

Diseño Conceptual



Temario: Diseño Conceptual

- Introducción.
 - Motivación.
 - Proceso de diseño.
- Modelos Multidimensionales.
- Estrategia basada en requerimientos.
 - Modelo CMDM.
 - Pautas de diseño.
 - Práctico.
- Estrategia basada en datos.
 - Metodología.
 - Práctico.
- Conclusiones.



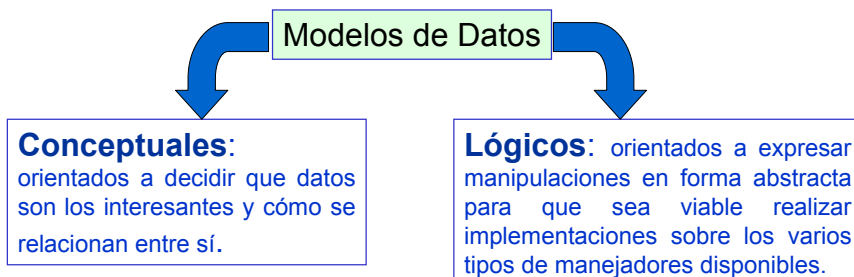
Introducción

- **Temas:**
 - Motivación.
 - Proceso de diseño.



Motivación

- ¿ Cuáles son las herramientas que necesita el diseñador para poder razonar sobre los datos y presentárselos al usuario ?

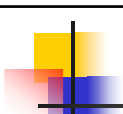




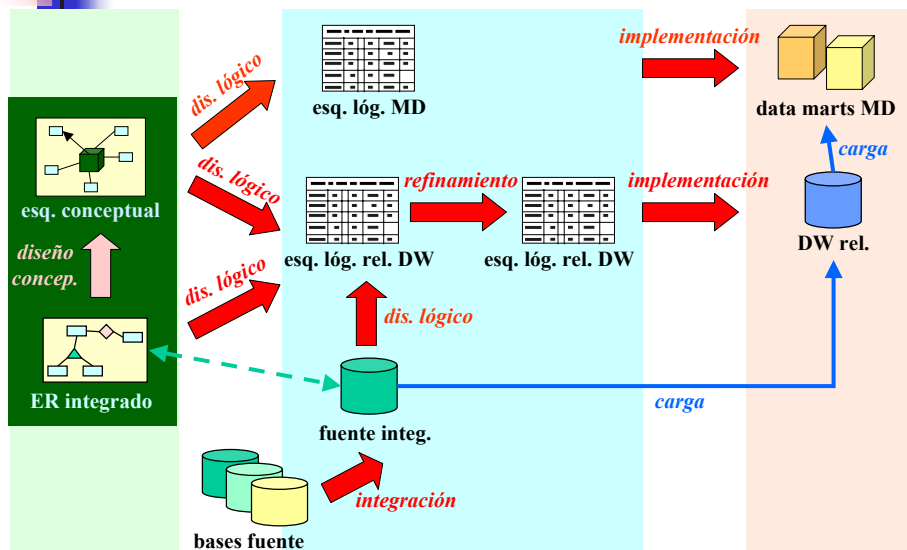
Modelos de Datos

■ Los niveles en diseño de BDs:

	Enfasis	Modelos de Registros	Multidimensional
Conceptual	Objetos y relaciones	Ent-Rel	- No hay estándares. - CMDM (InCo).
Lógico	Operaciones	Relacional	- MD específicos. - Star-Model (Rel).
Físico	Almacenamiento	Propietario DBMSs	Parámetros en servidores MD



Proceso de Diseño





Enfoques de Diseño Conceptual

- **Análisis desde requerimientos:**
 - Los requerimientos son el universo de información.
 - Las bases fuente se relacionarán luego.
 - Aplicable cuando se tienen Bases Fuentes complejas. (Se analizan con los requerimientos en mente).
 - Trabajos: [Car00], [Sap99], [Hus00], [Fra99].
- **Análisis desde datos:**
 - Datos fuentes son el universo de información.
 - El DW se obtiene transformando las fuentes.
 - Aplicable cuando los requerimientos están poco claros.
 - Trabajos: [Gol98a], [Cab98].



Etapas de Diseño Conceptual

- **Las principales etapas son:**
 - Definir un esqueleto de esquema:
 - Primer grupo de dimensiones medidas.
 - Establecer correspondencia entre requerimientos y datos fuentes.
 - Completar jerarquías en las dimensiones.
 - Especificar segundo grupo de medidas (calculadas).

iterar

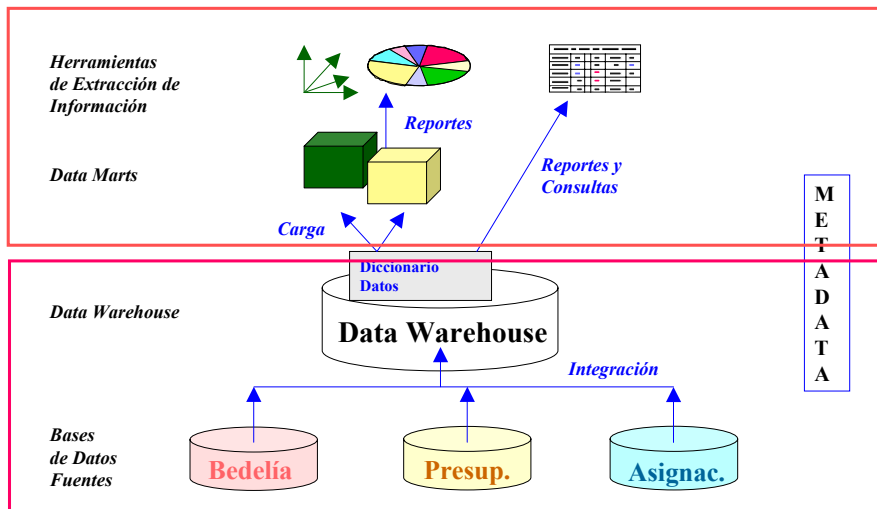


Modelos Multidimensionales

- **Temas:**
 - Motivación.
 - Estructuras en los MMD.
 - Operaciones en los MMD



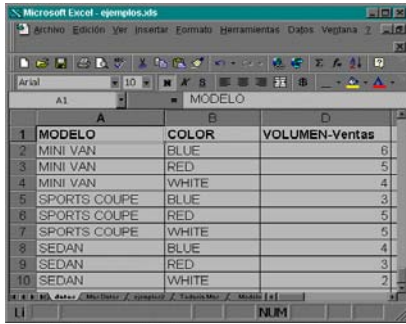
Motivación



Motivación

Representación Tabular

MODELO	COLOR	VOLUMEN-Ventas
MINI VAN	BLUE	6
MINI VAN	RED	5
MINI VAN	WHITE	4
SPORTS COUPE	BLUE	3
SPORTS COUPE	RED	5
SPORTS COUPE	WHITE	5
SEDAN	BLUE	4
SEDAN	RED	3
SEDAN	WHITE	2



Ventas de autos en función de Modelo y Color.

Motivación

Representación Matricial

M
O
D
E
L
O

Mini Van	6	5	4
Coupe	3	5	5
Sedan	4	3	2

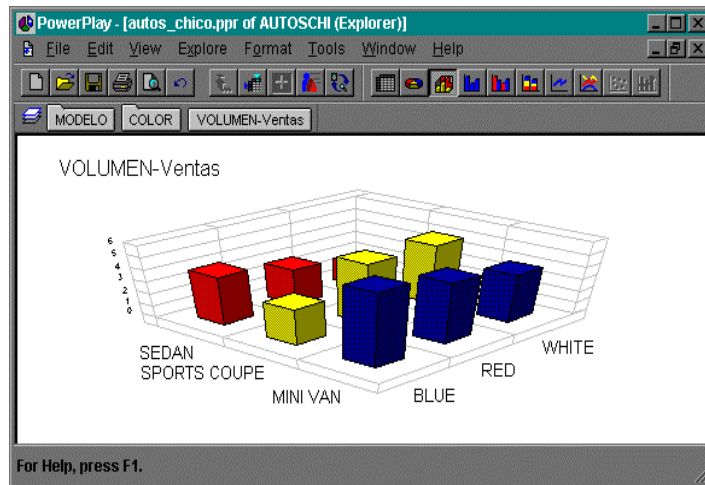
Blue Red White **COLOR**

MODELO	BLUE	RED	WHITE	COLOR
MINI VAN	6	5	4	15
SPORTS COUPI	3	5	5	13
SEDAN	4	3	2	9
MODELO	13	13	11	37

Ventas de autos en función de Modelo y Color.

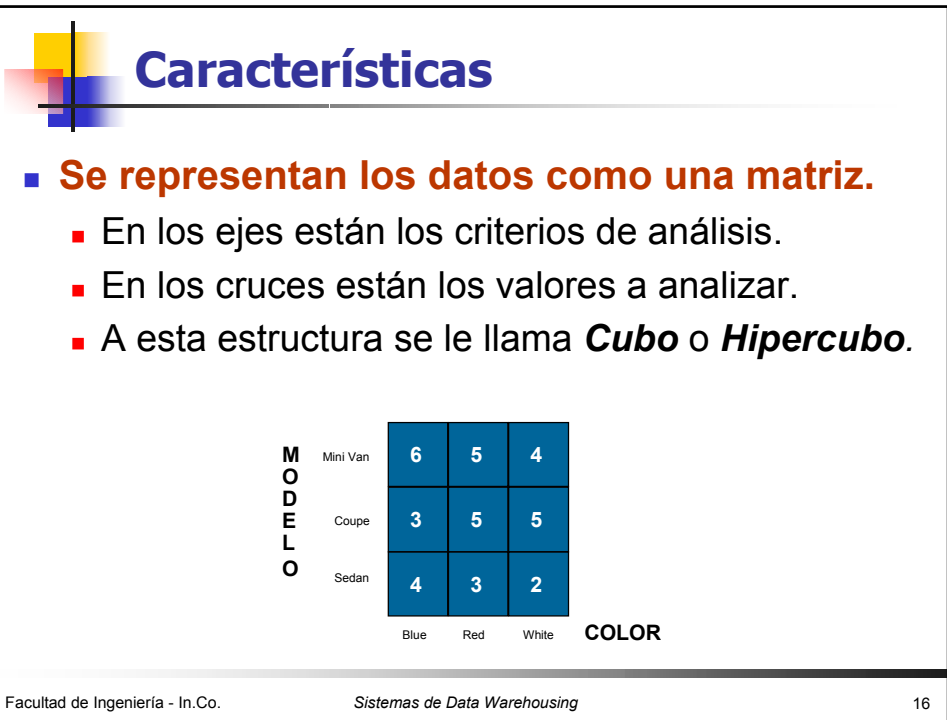
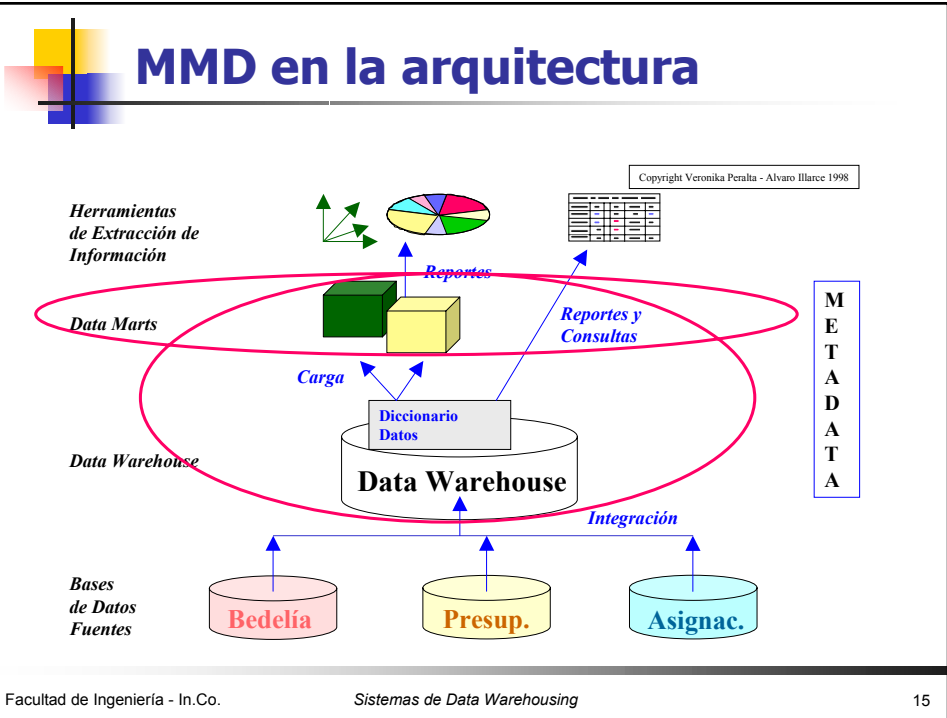


Motivación



Modelos Multidimensionales

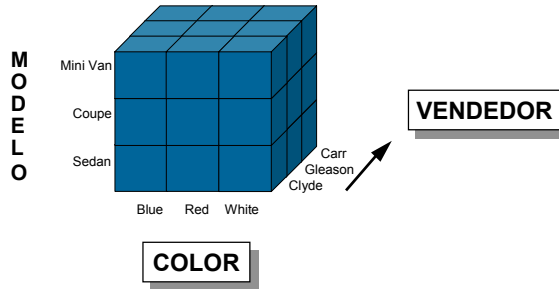
- **Qué tienen en común estas dos últimas representaciones?**
 - El usuario final (gerente) las entiende y maneja habitualmente.
- **Objetivos de los MMD:**
 - Representar los datos en forma cercana a la intuición del usuario.
 - Resolver problemas planteados en sistemas relacionales.





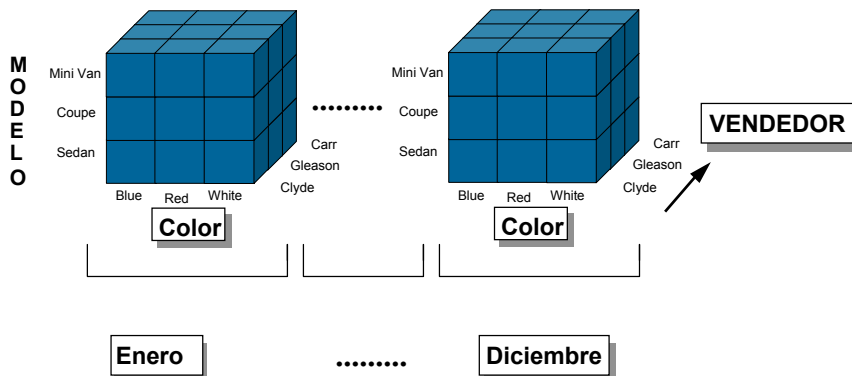
Características

- Agregando una 3a. dimensión:



Características

- Agregando una 4a. dimensión:





Estructuras básicas

- **Los Cubos o Hiper cubos constan de:**
 - Dimensiones:
 - Criterios de análisis de los datos.
 - Macro-objetos del problema.
 - Variables independientes.
 - Ejes en el hiper cubo.
 - Medidas
 - Valores o indicadores a analizar.
 - Datos asociados a relaciones entre los objetos del problema.
 - Variables dependientes.
 - Variables en la intersección de las dimensiones.



Estructuras básicas

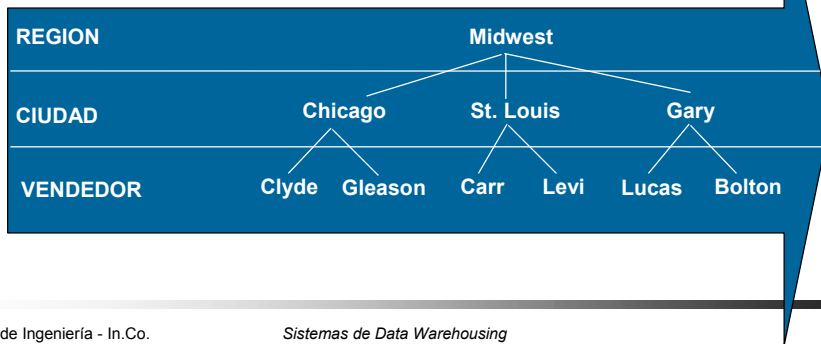
- **En el ejemplo anterior:**
 - Dimensiones:
 - Modelo
 - Color
 - Vendedor
 - Fecha
 - Medida:
 - Cantidad Vendida



Dimensiones

■ Jerarquías:

- Los valores se organizan en jerarquías (categorías).
- **Por ejemplo:** *Dimensión: Vendedores*



Dimensiones

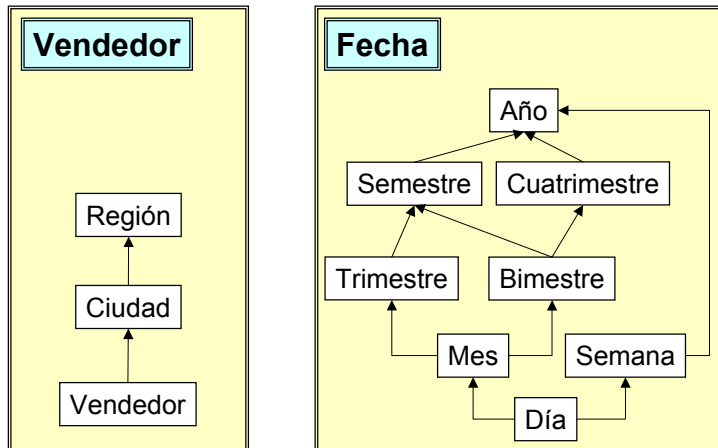
■ Jerarquías alternativas:

- Pueden haber varias jerarquías para una misma dimensión.
- **Por ejemplo:**
 - Dimensión Vendedores:
 - Región / Ciudad / Vendedor.
 - Sexo / Rango_Edad / Vendedor.



Dimensiones

■ Jerarquías Arbitrariamente Complejas



Medidas

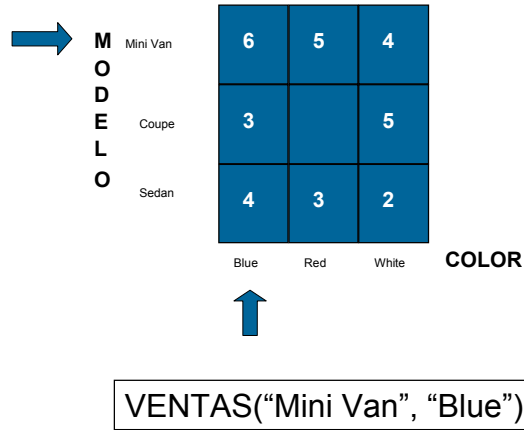
■ Propiedades:

- Se ubican en la intersección de algunos valores de las dimensiones. Dado un valor para cada dimensión se puede determinar un valor para la medida.

Definición: Se llama **coordenada** a una tupla formada por un valor de cada dimensión.



Medidas



Cubos

- **La realidad se modela como un conjunto de cubos.**
 - Cada cubo, esta formado por:
 - Un conjunto de *Dimensiones* organizadas en jerarquías.
 - Un conjunto de *Medidas* asociadas a cada *Coordenada*.
 - Es posible moverse en las jerarquías de las dimensiones y observar de esa forma, diferentes visiones de las medidas.

Caso de Estudio

- Dimensiones
 - Vendedores**
 - Región
 - Ciudad
 - Vendedor
 - Fechas**
 - Año
 - Semestre
 - Cuatrimestre
 - Trimestre
 - Bimestre
 - Mes
 - Semana
 - Día
 - Colores**
 - Color
 - Modelos**
 - Modelo
- Medidas
 - Ventas**
 - Cantidad

Diagram illustrating the dimensions and measures in a data warehouse model. The dimensions are categorized into Vendedores, Fechas, Colores, and Modelos. The measures are categorized into Ventas. The hierarchy for Vendedores is Vendedor -> Ciudad -> Región. The hierarchy for Fechas is Día -> Mes -> Trimestre/Bimestre -> Semestre/Cuatrimestre -> Año. The hierarchy for Colores is Color. The hierarchy for Modelos is Modelo. The measure Ventas is measured by Cantidad.

Operaciones

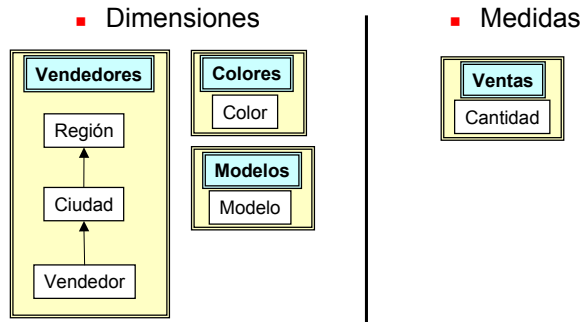
- Principales operaciones en modelos MD:
 - Slice.
 - Dice.
 - Rotación.
 - Drill-down.
 - Drill-up.
 - Roll-up.
 - Drill-across.
 - Drill-through.



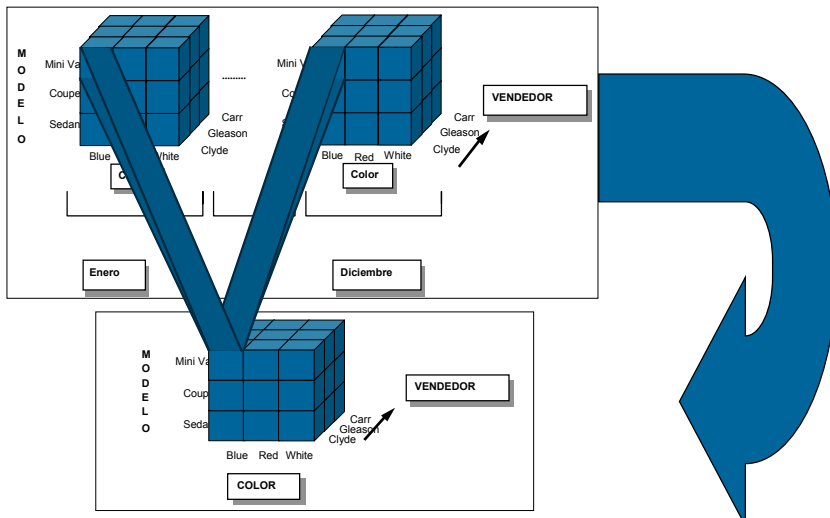
Operaciones: Slice

■ Seleccionar Dimensiones (Slice)

- Se define un subconjunto del hipercubo especificando sobre qué dimensiones interesa analizar qué medida.



Operaciones: Slice

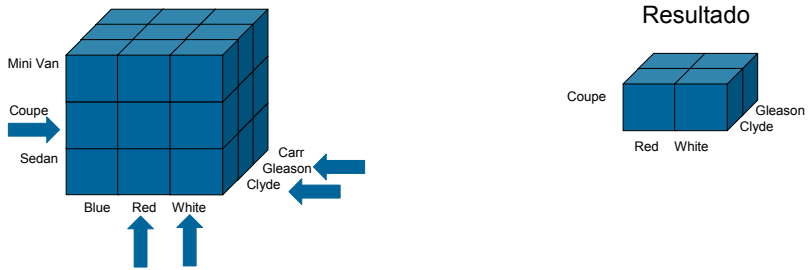




Operaciones: Dice

■ Filtrado (DICE)

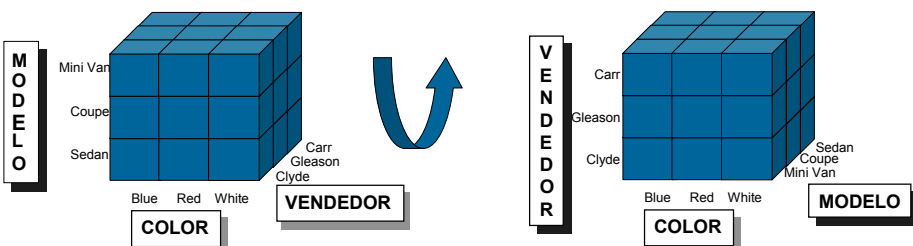
- Se fijan valores para algunas dimensiones.



Operaciones: Rotación

■ Rotación.

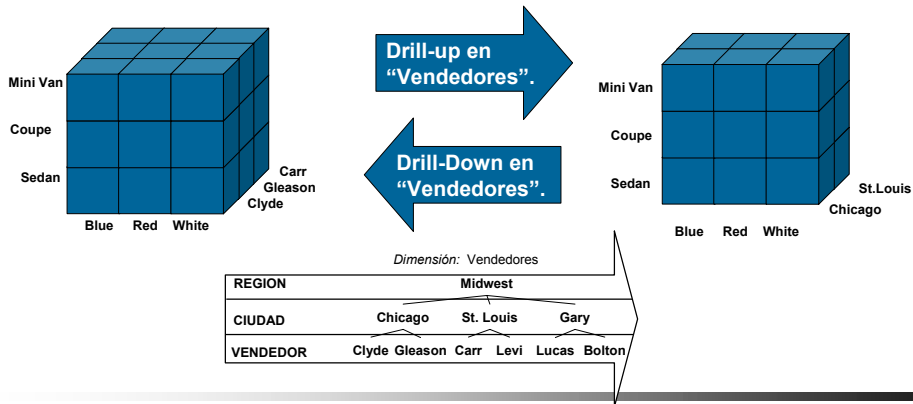
- Selecciona el orden de visualización de las dimensiones.





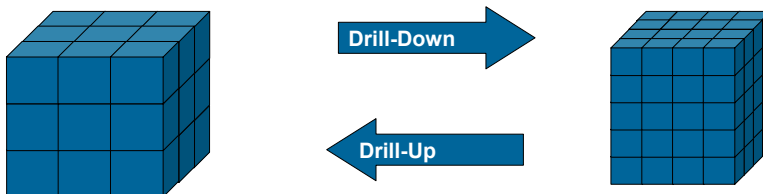
Operaciones: Drill-up, drill-down

- Movimientos en la Jerarquía de una Dimensión (Drill-up, Drill-down).



Operaciones: Drill-up, drill-down

- Drill-Up o Drill-Down pueden verse como ajuste en las escalas de los ejes.
- Son agrupamientos y des-agrupamientos.

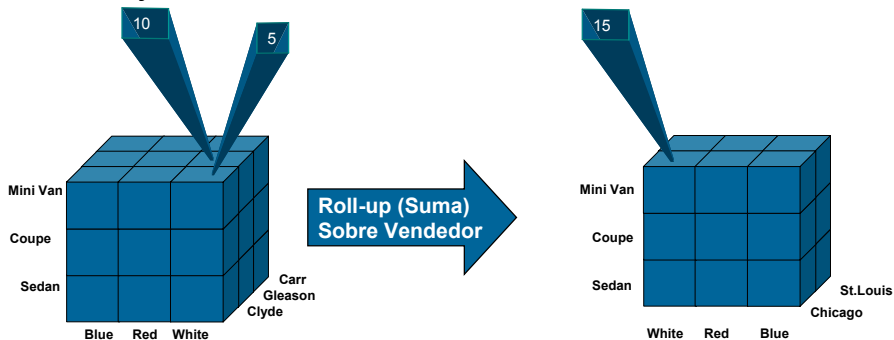




Operaciones: Roll-up

■ Consolidación (Roll-Up).

- Calcula las medidas en función de agrupamientos.
- Realiza el re-cálculo de la medida de acuerdo a los ajustes de escala.



Operaciones: Roll-up

■ Propiedades:

- Se debe especificar cuál es la operación que calcula el nuevo valor de la medida.
 - Esta operación puede ser: suma, promedio, etc.
 - Pueden haber medidas con comportamientos diferentes.
 - **Por ejemplo:**
 - Cantidades de productos vendidos se *acumulan*.
 - Notas en exámenes se *promedian*.



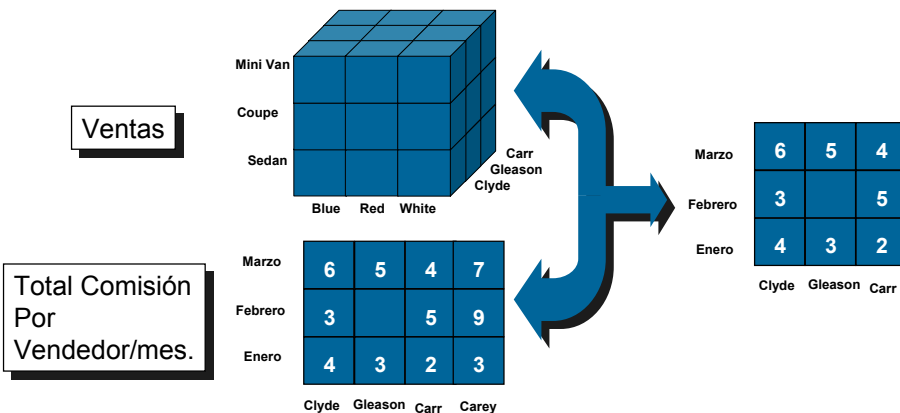
Operaciones: Roll-up

- En general cualquier operación de navegación en un cubo implica un nuevo cálculo de la medida.
- **Hay dos momentos posibles:**
 - Se asocia a la medida una operación por defecto.
 - En el momento de hacer un movimiento en la dimensión se especifica cómo se hacen los cálculos.



Operaciones: Drill-Across

- **Drill-Across**
 - Relaciona dos cubos.

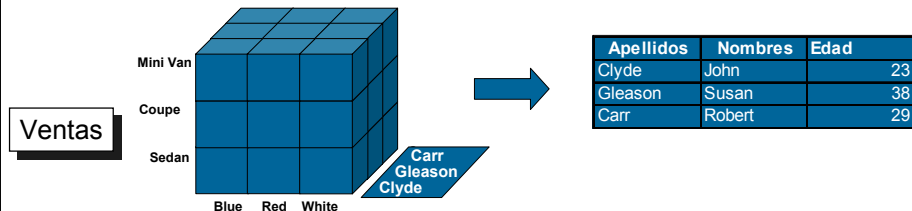




Operaciones: Drill-Through

■ Drill-Through.

- Accede a datos descriptivos.



Limites de los MMD

■ Los MMD no realizan adecuadamente ciertas operaciones:

- Salvo en algunos trabajos de investigación, no se habla de representación y consulta de datos *descriptivos* o *secundarios*.
- Tampoco se permiten diferentes RollUp's por medida.

Entonces:

- Se deben resolver con otras técnicas, o
- Se deben considerar en el diseño.



Características de los MMD

■ Resumen:

- Permiten describir una realidad en términos de matrices multidimensionales (Cubos).
- Desde el punto de Vista del DW, se utilizan para describir Datamarts o el DW Completo.
- Las Dimensiones pueden tener una o más jerarquías.



Características de los MMD

■ Resumen

- Las operaciones permiten construir diferentes visiones de los datos.
- Hay ciertas operaciones que no son totalmente resueltas por los productos actuales.



Pendientes...

- **¿Cómo se describen los aspectos Multidimensionales del DW?**
 - ¿Con qué lenguaje?
 - ¿Con qué criterios se construye esa descripción?
- **Respuestas Rápidas...**
 - No hay un lenguaje aceptado.
 - Hay criterios muy generales pero no hay criterios absolutos...

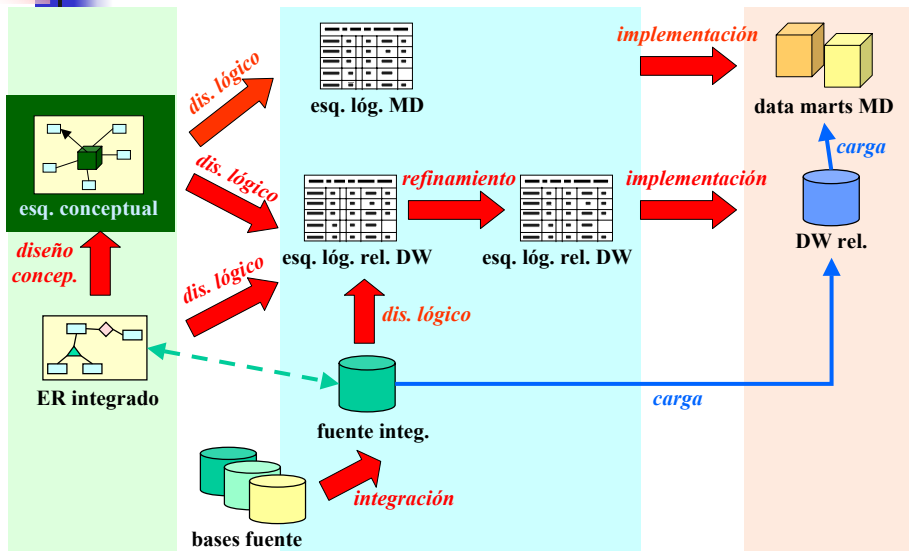


Diseño Conceptual

Estrategia basada en requerimientos

- **Temas:**
 - Modelo CMDM.
 - Pautas de Diseño.
 - Ejercicios.

Proceso de Diseño



Modelo CMDM

- **Objetivos principales:**
 - Representación directa, expresiva y precisa de esquemas multidimensionales.
- **Características:**
 - Basado en conceptos aceptados de MMD.
 - Independiente de las implementaciones.
- **Desarrollado por el grupo CSI (InCo).**
 - Tesis de maestrías [Car00].
 - Editor gráfico [Pic00].



Modelo CMDM

■ Estructuras básicas.

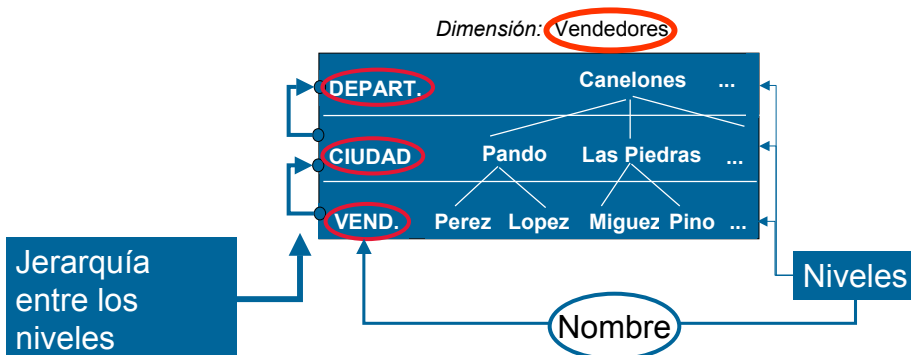
- Niveles.
- Dimensiones.
 - Con Jerarquías, formadas por Niveles.
 - Incluye Medidas (***Dimensionalidad Genérica***).
- Relaciones dimensionales.
- Cubos.
 - Cruzamientos específicos.



Modelo CMDM

■ Dimensiones:

- Una dimensión esta formada por:



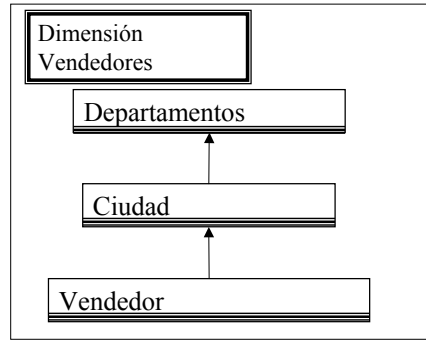
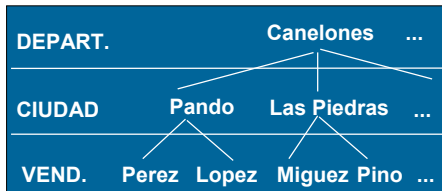


Modelo CMDM

■ Dimensiones:

■ Ejemplo:

Dimensión: Vendedores

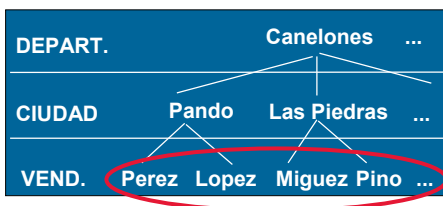


Modelo CMDM

■ Niveles:

■ Un nivel representa un conjunto de datos.

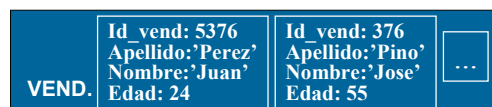
Dimensión: Vendedores



Los datos pueden ser no atómicos.



Nivel: Vendedor





Modelo CMDM

■ Niveles:

■ Ejemplo:

Nivel: Vendedor

VEND.	Id_vend: 5376 Apellido: 'Perez' Nombre: 'Juan' Edad: 24	Id_vend: 376 Apellido: 'Pino' Nombre: 'Jose' Edad: 55	...
-------	--	--	-----

Vendedor
Id_vendedor#: Integer
Apellido: String
Nombre: String
Edad: [18-60]



Modelo CMDM

■ Jerarquías:

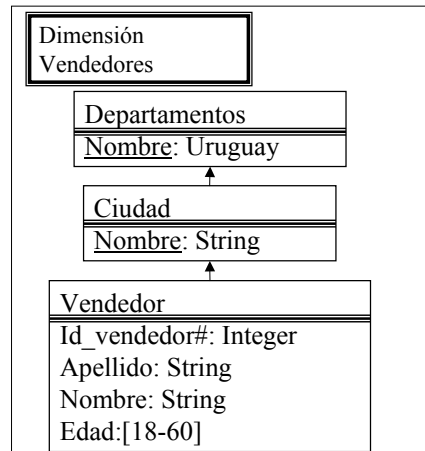
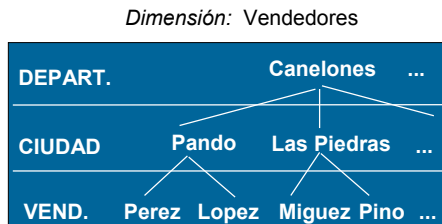
- Los niveles se organizan en **jerarquías**.
- Cada jerarquía está compuesta por uno o varios niveles.
- En cada jerarquía:
 - Se tiene una relación <1-n> entre objetos de nivel superior e inferior.



Modelo CMDM

■ Jerarquías:

■ Ejemplo:



Modelo CMDM

■ Relaciones Dimensionales:

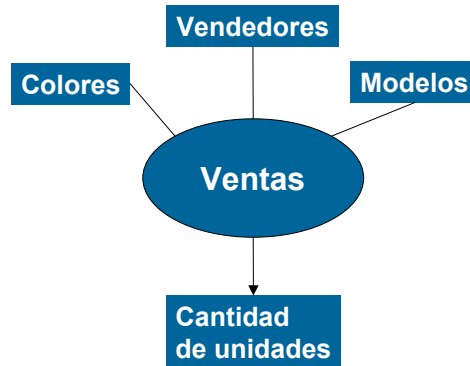
- Representan cruzamientos entre Dimensiones.
- Las Medidas participan como Dimensiones.
- Vista como una relación:
 - Se tiene un elemento en el conjunto relación si y solo si hay un cruzamiento.
 - Esto obliga a que las Dimensiones participantes realmente sean cruzables.



Modelo CMDM

■ Relaciones Dimensionales.

■ Ejemplo.



Modelo CMDM

■ Esquema Conceptual MD:

- Está formado por un conjunto de Relaciones Dimensionales.
- Las Relaciones Dimensionales pueden compartir Dimensiones.
 - Permite el Drill-Across.

■ Cubos:

- Dada una Relación Dimensional, un **cubo** representa un cruzamiento concreto entre niveles determinados de las Jerarquías de la RD.

Modelo CMDM

- **Cubos:**
 - Ejemplo.

The diagram illustrates a CMDM model. At the center is an oval labeled 'Ventas'. Four rectangular boxes represent dimensions: 'Vendedores' (top), 'Modelos' (right), 'Colores' (left), and 'Cantidad de unidades' (bottom). Arrows point from 'Ventas' to each of these boxes. Below this central structure are two 3D cubes. The left cube is labeled 'Carr' and has dimensions: 'Carr' (Mini Van, Coupe, Sedan), 'Colores' (Blue, Red, White), and 'Locations' (Clyde, Gleason, Carr). The right cube is labeled 'St.Louis' and has dimensions: 'Modelos' (Mini Van, Coupe, Sedan), 'Colores' (White, Red, Blue), and 'Locations' (Chicago, St.Louis). Red arrows point from the central 'Ventas' area towards these two example cubes.

Facultad de Ingeniería - In.Co. Sistemas de Data Warehousing 57

Modelo CMDM

- **Conclusiones sobre CMDM.**
 - Apunta a dar un modelo de especificación de esquemas conceptuales Multidimensionales.
 - Cumpliendo un rol equivalente al del Modelo ER.
 - Los esquemas conceptuales MD luego serán traducidos a esquemas lógicos MD.
 - Todavía en desarrollo, se trabaja sobre:
 - Mejoras de la herramienta CASE.
 - Un lenguaje de Restricciones de Integridad.
 - Una base formal para el modelo.

Facultad de Ingeniería - In.Co. Sistemas de Data Warehousing 58



Pautas de diseño

- **Preguntas básicas en diseño:**
 - ¿Qué es una Dimensión y qué es una Medida?
 - ¿Tiene sentido construir cualquier cruzamiento de niveles?
 - ¿Cuándo definir una Dimensión con varias jerarquias o varias Dimensiones?
 - ¿Con qué criterios definir una o varias Relaciones Dimensionales?
 - Dada una relación dimensional, ¿cualquier operación de roll-up es correcta?
 - ¿Cómo clasificar las medidas calculadas?



Def. de Dimensiones y Medidas

- **¿Qué items pueden ser medidas y cuáles dimensiones?**
 - No hay ningún criterio ni método efectivo para tomar la decisión.
 - Pueden haber “indicios” basados en:
 - Objetos del problema y/o criterios de análisis Vs. Indicadores y/o medidas.
 - Interés de totalizar según Drill-Ups.
 - Identificadores Vs. No Identificadores.
 - Estudio de distribución de valores.
 - Estudio de independencia de variables.



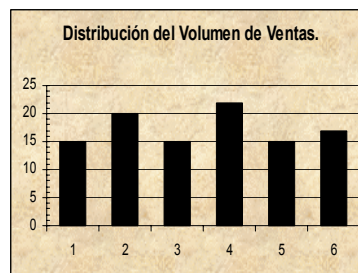
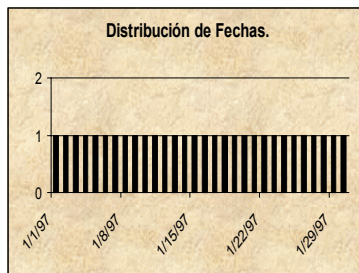
Def. de Dimensiones y Medidas

- **Separar objetos de medidas o Identificadores Vs. no identificadores.**
 - Ejemplos:
 - Los vendedores, son un objeto del problema, por lo tanto tienen posibilidades de funcionar bien como una dimensión.
 - El tiempo es típicamente, un criterio de análisis.
 - El promedio de autos vendidos por vendedor y por fecha, es casi seguro una medida.
 - Contraejemplo:
 - Los autos vendidos pueden ser considerados una medida... en realidad las cantidades de autos vendidos.



Def. de Dimensiones y Medidas

- **Estudios de ocurrencia de los valores de los atributos:**
 - Distribución “fija” / “discreta”: dimensión.
 - Distribución “interesante” / “continua”: medida.





Def. de Dimensiones y Medidas

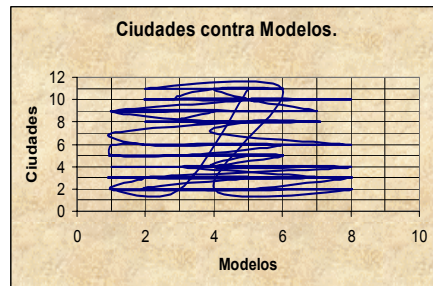
- **Estudio de Independencia de las variables.**
 - **Variables Independientes:** Las dos pueden ser dimensiones simultáneamente.
 - **Variables Dependientes:** Dos niveles en la misma jerarquía.

No debieran existir niveles distintos de la misma jerarquía que sean independientes entre sí.



Def. de Dimensiones y Medidas

- **Cómo hacer el estudio de Independencia de Variables?**





Jerarquías en dimensiones

- **Criterios para construir jerarquías:**
 - Definir caminos de navegación (drill-down/up).
 - Definir agrupamientos naturales de los datos (reportes).
 - Asociar diferentes niveles de cálculo de indicadores.
- **Atención !**
 - Cuidado con el comportamiento de las medidas cuando se hace roll-up.



Jerarquías en dimensiones

- **Ejemplos:**
 - Los supervisores de ventas estudian zonas, que son conjuntos de departamentos.
 - El pago de comisiones a vendedores depende de su antigüedad, cada 5 años cambian los porcentajes.
 - Los jefes de sección deben presentar informes trimestrales y anuales de ventas.
 - Cada cliente puede realizar compras en diferentes ciudades del país.



Jerarquías en dimensiones

- **Disyuntiva:**
 - Definir dimensiones “complejas” o “simples”.
- **Criterios para construir jerarquías alternativas:**
 - Tienen consultas en común.
 - Conceptualmente representan una misma idea.



Jerarquías en dimensiones

- **Ejemplos:**
 - Los supervisores de ventas estudian las ventas según la ubicación geográfica de los clientes.
 - Los analistas de marketing estudian los clientes según segmentos de mercado.
 - Se hacen reportes de distribución geográfica de los segmentos de mercado.
 - La sección depósito maneja fechas calendario, en cambio contaduría maneja años fiscales.



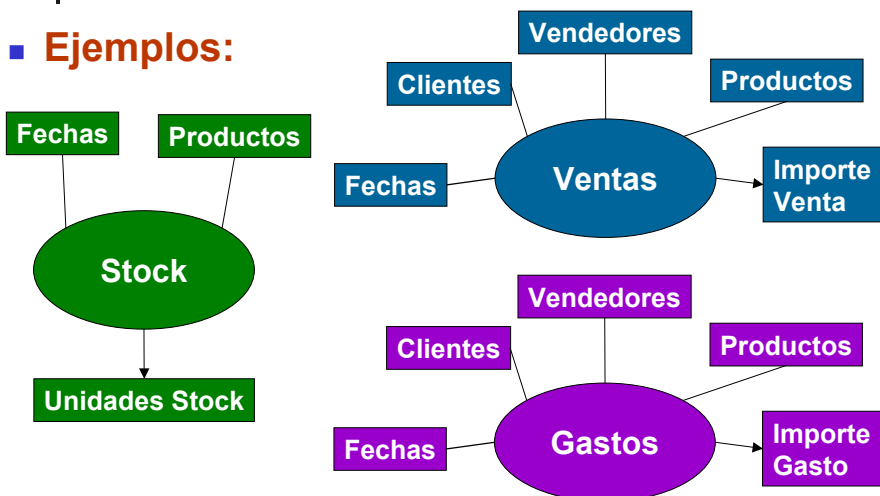
Relaciones dimensionales

- **Disyuntiva:**
 - Definir una o varias Relaciones Dimensionales
- **Criterios para construir relaciones dimensionales:**
 - Cada relación representa un concepto del problema.
 - Las dimensiones involucradas se cruzan entre sí.
 - Las medidas están relacionadas (cálculos).



Relaciones dimensionales

■ Ejemplos:



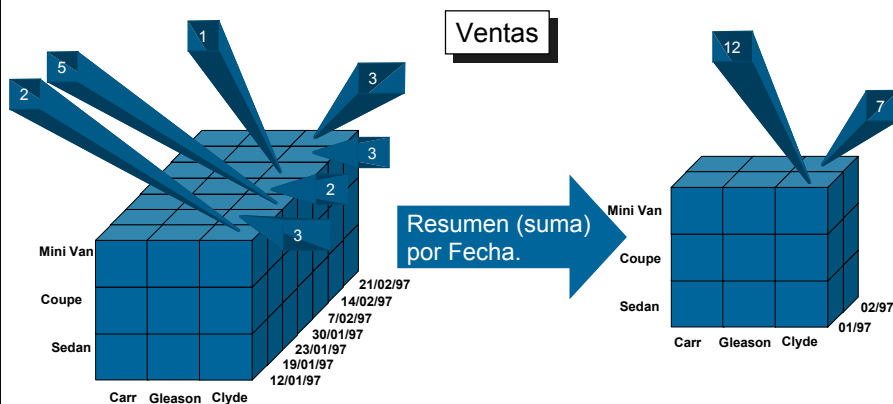


Funciones Roll-Up

- **Dada una Relación Dimensional y una medida:**
 - ¿Cualquier función de roll-up es válida?
- **NO !**
 - En particular no siempre se puede sumar valores.



Aditividad





Aditividad

- Para Ralph Kimball, *The Data Warehouse Toolkit*. [Kim96]
 - “The most useful facts in a fact table are numeric and ADDITIVE”.
 - “Additivity is crucial because data warehouse applications almost never retrieve a single fact table record And the only useful thing to do with so many records is to add them up”
- Para Lenz & Shoshani, *Summarizability in OLAP and Statistical Data Bases*. [Len97]
 - “El resumen es una propiedad importante. Un resumen incorrecto puede llevar a análisis, conclusiones y decisiones erróneas”.
- Principales avances en el tema de aditividad :
 - Proviene de las Bases de Datos Estadísticas.



Problema: Disjuntez

Medida: Cantidad de vendedores diferentes que vendieron determinado modelo en determinada fecha.

$\sum_{mes=1}^{12} \wedge año=97 (CantVendedores("mini-van", fecha))$

Mini Van	2	0	2	2	1	1	1
Coupe	0	2	3	0	1	1	1
Sedan	1	1	0	1	2	1	1
	12/01/97	19/01/97	23/01/97	30/01/97	7/02/97	14/02/97	21/02/97

RollUp (suma por Fecha ????)

¿Cómo eliminar repetidos?

6	3
5	3
3	4
01/97	02/97



Problema: Disjuntez

- **Si no hay disjuntez por ninguna dimensión:**
 - Dimensiones: materias, fechas, institutos.
 - Medida: cantidad de aprobados.
 - Un alumno puede dar exámenes en varias fechas, de varias materias, de varios institutos.



Problema: Disjuntez

- **Si la disjuntez es sólo para algunas dimensiones:**
 - Dimensiones: fechas, geografía, rubro, producto.
 - Medida: cantidad de clientes que compraron.
 - Es disjunta para las dimensiones que clasifican a los clientes.
 - Hay intersección para las otras dimensiones.
 - Un cliente sólo puede estar en un rubro y un lugar geográfico, pero puede comprar varios productos en varias fechas.



Problema: Disjuntez

- **Si la disjuntez se da para algunos niveles de las dimensiones:**
 - Dimensiones: vendedores, fechas, productos (rama, producto).
 - Medida: cantidad de vendedores distintos.
 - Cada vendedor se especializa en una sola rama.



Problema: Disjuntez

- **Debemos estudiar:**
 - ¿Cómo se cuentan los objetos respecto a las dimensiones?
 - Si sólo puede estar en una rama jerárquica.
 - Si puede haber intersecciones: en toda la dimensión o en algunos niveles.

Problema: Cálculos

Medida: **Índice de venta por vendedor**
 Definición: Cantidad vendida / Cantidad de vendedores

$\sum_{mes=1 \wedge año=97} (\text{IndiceVenta}(\text{"mini-van"}, \text{fecha}))$

Mini Van	1	1	4	1	2	1	1
Coupe	0	2	1	0	4	1	1
Sedan	3	4	0	3	3	1	1

RollUp (suma) por Fecha ????

7	4
3	6
10	5

total ventas / total de vendedores

Facultad de Ingeniería - In.Co. Sistemas de Data Warehousing 79

Problema: Cálculos

- **Cuándo hacer el roll-up**
 - Qué debo hacer?
 - Aplicar el roll-up a cada medida y luego el cálculo.
 - Hacer el cálculo y luego aplicar algún roll-up.
 - Los dos casos pueden ser válidos en distintos contextos.

Facultad de Ingeniería - In.Co. Sistemas de Data Warehousing 80



Problema: Cálculos

■ Ejemplo:

- Nota Promedio= suma notas / total materias
 - 1o (matematicas 3, fisica 5). Prom=4.
 - 2o (programacion 10). Prom=10. } Prom=7
 - Cantidad=3, Suma=18, Prom=6.
- Caudal promedio
 - Me interesa promediar los promedios de cada dia, no importando la cantidad de muestras.
 - Lunes (10,11,12,11,11), martes (15,15), miercoles(14,14). Prom=13,3.
 - Cantidad=9, Suma=113, Prom=12,6.



Tipos de Medidas

■ Flow o aditivas

- Conserva la semántica al aplicar la suma como RollUp, realizando DrillUp por cualquier dimensión que la defina.
- Refieren a un evento o periodo, y son registradas al final del mismo.
- Ejemplos:
 - \$ en una factura, cantidad de nacimientos por mes, expedientes procesados en una semana, cantidad de ventas en el día.



Tipos de Medidas

■ Stock o semi_aditivas

- Conserva la semántica al aplicar la suma como RollUp, al realizar DrillUp, en todas las dimensiones menos el tiempo.
- Son registradas en un punto específico del tiempo y refieren a ese instante.
- Ejemplos:
 - Inventarios (de cualquier tipo).
 - Saldos de cuentas.
 - Censos de ciudadanos.



Tipos de medidas

■ Value_per_unit o no aditivas

- No conserva la semántica al aplicar la suma como RollUp, al realizar DrillUp, en varias o ninguna dimensión.
- Son determinados en un momento dado pero su unidad de medida las diferencia de stock.
- Ejemplos:
 - Precio por ítem, tasa de cambio, edades, notas.
 - Mediciones: temperaturas, caudales.
 - Anomalías de disjuntéz.



Condiciones para Roll-up

- **Para que la sumarización sea correcta [Len97]:**
 - Los agrupamientos de objetos entre niveles consecutivos de la jerarquías de las dimensiones deben formar subconjuntos disjuntos sobre el conjunto global.
 - Los agrupamientos de individuos entre niveles de la jerarquía debe ser completo.
 - Compatibilidad de las 3 partes del resumen:
 - La temporalidad o no de la categoría o dimensión.
 - El tipo de la medida a sumar (flow o aditiva, stock o semi_aditiva, vpu o no_aditiva).
 - La función asociada (min, max, sum, avg, range).



Condiciones para Roll-up

Temporal

	Stock	Flow	Vpu
Min	Ok	Ok	Ok
Max	Ok	Ok	Ok
Sum	No ok	Ok	No ok
Avg	Ok	Ok	Ok
Range	Ok	Ok	Ok

No Temporal

	Stock	Flow	Vpu
Min	Ok	Ok	Ok
Max	Ok	Ok	Ok
Sum	Ok	Ok	No ok
Avg	Ok	Ok	Ok
Range	Ok	Ok	Ok



Medidas Calculadas

- **Tipos de fórmulas [Tho97]:**
 - Según el rol en el análisis:
 - Descriptivas, Explicativas, Predictoras, Prescriptivas.
 - Según la transformación sobre los datos:
 - Fórmulas que mantienen granularidad.
 - Fórmulas de agrupamiento.

- **Problemas técnicos:**
 - Tener en cuenta precedencia de operadores.
 - Evaluar la materialización de cálculos.



Tipos de fórmulas

- **Fórmulas descriptivas:**
 - Datos fuentes o agregación de datos fuentes.
 - Ej:
 - Cantidad de unidades en stock.
 - Costo del ítem
 - Último precio de venta
 - Cantidad de unidades salientes
 - Margen de ganancia
 - Promedio diario de unidades en stock.



Tipos de fórmulas

■ Fórmulas explicativas:

- Relaciones entre datos fuentes (o agregaciones).

- Ejemplo:

- Rotación (stock) =
$$\frac{\text{Total de unidades salientes}}{\text{Promedio diario de unidades}}$$

- % de ganancia =
$$\frac{\text{último precio venta} - \text{costo}}{\text{último precio venta}}$$

- retorno (inventario) = rotación * % de ganancia



Tipos de fórmulas

■ Fórmulas predictoras:

- Predicción sobre valores que aún no han sido medidos (por pertenecer al futuro). Es la combinación de una relación conocida con una extrapolación.

- Ejemplo:

- Predecir variaciones sobre las ventas (y/o ganancias) a partir del retorno de inventario.

Si el margen de retorno del inventario es x las ventas tendrán una variación futura de $y = f(x)$



Tipos de fórmulas

- **Fórmulas prescriptivas:**
 - Indican acciones a tomar.
 - Son como “recetas” que se calculan con fórmulas complejas, a partir de los valores medidos.
 - Ejemplo:
 - ¿Cuántas unidades debo producir de cada producto?
 - ¿A qué precio debo vender?
 - ¿En cuánto debo reducir los costos para mejorar ganancias / competencia?



Ejemplo

- **Siguiendo las etapas de diseño:**
 - Definir un esqueleto de esquema:
 - Primer grupo de dimensiones medidas.
 - Establecer correspondencia entre requerimientos y datos fuentes.
 - Completar jerarquías en las dimensiones.
 - Especificar segundo grupo de medidas (calculadas).
- iterar**
-



Ejemplo (cont.)

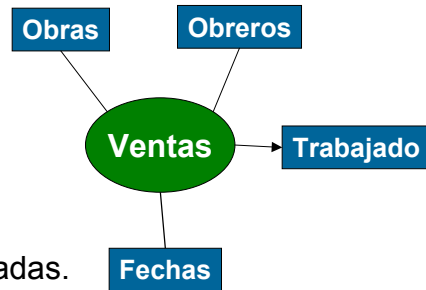
- Se desea analizar la actividad de los obreros de una empresa constructora.
- Primer esquema MD:

- Dimensiones:

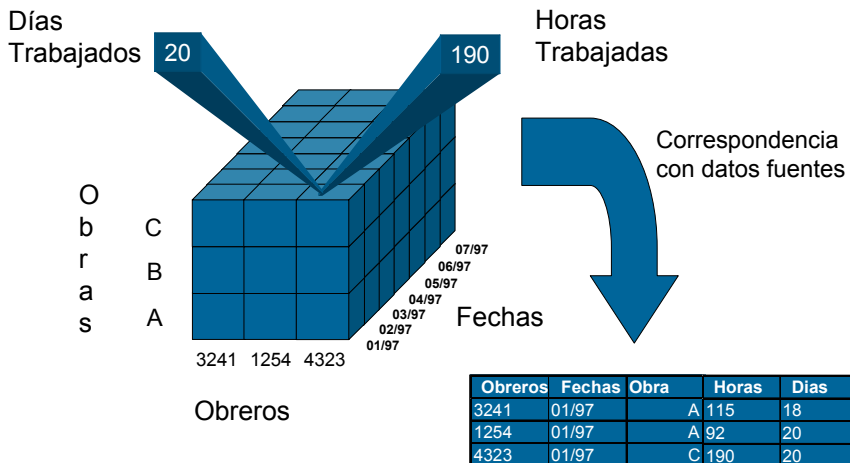
- Obras.
- Fechas.
- Obreros.

- Medidas:

- Cantidad de horas trabajadas.
- Días Trabajados.



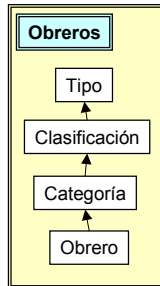
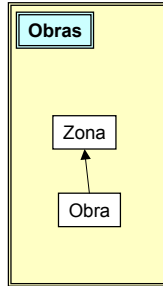
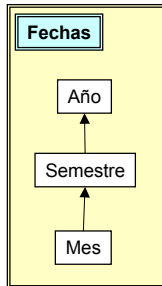
Ejemplo (cont.)



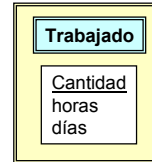


Ejemplo (cont.)

■ Dimensiones



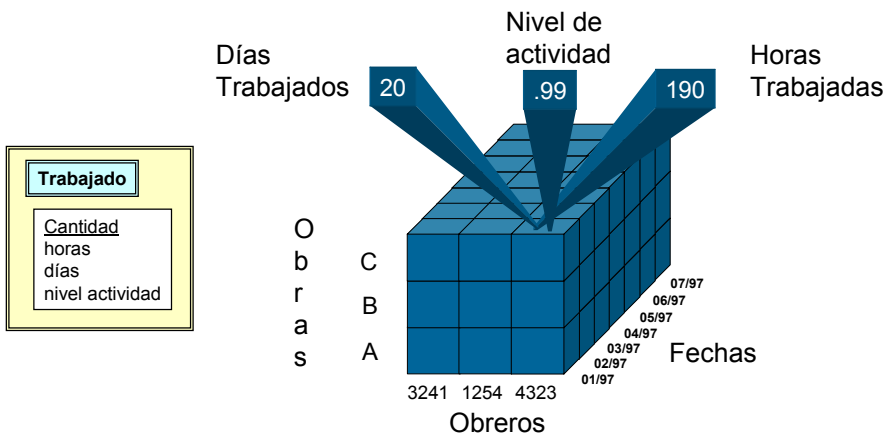
■ Medidas



Ejemplo (cont.)

■ Se agregan medidas calculadas:

- Nivel de Actividad = $\text{Horas/Días} * 9.6$.





Práctico

- **Se siguen las etapas de diseño:**
 - Definir un esqueleto de esquema:
 - Primer grupo de dimensiones medidas.
 - Establecer correspondencia entre requerimientos y datos fuentes.
 - Completar jerarquías en las dimensiones.
 - Especificar segundo grupo de medidas (calculadas).
- iterar**
-



Esqueleto del Esquema

- **Definir el esqueleto del esquema:**
 - Primer grupo de dimensiones.
 - Primer grupo de medidas.
- **Surgen de:**
 - Visión general del negocio y los requerimientos.
- **Documentación:**
 - Descripción de objetos del negocio.
 - Formalización de los requerimientos.
 - Primer versión del esquema conceptual.

Objetos del Negocio

Listado:

- Objeto.
- Descripción.
- Medida?
- Dimensión?
- Opcional?

Referencia: [Hus00]

Attribute	Description	M	D	O
effectiveDay	data import date	no	yes	no
month	time aggregation	no	yes	no
quarter	time aggregation	no	yes	no
year	time aggregation	no	yes	no
accountID	account key	no	yes	no
balance	balance at effective day	yes	no	no
balanceClass	balance classification	no	yes	no
turnover	turnover at effective day	yes	no	no
turnoverClass	turnover classification	no	yes	no
creditlimit	creditlimit of the account	yes	no	no
interest	interest rate	yes	no	no
custID	customer key	no	yes	no
custName	customer name	no	yes	no
custAge	age of a private customer	no	yes	yes
customerType	classification of customers	no	yes	no
profession	profession of a private customer	no	yes	yes
branch	branch of a business customer	no	yes	yes
productID	product description	no	yes	no
productType	classification of products	no	yes	no
orgID	attending organizational unit	no	yes	no
orgName	name of an organizational unit	no	yes	no
orgGroup	grouping of organizational units	no	yes	no
orgType	classification of organizational units	no	yes	no
businesssector	classification of orgGroup and orgType	no	yes	no

Requerimientos

- Dimensiones y medidas involucradas en cada requerimiento.

Referencia: [Bal98]

Dimensions and Measures	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
<i>Dimensions</i>									
Sales		X		X	X			X	X
Manufacturing	X		X						
Product	X	X	X		X	X	X	X	X
<i>Measures</i>									
Average quantity on hand	X								
Total cost		X	X				X		
Total revenue		X	X				X		
Quantity sold							X		
Percentage of models eligible for discount				X					
Percentage of models eligible for discount that are actually discounted				X					
Percentage of a model sold through a retail outlet					X				
Percentage of a model sold through a corporate sales office order desk					X				
Percentage of a model sold through a sales person					X				



1a versión del esq. conceptual

- **Definir:**
 - Jerarquías para las dimensiones.
 - Medidas.
 - Relaciones dimensionales.

- **Se utiliza un modelo conceptual:**
 - Por ejemplo CMDM.



Correspondencias

- **Correspondencias entre:**
 - Esquema conceptual.
 - Bases fuentes.
- **Objetivo:**
 - Ubicar en las fuentes los objetos conceptuales.
 - Verificar que los datos existen.
- **Documentación:**
 - Cuadros de correspondencias.
 - Parejas: ítem de un nivel, atributo de una tabla o cálculo.



Refinamiento del Esquema

- **Aplicar pautas de diseño.**
 - Refinar jerarquías de dimensiones.
 - Medidas calculadas.
 - Funciones de roll-up.
 - Estudio de aditividad (roll-up suma):
 - Para cada dimensión.
 - Para los niveles de las dimensiones.
 - Restricciones de integridad.
- **Documentación:**
 - Esquema conceptual.
 - Cuadros de roll-up.



Estudio de Aditividad

■ Cuadros de Roll-Up.

	fecha	estudiante	departam	materia
rendimiento	+	+	+	+
avance	⇒	+		
#materias	PR	NA	NA	+

Referencias:

+ suma
⇒ último período
PR promedio
NA no aditiva



Conclusiones

- **Técnica de diseño**
 - Guiada por los requerimientos de información.
 - Aparecen tareas de mapping.
- **Introducción de un modelo conceptual.**
 - Se busca una notación abstracta y efectiva.
 - Representación de conceptos multidimensionales
- **Pautas de buen diseño.**
 - Construcción de dimensiones y jerarquías.
 - Construcción de relaciones dimensionales.
 - Aditividad en medidas.



Perspectivas

- **Pautas de buen diseño:**
 - Tema en investigación.
 - Formas Normales MD.
 - Criterios de BD Estadísticas.
 - Trabajo en el InCo - CSI.
- **Herramientas CASE:**
 - Sistemas comerciales orientados a diseño lógico.
 - Carencias en nivel conceptual.
 - Trabajo en el InCo - CSI.



Estrategia basada en datos

- **Temas:**
 - Modelo Dimensional Fact (DF).
 - Metodología.
 - Ejercicios.



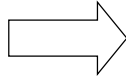
- **Principios**
 - Un diseño conceptual es una actividad necesaria
 - buena documentación
 - cumplimiento de los requerimientos
 - Un dw se apoya en conceptos multidimensionales
- **Sin embargo,**
 - No hay demasiados trabajos sobre cómo realizar un diseño conceptual de un dw.
 - El modelo ER no es un modelo conceptual aceptado como adecuado para dw.



Enfoque de la solución

- **Proponer un modelo conceptual para dws.**
- **Proponer una metodología semi-automática:**

bd operacional
descrita por un
esquema ER



modelo conceptual
de un dw

- La metodología puede también aplicarse a partir de un esquema de bd relacional con info. sobre las FKs.



El modelo conceptual DF

- **Modelo conceptual de un dw [GoI98a]**
= { esquema hecho }
- **Componentes básicos de un esquema hecho:**
 - Hecho
 - Dimensiones
 - Jerarquias



DF / Hecho

■ **Hecho**

- Corresponde a **eventos** que ocurren dinámicamente asociados a aquello que se desea analizar.
- Ejemplos:
 - En un comercio: venta de productos
 - En un hospital: admision de pacientes
 - En un operador telefónico: llamada telefónica



DF / Dimensiones – Jerarquías

■ **Dimensiones**

- Una dimensión define la granularidad elegida para representar hechos.

■ **Jerarquías**

- Una jerarquía define *la manera relevante para el proceso de toma de decisiones* en que las instancias hecho pueden ser seleccionadas y agrupadas (aggregates).

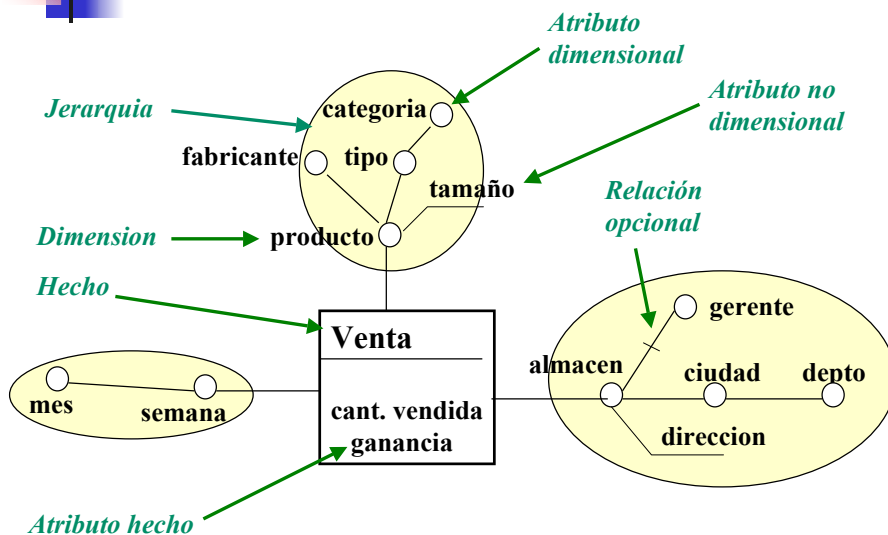


DF / Representación gráfica

- Un *esquema hecho* se estructura como un "árbol".
- La raíz es un *hecho*
 - Puede contener *atributos hecho o medidas*
Atributos numéricos que "miden" el hecho
 - Sin medidas, representa eventos.



DF / Representación gráfica





DF / Representación gráfica

- **Nodo círculo** representa *atributo dimensión*.
- **Nodo directamente conectado al hecho** representa una *dimensión*.
 - Un hecho representa una relación N:N entre las dimensiones.
- **"Subarbol" cuya raíz es una dimensión** representa una *jerarquía*.
 - Los arcos representan relaciones N:1
 - La dimensión define la granularidad "más fina" de la jerarquía.



DF / Representación gráfica

- **Atributos no dimensionales**
 - Representa info. adicional sobre un atributo de una jerarquía.
 - No pueden ser utilizados para agrupar
 - Se representan con una línea
- **Relaciones de opcionalidad**
 - Entre pares de atributos
 - Se representan con arcos con una marca



Aditividad de medidas

- **Patrón de consultas típicas en olap**

Buscar *datos resumidos/agrupados* para llenar un reporte estructurado para ser analizado con propósitos estadísticos o de toma de decisiones.

- **Resumir/agrupar (aggregates) requiere:**

1. *criterio de agrupamiento*

Olap utiliza criterios basados en el uso de las dimensiones y jerarquias.

2. *un operador* para combinar los valores de medidas de cada instancia hecho en valores de cada grupo.



Aditividad de medidas

- **No todas las combinaciones entre *criterios de agrupamiento* y *operadores* resultan en *resúmenes "consistentes"*.**

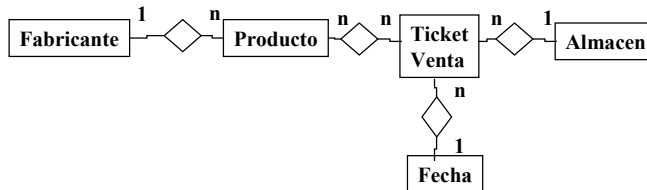
- **Algunas causas de combinaciones problemáticas:**

- Significado de la medida
- Cardinalidad de la relación entre el hecho y una dimensión

Aditividad de medidas

- **Cardinalidad de la relación entre los correspondientes en oltp al hecho y a una dimensión**

E.g. Cant. de clientes (estimada contando la cantidad de tickets)
Dimensiones: Producto, Almacen, Fecha

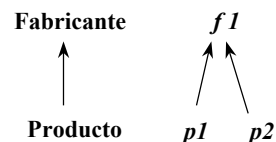


Agrupar por fabricante, sumando la cantidad de clientes produciría un resultado inconsistente.

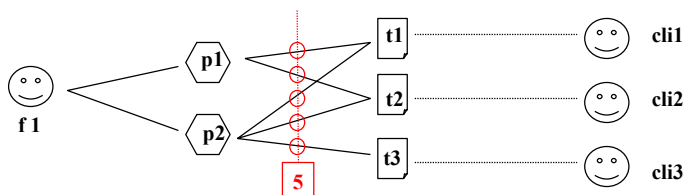
- El mismo ticket puede incluir varios productos.

Aditividad de medidas

Producto	Almacen	Fecha	CantClientes
<i>p1</i>	<i>a1</i>	<i>d1</i>	2
<i>p2</i>	<i>a1</i>	<i>d1</i>	3



Para *f1*, 5 tickets puede ser inconsistente ya que el "2" y el "3" pueden ser el resultado de contar más de una vez 1 ticket de venta con los productos *p1* y *p2*.





Aditividad de medidas

- **Aditiva**
 - Una medida es aditiva si el operador *sum* puede ser aplicado para resumir sobre todas las jerarquías.
 - E.g. cant. vendida
- **Semi-aditiva**
 - Una medida es semi-aditiva si no es aditiva para una o más dimensiones.
 - E.g. nivel de inventario
- **No aditiva**
 - E.g. temperatura
- **Las medidas semi-aditivas y no-aditivas podrían ser resumidas usando otros operadores (e.g. avg, min).**

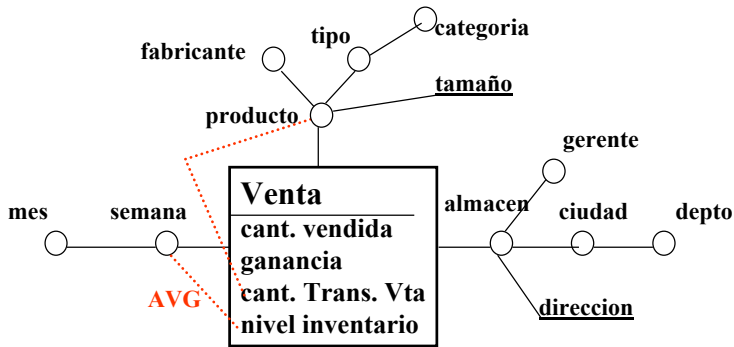


DF / Aditividad

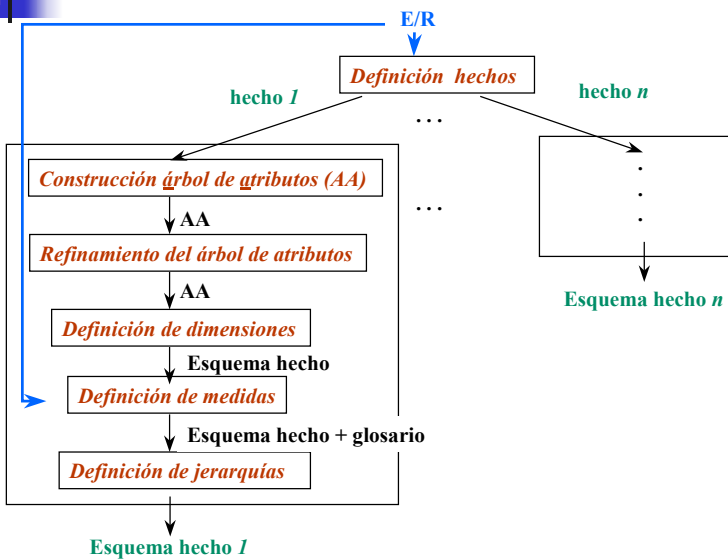
- **DF permite especificar la aditividad de las medidas.**
 - Por defecto, los medidas son aditivas.
 - Se incorporan elementos gráficos para representar las semi-aditivas y no-aditivas



DF / Aditividad



Metodología de diseño conceptual



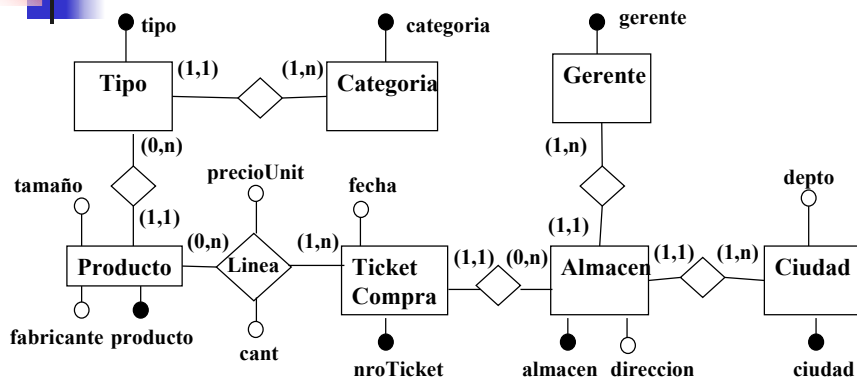


E/R de partida

- **Puede contener jerarquias de generalización**
 - Se consideran equivalentes a relaciones 1:1 entre la super-entidad y cada sub-entidad.
- **Atributos simples.**



Ejemplo / ER de partida



- **C/instancia de *Linea* representa al vínculo de un único producto dentro de un ticket de compra.**



Definición de hechos

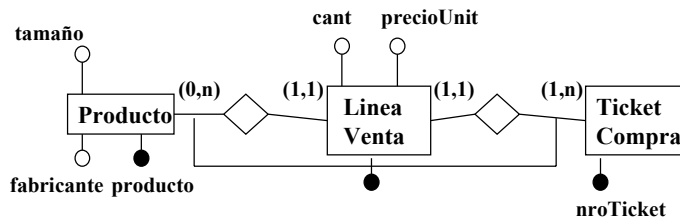
- **Un hecho**
 - es un concepto de interés para el proceso de toma de decisiones,
 - corresponde a eventos que ocurren dinámicamente en el negocio de la empresa
- **Puede ser representado en el E/R por:**
 - una *entidad F*
 - *relación n-aria entre entidades* (Sugerencia: transformar relación en entidad)



Definición de hechos

- **Guía general**
 - Entidades o relaciones que representan registros actualizados frecuentemente son buenos candidatos (e.g. *Linea de Venta*).
 - Aquellos representando propiedades estáticas del dominio no lo son (e.g. *Producto*).
 - *Nota:* Esta guías son independientes de los req. del usuario.
- **C/hecho identificado en el E/R se transforma en la raíz de un esquema hecho diferente.**

Def. de hechos / Transformación del E/R



Construcción del árbol de atributos

■ Def. árbol de atributos (AA)

Dada una porción $P_{E/R}$ de interés de un esquema E/R y una entidad $F \in P_{E/R}$, se llama **árbol de atributos** a un árbol tal que:

- c/nodo se corresponde con un atributo de $P_{E/R}$,
- la raíz se corresponde con el identif. de F ,
- para c/ nodo n , el atributo determina funcionalmente a todos los atributos correspondientes a los descendientes de n .

Construcción del árbol de atributos (2)

- Dada una entidad F elegida para representar un hecho, el árbol de atributos puede ser construido automáticamente mediante $translate(F, id(F))$.

translate (E,v)

{ for each atributo $a \in E / a \neq id(E)$

do addchild(v,a);

for each entidad G conectada a E por una relación

R x-to-1 do

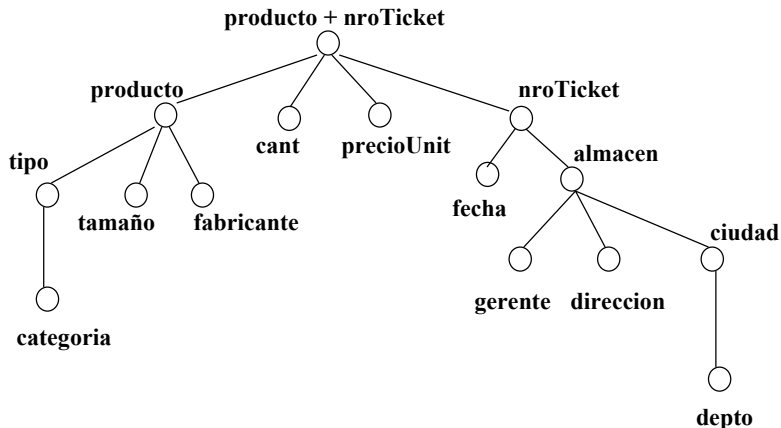
{ for each atributo $b \in R$ do addchild(v,b);

addchild(v, id(G));

translate(G, id(G)); }

}

Arbol de atributos / Ejemplo





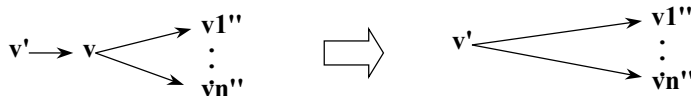
Afinado del árbol de atributos

- **No todos los atributos del árbol obtenido pueden ser de interés.**
 - Se requiere información de los requerimientos del usuario o del dominio de la aplicación.
- **Se proponen 2 operaciones:**
 - Podar (pruning)
 - Injertar (grafting)
- **Podar consiste en eliminar un subárbol desde la raíz.**
 - E.g. *ciudad*.



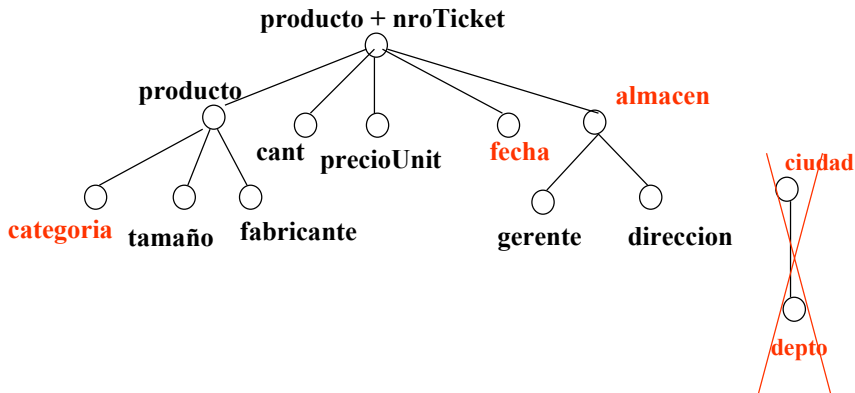
Afinado del AA / Injertar

- **Injertar en v consiste en:**



- Se aplica cuando a pesar de que un nodo del árbol contenga info. no relevante, los descendientes se deben mantener.
 - E.g. 1 - interesa agrupar por *categoría* pero no por *tipo* de productos.
 - E.g. 2 - no interesa el detalle a nivel de *ticket*

Afinado del AA / Ejemplo



Consecuencias de Injertar

- **Injertar en un nodo hijo h de la raíz de un esquema hecho, corresponde a "disminuir" el nivel de granularidad de las instancias hecho.**
 - Si el nodo h tiene **más de un hijo**, injertar **aumenta la cantidad de dimensiones del esquema hecho.**
- **Cuando el nodo donde injertar tiene una relación de opcionalidad, todos sus hijos heredan la relación de opcionalidad**



Definición de dimensiones

- La **elección de las dimensiones** es **crucial en el diseño de un dw** ya que determinan la **granularidad** de las instancias hecho.
 - Necesidad de información a partir de requerim.
- Las **dimensiones** determinan la manera en que las instancias hecho pueden ser **resumidas** (aggregated) en el **proceso de toma de decisiones**.



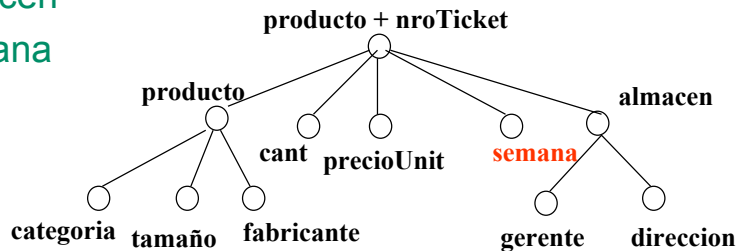
Definición de dimensiones (1)

- La dimensión **tiempo** es una dimensión clave para dws.
- **Diseño de dw a partir de un E/R temporal:**
 - **tiempo** aparece explícito y es candidato directo a dimensión.
 - puede requerir afinamiento del AA mediante la operación injertar para que sea hijo de la raíz.
- **Diseño de dw a partir de un E/R snapshot:**
 - **tiempo** debería ser agregado como dimensión



Dimensiones / Ejemplo

- Asumamos que los análisis más detallados son por semanas.
- Las dimensiones serán:
 - Producto
 - Almacén
 - Semana



Definición de medidas

- Ejemplos de *medidas*:
 - *contar* la cantidad de instancias de una entidad F representando a un hecho.
 - *sum/avg/min/max* de expresiones sobre atributos del AA no dimensiones.
 - un hecho puede no tener medidas.



Def. de medidas / Glosario

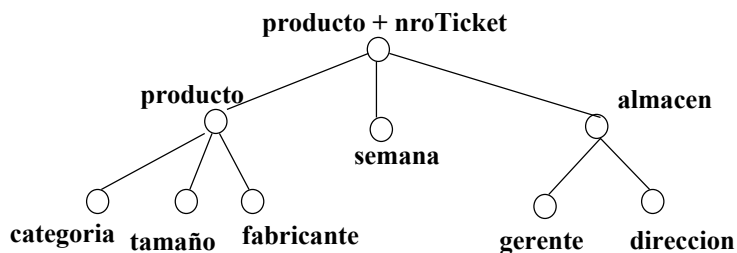
- **Definición de un glosario:**
 - Asocia a c/medida una expresión describiendo el cálculo a partir de los atributos del esquema E/R.
- **Ejemplos**
 - $cantVendida = sum(LineaV.cant)$
 - $ganTotal = sum(LineaV.cant * LineaV.precioUnit)$
 - $cantTransVenta = count(LineaV)$

Estos operadores se interpretan como aplicados sobre todas las instancias de la entidad *LineaV* del E/R que relacionan la misma *semana*, *almacen* y *producto*



Definición de medidas

- **Podar atributos numéricos no relevantes y sin sentido.**
 - E.g. *cant* y *precioUnit*





Definición de jerarquías

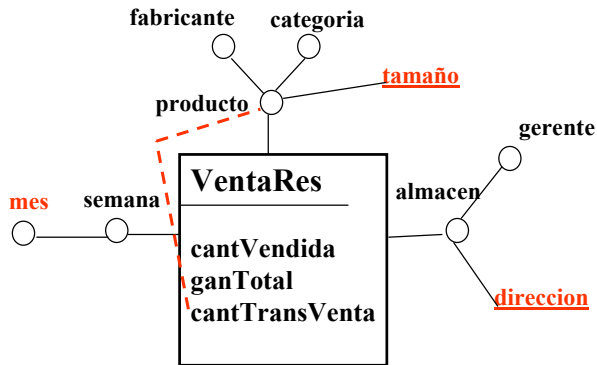
- **El AA muestra posibles jerarquías.**
- **El AA puede ser refinado.**
 - Podar e injertar para eliminar detalles no relevantes.
 - E.g. relaciones 1:1.
 - Agregar nuevos niveles de resumen definiendo rangos para los atributos numéricos.
 - E.g. en la dimensión *semana*, agregar el atributo *mes*.



Definición de jerarquías

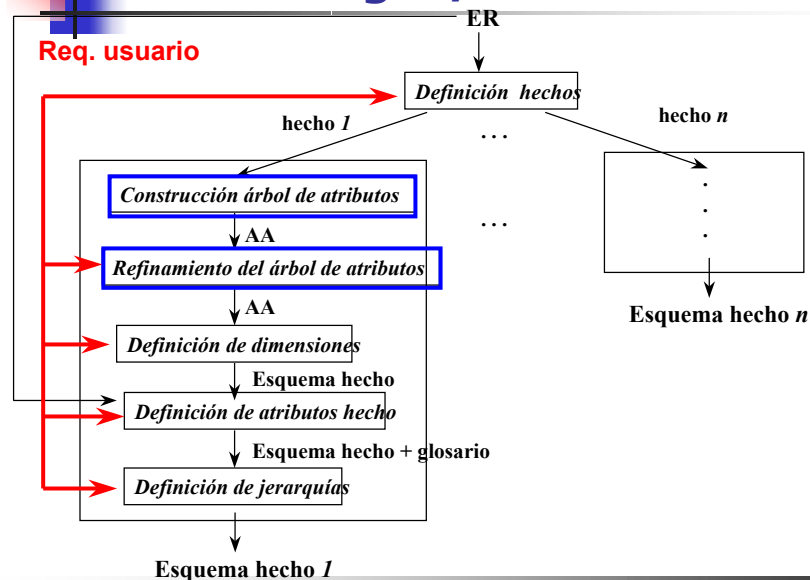
- **Identificar los *atributos no dimensionales***
 - No usados para resumir sino informativos
 - E.g. dirección, tamaño
- **Especificar aditividad de las medidas.**

Def. Jerarquías / Ejemplo



$cantVendida = \text{sum}(Venta.cant)$
 $ganTotal = \text{sum}(Venta.cant * Venta.precioUnit)$
 $cantTransVenta = \text{count}(LineaV)$

Metodología / Resumen



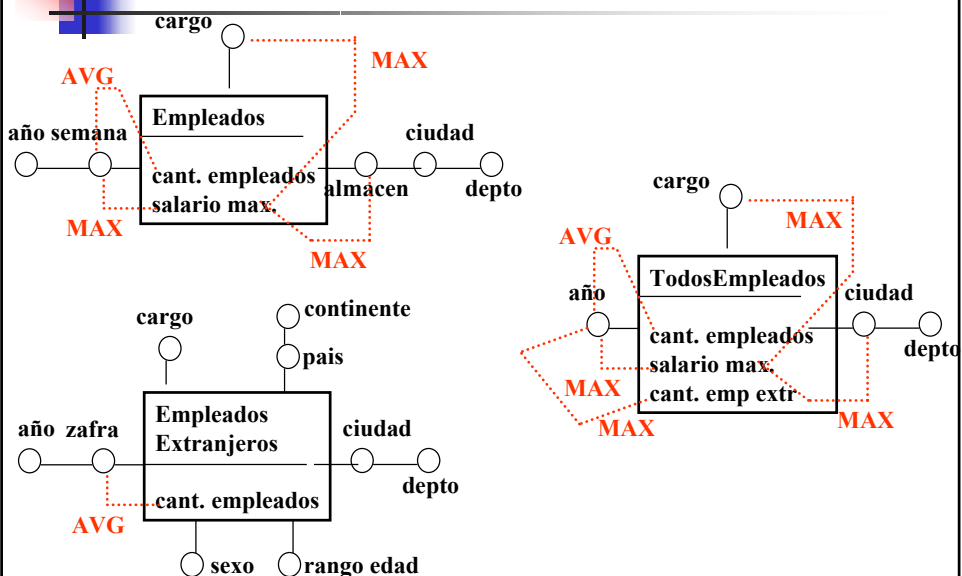


DF / Sobreposición de esquemas hecho

- En DF, hechos diferentes son representados por esquemas hecho diferentes.
- Consultas "*drill accross*"
 - E.g. comparar medidas de esquemas distintos pero relacionados.
- Propuesta:
 - Esquemas hecho compatibles
 - Comparten al menos una dimensión.
 - Definir el resultado de sobreponer dos esquemas hecho compatibles.



DF / Sobreposición



Resumen

- **Metodología de diseño conceptual**

The diagram illustrates the process of conceptual design for a data warehouse. It is organized into three horizontal layers separated by dotted lines. The top layer, labeled 'Req. Usuario + Meta-info.', shows an arrow pointing from the user requirements to the 'Metodología de diseño lógico' (Logical Design Methodology) in the middle layer. The middle layer, labeled 'DW', contains a box for 'fact schemes' and the 'Metodología de diseño lógico' oval, with an arrow pointing from 'fact schemes' to the methodology and another from the methodology to the 'DW lógico (relacional, multidimensional)' (Logical DW (relational, multidimensional)) box. The bottom layer, labeled 'BDs Operacionales' (Operational Databases), contains a box for 'ER' (Entity-Relationship) and the 'Metodología de diseño conceptual' (Conceptual Design Methodology) oval. Arrows show the flow from 'ER' to the conceptual methodology, from the conceptual methodology to 'fact schemes', and from 'Req. Usuario + Meta-info.' to both 'fact schemes' and 'Metodología de diseño lógico'.

Facultad de Ingeniería - In.Co. *Sistemas de Data Warehousing* 149

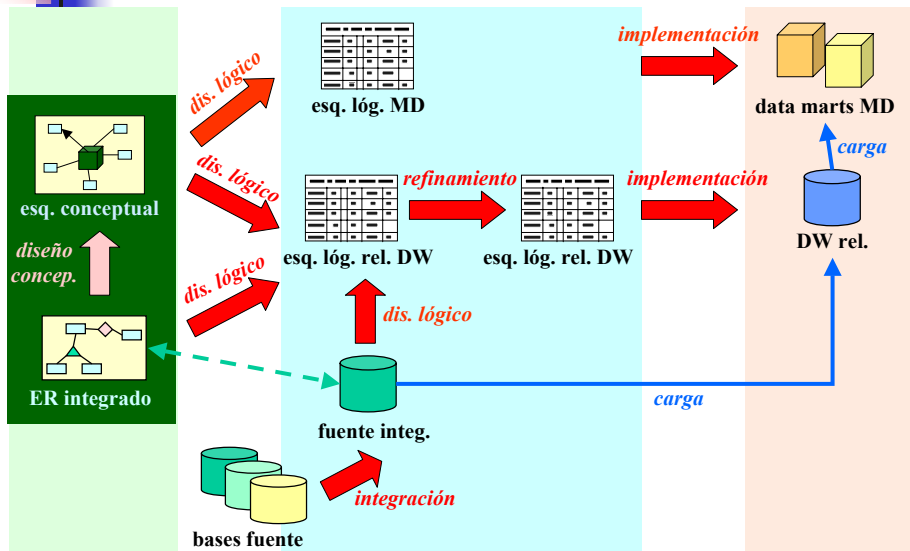
Diseño Conceptual

Conclusiones

- **Temas:**
 - Proceso de diseño.
 - Enfoques de diseño conceptual.
 - Bibliografía.

Facultad de Ingeniería - In.Co. *Sistemas de Data Warehousing* 150

Proceso de Diseño



Enfoques de diseño conceptual

- **Análisis desde requerimientos:**
 - Aplicable cuando se tienen Bases Fuentes complejas. (Se analizan con los requerimientos en mente).
 - Trabajos: [Car00], [Sap99], [Hus00], [Fra99].
- **Análisis desde datos:**
 - Aplicable cuando los requerimientos están poco claros.
 - Trabajos: [Gol98a], [Cab98].



Diseño Conceptual

- **Las principales etapas son:**
 - Definir un esqueleto de esquema:
 - Primer grupo de dimensiones medidas.
 - Establecer correspondencia entre requerimientos y datos fuentes.
 - Completar jerarquías en las dimensiones.
 - Especificar segundo grupo de medidas (calculadas).

iterar



Bibliografía

- **[Bal98]** Ballard, C. Herreman, D. Schau, D. Bell, R. Kim, E. Valncic, A.: "Data Modeling Techniques for Data Warehousing". SG24-2238-00. IBM Red Book. 1998.
- **[Cab98]** Cabibbo, L. Torlone, R.: "A Logical Approach to Multidimensional Databases", EDBT, 1998.
- **[Car00]** Carpani, F.: "CMDM: A conceptual multidimensional model for Data Warehouse". Master Thesis. Advisor: Ruggia. InCo - Pedeciba, UdelaR, Uruguay, 2000.
- **[Fra99]** Franconi, E. Sattler, U.: "A Data Warehouse Conceptual Data Model for Multidimensional Aggregation", DMDW'99, Germany, 1999.
- **[Gol98a]** Golfarelli, M. Maio, D. Rizzi, S.: "Conceptual Design of Data Warehouses from E/R Schemes.", HICSS'98, IEEE, Hawaii, 1998.



Bibliografía

- **[Hus00]** Hüsemann, B. Lechtenbörger, J. Vossen, G.: "Conceptual Data Warehouse Design". DMDW'00, Sweden, 2000.
- **[Ken96]** Kenan Technologies: "An Introduction to Multidimensional Databases". White Paper, Kenan Technologies, 1996.
- **[Kim96]** Kimball, R.: "The Datawarehouse Toolkit ". John Wiley & Son, Inc., 1996.
- **[Pic00]** Picerno, A. Fontan, M.: "Un editor para CMDM". Undergraduate Project. Advisor: Carpani. InCo, UdelaR, Uruguay. 2000.
- **[Sap99]** Sapia, C. Blaschka, M. Höfling, G. Dinter, B.: "Extending the E/R Model for the Multidimensional Paradigm". DWDM'98, Singapure, 1998.
- **[Tho97]** Thomsen, E.: "OLAP Solutions. Building Multidimensional Information". John Wiley & Sons, Inc. 1997.