Almacenes de datos (Data Warehouse - DW)

Dr. Luis Gustavo Esquivel Quirós

Este material está basado en documentos desarrollados por Elzbieta Malinowski y Esteban Zimányi

Caso de estudio

- Pensemos en una cadena de tiendas:
  - Cada tienda mantiene sus propios registros de clientes y registros de ventas.
- Pregunta difícil de responder: "encontrar las ventas totales del producto X de tiendas en Canadá".
- Problemas interesantes: El mismo cliente puede ser visto como diferentes clientes para diferentes tiendas; esto hace difícil detectar información duplicada del cliente.
- Problemas comunes:
  - Datos imprecisos o faltantes en las direcciones de algunos clientes.
  - Registros de compras se almacenan por tiempo limitado (por ejemplo, 12 meses); luego se eliminan o se archivan.
  - El mismo “producto” puede tener diferentes precios, o diferentes descuentos en diferentes tiendas.

Problemas al realizar análisis de datos

- Los mismos datos pueden ser encontrados en varios sistemas diferentes , por ejemplo: datos de clientes en diferentes tiendas o un mismo concepto o producto se define de forma diferente.
- Fuentes heterogéneas de datos, por ejemplo: datos relacionales, datos no estructurados en archivos (por ejemplo, MS Word) o sistemas heredados, entre otros.
- La calidad de los datos es mala, por ejemplo: datos faltantes, datos imprecisos o con diferente uso en los sistemas.
- Datos "volátiles", por ejemplo: datos eliminados (tiempo de retención o espacio) o cambios en los datos a lo largo del tiempo (sin información histórica).

EL AUMENTO ESPECTACULAR DEL VOLUMEN DE DATOS HACE EVIDENTE LA NECESIDAD DE UNA INFRAESTRUCTURA PARA LA LÓGICA DE INFORMACIÓN.

SURGE COMO RESPUESTA A LA PROBLEMÁTICA DE EXTRAER INFORMACIÓN SINTÉTICA A PARTIR DE DATOS ATÓMICOS ALMACENADOS EN BD DE PRODUCCIÓN.

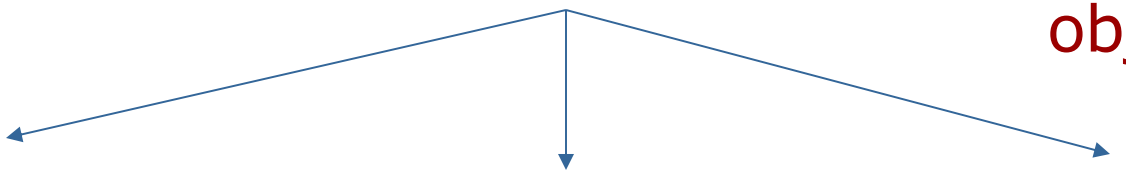
ALMACÉN DE DATOS(DW)

Motivación

Disponer de Sistemas de Información de apoyo a la toma de decisiones

Disponer de BD que permitan extraer conocimiento de la información histórica almacenada en la organización

objetivos



Análisis de la organización

Previsiones de evolución

Diseño de estrategias

Ejemplo

Organización: Cadena de supermercados

Actividad objeto de análisis: ventas de productos

Objetivo: aumentar ventas con publicidad adecuada

Problemas:

Necesitamos sólo los datos necesarios de la BD

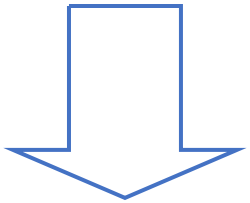
Fuentes de datos diversas (BDs diferentes, ficheros de texto, ficheros XML…)

Fuentes de datos externas

Demasiados datos

Análisis en tiempo real

LA BD Relacional NO BASTA!!
SE NECESITA OTRA COSA



Almacén de datos o Data Warehouse

Data Warehouse

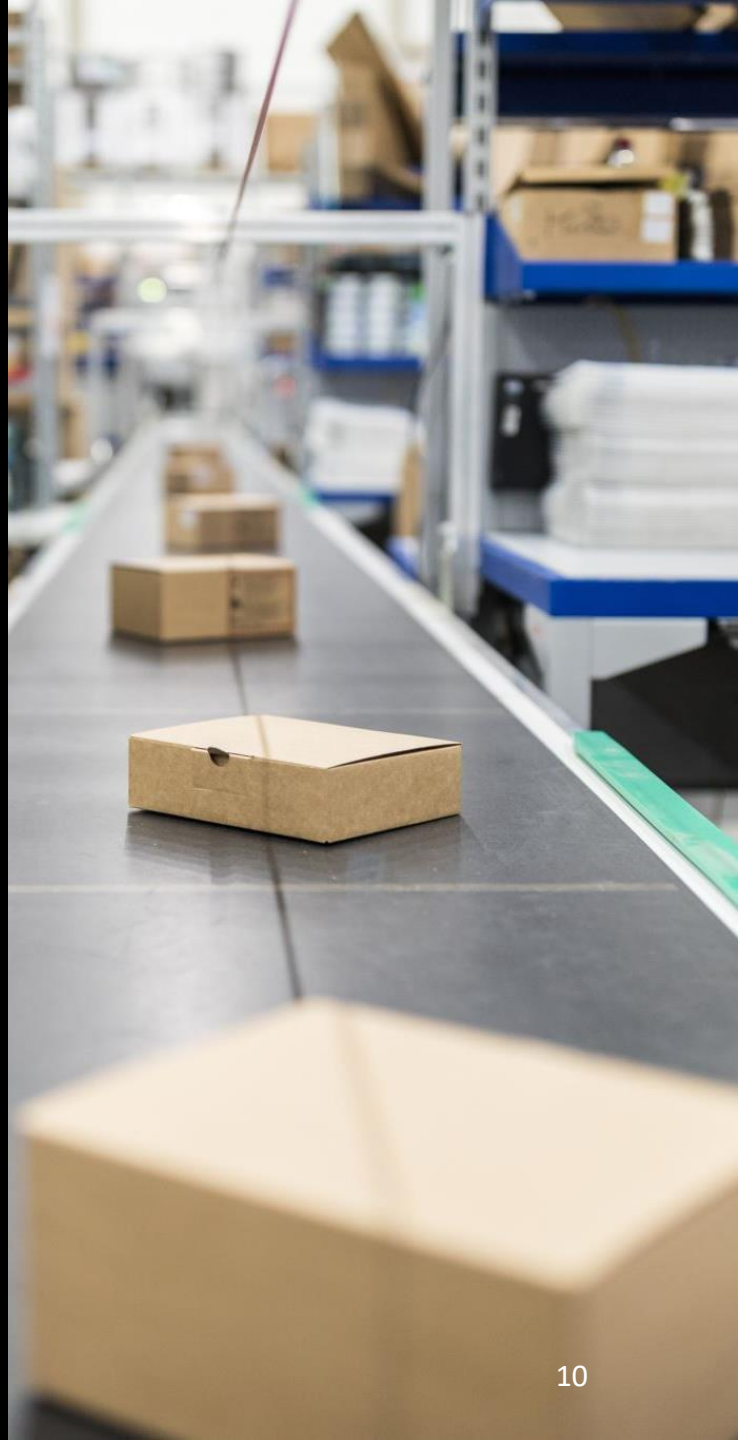
- El concepto de DW se remonta a principios de los 80, cuando la empresa Teradata introdujo un sistema de base de datos diseñado específicamente para el soporte de decisiones.
- A finales de los 80 (1988), los investigadores de IBM, Barry Devlin y Paul Murphy, desarrollaron el "almacén de datos empresariales".
- Sin embargo la literatura menciona a Inmon como padre del término "data warehousing".
- La expansión y popularidad comenzó con:
  - La publicación de libros de Bill Inmon (1991) y Kimball (1996). En estos libros se da un manejo más extendido de los problemas de temas de diseño e implementación. Esto fomentó el creciente interés de empresas y proveedores.

Data Warehouse

- Mayor interés del mundo académico
- Herramientas comerciales dedicadas para el almacenamiento de datos.
- Una brecha significativa entre investigadores y profesionales:
  - Los investigadores pasan por alto problemas prácticos
  - Poca aceptación de los resultados de la investigación por parte del mundo industrial
  - Problemas y fallas:
    - No hay modelo conceptual
    - No hay metodología de diseño
    - No hay estándares para metadatos
    - No hay soluciones para ETL

¿QUÉ ES UN DATA WAREHOUSE?

- DW es un conjunto de tecnologías, NO ES UN PRODUCTO.
- Es una arquitectura que debe construirse de acuerdo a las necesidades y entorno específico de los clientes. Debe construirse de manera *iterativa,* para consolidar y administrar datos de varias fuentes con el propósito de conseguir resultados en un periodo de tiempo aceptable:
    - *Ayudar a la toma de decisiones(*Decision Support System - *DSS).*
    - *Descubrir conocimiento (Data Mining->minería de datos).*
    - *Responder preguntas de negocio (OLAP->análisis de datos).*



¿QUÉ ES UN DATA WAREHOUSE?

- Para poder utilizar esta tecnología es necesario comprender:
  1. ¿Por qué las bases de datos operativas (transaccionales) no son la mejor opción para respaldar el proceso de toma de decisiones?
  2. ¿Qué modelos y componentes se utilizan actualmente para el diseño de almacenes de datos?
  3. La importancia de las fuentes de datos , calidad de datos y transformación de datos.
  4. La relación existente entre modelos multidimensionales de Almacenes de Datos (DW) y el Procesamiento Analítico en Línea (OLAP).
  5. Metodología de diseño DW.

¿QUÉ ES UN DATA WAREHOUSE?

- Según Inmon:
  - Un DW es una colección de datos que varían en el tiempo
    - Orientados a un tema
    - Integrados
    - No volátiles
    - Generalmente comprenden distintos intervalos en el tiempo
  - Sirven para respaldar una o varias decisiones de gestión

¿QUÉ CONOCEMOS HASTA ESTE MOMENTO?

Las bases de datos operativas (sistemas de procesamiento de transacciones en línea u OLTP) no son adecuadas para el análisis de datos. Estas contienen datos detallados, no incluyen datos históricos, funcionan mal para consultas complejas debido a la normalización.

Los DW abordan los requisitos de los usuarios que toman decisiones

Un almacén de datos es una colección de datos orientados al tema, integrados, no volátiles y que varían en el tiempo; los cuales permiten respaldar las decisiones de administración de datos.

¿QUÉ ES UN DATA WAREHOUSE?

ALMACEN DE DATOS(DW)

Base de Datos diseñada para el objetivo de exploración distinto que al de las BD de los sistemas operacionales

Sistema Operacional → BD orientada al proceso

Sistema de almacén de datos(DW) → BD orientada al análisis

¿QUÉ ES UN DATA WAREHOUSE?

ALMACEN DE DATOS(DW)

Colección de datos diseñada para dar apoyo a los procesos en la toma de decisiones

características

- Orientada hacia la información relevante de la organización.
- Integrada
- Variable en el tiempo
- No volátil

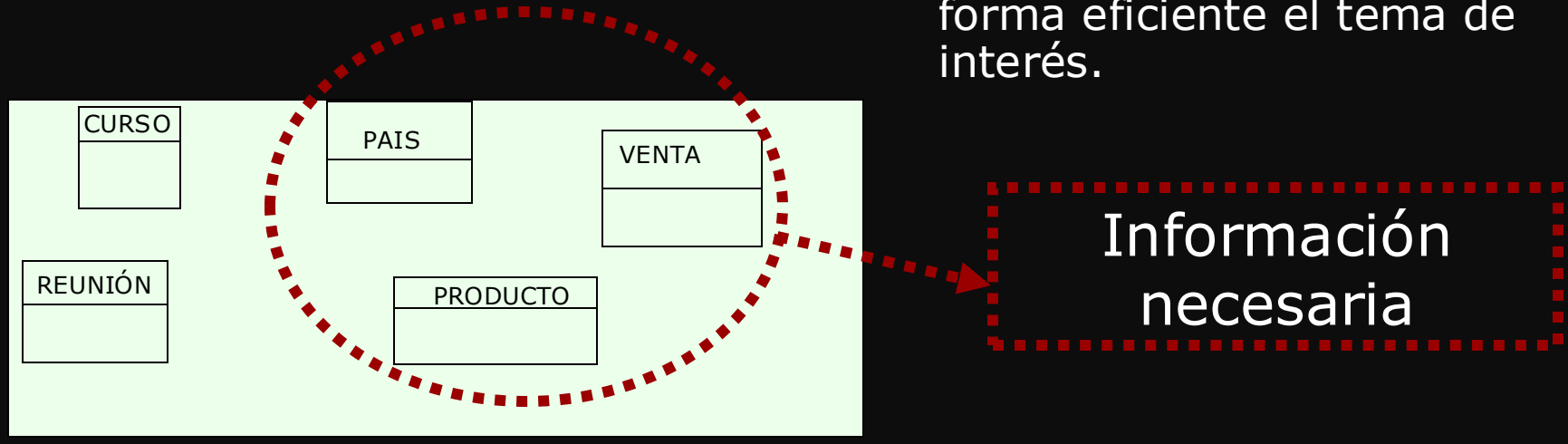
¿QUÉ ES UN DATA WAREHOUSE?

Orientado al tema

Se enfoca en temas particulares de análisis (por ejemplo, compra, inventario y ventas en una empresa minorista).

Las bases de datos que conocemos hasta el momento se centran en funciones específicas de una aplicación.

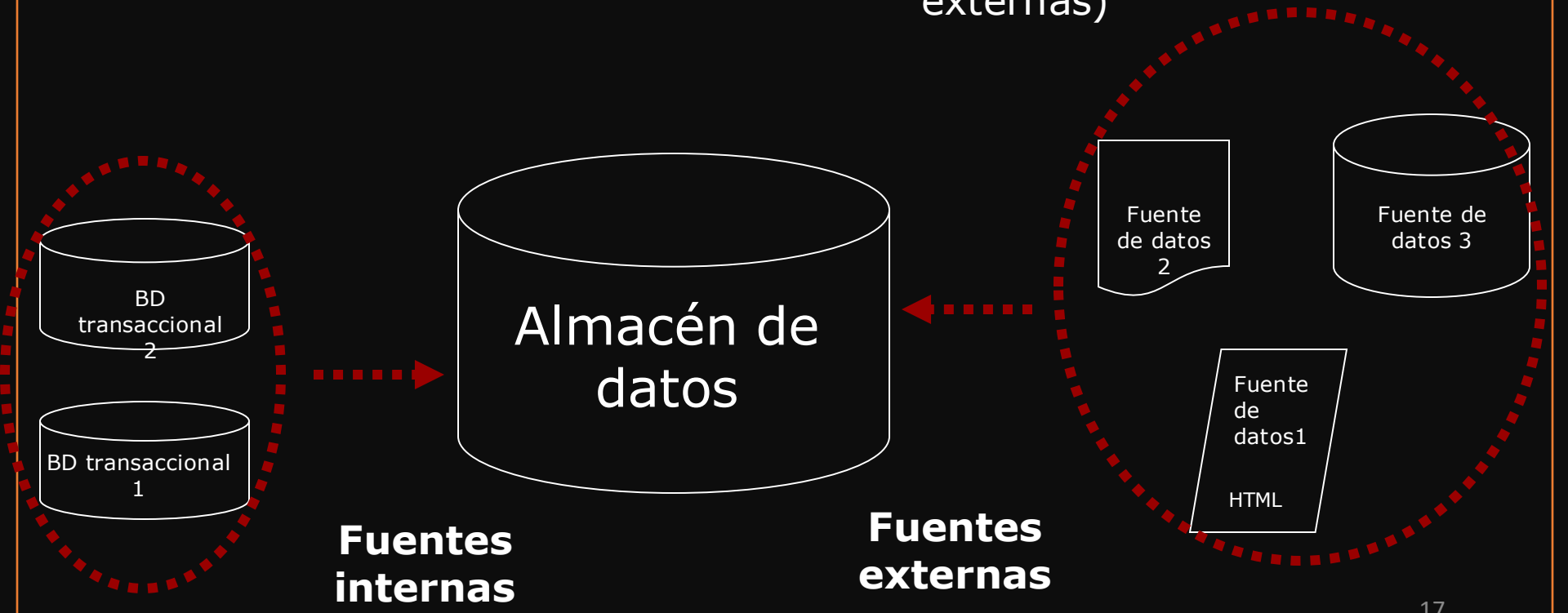
Se diseña para consultar de forma eficiente el tema de interés.



¿QUÉ ES UN DATA WAREHOUSE?

Integrada

Integra datos de varios sistemas operacionales de la organización(y/o fuentes externas)



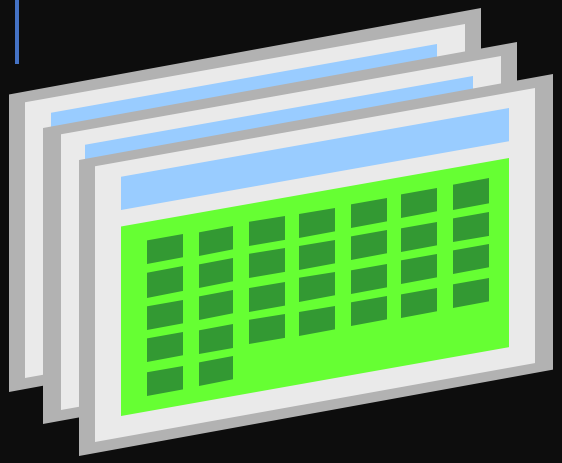
¿QUÉ ES UN DATA WAREHOUSE?

Variable en el tiempo

Se mantienen diferentes valores para la misma información y el momento en que ocurrieron los cambios en estos valores

Los datos son almacenados como fotos (snapshots) correspondientes a periodos de tiempo.

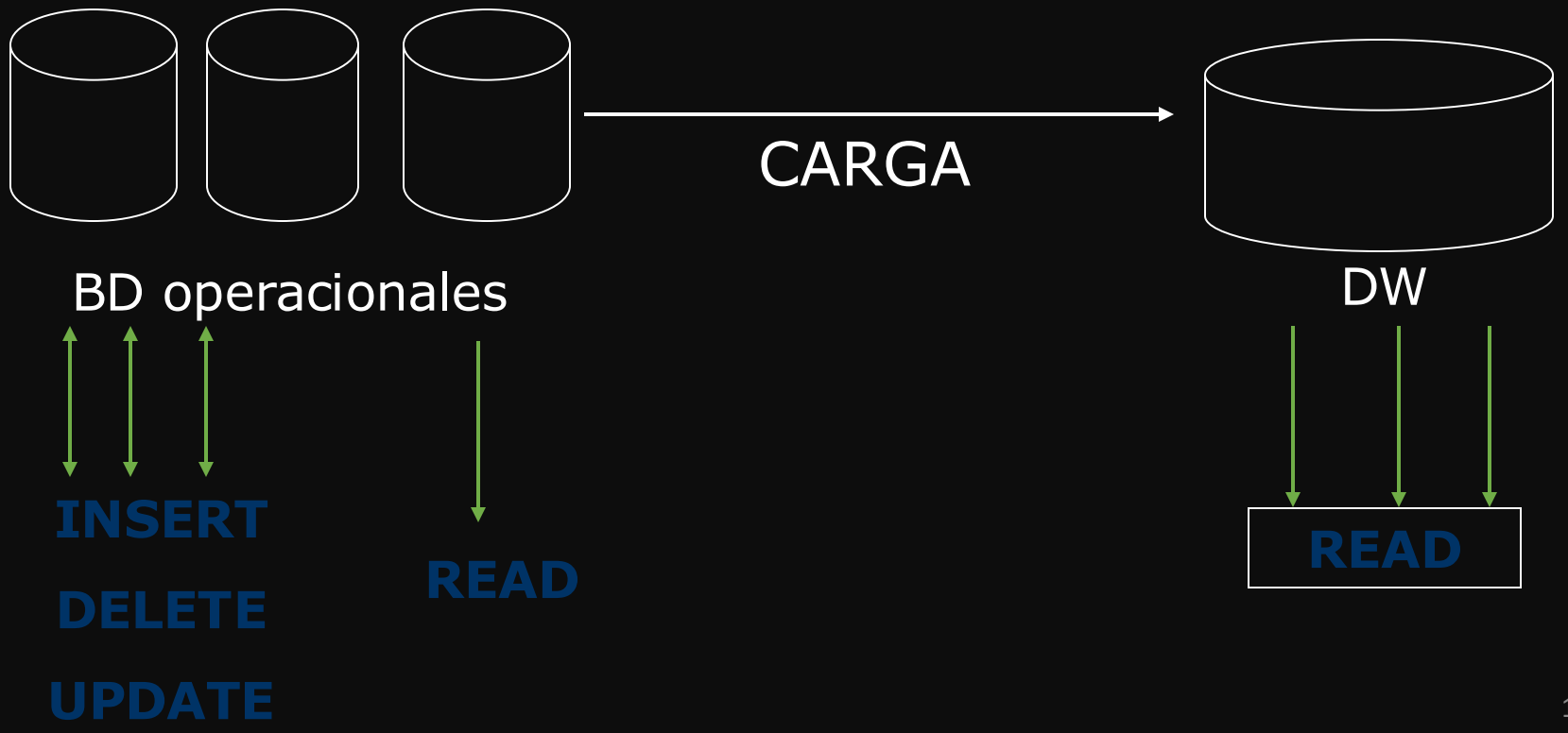
Tiempo	Datos
01/2022	Datos de Enero
02/2022	Datos de Febrero
03/2022	Datos de Marzo



¿QUÉ ES UN DATA WAREHOUSE?

No volátil

La modificación y eliminación de datos no está permitida. El DW es de un período de tiempo prolongado.



Data Warehouse

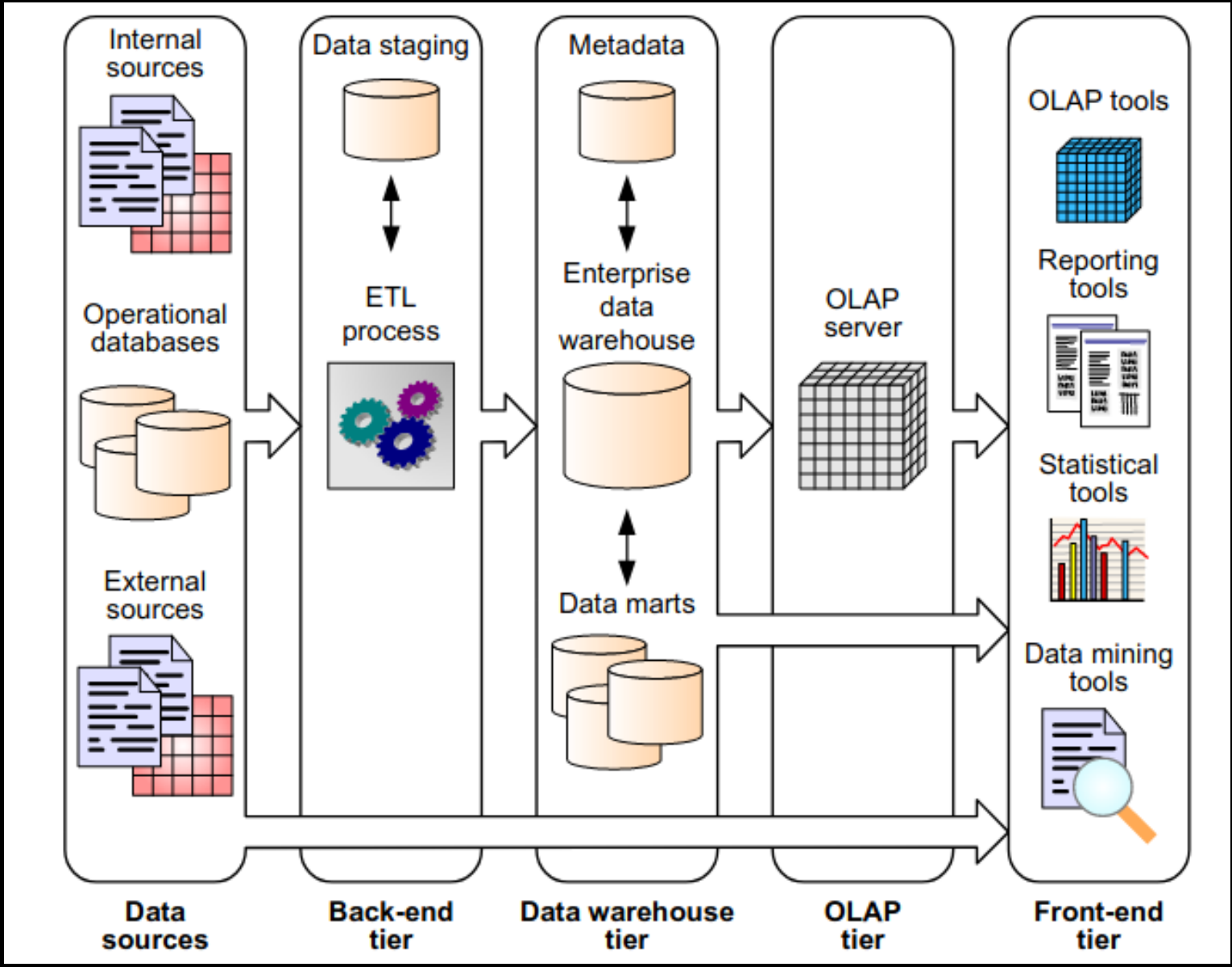
Se centran en los requisitos analíticos de los gerentes en varios niveles del proceso de toma de decisiones

- Estos temas varían según el tipo de actividades realizadas por la organización
  - Análisis de ventas en una empresa minorista
  - Análisis del comportamiento de los clientes en el uso de servicios bancarios
  - Análisis de la utilización de un sistema ferroviario en una empresa de transporte

BD Transaccional vs DW

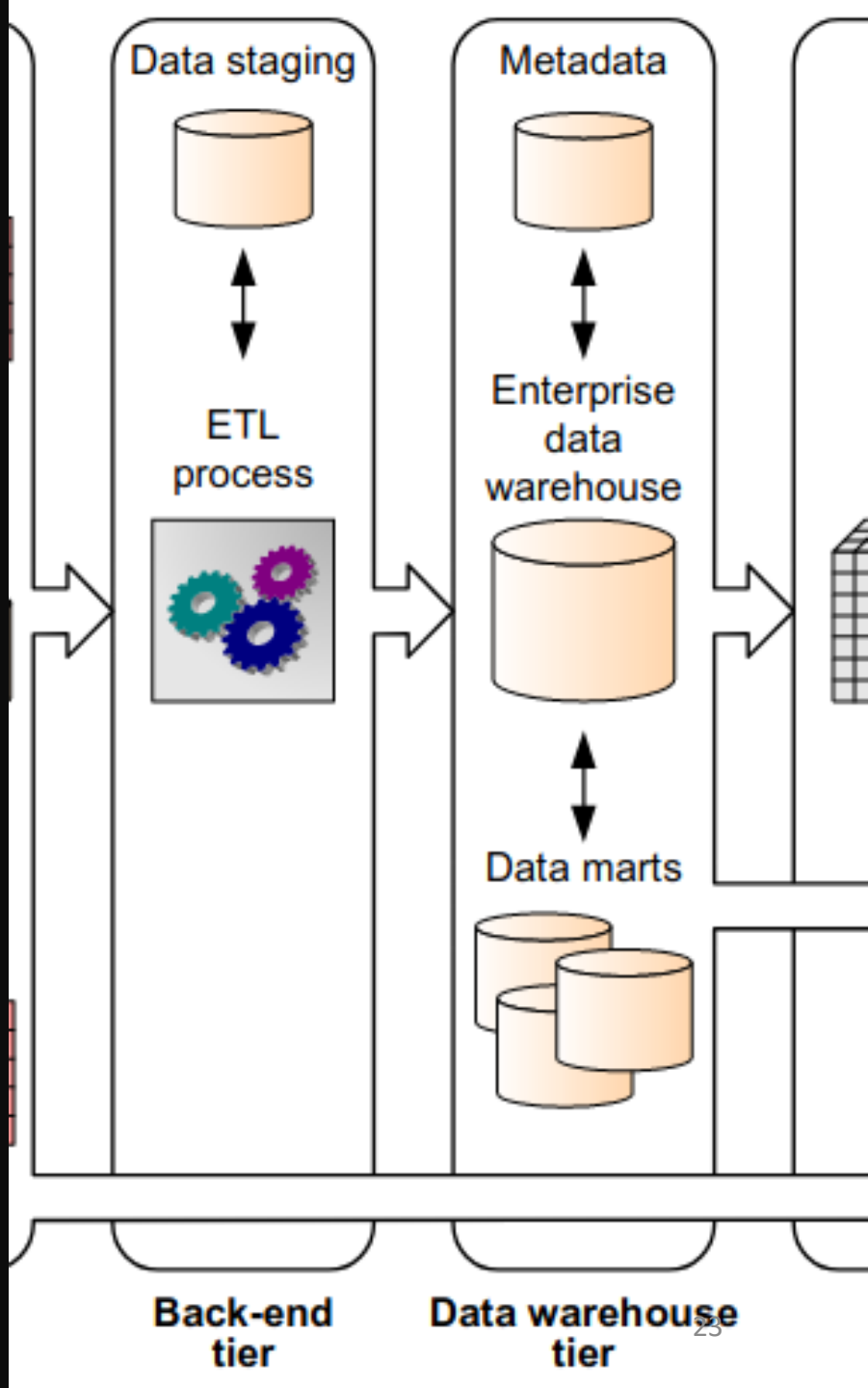
Aspect	Operational databases	Data warehouses
User type	Operators, office employees	Managers, high-ranking executives
Usage	Predictable, repetitive	Ad hoc, nonstructured
Data content	Current, detailed data	Historical, summarized data
Data organization	According to operational needs	According to the analysis problem
Data structures	Optimized for small transactions	Optimized for complex queries
Access frequency	High	From medium to low
Access type	Read, update, delete, insert	Read, append only
Number of records per access	Few	Many
Response time	Short	Can be long
Concurrency level	High	Low
Lock utilization	Necessary	Not necessary
Update frequency	High	None
Data redundancy	Low (normalized tables)	High (unnormalized tables)
Data modeling	ER model	Multidimensional model
Modeling and implementation	Entire system	Incremental

Arquitectura



ARQUITECTURA

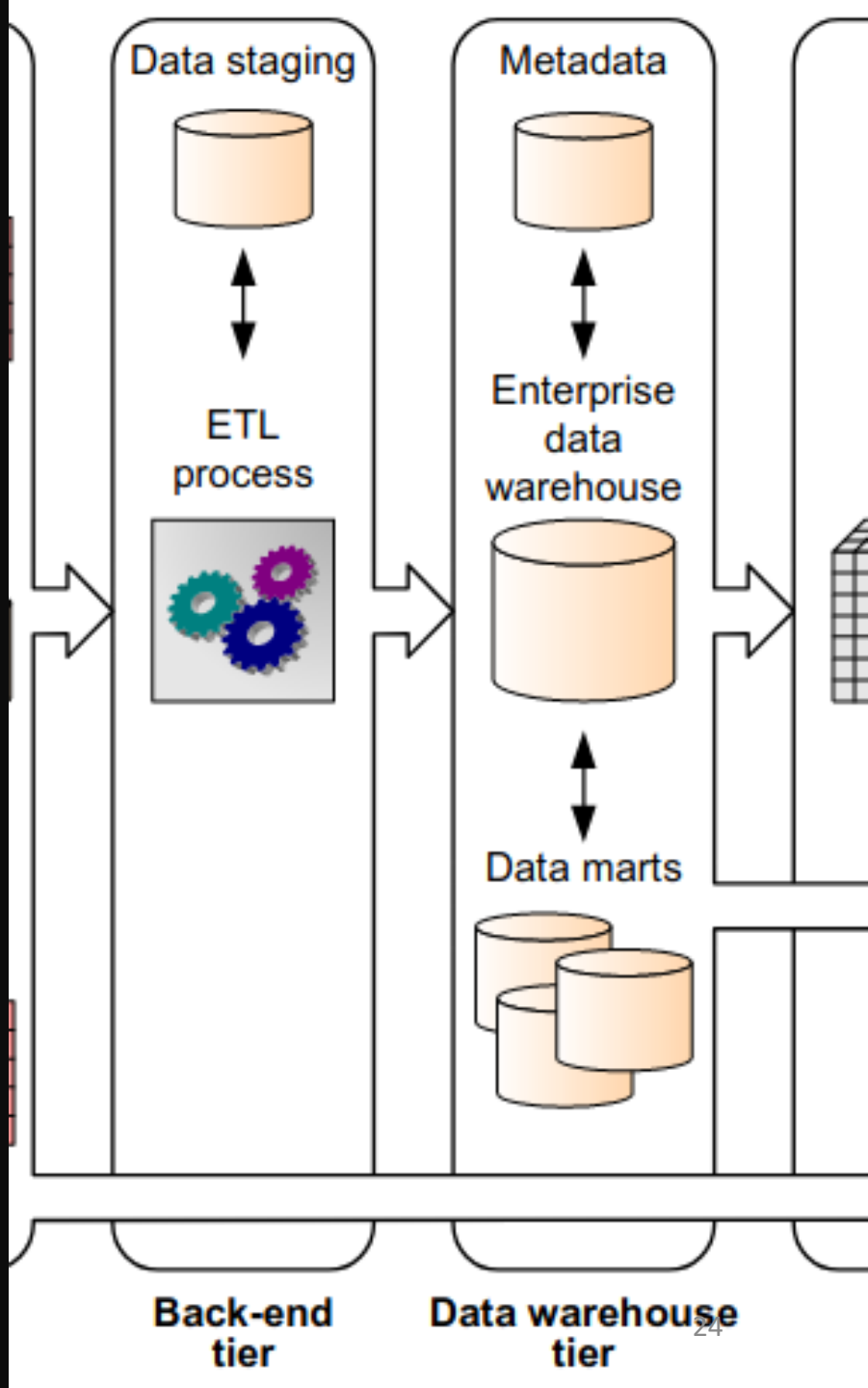
Extraction-transformation-loading (ETL):Realiza las funciones de *extracción* de las fuentes de datos (transaccionales o externas), *transformación* (limpieza,consolidación..) y *carga* del DW.



ARQUITECTURA

ETL:

- Incluye herramientas utilizadas para alimentar datos de bases de datos operativas y otras fuentes de datos
  - Problema de interoperabilidad: uso de ODBC, OLEDB, JDBC, etc.
  - Limpieza, integración y agregación
  - Carga inicial y posterior actualización
- Un área de preparación de datos puede ser una base de datos intermedia donde todos los procesos de integración y transformación de datos se ejecutan antes de la carga de los datos en el almacén de datos.

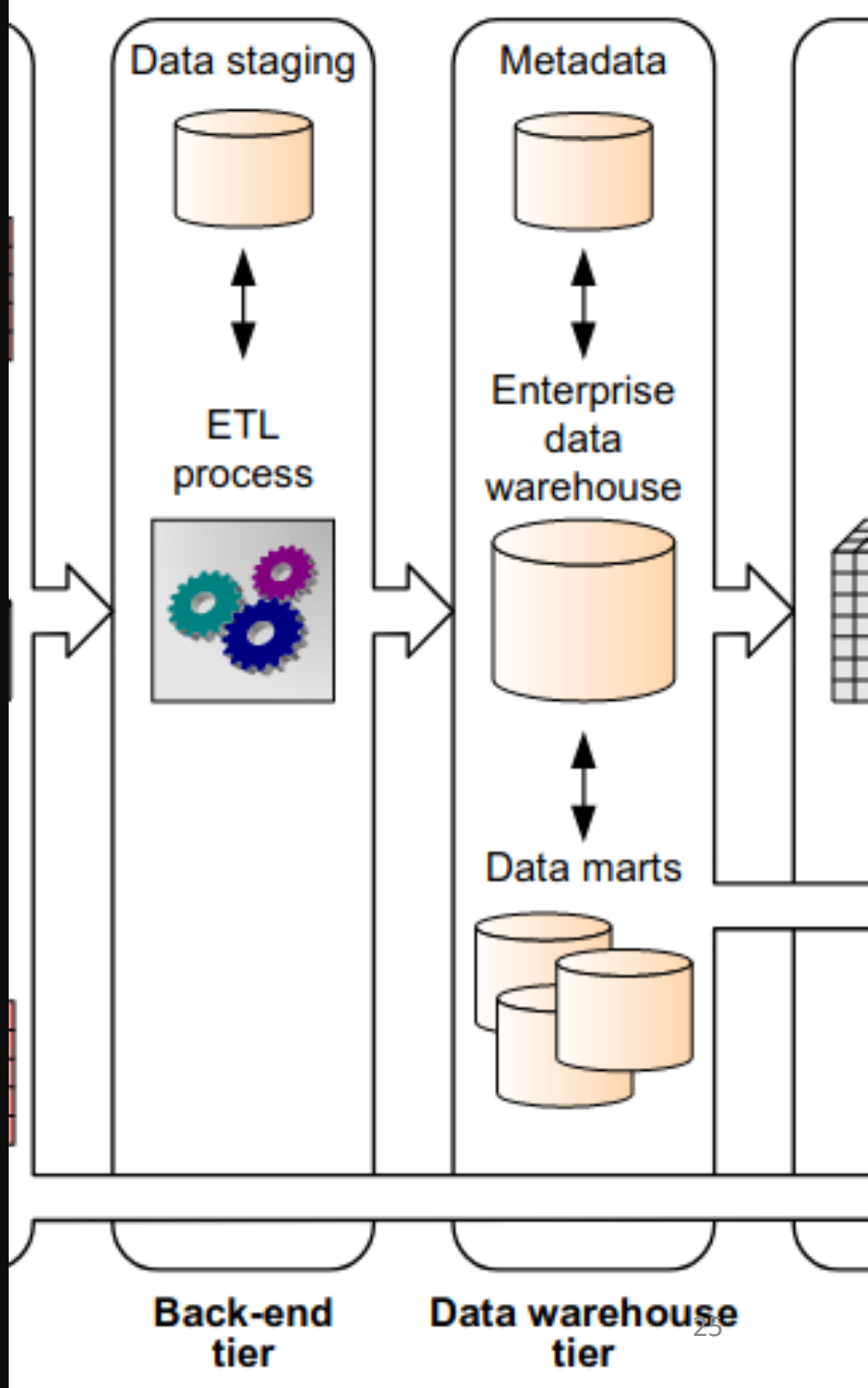


ARQUITECTURA

Un almacén de datos empresarial es un almacén de datos centralizado que incluye todas las áreas funcionales o departamentales de una organización

Varios mercados de datos (data marts):

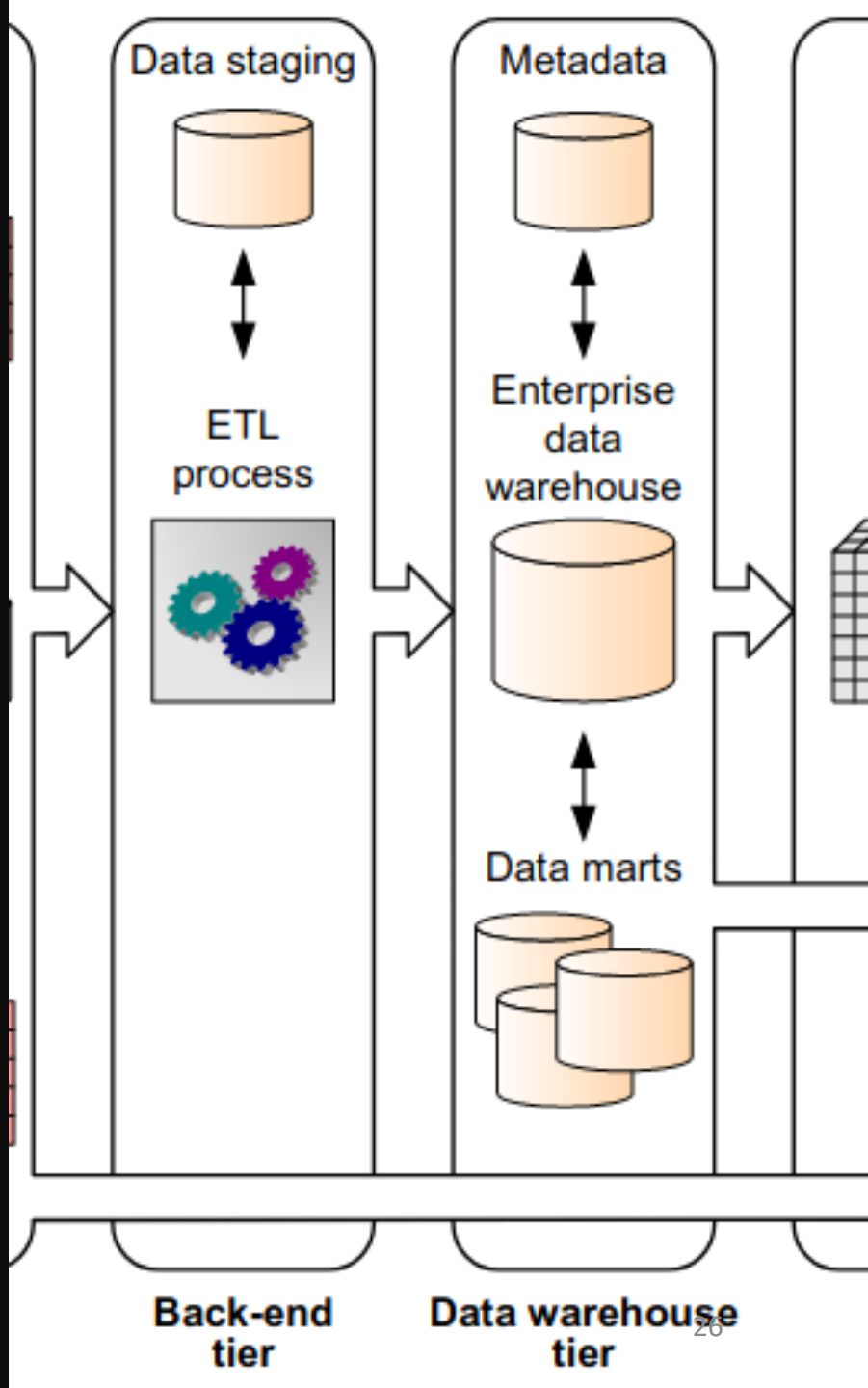
- Pequeño almacén de datos especializado relacionado con áreas comerciales específicas o con necesidades particulares de un grupo de usuarios del conocimiento que requieren solo un subconjunto de los datos contenidos en un DW
- Los datos se pueden derivar de DW o de los sistemas de origen
- Los data marts pueden estar físicamente colocados con el DW o pueden tener su propia plataforma separada



ARQUITECTURA

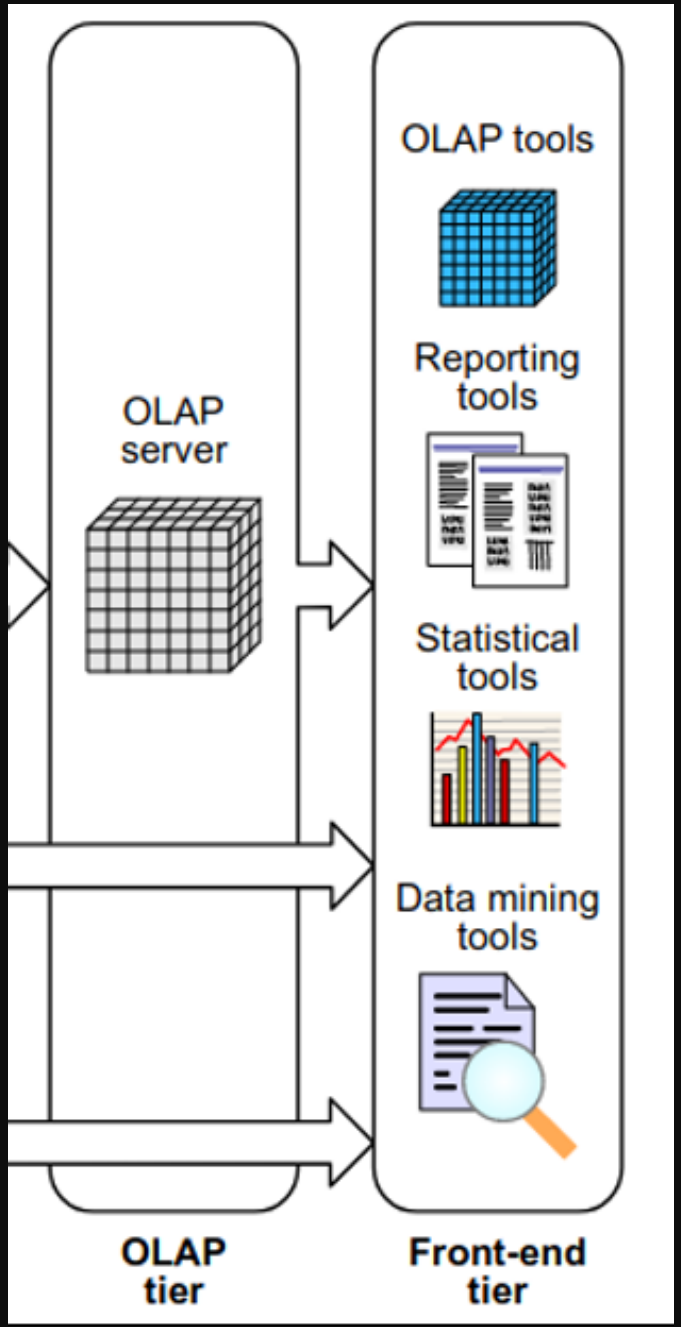
Un data mart también es un repositorio de metadatos que almacena información sobre el almacén de datos y su contenido

- Metadatos comerciales: describen el significado (o semántica) de los datos y las reglas, políticas y restricciones de la organización
- Metadatos técnicos: describen cómo se estructuran y almacenan los datos en un sistema informático, y las aplicaciones y procesos que manipulan los datos
    - Descripción de la estructura del DW y los data marts a nivel lógico y físico, así como información relacionada con la seguridad y monitoreo
    - Descripción de las fuentes de datos incluyendo su titularidad, frecuencias de actualización, limitaciones legales, métodos de acceso y otros
    - Descripción del proceso ETL que incluye tipos de datos, extracción de datos, limpieza, reglas de transformación y valores predeterminados, reglas de actualización y depuración de datos, algoritmos para resumir o unir datos, etc.



ARQUITECTURA

Interfaces y Operaciones de Consulta: Permiten acceder a los datos y sobre ellos se conectan herramientas más sofisticadas (OLAP, EIS - Executive Informations System, minería de datos).

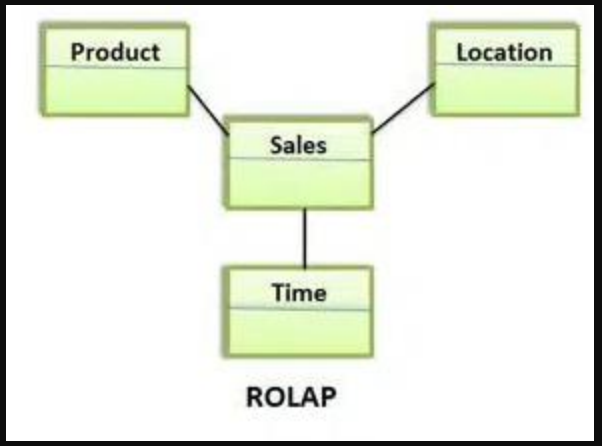


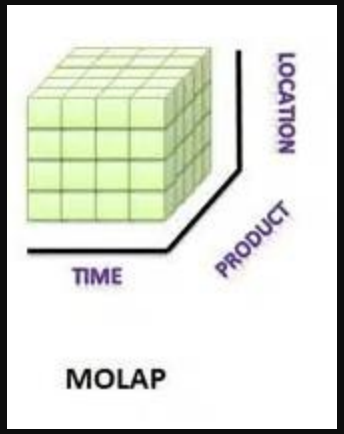
ARQUITECTURA

Un servidor OLAP que admite operaciones y datos multidimensionales

Diferentes modelos de datos físicos:

- ROLAP (OLAP relacional): almacena datos en bases de datos relacionales y admite extensiones de SQL
- MOLAP (OLAP multidimensional) almacena directamente datos en estructuras de datos especiales (por ejemplo, matrices) e implementa las operaciones OLAP sobre esas estructuras de datos.
- HOLAP (OLAP híbrido): combina ambas tecnologías, beneficiándose de la capacidad de almacenamiento de ROLAP y las capacidades de procesamiento de MOLAP





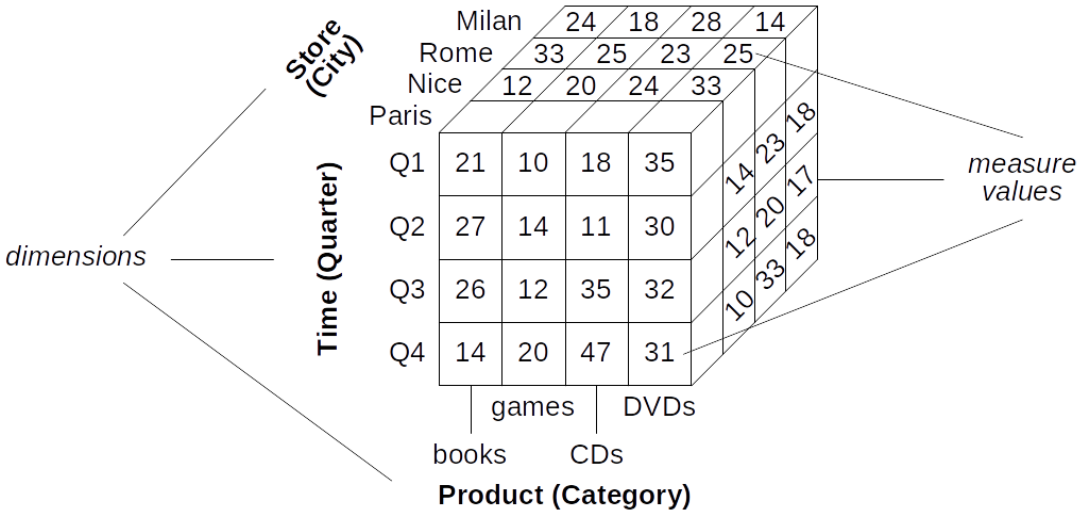
Modelo multidimensional

Los DW y OLAP utilizan una vista multidimensional de datos

Representado a nivel abstracto como un cubo de datos o un hipercubo

- Dimensiones: perspectivas para el análisis de datos, define las dimensiones naturales de los datos

  Celdas (hechos): contienen medidas, valores que se analizaran, valores que fueron medidos



¿Qué es el modelo multidimensional?

Vista lógica de la empresa

Muestra las principales entidades del negocio empresarial y las relaciones entre ellas

No está vinculado a tablas y bases de datos físicas

No es un diagrama ER

¿Qué es el modelo multidimensional?

"Simply speaking, the database (operational) systems are where you put the data in, and the Data warehouse (Business Intelligence) system is where you get the data out." — Ralph Kimball

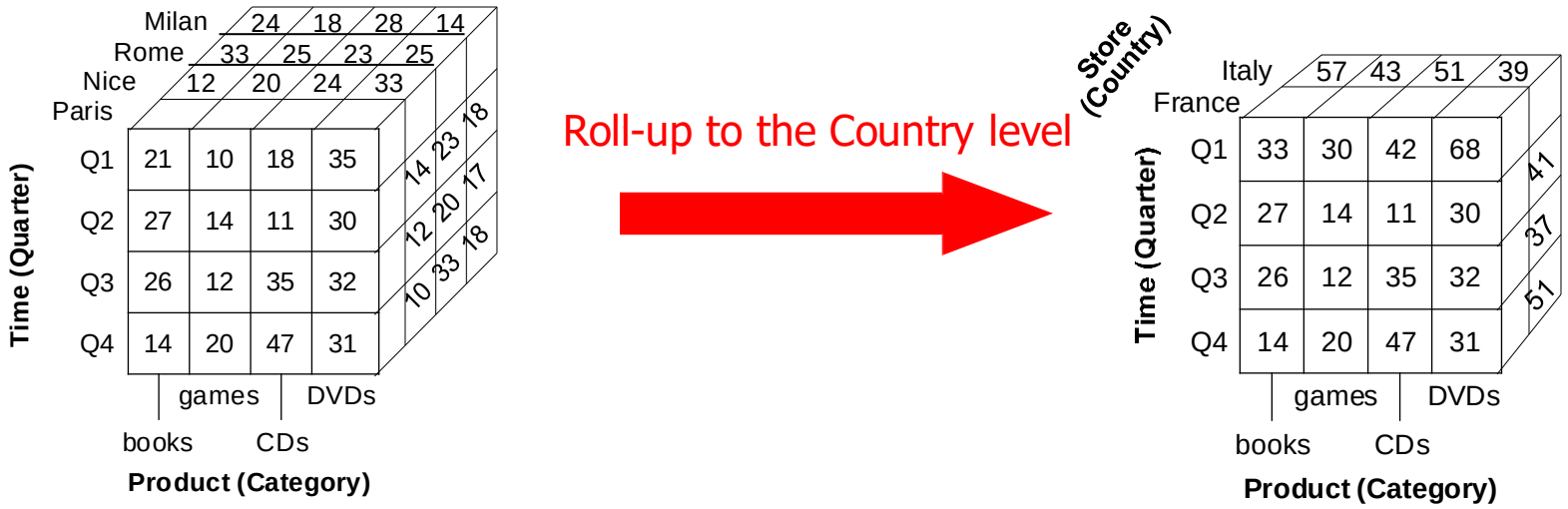
Modelo Multidimensional

El esquema multidimensional está especialmente diseñado para modelar sistemas de almacenamiento de datos.

Los esquemas están diseñados para abordar las necesidades únicas de bases de datos muy grandes diseñadas con fines analíticos (OLAP).

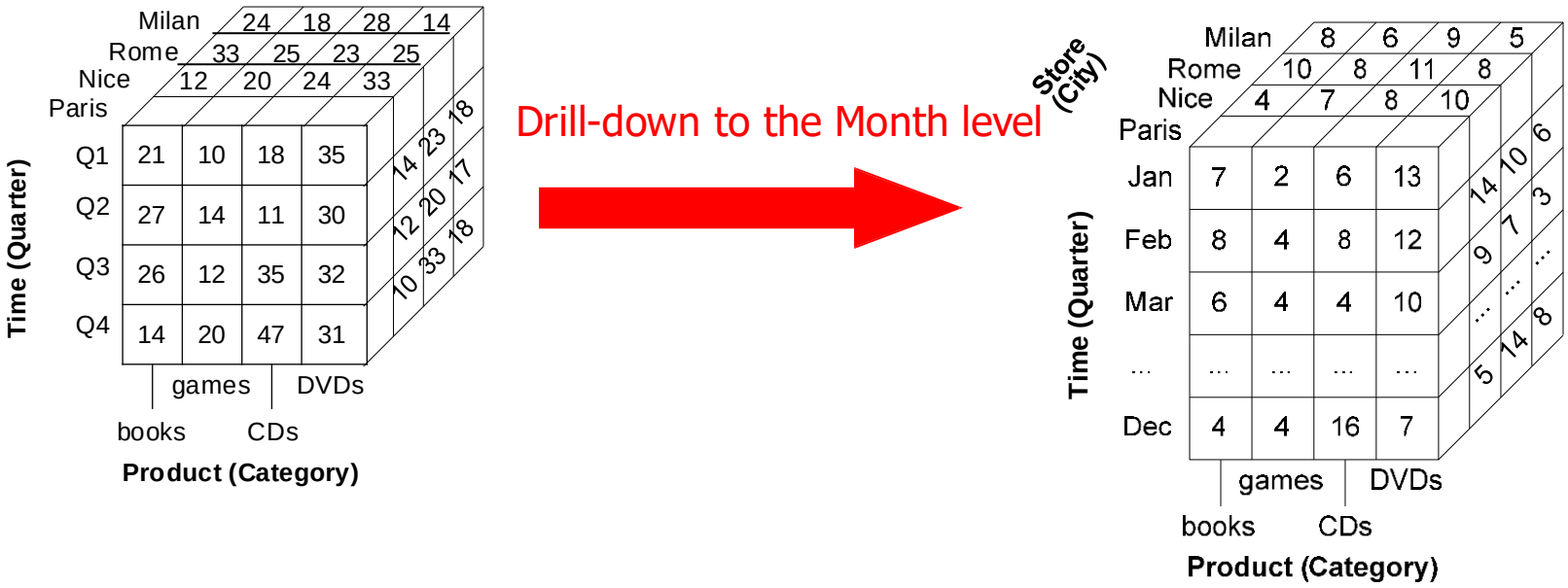
Modelo Multidimensional

- En un cubo se definen varias operaciones:
  - Roll-up: transforma medidas detalladas en medidas resumidas cuando se asciende en una jerarquía



Modelo Multidimensional

- En un cubo se definen varias operaciones:
  - Drill-down: realiza la operación opuesta a la operación roll-up, es decir, se mueve de un nivel más general a un nivel detallado en una jerarquía

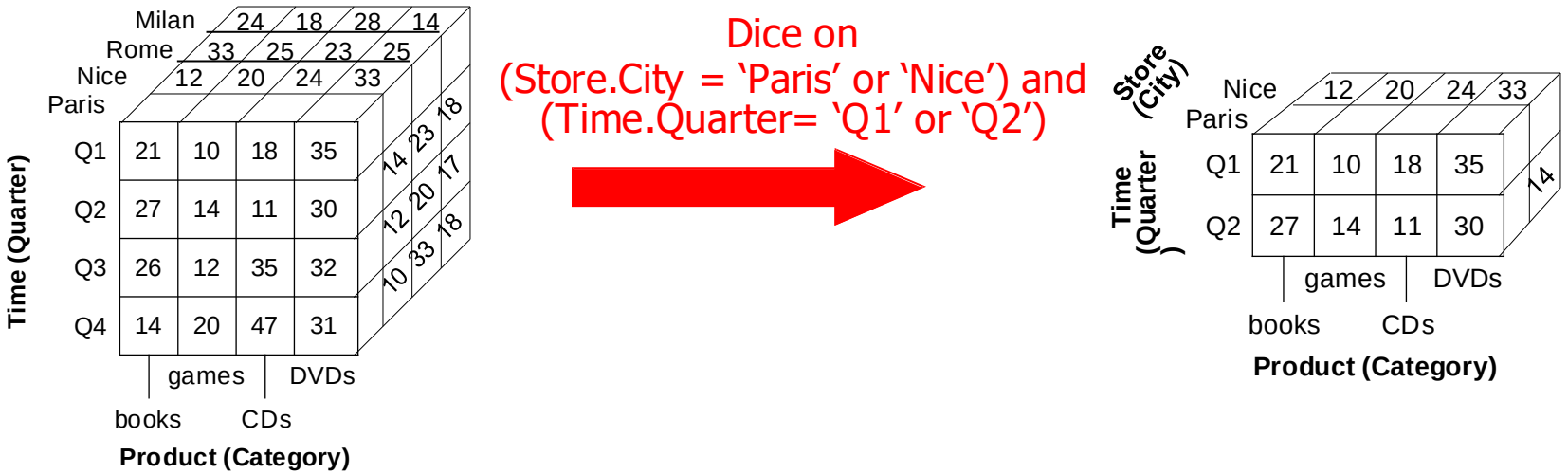


Modelo Multidimensional

- En un cubo se definen varias operaciones:
  - Slice: realiza una selección en una dimensión de un cubo, lo que da como resultado un subcubo

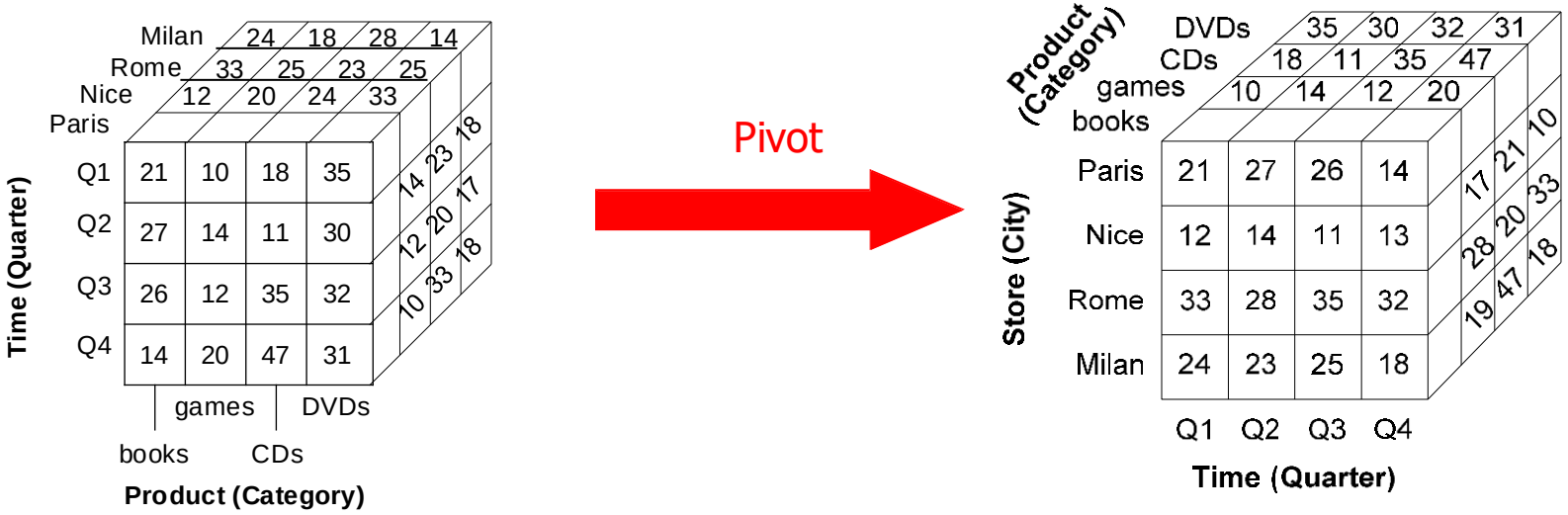
Modelo Multidimensional

- En un cubo se definen varias operaciones:
  - Dice: define una selección en dos o más dimensiones, definiendo así nuevamente un subcubo



Modelo Multidimensional

- En un cubo se definen varias operaciones:
  - Pivot (Rotate): gira los ejes de un cubo para proporcionar una presentación alternativa de datos



Modelo Multidimensional

Cuenta de Estudiantes	Sexo		
Distribución geográfica	Femenino	Masculino	Total general
⊞ALAJUELA	1	22	23
⊞CARTAGO	3	17	20
⊞GUANACASTE	2	2	4
⊞HEREDIA	3	15	18
⊞LIMÓN		3	3
⊞PUNTARENAS	1	3	4
⊞SAN JOSÉ	12	61	73
Total general	22	123	145

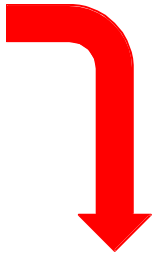
Drill-down

Roll-up

Cuenta de Estudiantes	Sexo		
Distribución geográfica	Femenino	Masculino	T
⊞ALAJUELA	1	22	
⊞CARTAGO	3	17	
⊟GUANACASTE	2	2	
ACADEMIA TEOCALI		1	
COLEGIO EXP. BILINGUE DE SANTA CRUZ		1	
CTP DE HOJANCHA	1		
LICEO MAURILIO ALVARADO (TILARAN)	1		
⊞HEREDIA	3	15	
⊟LIMÓN		3	
COL.EXPERIMENTAL BILINGUE DE POCOCI		1	
COLEGIO VALLE DEL SOL		1	
LICEO EXPER. BILINGÜE DE RIO JIMENEZ		1	
⊞PUNTARENAS	1	3	
⊞SAN JOSÉ	12	61	
Total general	22	123	

Modelo Multidimensional

Cuenta de Estudiantes	Sexo		
Distribución geográfica	Femenino	Masculino	Total general
ALAJUELA	1	22	23
CARTAGO	3	17	20
GUANACASTE	2	2	4
HEREDIA	3	15	18
LIMÓN		3	3
PUNTARENAS	1	3	4
SAN JOSÉ	12	61	73
Total general	22	123	145



Slice

Cuenta de Estudiantes	Sexo		
Distribución geográfica	Femenino	Masculino	Total general
ALAJUELA	1	22	23
LIMÓN		3	3
Total general	1	25	26

Modelo Multidimensional

Cuenta de Estudiantes	Sexo		
Distribución geográfica	Femenino	Masculino	Total general
⊞ALAJUELA	1	22	23
⊞CARTAGO	3	17	20
⊞GUANACASTE	2	2	4
⊞HEREDIA	3	15	18
⊞LIMÓN		3	3
⊞PUNTARENAS	1	3	4
⊞SAN JOSÉ	12	61	73
Total general	22	123	145

Dice

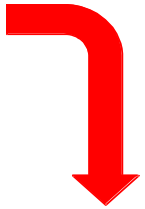
Cuenta de Estudiantes	Sexo	
Distribución geográfica	Masculino	Total general
⊞ALAJUELA	22	22
⊞LIMÓN	3	3
Total general	25	25

Dice with drill-down

Cuenta de Estudiantes	Sexo	
Distribución geográfica	Masculino	Total ge
⊟ALAJUELA	22	
COLEGIO BILINGUE DE SAN RAMON	1	
COLEGIO BILINGÜE SANTA FE	1	
COLEGIO DE NARANJO	1	
COLEGIO EL CARMEN	1	
COLEGIO EXP. BILINGUE DE PALMARES	1	
COLEGIO MARISTA	3	
COLEGIO PATRIARCA SAN JOSE	1	
COLEGIO REDENTORISTA SAN ALFONSO	5	
CTP JESUS OCAÑA ROJAS	1	
Cualquier Colegio del Exterior	1	
INSTIT. CENTROAMERICANO ADVENTISTA	1	
INSTITUTO SUPERIOR JULIO ACOSTA	1	
LICEO DE ATENAS	1	
LICEO LEON CORTES CASTRO	1	
LICEO OTILIO ULATE BLANCO	2	
⊞LIMÓN	3	
Total general	25	

Modelo Multidimensional

Cuenta de Estudiantes	Sexo		
Distribución geográfica	Femenino	Masculino	Total general
ALAJUELA	1	22	23
CARTAGO	3	17	20
GUANACASTE	2	2	4
HEREDIA	3	15	18
LIMÓN		3	3
PUNTARENAS	1	3	4
SAN JOSÉ	12	61	73
Total general	22	123	145

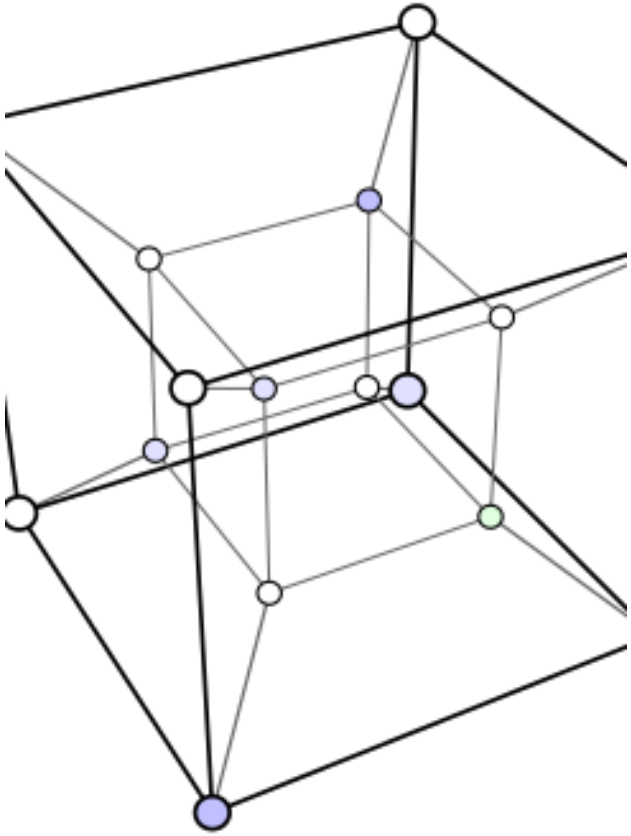


Pivot

Cuenta de Estudiante	Distribución geográfi							
	ALAJUELA	CARTAGO	GUANACASTE	HEREDIA	LIMÓN	PUNTARENAS	SAN JOSÉ	Total gener
Sexo								
Femenino	1	3	2	3		1	12	
Masculino	22	17	2	15	3	3	61	1
Total general	**23**	**20**	**4**	**18**	**3**	**4**	**73**	**1**

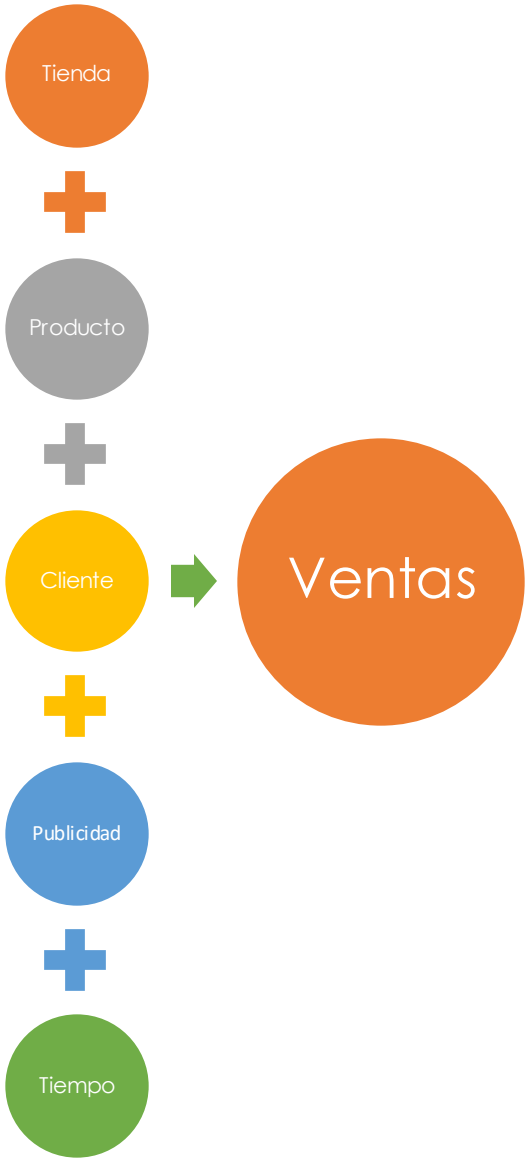
Modelo Multidimensional

- Sin embargo, los almacenes de datos y los cubos OLAP pueden tener muchas dimensiones (hipercubo), ¿cómo representarlo?
  - Teseracto – es el análogo en cuatro dimensiones del cubo
- Además, los almacenes de datos también se implementan generalmente como bases de datos relacionales.
- La práctica habitual es utilizar esquemas de estrella o copo de nieve que permiten distinguir
  - El objeto de análisis
  - Las diferentes perspectivas de análisis.



Modelo Multidimensional

Modelo Multidimensional



Modelo Multidimensional

- Existen diferentes propuestas de esquemas multidimensionales, entre ellas tenemos:
  - Esquema de estrella (Start schema)
  - Esquema de copo de nieve (Snowflake Schema)

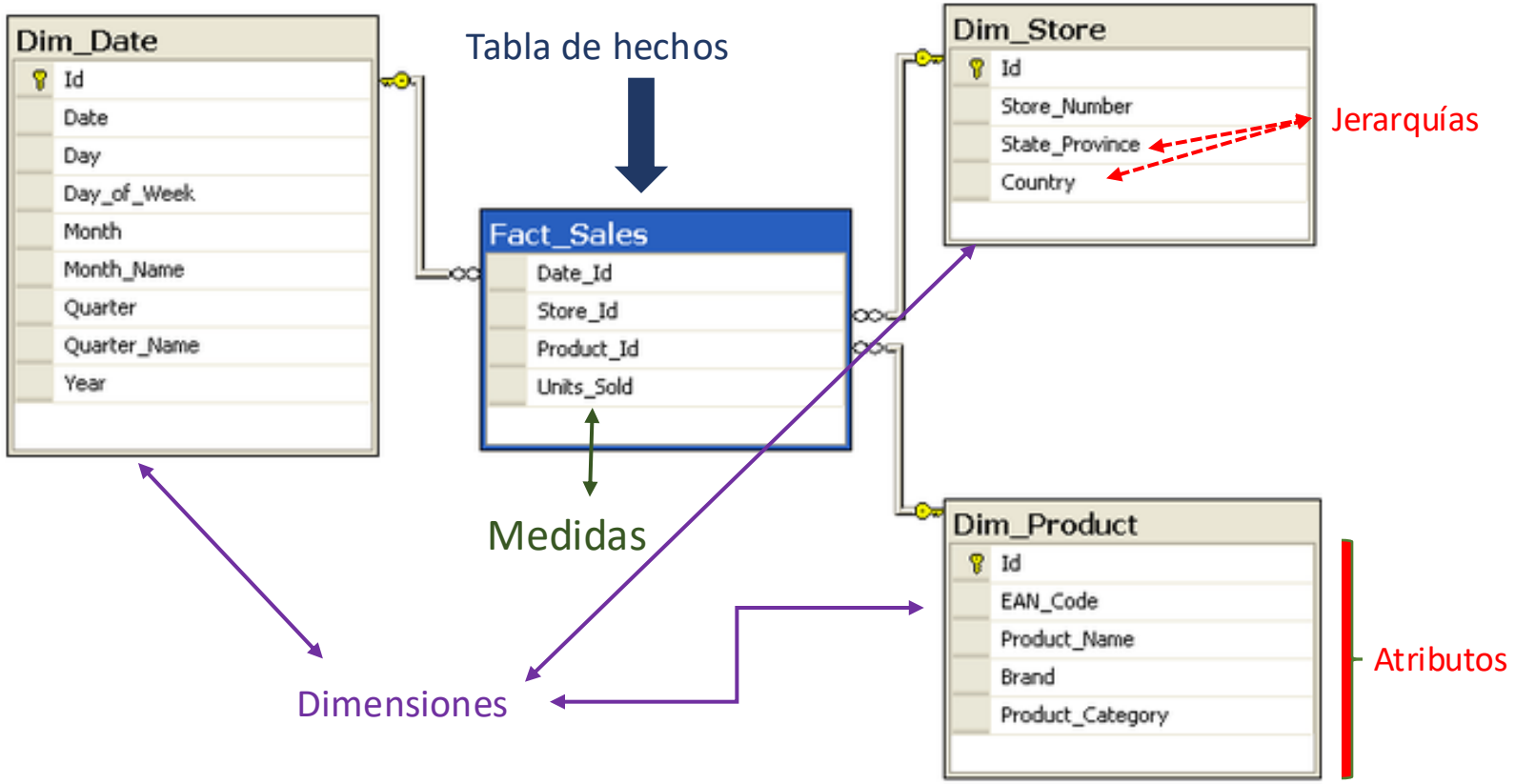
Modelo Multidimensional

- Los esquemas de estrella y copo de nieve incluyen
  - Tabla de hechos (fact) con atributos llamados medidas
  - Tablas de dimensiones (dimension) con
    - Atributos descriptivos
    - Atributos que forman jerarquías

Esquema estrella

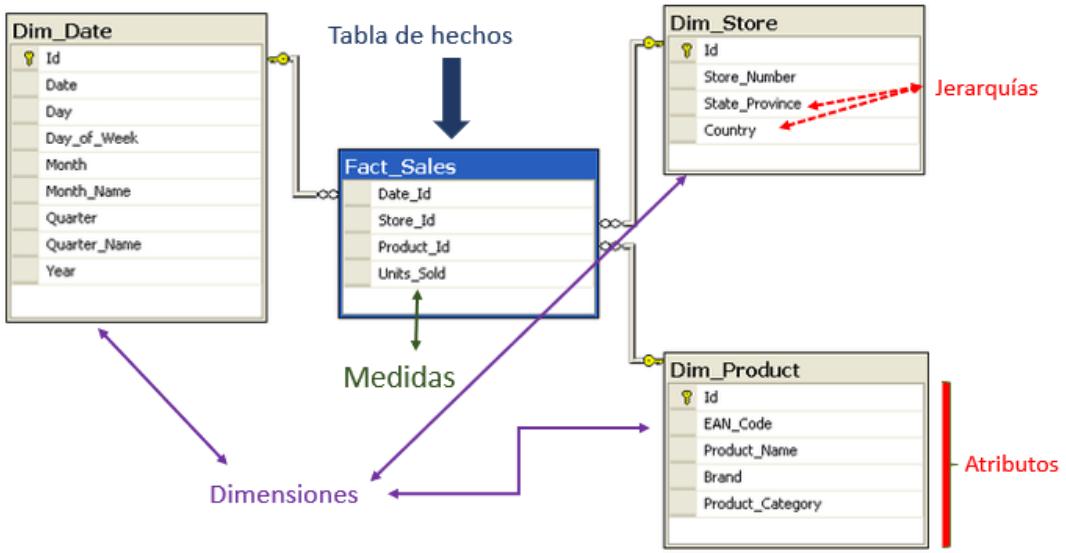
- Normalmente en el centro de la estrella puede tener una tabla de hechos y varias tablas de dimensiones asociadas. Se le conoce como esquema de estrella porque su estructura se asemeja a una estrella.
- Es el tipo más simple de esquema de almacén de datos. También se conoce como esquema de unión en estrella y está optimizado para consultar grandes conjuntos de datos.

Esquema estrella



Esquema estrella

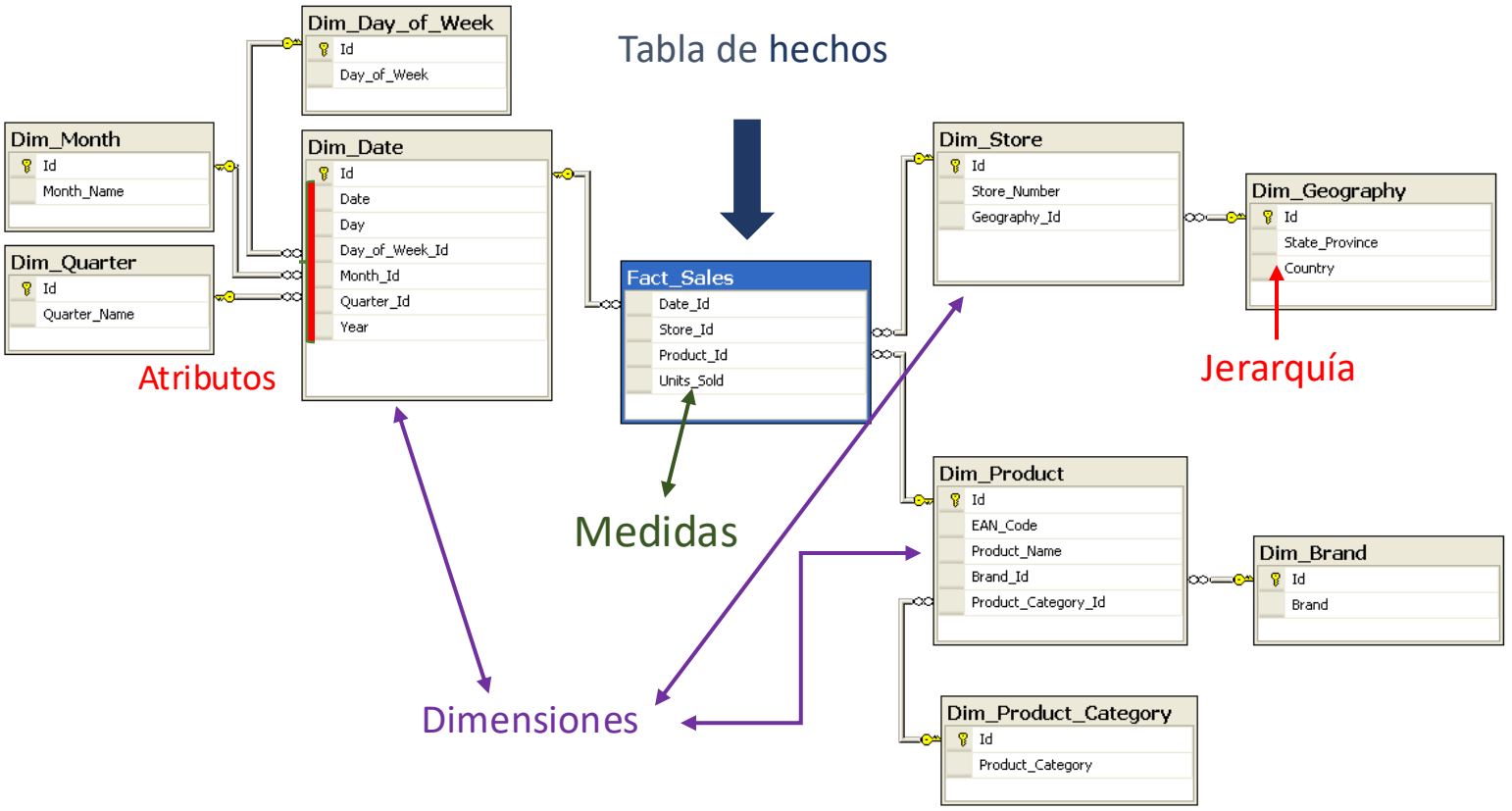
- Cada dimensión en un esquema en estrella se representa con la única tabla de una dimensión
- La tabla de dimensiones debe contener el conjunto de atributos
- La tabla de dimensiones se une a la tabla de hechos mediante una clave externa
- Las tablas de dimensiones no están unidas entre sí
- La tabla de hechos contendría la clave y la medida



Esquema copo de nieve

- Es una disposición lógica de tablas en una base de datos multidimensional de manera que el diagrama ER se asemeja a la forma de un copo de nieve
- Un esquema de copo de nieve es una extensión de un esquema de estrella y agrega dimensiones adicionales
- Las tablas de dimensiones están normalizadas, lo que divide los datos en tablas adicionales.

Esquema copo de nieve

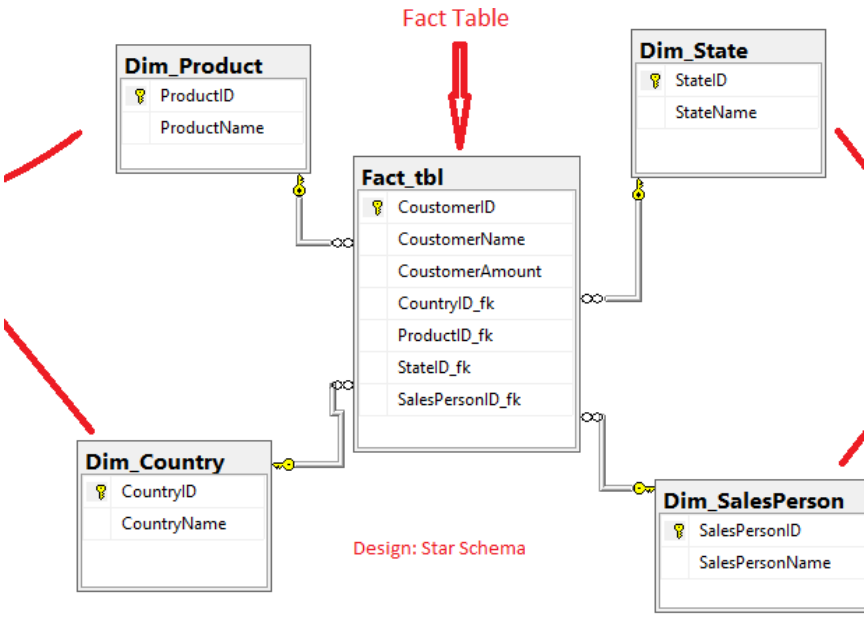


Esquema copo de nieve

- El principal beneficio del esquema de copo de nieve es que utiliza menos espacio en disco
- Se agrega una dimensión al esquema que resulta muy fácil de implementar
- Debido a las múltiples tablas, el rendimiento de las consultas se reduce
- El principal desafío al que se enfrentará al utilizar el esquema de copo de nieve es que debe realizar más esfuerzos de mantenimiento debido a la mayor cantidad de tablas de búsqueda

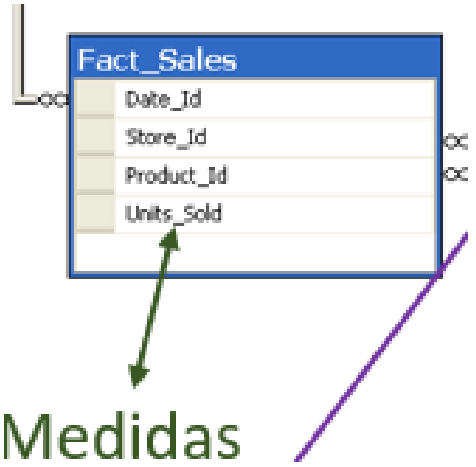
Tabla de hechos (Fact table)

- Vincula varias dimensiones
- Representa el foco de análisis
- Ejemplo: ventas en tiendas, cuentas de clientes, accidentes en carretera



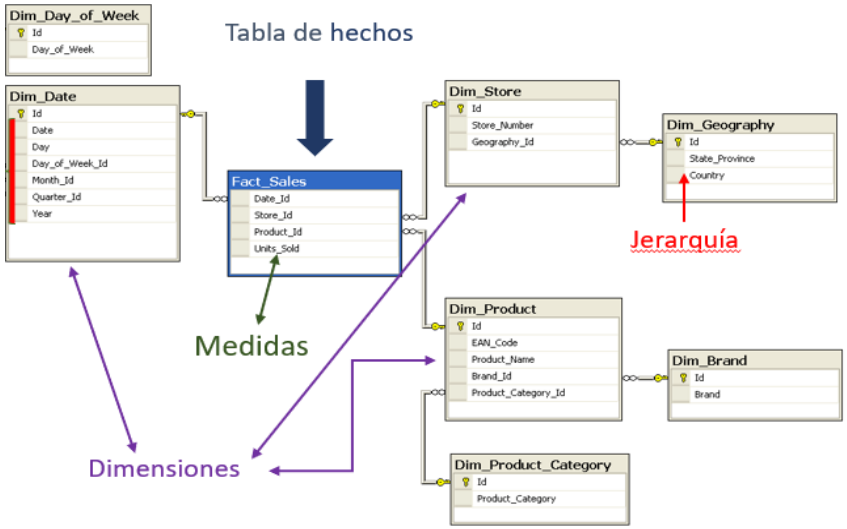
Medidas (Measures)

- Por lo general son valores numéricos utilizados para realizar una evaluación cuantitativa de varios aspectos de una organización
- Ejemplo: monto o cantidad de ventas, saldo promedio de la cuenta, monto del seguro
- Pueden ser de diferentes tipos:
  - Aditivo: se puede sumar, p. Ej., Cantidad de ventas
  - Semiaditivo: no siempre se puede sumar, p. Ej., Niveles de inventario durante diferentes períodos de tiempo
  - No aditivo: no se puede sumar, p. Ej., Precio unitario
- Se puede calcular
- Puede estar ausente de la tabla de hechos



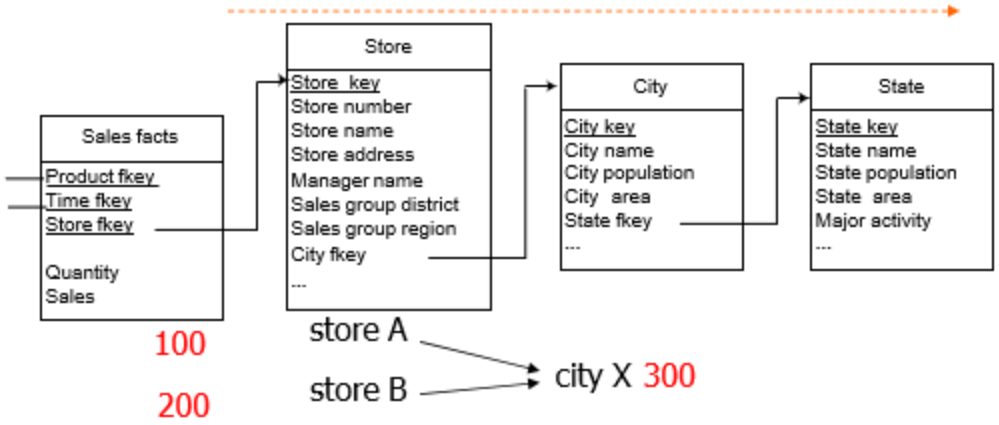
Dimensiones (dimensions)

- Se utiliza para ver las medidas desde diferentes perspectivas
- Ejemplo:
  - Se puede usar una dimensión de tiempo para analizar cambios en las ventas durante varios períodos de tiempo
  - Se puede usar una dimensión de ubicación para analizar las ventas de acuerdo con la distribución geográfica de las tiendas
- Deben incluir valores descriptivos (sin codificación)
- Pueden requerir representar cambios en el tiempo: los usuarios pueden combinar varias perspectivas de análisis diferentes (es decir, dimensiones) de acuerdo con sus necesidades
- Ejemplo:
  - Un usuario puede requerir información sobre las ventas de accesorios de computadora (la dimensión del producto) en agosto de 2022 (la dimensión del tiempo) en todas las ubicaciones de la tienda (la dimensión de la tienda)



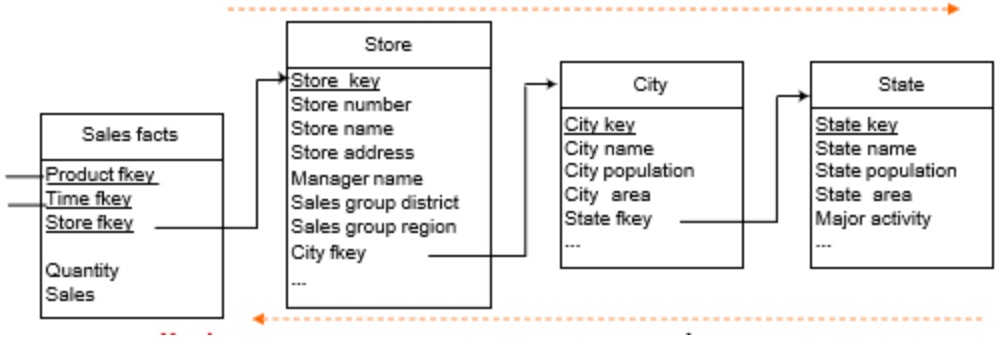
Jerarquías (Hierarchies)

- Cada dimensión puede tener varias jerarquías
- Dos operaciones requieren jerarquías:
  - Roll-up: transforma medidas detalladas en datos resumidos



Jerarquías (Hierarchies)

- Cada dimensión puede tener varias jerarquías
- Dos operaciones requieren jerarquías:
  - Roll-up: transforma medidas detalladas en datos resumidos



  - Drill-down: convierte medidas agregadas en datos más detallados

Jerarquías (Hierarchies)

- Cada dimensión puede tener varias jerarquías
- Ejemplos: jerarquías en la dimensión Tienda:
  - Nombre de la tienda – Nombre de la ciudad – Nombre del estado
  - Nombre de la tienda – Distrito del grupo de ventas – Región del grupo de ventas
- Las jerarquías se pueden incluir en esquemas de estrella y copo de nieve
- Permiten a los usuarios explotar medidas en varios niveles de detalle

Diferencias entre Estrella y copo de nieve

Star Schema

- Las jerarquías de las dimensiones se almacenan en la tabla de dimensiones
- Contiene una tabla de hechos rodeada de tablas de dimensiones
- Alto nivel de redundancia de datos
- El procesamiento de cubos es más rápido

Snowflake Schema

- Las jerarquías se dividen en tablas independientes
- Una tabla de hechos rodeada por una tabla de dimensiones que a su vez está rodeada por una tabla de dimensiones
- Redundancia de datos muy baja
- El procesamiento del cubo puede ser lento debido a "joins" complejos

Ventajas del esquema estrella

- El esquema es fácil de entender para la formulación de consultas
- Puede accederse fácilmente
- Se necesitan pocas uniones para expresar consultas, debido al alto nivel de desnormalización

Desventajas del esquema estrella

- No permite modelar adecuadamente las jerarquías, ya que la estructura de la jerarquía no es clara y se requiere información adicional para representarla, es decir, conocimiento semántico sobre la aplicación.
- Es difícil asociar claramente los atributos dentro de sus niveles correspondientes
- En las jerarquías que tienen muchos niveles, la cantidad de atributos es al menos tan grande como la profundidad de la jerarquía, lo que dificulta la comprensión de la estructura de la jerarquía
- Está desnormalizado, incurriendo en un alto nivel de redundancia de datos

Ventajas del esquema copo de nieve

- Representa mejor las estructuras jerárquicas
- Los niveles se pueden reutilizar entre diferentes jerarquías
- Altamente normalizado que conduce a ninguna repetición de datos
- Puede gestionar fácilmente la heterogeneidad entre niveles, es decir, permite que diferentes niveles de una jerarquía incluyan atributos específicos

Desventajas del esquema copo de nieve

- Requiere operaciones de unión para recuperar diferentes niveles de jerarquía
- No permite a los diseñadores representar diferentes tipos de jerarquías existentes en situaciones del mundo real, por ejemplo, no se puede usar sin modificaciones si hay heterogeneidad dentro de un nivel, es decir, diferentes características de los productos

Conceptos de diseño: granularidad

- Granularidad de la medida:
  - El nivel de detalle en el que se representan las medidas.
  - Está relacionado con la granularidad de los datos de dimensión involucrados en la tabla de hechos
    - Por ejemplo: si las medidas deben considerarse diariamente, la dimensión de tiempo debe incluir registros que representen cada día
  - Es importante porque determina
    - El volumen de datos
    - El tiempo de respuesta de la consulta
    - El tipo de consultas

Conceptos de diseño: granularidad

- La decisión sobre el nivel de granularidad depende de
  - Tipo de consulta y frecuencia
  - Limitaciones de almacenamiento
  - Tiempo de respuesta de la consulta
- Muy a menudo es necesario considerar diferentes niveles de granularidad e incluir datos detallados y agregados previamente.

Ventajas del modelo multidimensional

- Garantiza una mejor comprensión de los datos para fines de análisis:
  - Representa claramente el foco de análisis y los datos numéricos (medidas) de interés
  - Permite distinguir datos que cambian con frecuencia (medidas) de objetos más estáticos (dimensiones)
- Tiene una estructura más simple que el modelo relacional tradicional
- Facilita la formulación de consultas y la agregación de datos
- No hay conexiones ambiguas entre objetos (tablas) que pueden dar resultados diferentes según las rutas elegidas entre tablas al realizar operaciones de “join”

Otros enfoques

- Inmon se opone al uso de un modelo multidimensional para el diseño de almacenes de datos y utiliza el modelo Entidad Relación clásico o técnicas de normalización para bases de datos relacionales
- Luego, los mercados de datos (data marts) pueden diseñarse con base a un modelo relacional multidimensional o convencional
- Hay una discusión “fuerte” entre ambas practicas

Modelos conceptuales multidimensionales

- En la comunidad de bases de datos, se ha reconocido durante varias décadas que los modelos conceptuales
  - Permiten una mejor comunicación entre diseñadores y usuarios con el fin de comprender los requisitos de la aplicación.
  - Son más estables que un esquema (lógico) orientado a la implementación, que debe cambiarse cada vez que cambia la plataforma de destino
  - Proporciona un mejor soporte para las interfaces de usuario visuales
- Sin embargo, normalmente se muestra un escaso interés en el modelado conceptual multidimensional

Modelos conceptuales multidimensionales

- Actualmente no existe un modelo conceptual bien establecido para datos multidimensionales, aunque ha habido varias propuestas basadas en UML, en el modelo ER o utilizando notaciones específicas
- Estos modelos incluyen conceptos multidimensionales básicos, sin embargo, carecen de definición de diferentes tipos de jerarquías y mapeo a la plataforma de implementación

Modelos conceptuales multidimensionales

- Como consecuencia, en el estado actual de las cosas, los almacenes de datos se diseñan utilizando en su mayoría modelos lógicos (esquemas de estrella y copo de nieve)
- Los usuarios tienen dificultades para expresar sus requerimientos, ya que se requiere un conocimiento especializado relacionado con temas técnicos
- Estos modelos limitan a los usuarios a definir solo aquellos elementos que los sistemas de implementación subyacentes pueden administrar
    - Ejemplo: los usuarios están obligados a usar solo las jerarquías simples que se implementan en muchas herramientas de almacenamiento de datos actuales