apoya la consulta, el análisis estadístico básico, y la notificación utilizando tablas cruzadas, tablas, cuadros y gráficos

- **Procesamiento analítico**
  análisis multidimensional de datos de almacenamiento de datos
  apoya las operaciones básicas de OLAP,

- **Minería de datos**
  descubrimiento de conocimiento a partir de patrones ocultos
  asociaciones, clasificación, predicción,

12 reglas de un Data Warehouse

- Data Warehouse y entornos operativos se Separados
- Los datos se integran
- Contiene datos históricos durante un largo período de tiempo
- Datos son un conjunto de datos capturados en un punto dado en el tiempo
- Datos son orientados al sujeto

12 reglas de un Data Warehouse

- Principalmente sólo lectura con actualizaciones periódicas por lotes
- Ciclo de Vida de Desarrollo sigue enfoque dirigido por los datos frente al enfoque basado en procesos tradicional
- Datos contiene varios niveles de detalle: Actual, Viejo, ligeramente resumido, muy resumido

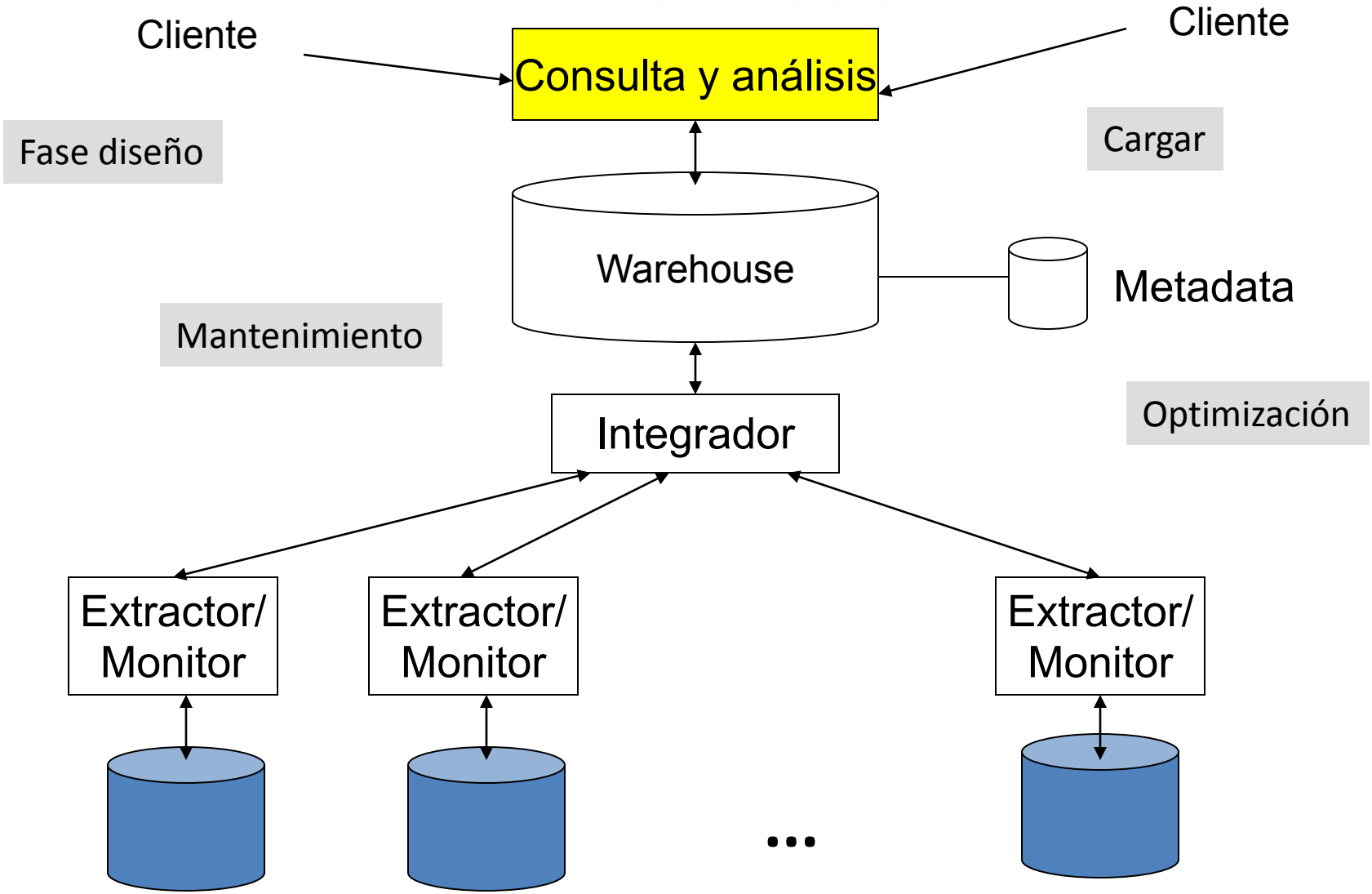
12 reglas de un Data Warehouse

- Medio Ambiente se caracteriza por transacciones de sólo lectura a conjuntos de datos muy grandes
- Sistema rastrea las fuentes de los datos, transformaciones y almacenamiento
- Los metadatos son un componente crítico Fuente, transformaciones, integración, almacenamiento, relaciones, historia, etc.
- Contiene un mecanismo retorno para el uso óptimo de los recursos

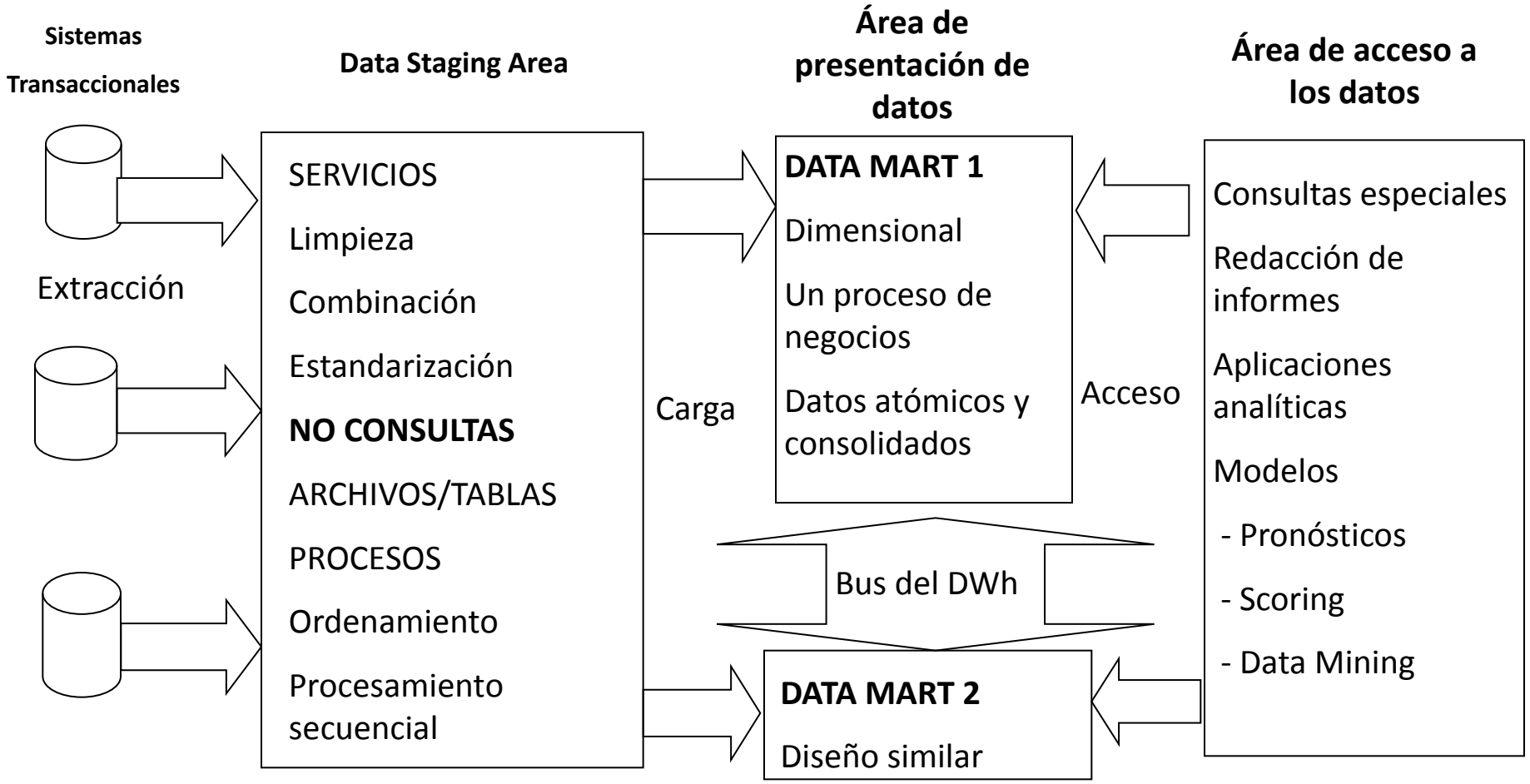
Arquitectura Data Warehouse

- Componentes
  - Fuentes de datos
  - Organización de los datos (Data Staging)
  - Presentación de datos
  - Acceso a los datos

Arquitectura genérica Data Warehouse



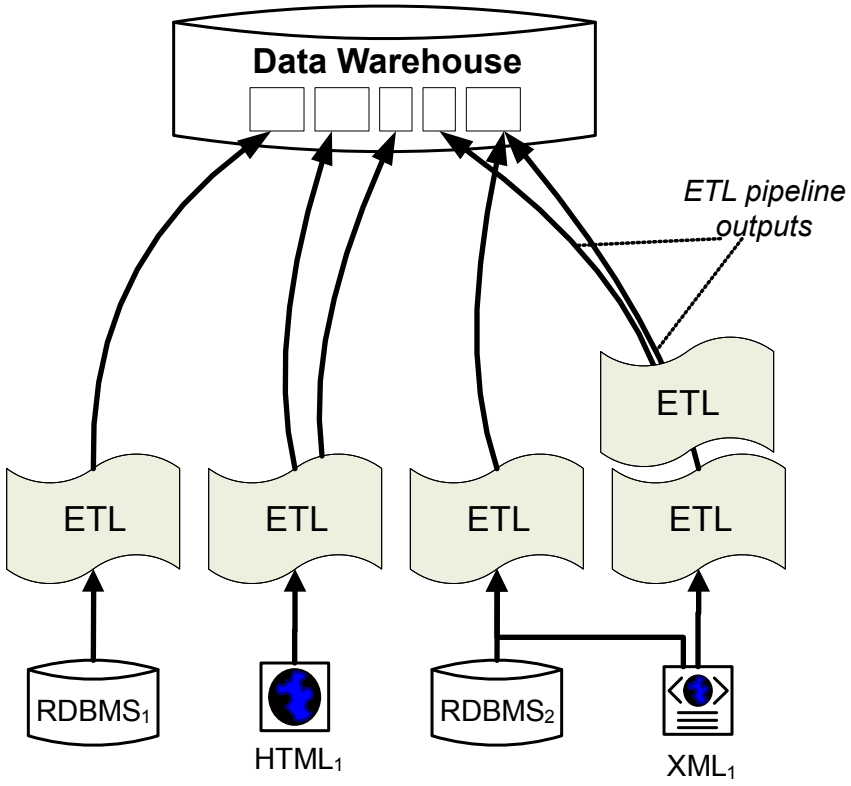
Arquitectura genérica Data Warehouse con Data Mart



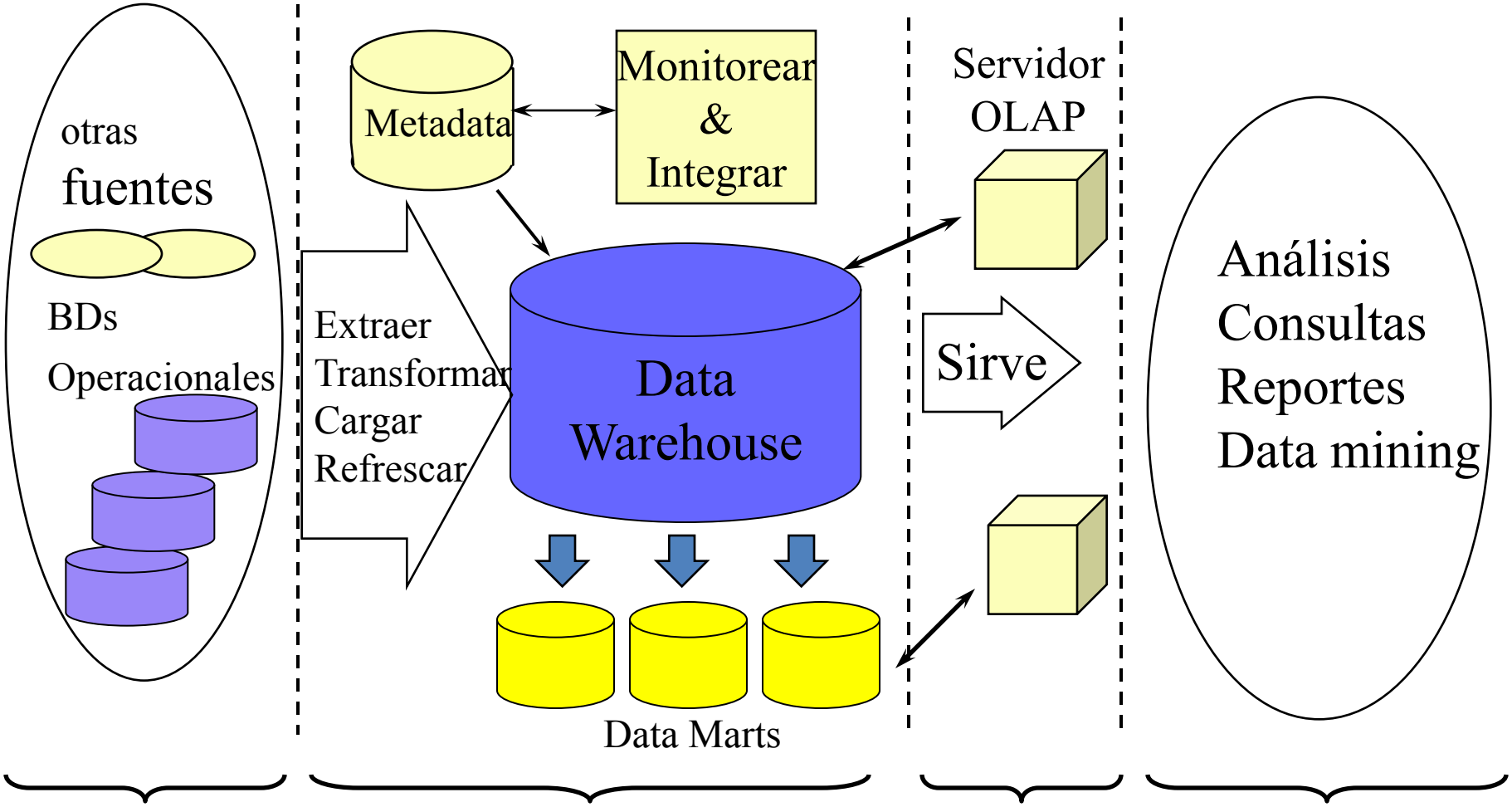
Arquitectura Data Warehouse primer modelo

- En la parte superior - una base de datos centralizada
  - Generalmente configurado para consultas y agrega – no transacciones
  - Muchos índices, vistas , etc.

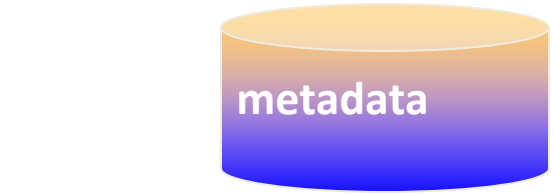
- Los datos se cargan y actualizan periódicamente a través de ETL (Extract / Transform / Load)



Arquitectura varios niveles

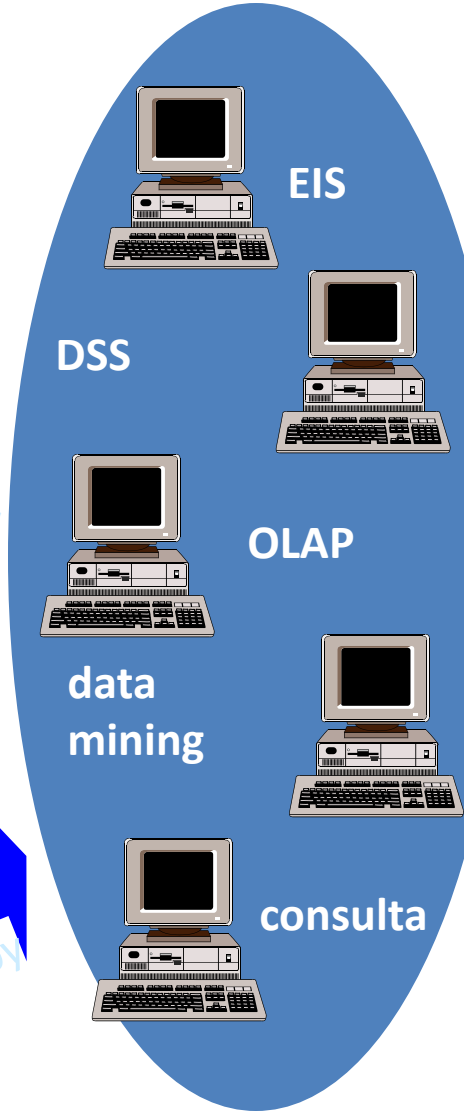


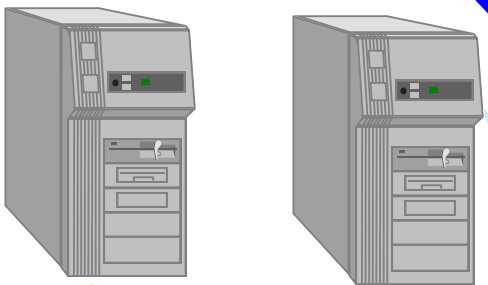
Arquitectura varios niveles



extraction
cleaning
validation
summariz.

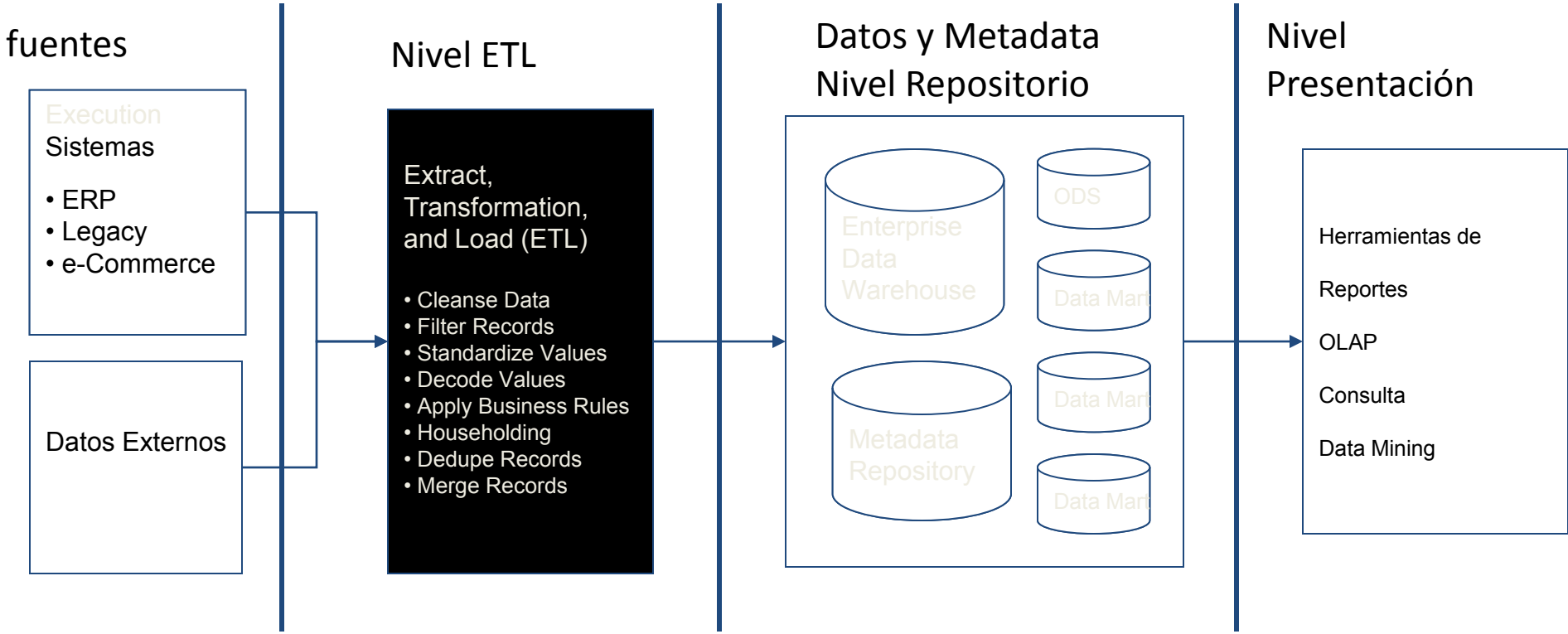
used by



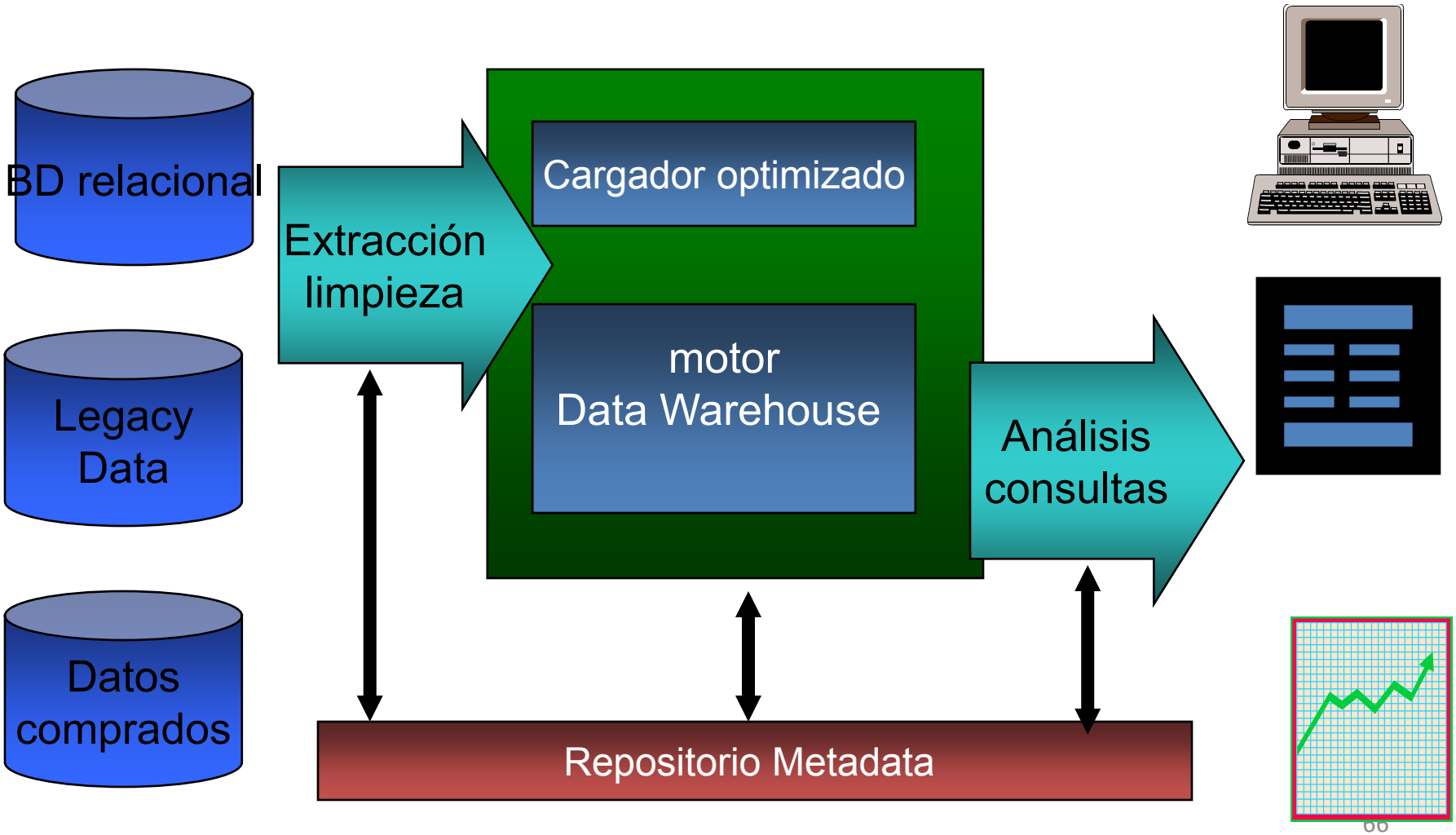


used by

Arquitectura varios niveles

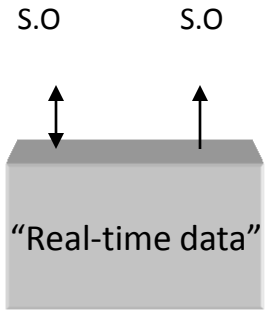


Arquitectura varios niveles

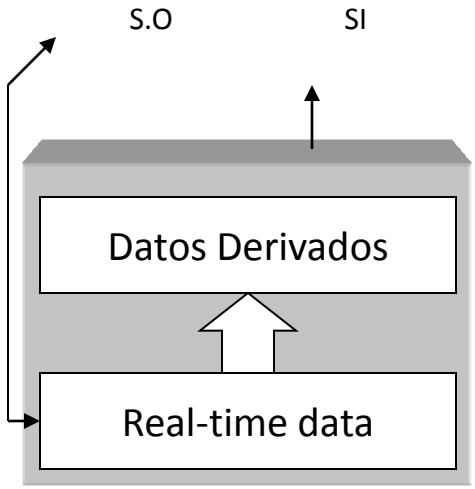


Arquitectura Data Warehouse: Vista Conceptual

- Una sola capa
  - Cada elemento de datos sólo se almacena oz almacén virtual

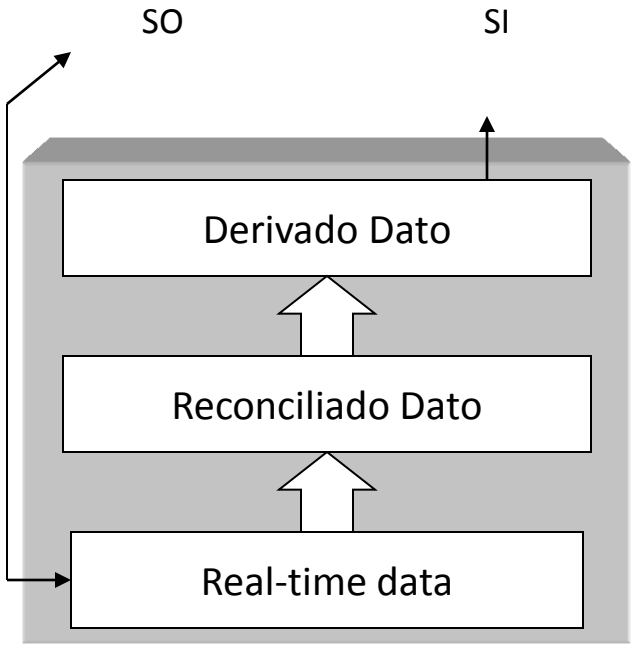


- De dos capas
  - tiempo real + datos Derivados
  - Mayor uso industrial hoy en día



Arquitectura de tres capas: Vista Conceptual

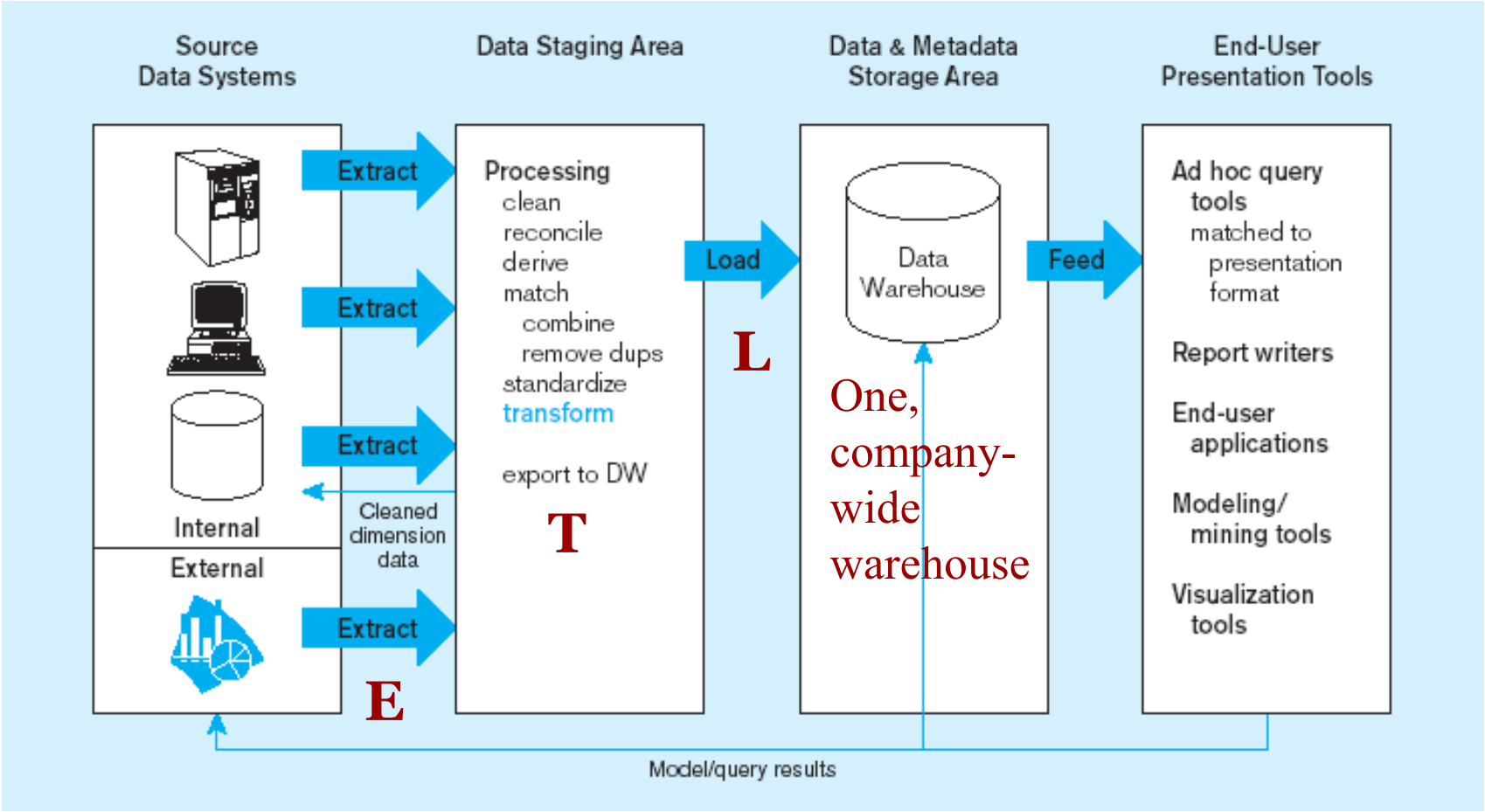
Procesamiento de datos en tiempo real a los datos derivados



Nivel de Visualización "Particular informativo necesidades "

Implementación física de Data Warehouse

Genérica arquitectura de dos niveles de data warehousing

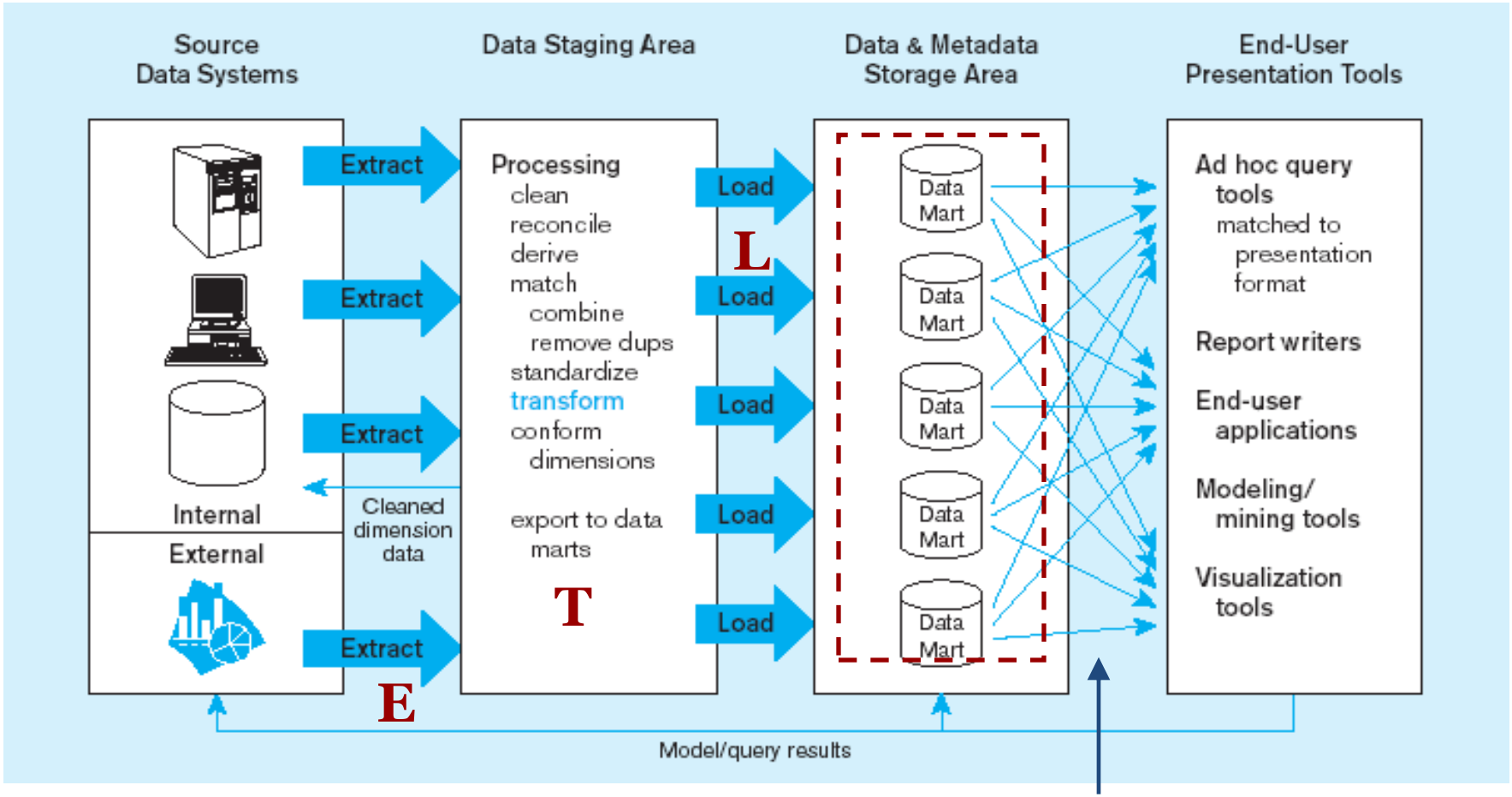


Extracción de datos de periódica

Datos no están completamente al día

Arquitectura independiente data mart y data warehousing

Data marts: Mini-warehouses,

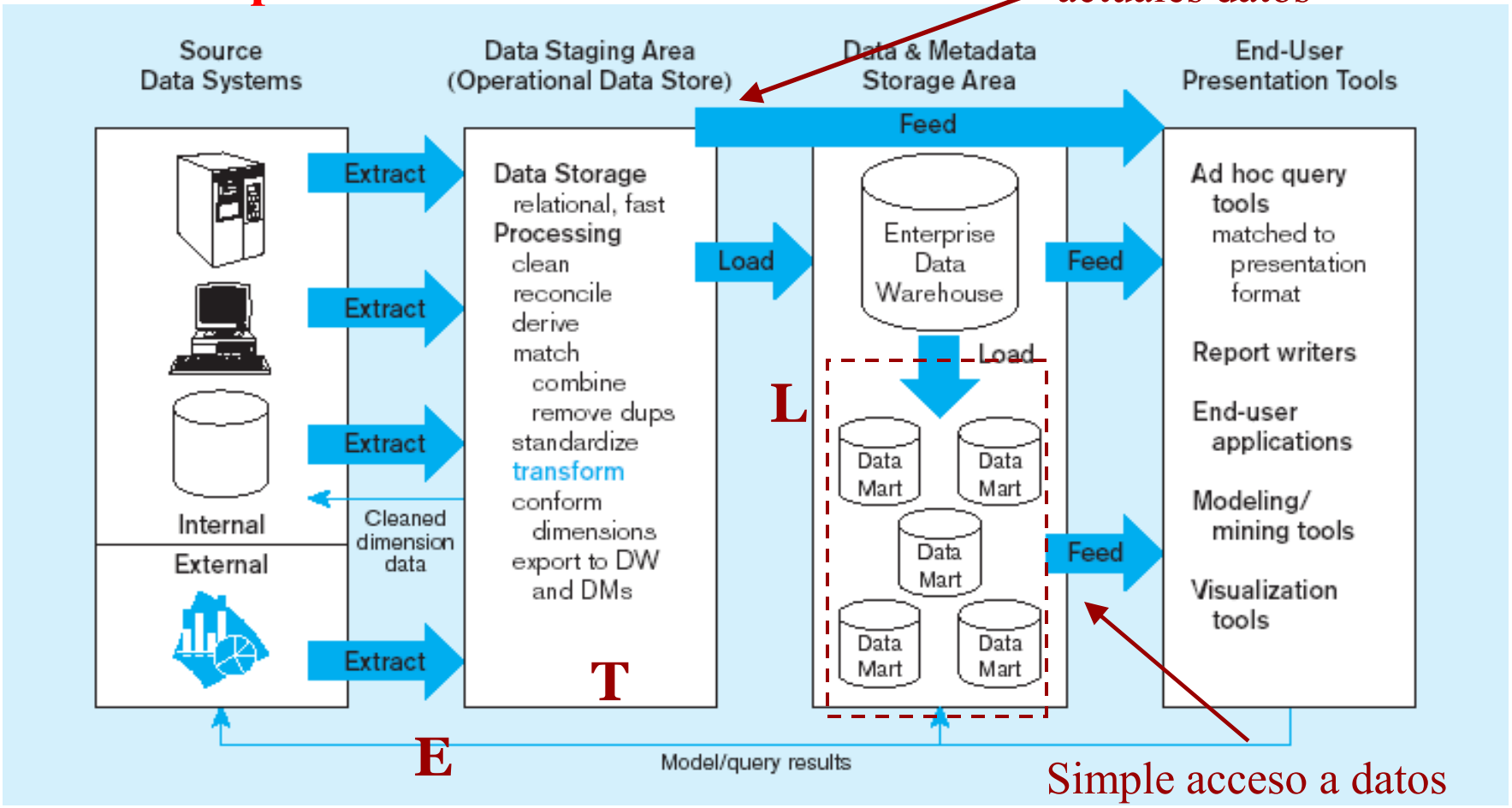


Separado ETL para c/data mart

Acceso datos complejo debido a ***multiples*** data marts

Dependiente data mart con datos operacionales 3 niveles

Se pueden obtener actuales datos

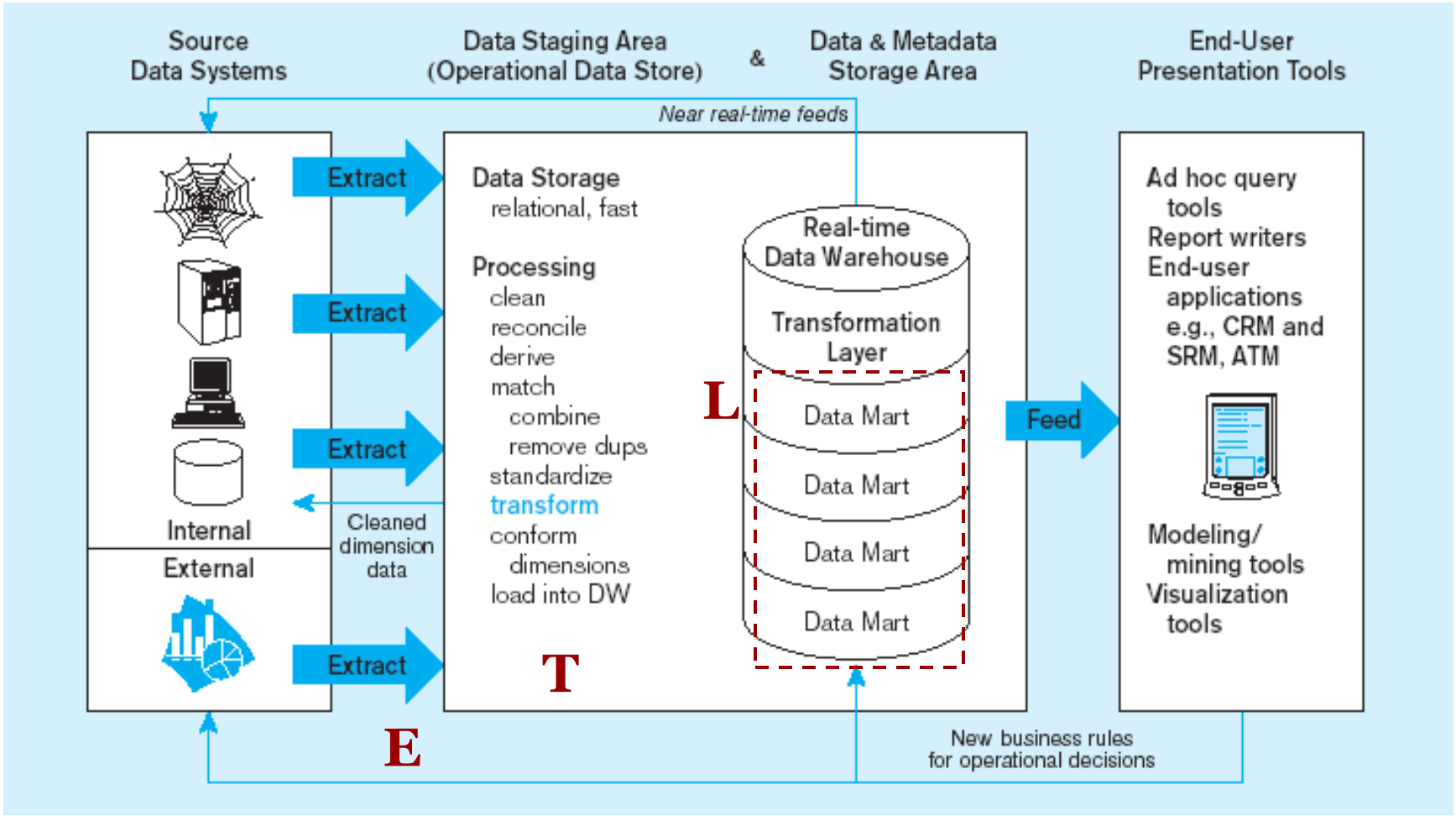


Simple acceso a datos

Simple ETL para ***enterprise data warehouse (EDW)***

Dependiente data marts cargado desde EDW

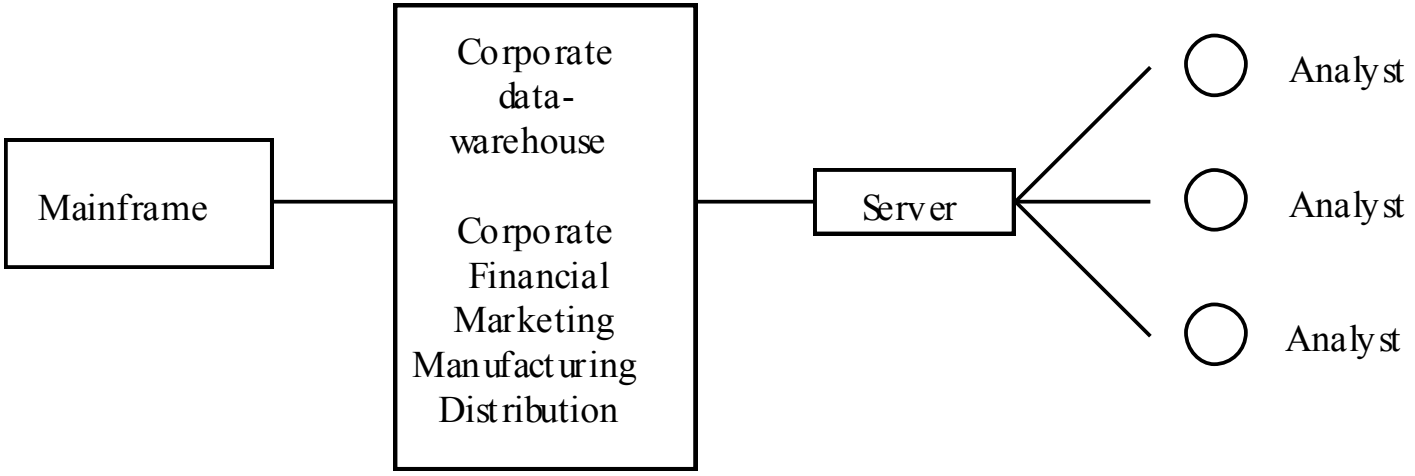
data mart Lógico y warehouse RT



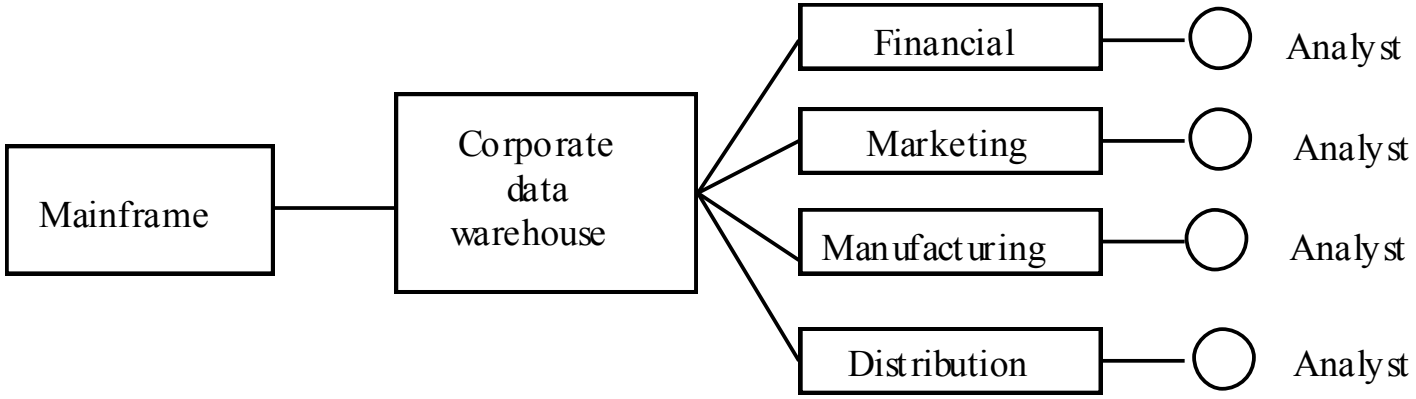
ETL RT para
Data Warehouse

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

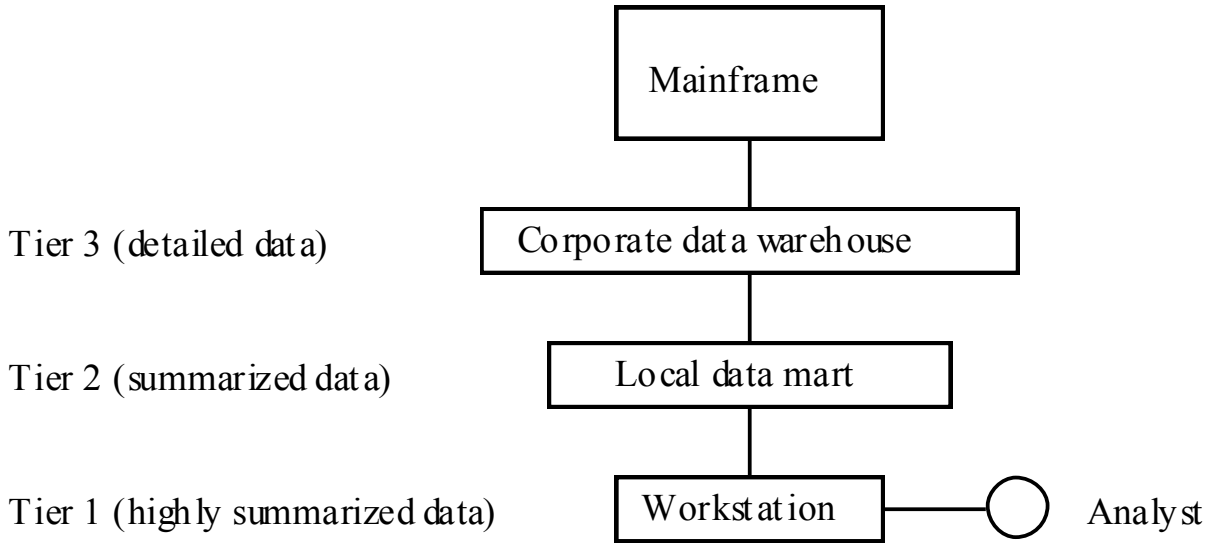
Centralizado data warehouse



Federado data warehouse

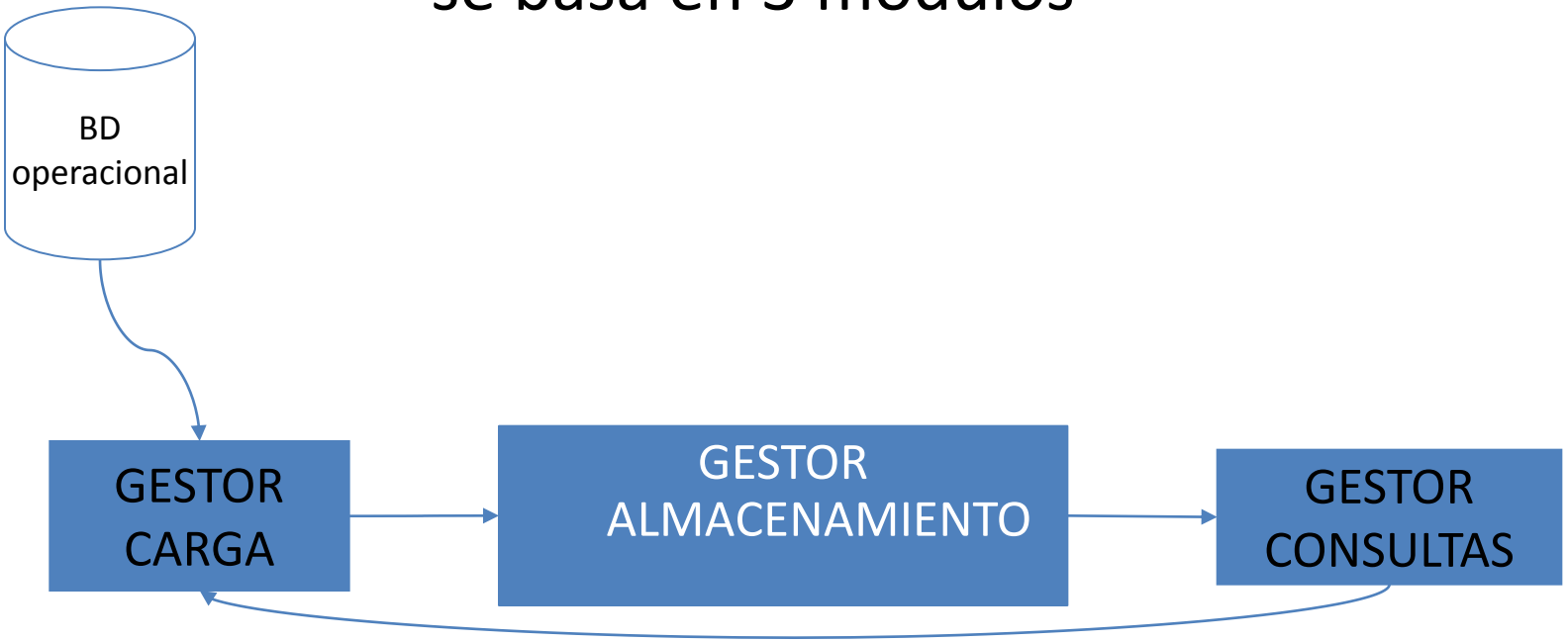


Multicapas data warehouse



ARQUITECTURA PLANA DW

se basa en 3 módulos



GESTOR DE CARGA

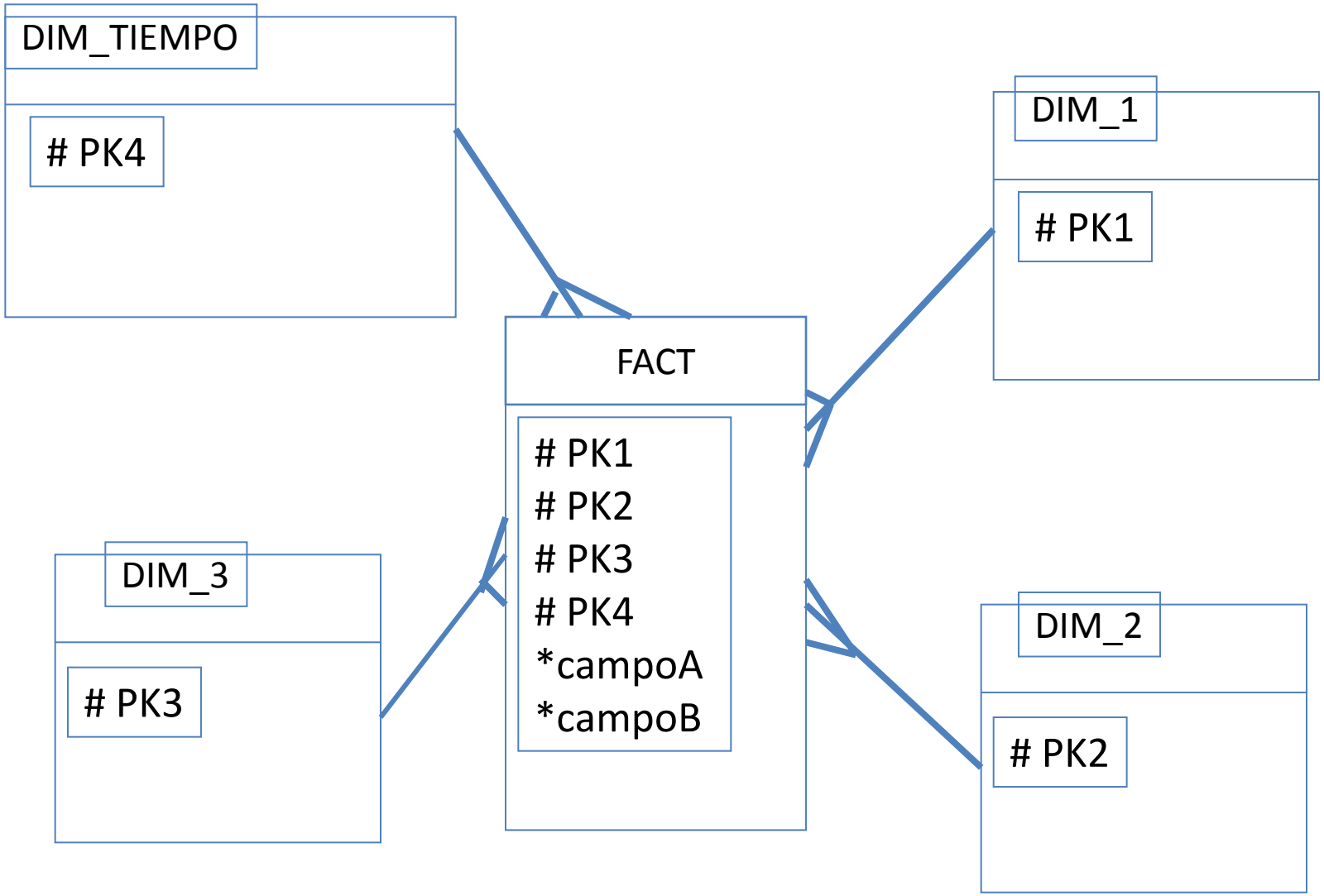
- Permite hacer la carga.

- Como dificultades nos podemos encontrar:
  - La integración de los datos
  - Elección del momento de la carga
  - El tiempo de carga sea el mínimo posible
  - 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 STAR

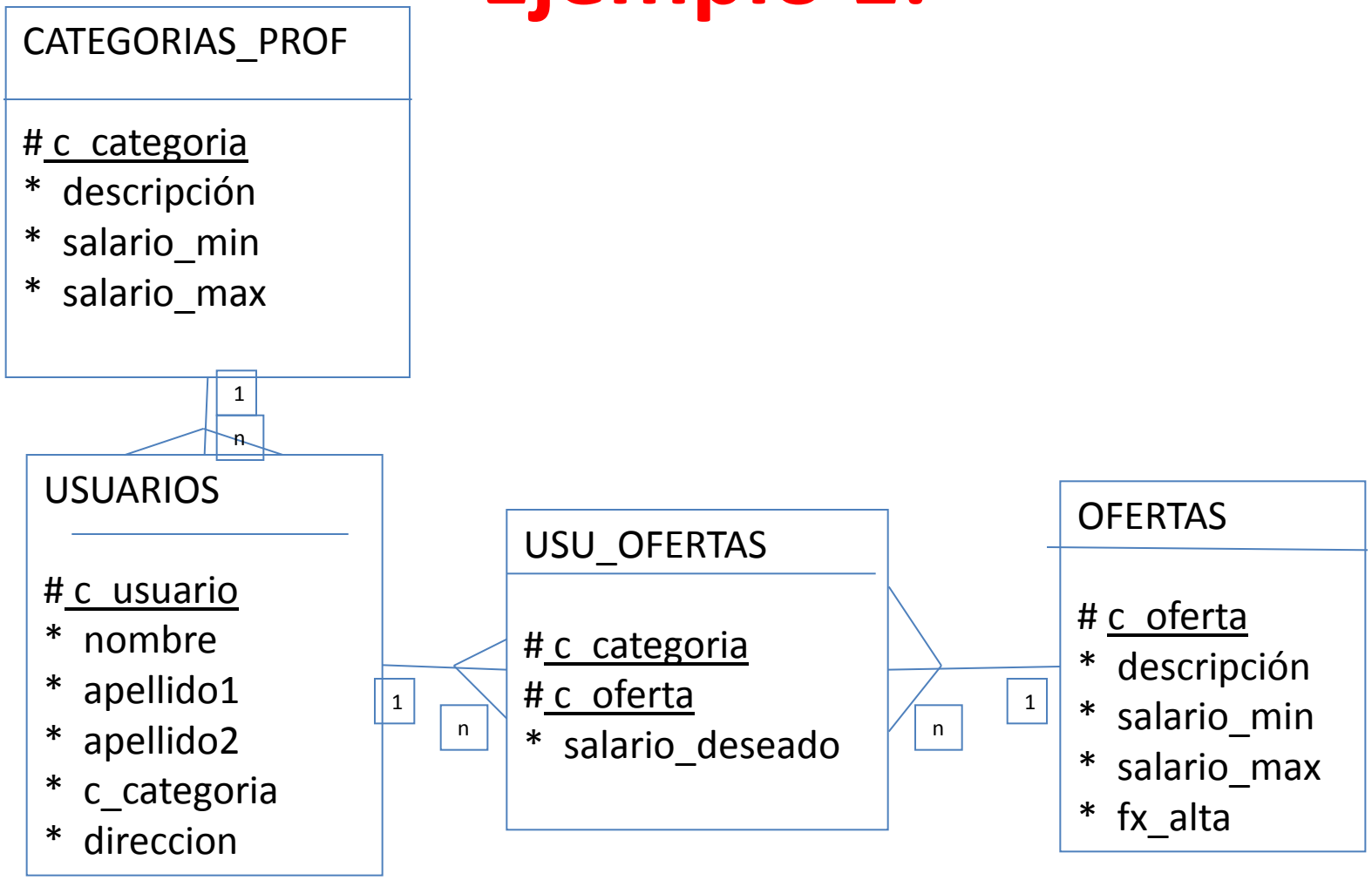
MODELO STAR


DIM_TIEMPO
PK4
DIM_1
PK1
FACT
PK1
PK2
PK3
PK4
*campoA
*campoB
DIM_3
PK3
DIM_2
PK2

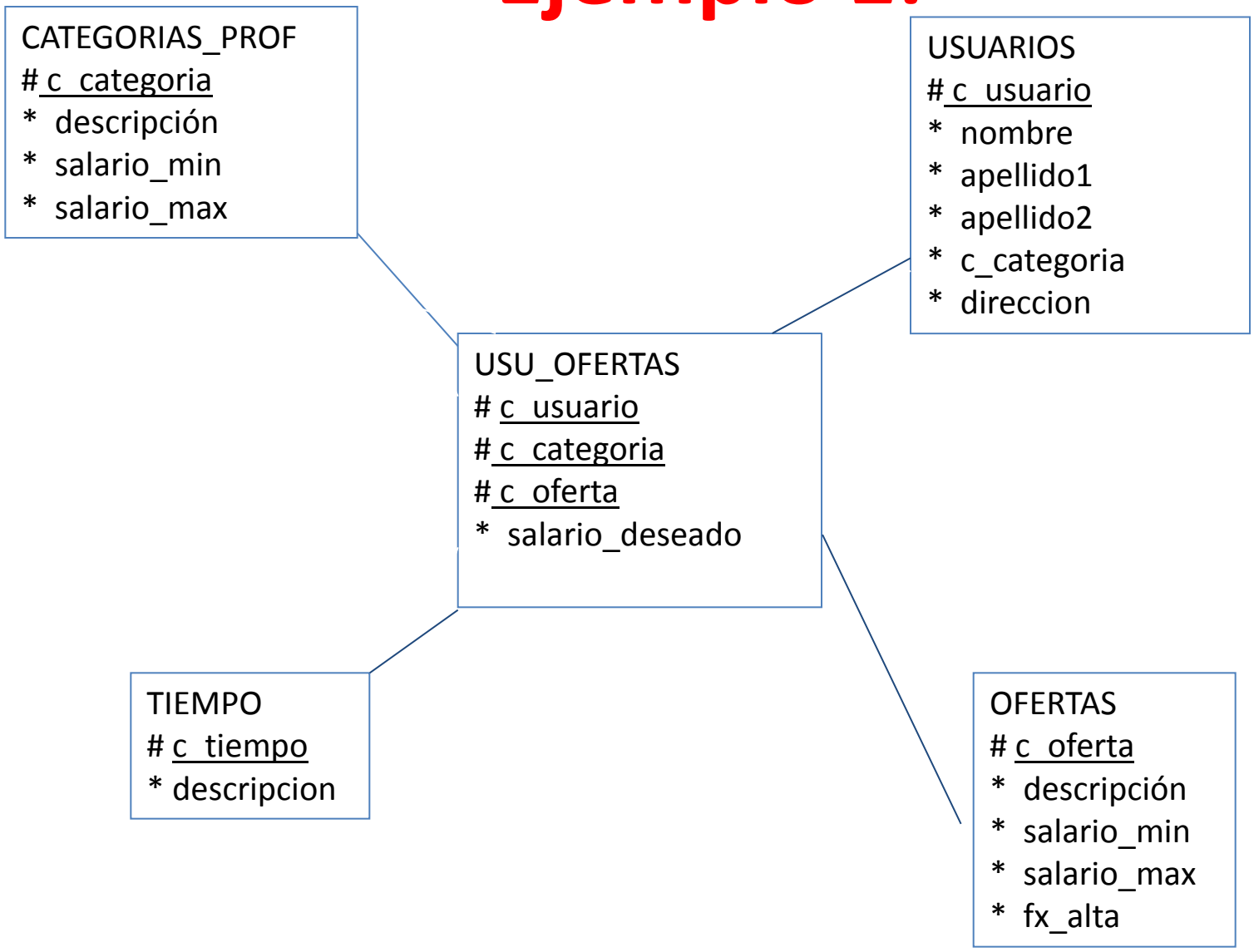
GESTOR DE CONSULTAS

- Las consultas se hacen sobre la tabla FACT.
- También se encarga de los perfiles, pues las consultas (reports) serán diferentes dependiendo del usuario y sus necesidades.

Ejemplo 1:


CATEGORIAS_PROF
c_categoria
* descripción
* salario_min
* salario_max
1
n
USUARIOS
c_usuario
* nombre
* apellido1
* apellido2
* c_categoria
* direccion
1
n
USU_OFERTAS
c_categoria
c_oferta
* salario_deseado
n
1
OFERTAS
c_oferta
* descripción
* salario_min
* salario_max
* fx_alta

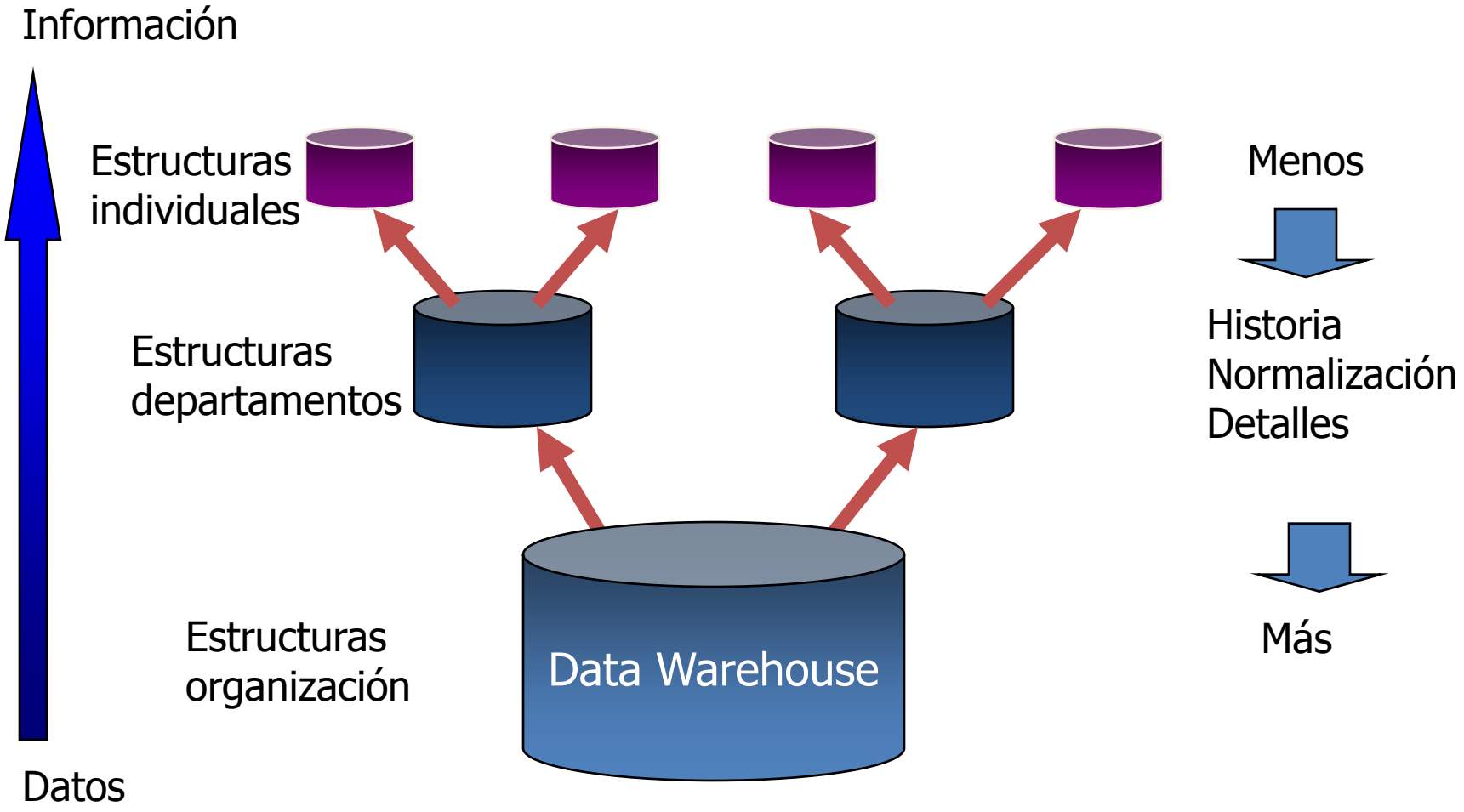
Ejemplo 1:


CATEGORIAS_PROF
c_categoria
* descripción
* salario_min
* salario_max
USUARIOS
c_usuario
* nombre
* apellido1
* apellido2
* c_categoria
* direccion
USU_OFERTAS
c_usuario
c_categoria
c_oferta
* salario_deseado
TIEMPO
c_tiempo
* descripcion
OFERTAS
c_oferta
* descripción
* salario_min
* salario_max
* fx_alta

Tres modelos de Data Warehouse

- **Warehouse de la empresa**
  - recoge toda la información sobre temas que abarcan toda la organización

- **Data Mart**
  - un subconjunto de datos de toda la empresa que es de valor para un grupo específico de usuarios. Su ámbito de aplicación se limita a grupos específicos, seleccionado, como el marketing data mart

- **Virtual warehouse**
  - Un conjunto de vistas de bases de datos operacionales
  - Sólo revisa algunas de las vistas

Desde Data Warehouse a Data Marts