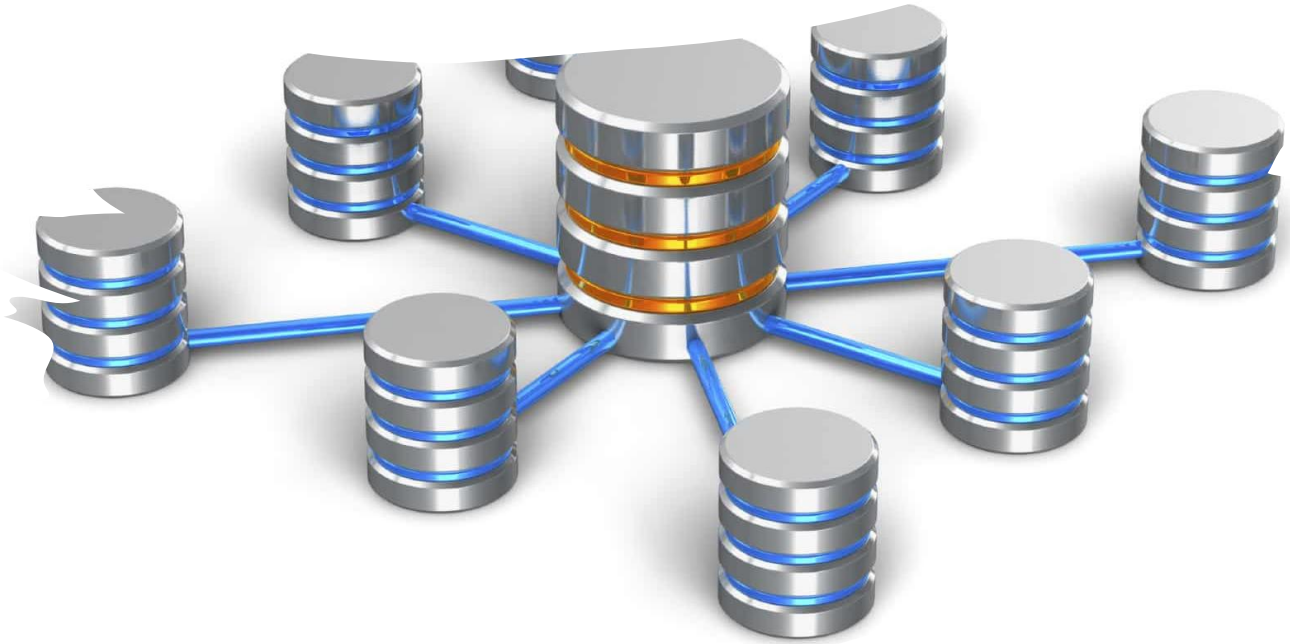


Almacenes de datos (Data Warehouse - DW)

Dr. Luis Gustavo Esquivel Quirós

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

Modelo conceptual multidimensional

- En la comunidad de bases de datos se reconoce que los modelos conceptuales:
  - Permiten una mejor comunicación entre diseñadores y usuarios, con la finalidad 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
  - Proporcionan un mejor soporte para las interfaces de usuario visuales
- Sin embargo, existe escaso interés en el modelado conceptual multidimensional

Modelo conceptual multidimensional

- Actualmente no existe un modelo conceptual bien establecido, 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 (herramientas automatizadas)

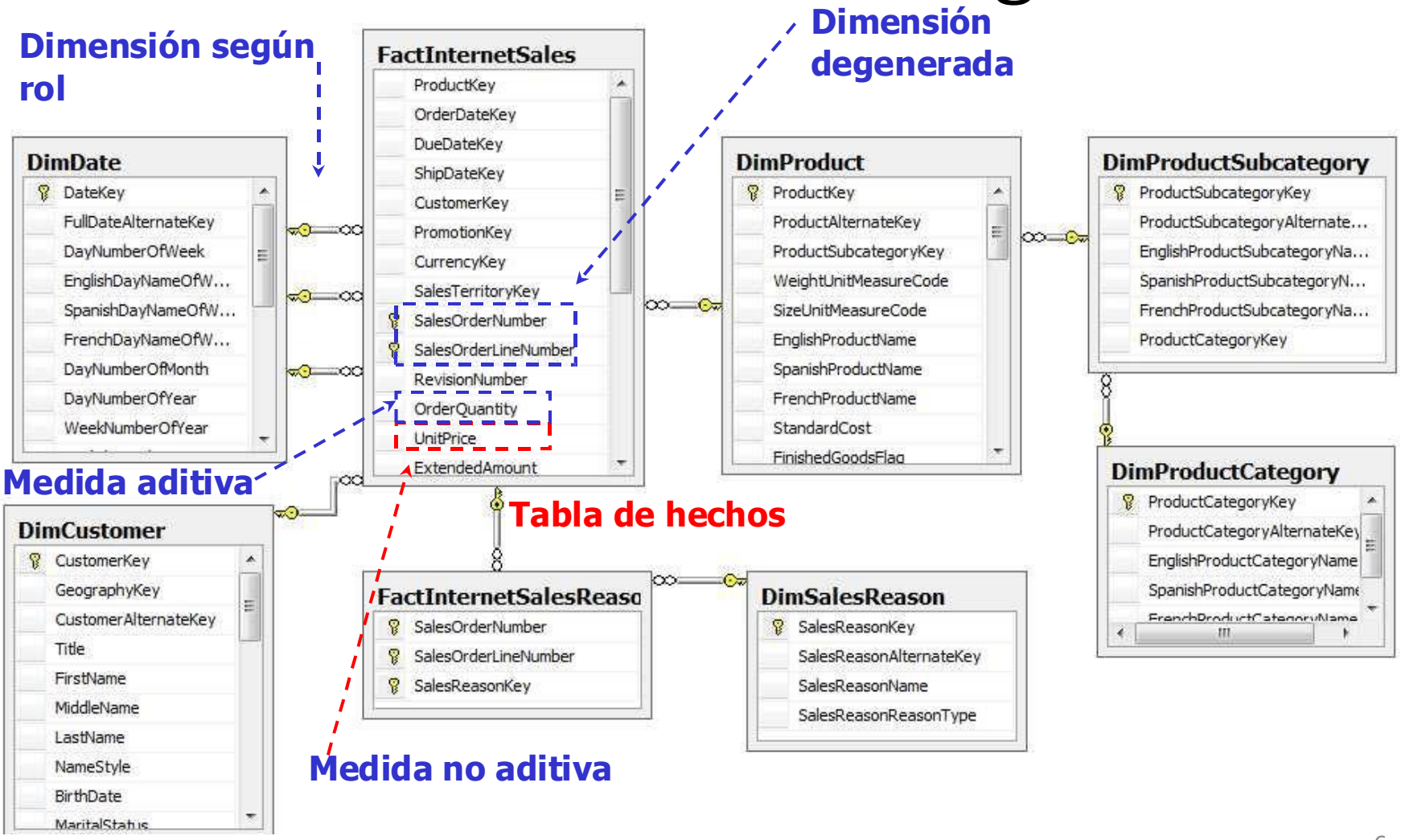
Modelo conceptual multidimensional

- En la actualidad los almacenes de datos se diseñan utilizando en su mayoría modelos lógicos (esquemas de estrella y copo de nieve)
- Esto produce que los usuarios tengan dificultades para expresar sus requerimientos, ya que se requieren conocimientos especializados relacionados con temas técnicos

Modelo conceptual multidimensional

- 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

Diferentes tipos de medidas y dimensiones: modelo lógico



Diferentes tipos de jerarquías: Múltiples jerarquías

- Se puede crear varias jerarquías para una dimensión cuando desea organizar los miembros de la dimensión de diferentes maneras.
  - Por ejemplo, en una dimensión de tiempo, puede crear jerarquías para el año calendario y el año fiscal. Debido a que los miembros de dimensión en jerarquías separadas se pueden usar para representar la misma entidad, cada jerarquía debe contener los mismos miembros de nivel más bajo.
    - Por ejemplo, en una dimensión de tiempo, la jerarquía del calendario puede tener niveles de año, mes y día. La jerarquía fiscal puede tener niveles de año, trimestre y día. El nivel más bajo en ambas dimensiones es el nivel Día.

Diferentes tipos de jerarquías: jerarquías balanceadas

- Todas las ramas de la jerarquía descienden al mismo nivel. El padre de cada miembro proviene del siguiente nivel más alto.
- Se puede usar una jerarquía equilibrada para representar el tiempo donde el significado y la profundidad de cada nivel, como Año, Trimestre y Mes, son consistentes. Son coherentes porque cada nivel representa el mismo tipo de información y cada nivel es lógicamente equivalente.

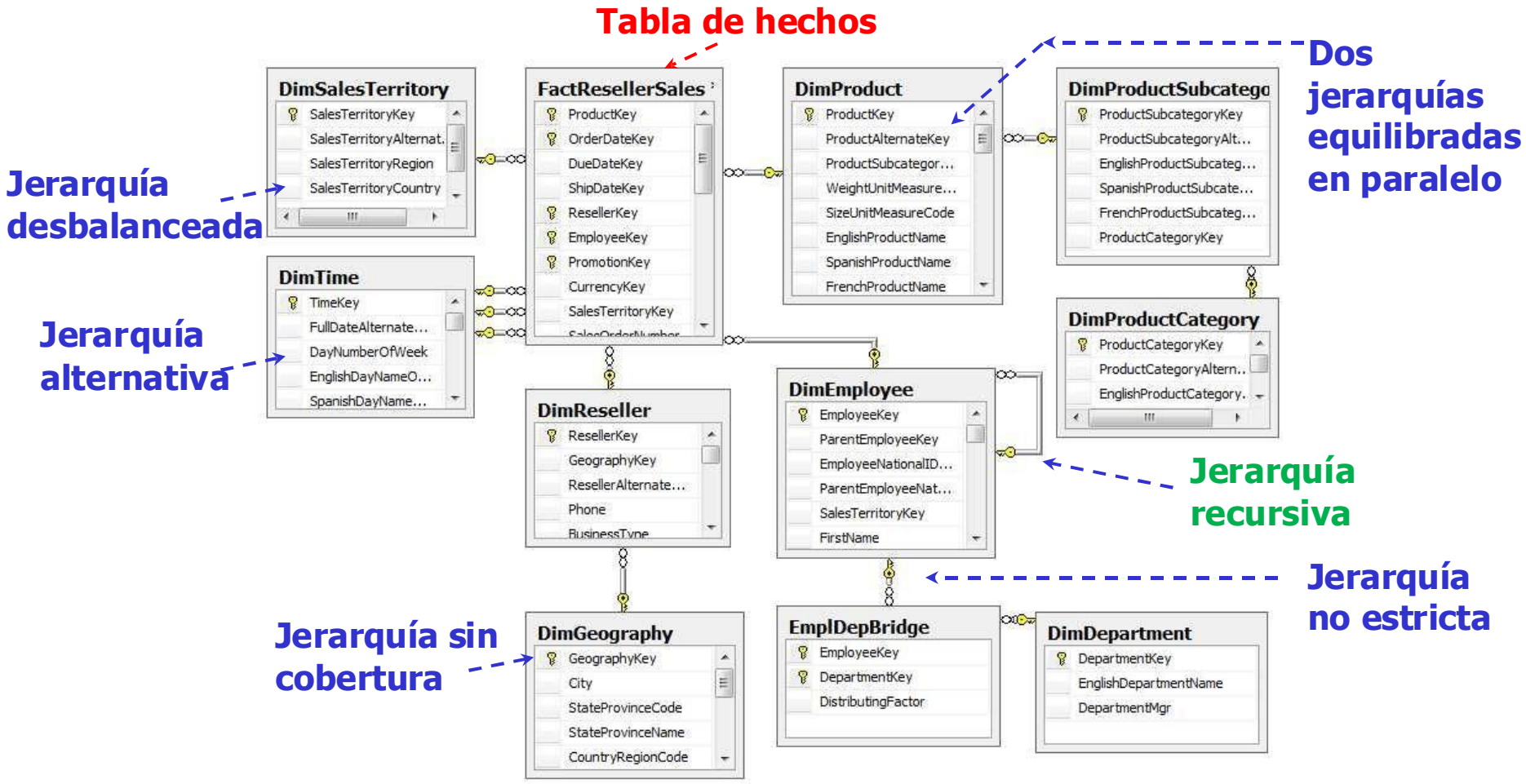
Diferentes tipos de jerarquías: jerarquías desbalanceadas

- Incluyen niveles que son lógicamente equivalentes, pero cada rama de la jerarquía puede descender a un nivel diferente. En otras palabras, una jerarquía desbalanceada contiene miembros hoja en más de un nivel. El padre de cada miembro proviene del nivel inmediatamente superior.
- Un ejemplo son las relaciones jerárquicas entre los empleados de una organización. Los niveles dentro de la estructura organizacional normalmente están desequilibrados, entonces algunas ramas en la jerarquía tienen más niveles que en otras.

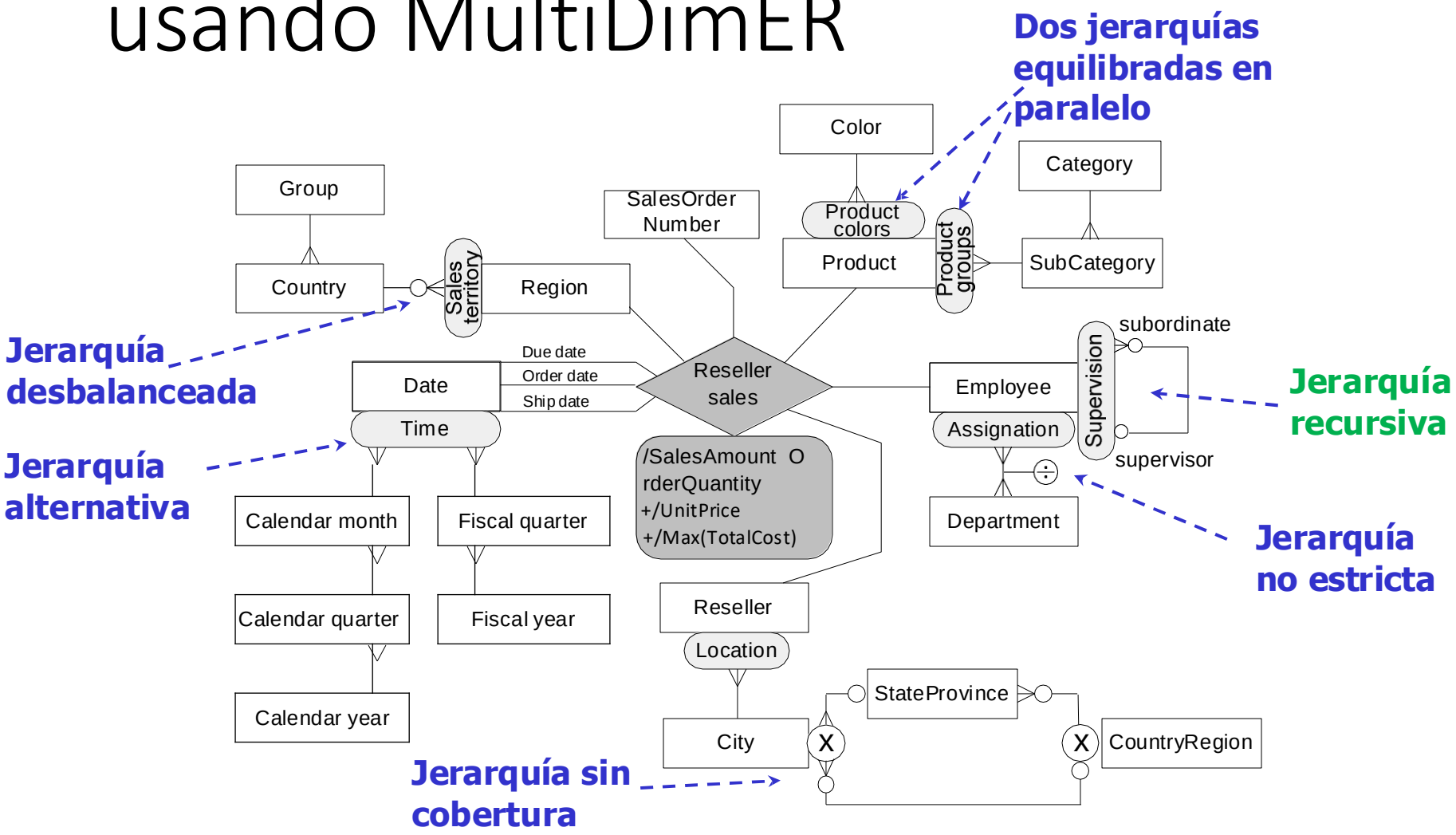
Diferentes tipos de jerarquías: jerarquías sin cobertura o irregular

- En una jerarquía irregular, el padre de al menos un miembro no proviene del nivel inmediatamente superior, sino de un nivel superior.
- Por ejemplo, en una jerarquía geográfica con los niveles de continente, región, estado y ciudad definidos. Una rama tiene América del Norte como continente, Canadá como región, Ontario como estado y Toronto como ciudad. Otra rama tiene Europa como continente, Grecia como región y Atenas como ciudad, pero no tiene entrada para el nivel estatal porque este nivel no es aplicable. El padre de Atenas está a nivel de región en lugar de a nivel estatal, lo que crea una jerarquía irregular.

Diferentes tipos de jerarquías: modelo lógico



Representación conceptual usando MultiDimER

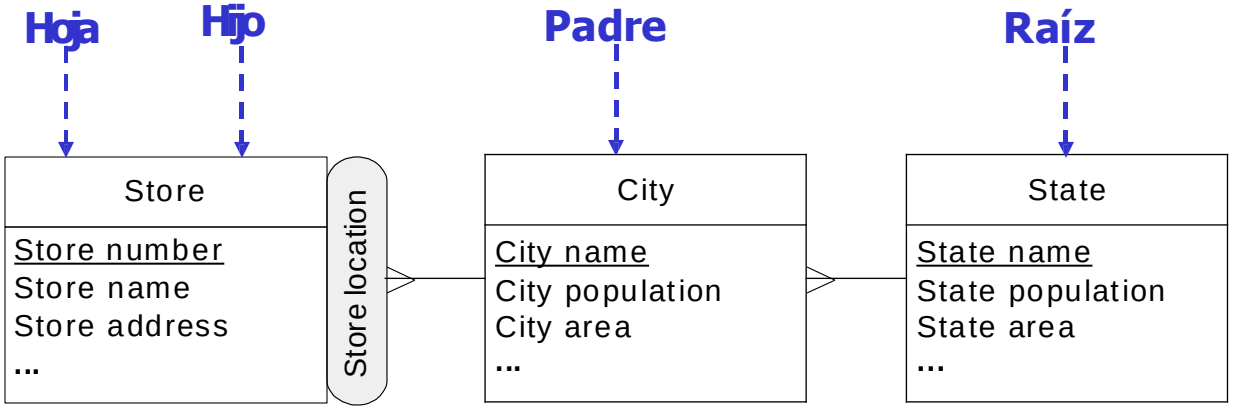


Modelo multidimensional conceptual MultiDimER

- El modelo MultiDimER
  - Se basa en notaciones ER
  - Incluye conceptos como hecho, medida, dimensión y jerarquía
  - Propone clasificaciones y notaciones para diferentes tipos de jerarquías existentes en aplicaciones del mundo real
  - Se puede asignar a estructuras relacionales de estrella o copo de nieve

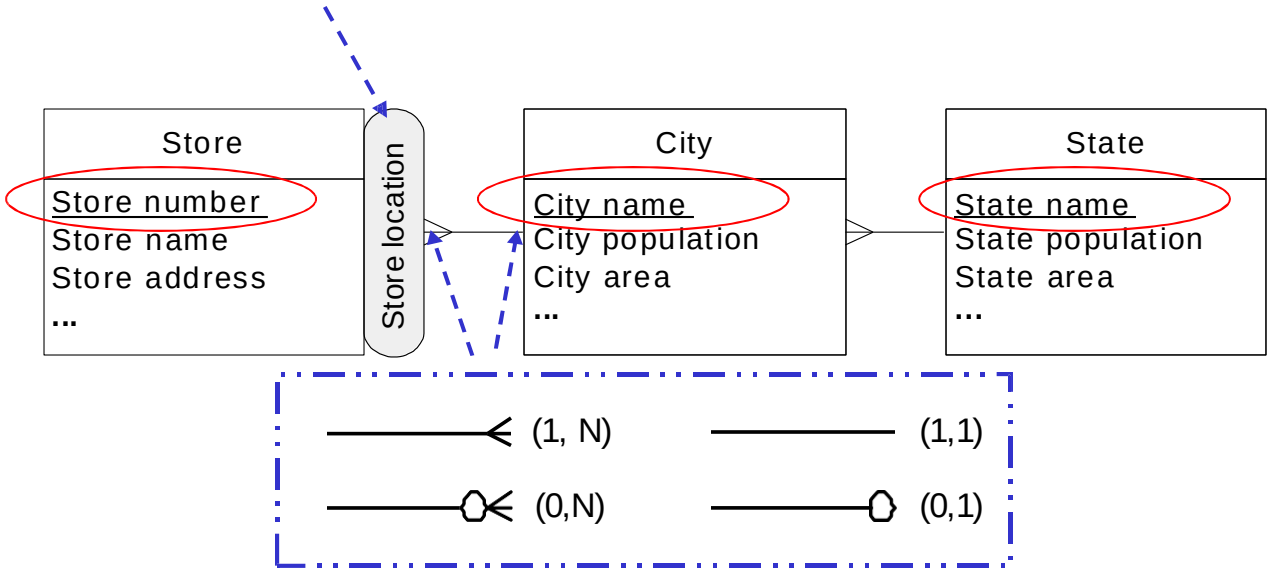
Modelo MultiDimER: notación

- Dimensión: nivel o una o más jerarquías
- Jerarquía: varios niveles relacionados
- Nivel - tipo de entidad
- Miembro: cada instancia de un nivel
- Niveles de hoja y raíz: primer y último nivel en una jerarquía
- Niveles de hijos y padres: los niveles inferior y superior



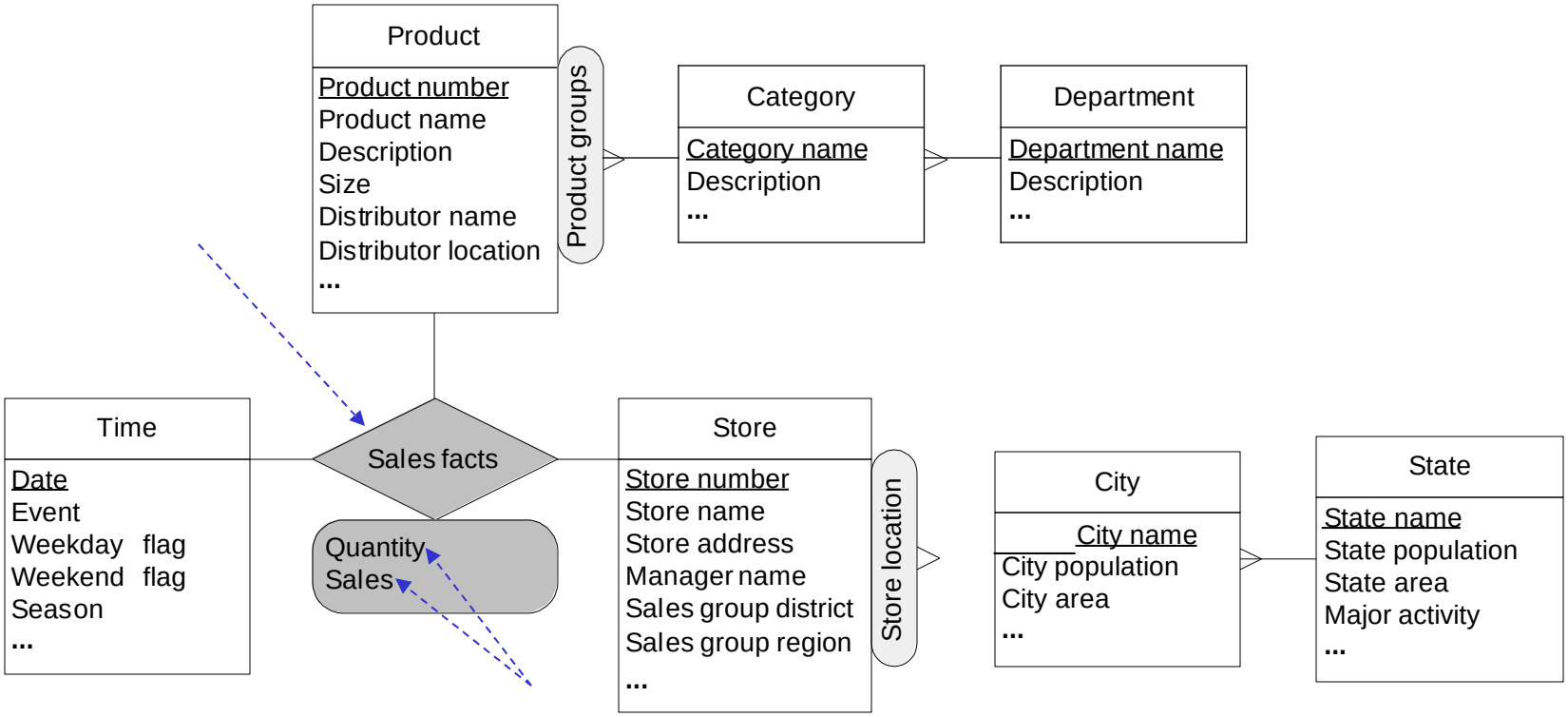
Modelo MultiDimER: notación

- Cardinalidad: el número mínimo y máximo de miembros en un nivel en relación con los miembros en otro nivel
- Criterio: expresa diferentes estructuras jerárquicas utilizadas para el análisis
- Atributo clave: indica cómo se agrupan los miembros secundarios
- Atributos descriptivos: describen las características de los miembros



Modelo MultiDimER: notación

- Relación de hechos - relación n-aria entre niveles de hojas
- Medidas - atributos usados para agregaciones



Simplificación de los pasos de diseño

- Seleccione el tema de análisis
- Seleccione las dimensiones
- Seleccione el nivel de granularidad
- Seleccione las medidas
- Distinga las jerarquías

Supermercado: descripción

- Un ejemplo del libro de Kimbal usando notaciones MultiDim
- La sede de una gran cadena de supermercados desea analizar los datos disponibles
- Hay 500 grandes tiendas de comestibles repartidas en diferentes estados
- Cada tienda es un supermercado moderno típico que incluye comestibles, alimentos congelados, productos lácteos, carnes, productos agrícolas, panadería, flores, productos duraderos, licores y medicamentos
- Cada tienda tiene aproximadamente 60000 productos individuales
- Los productos tienen códigos de barras, 40 000 de ellos provienen de fabricantes externos y 20 000 provienen de departamentos como el de carnes, panadería o flores

Supermercado: descripción

- La información se recoge en los puntos de venta (caja registradora)
- La ganancia proviene de cobrar tanto como sea posible por cada producto, reducir los costos de adquisición del producto y los gastos generales, así como atraer la mayor cantidad de clientes posible.
- Las decisiones gerenciales más importantes tienen que ver con precios y promociones
- Las promociones son de diferentes tipos: reducción temporal de precios, anuncios en periódicos, exhibiciones en las tiendas de comestibles (exhibiciones en estantes, exhibiciones en pasillos y cupones)

Supermercado: descripción

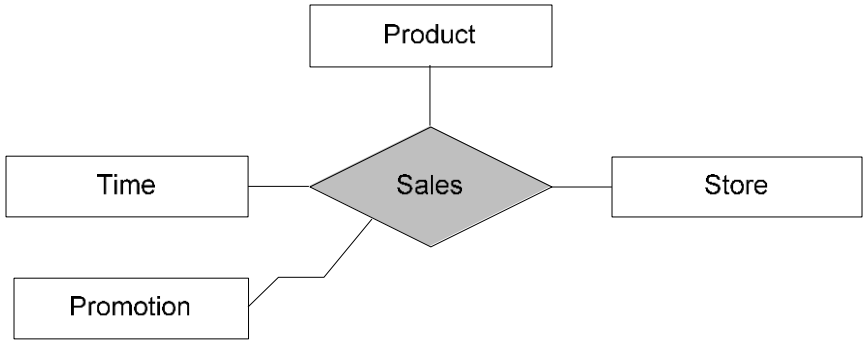
- El almacén de datos debe incluir datos que permitan analizar qué productos se venden en qué tienda, a qué precios, en qué días y durante qué promoción
- La información sobre el comportamiento de compra del cliente no se puede considerar ya que no hay identificación del cliente en el sistema

Supermercado: elementos multidimensionales

- Foco de análisis: ventas de productos
- Dimensiones: Tiempo, Producto, Tienda, Promoción
- Granularidad:
  - ventas diarias
  - cada producto
  - cada tienda
  - cada promoción
- Medidas: cantidad y monto de las ventas

Esquema conceptual del supermercado: primera aproximación

- Representación gráfica mediante una relación de hechos y dimensiones.

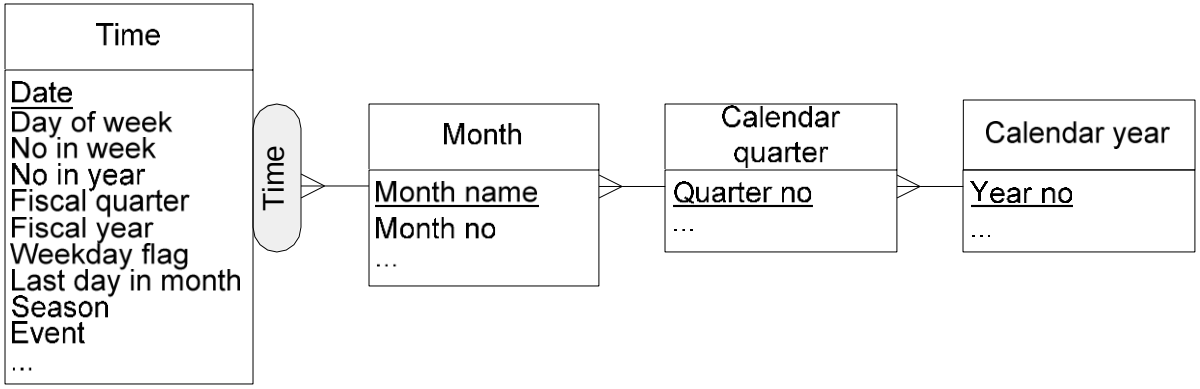


Esquema conceptual: refinando las dimensiones

- Para cada dimensión
  - Defina jerarquías
  - Para cada nivel de jerarquía, defina atributos clave y descriptivos

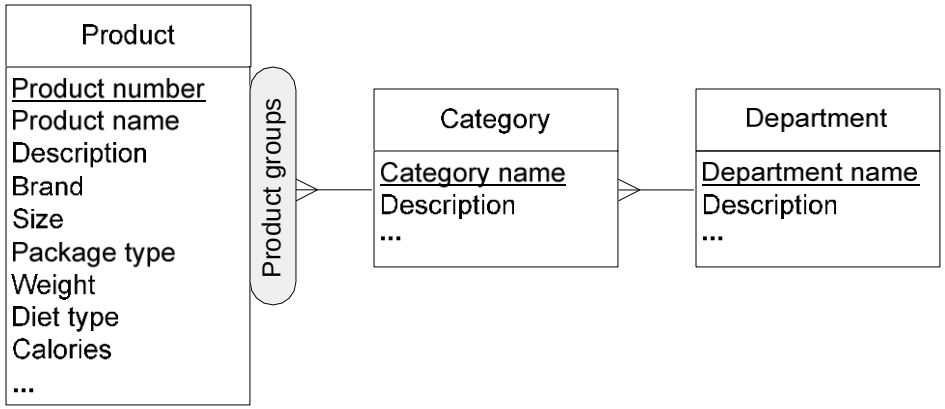
Dimensión de tiempo

- Existe en todos los DW
- Se puede construir por adelantado
- No requiere mucho espacio (10 años = 3650 registros)
- Granularidad diaria para el esquema de supermercado
- Incluye una jerarquía: Día-Mes-Trimestre-Año
- Puede incluir otras jerarquías (calendario fiscal) que en el esquema anterior se usa solo con fines descriptivos
- ¿Por qué es necesario tener una dimensión temporal separada?



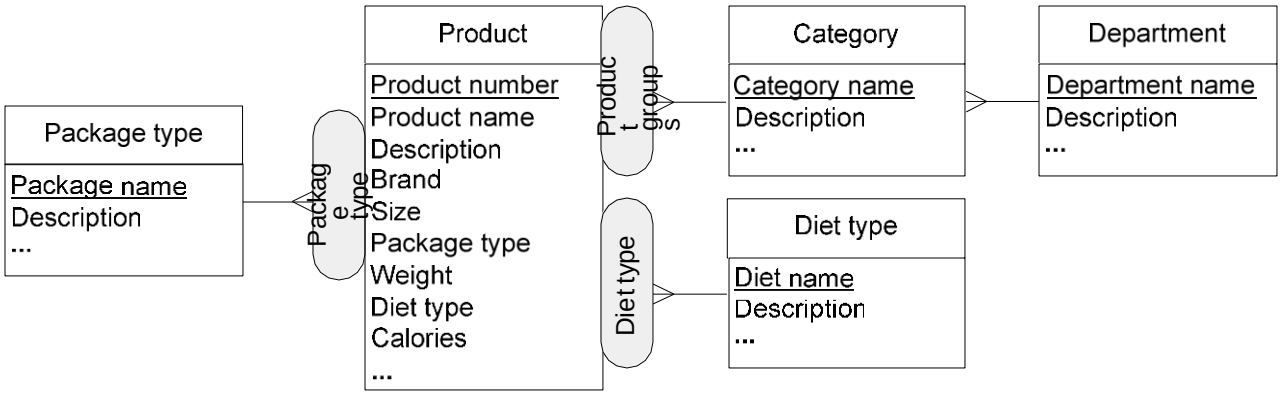
Dimensión del producto

- Contiene datos del archivo maestro del producto
- Debe incluir campos de descripción, sin codificación



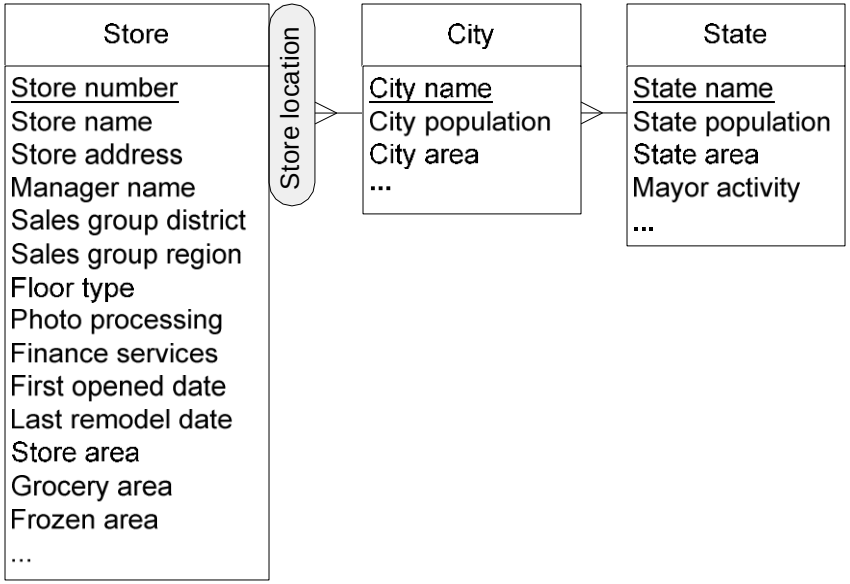
Dimensión del producto

- Ejemplo de varias jerarquías adjuntas a la dimensión Producto
  - Producto – Categoría – Departamento
  - Producto – Tipo de dieta
  - Producto: tipo de paquete (no se muestra)
- Es importante incluir tantos atributos y jerarquías como sea posible

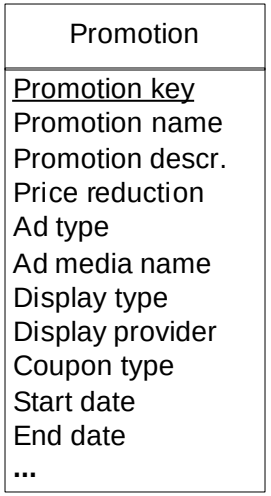


Dimensión de la tienda

- Representa una dimensión geográfica (localización, tipo de clima, etc.)
- Puede incluir diferentes jerarquías según las divisiones administrativas (tienda-ciudad-estado) y organizacionales (división de ventas)
- Puede incluir valores aditivos numéricos (área)



Dimensión de promoción



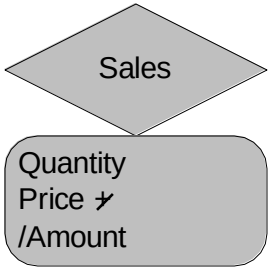
- Dimensión causal
- Hay diferentes tipos de promociones:
  - Reducción temporal de precio
  - Publicación en el periódico
  - Pantalla de pasillo
  - Cupones
- Esta dimensión puede ayudar en diferentes tipos de análisis.
  - Levante – aumento de las ventas
  - Cambio de horario: disminución de las ventas después de la promoción
  - Canibalización: un producto aumenta las ventas pero otros productos relacionados disminuyen las ventas
  - Otros

Dimensión de promoción

- La dimensión de promoción se puede representar como cuatro dimensiones: una para cada tipo de promoción
- Preguntas:
  - ¿Cuáles son las ventajas de utilizar una o varias dimensiones?
  - ¿Sería mejor crear registros separados para cada promoción?
  - ¿Podemos presentar la situación de que un mismo producto tiene tres promociones diferentes en una tienda y dos promociones en otra tienda?

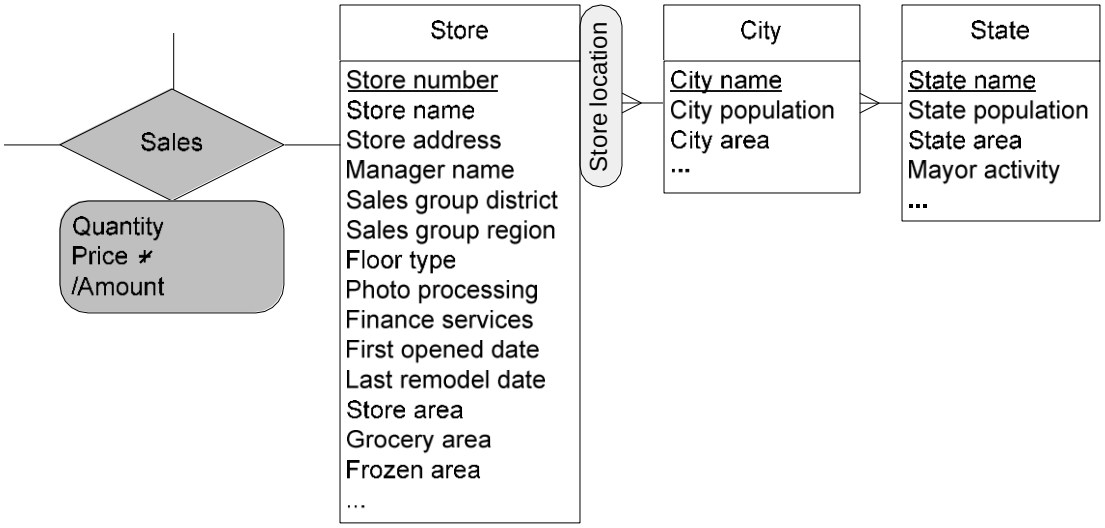
Medidas

- Medida de interés:
  - Cantidad
  - Monto
- La cantidad es una medida calculada (derivada) que utiliza Precio y Cantidad como atributos base
- El precio es una medida no aditiva: no se puede sumar al atravesar jerarquías



Clasificación de medidas

- Las medidas se agregan al atravesar las jerarquías de las dimensiones
- De manera predeterminada, se usa el operador de suma, por ejemplo, "moviéndose" del nivel de Tienda a Ciudad, las cifras de ventas para todas las ciudades en la misma ciudad se agregarán usando el operador de suma



Clasificación de medidas

- Para asegurar la correcta agregación de medidas, las condiciones de sumarización deben cumplirse.
- La capacidad de sumarizar se refiere a la agregación correcta de medidas en un nivel superior de la jerarquía (ejemplo, nivel Estado en…) teniendo en cuenta las agregaciones existentes en un nivel inferior de la jerarquía (ejemplo, nivel de ciudad)
- Las condiciones de sumarización incluyen las siguientes:
  - Disjunción de instancias: la agrupación de instancias en un nivel con respecto a su padre en el siguiente nivel debe dar como resultado subconjuntos disjuntos, por ejemplo, una ciudad no puede pertenecer a dos estados
  - Completitud: todas las instancias están incluidas en la jerarquía y cada instancia está relacionada con un padre en el siguiente nivel, por ejemplo, la jerarquía de ubicación de la tienda contiene todas las tiendas y cada tienda está asignada a una ciudad
  - Uso correcto de las funciones de agregación: las medidas pueden ser de varios tipos, y esto determina el tipo de función de agregación que se puede aplicar.

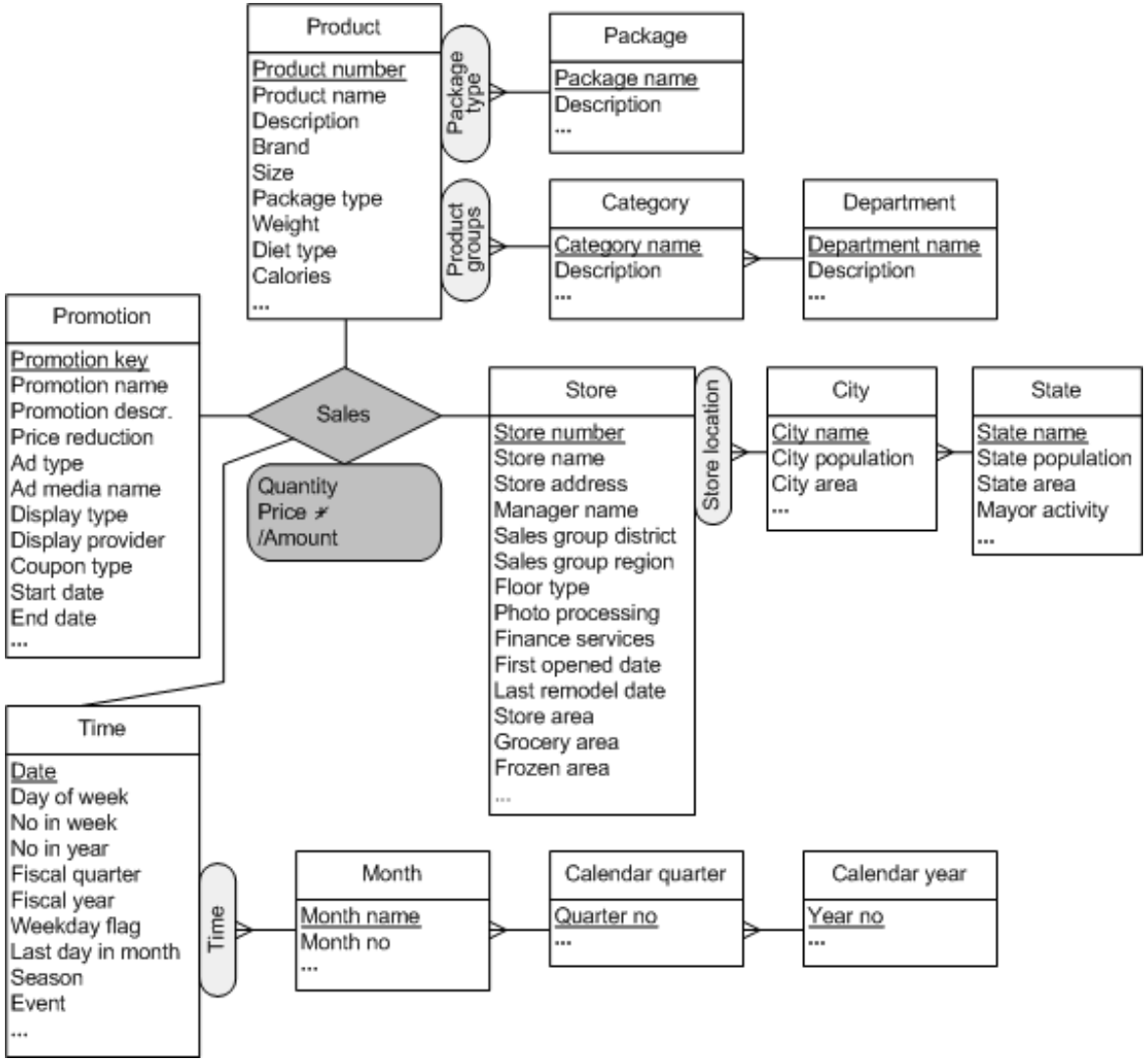
Clasificación de medidas

- Tipos de medida
  - Aditivos (flujo o tasa) cuando se pueden resumir de manera significativa usando la suma a lo largo de todas las dimensiones, por ejemplo, cantidades de productos vendidos
  - Semiaditivos (existencias o nivel) cuando se pueden resumir de manera significativa usando la suma a lo largo de algunas, pero no todas, las dimensiones, por ejemplo, las cantidades de inventario no se pueden agregar a lo largo de la dimensión de tiempo
  - No aditivos (valor por unidad) cuando no se pueden resumir de manera significativa mediante la suma en cualquier dimensión, por ejemplo, precio unitario, tipo de cambio

Clasificación de medidas

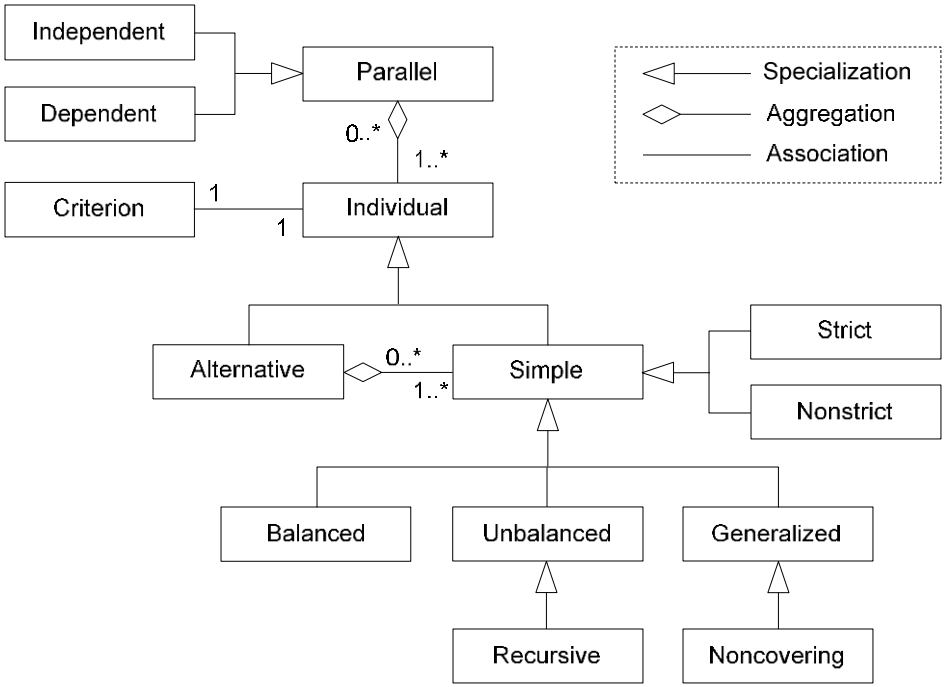
- Los diferentes tipos de medidas exigen un uso adecuado de las funciones para ser agregadas
- Esto es particularmente importante en el caso de medidas semiaditivas y no aditivas
  - Medida de cantidad de inventario
    - Es una medida semiaditiva: se puede agregar en la dimensión Tienda, pero no se puede agregar en la dimensión Tiempo
    - Para garantizar los valores correctos cuando se mueve en la jerarquía de tiempo, se puede usar el promedio
  - Medida de precio
    - Es una medida no aditiva: no se puede agregar en ninguna dimensión
    - Para garantizar los valores correctos cuando se mueve en la jerarquía de tiempo, se puede usar el promedio u otra función como mínimo, máximo
- El modelo MultiDimER proporciona notaciones para indicar diferentes tipos de medidas

Esquema conceptual para el análisis de ventas



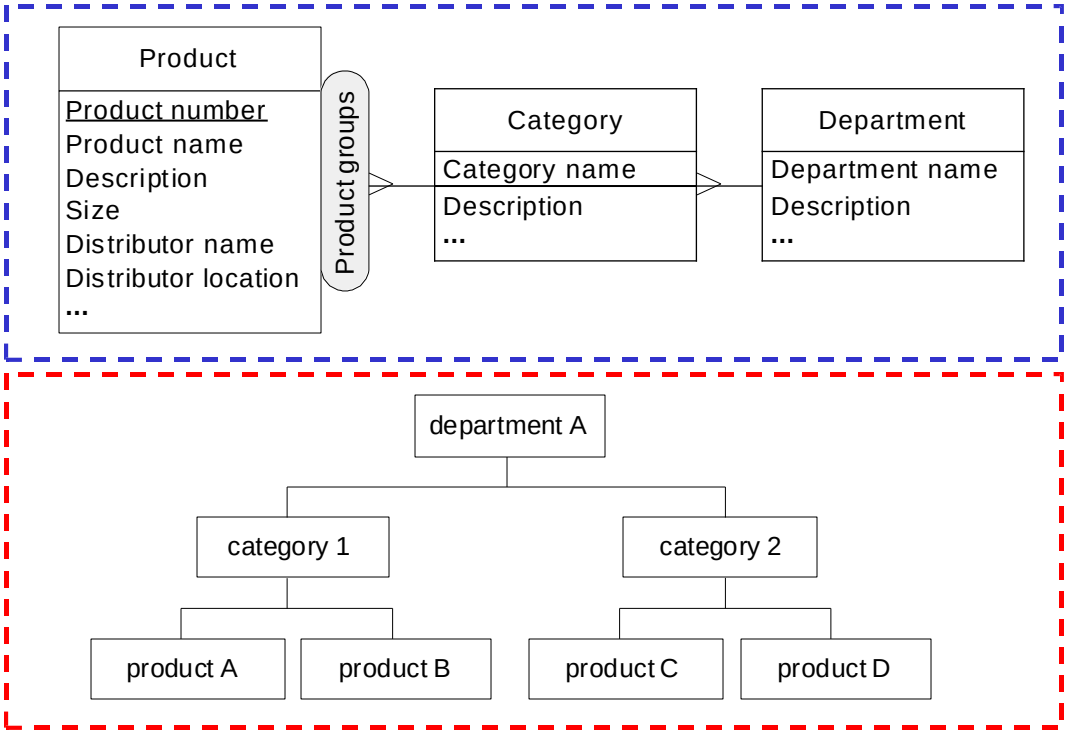
Clasificación de jerarquía

- El modelo MultiDimER incluye clasificación de jerarquías a nivel de esquema e instancia y propone su notación gráfica
- En este curso solo nos referimos a jerarquías equilibradas y recursivas.



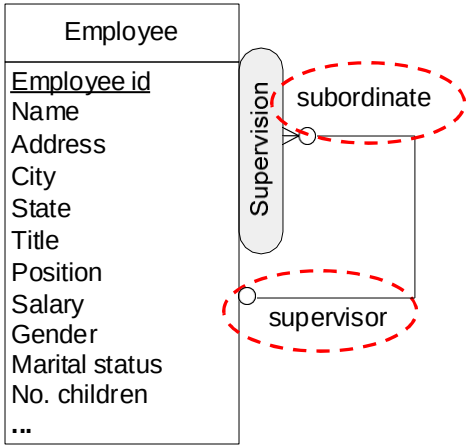
Jerarquía simple: Equilibrada

- A nivel de esquema: solo una ruta
- A nivel de instancia: los miembros forman un árbol equilibrado



Jerarquía simple: recursiva

- El mismo nivel está vinculado por los dos roles de una relación padre-hijo
  - Se utiliza principalmente cuando todos los niveles de la jerarquía expresan la misma semántica, es decir, cuando las características del padre y el hijo son similares (o las mismas)

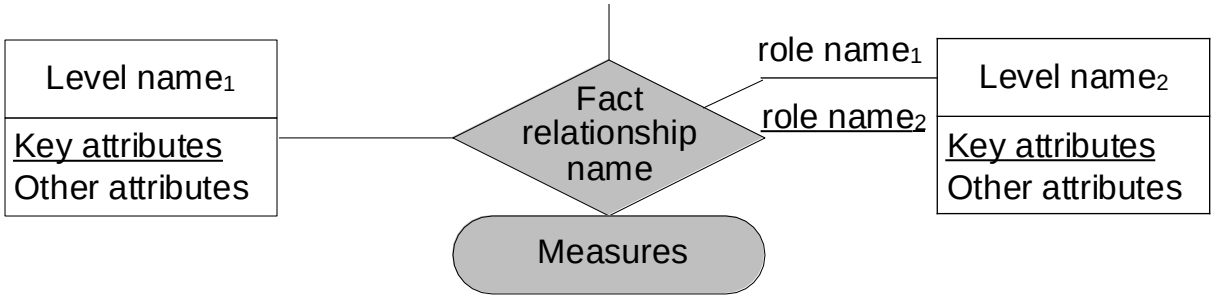


Dimensiones

- En la vida real uno puede encontrar diferentes tipos de dimensiones
  - En este curso solo nos referimos a dimensiones según rol y dimensiones degeneradas.
- En el modelo MultiDimER, una relación de hecho representa una relación n-aria entre niveles de hoja
- Un miembro hoja participa en una relación de hechos de cero a muchas veces
  - Ejemplo:
    - Si algún producto no se ha vendido (todavía), no se hará referencia a él en la relación de hechos de ventas.
    - Si un producto ha sido vendido en diferentes tiendas o en diferentes fechas, participará varias veces en la relación de hechos, cada vez con los valores correspondientes de las medidas Cantidad y Monto.

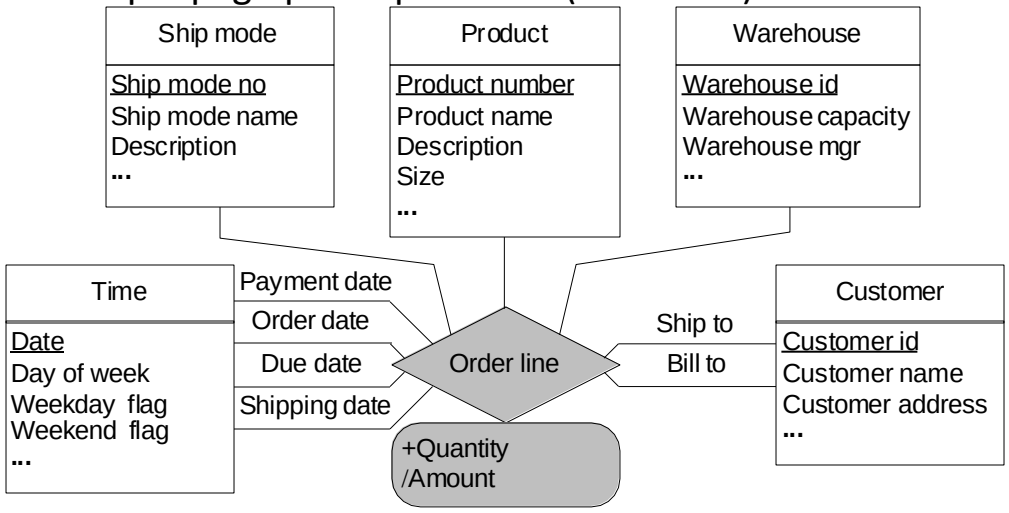
Dimensiones de juego de roles o según rol

- En algunas situaciones, una dimensión puede ser requerida varias veces en la misma relación de hechos
- En lugar de utilizar varias dimensiones, definimos los roles que pueden desempeñar
- En el modelo MultiDimER, para indicar estos diferentes roles, incluimos en el esquema enlaces adicionales entre el nivel de hoja y la relación de hechos, cada uno con el nombre de rol correspondiente



Dimensiones de juego de roles o según rol

- Ejemplo del envío DW:
  - Diferentes roles de la dimensión Tiempo
    - Hora en que un cliente pide un producto (Fecha de pedido)
    - Hora en que se envía el producto (fecha de envío)
    - Hora en que se debe entregar el producto (fecha de vencimiento)
    - Momento en que se paga efectivamente (Fecha de pago)
  - Diferentes roles de la dimensión Cliente
    - Cliente a quien se envía el producto (Envie a)
    - Cliente que pagó por el producto (facturar a)



Dimensiones de juego de roles o según rol

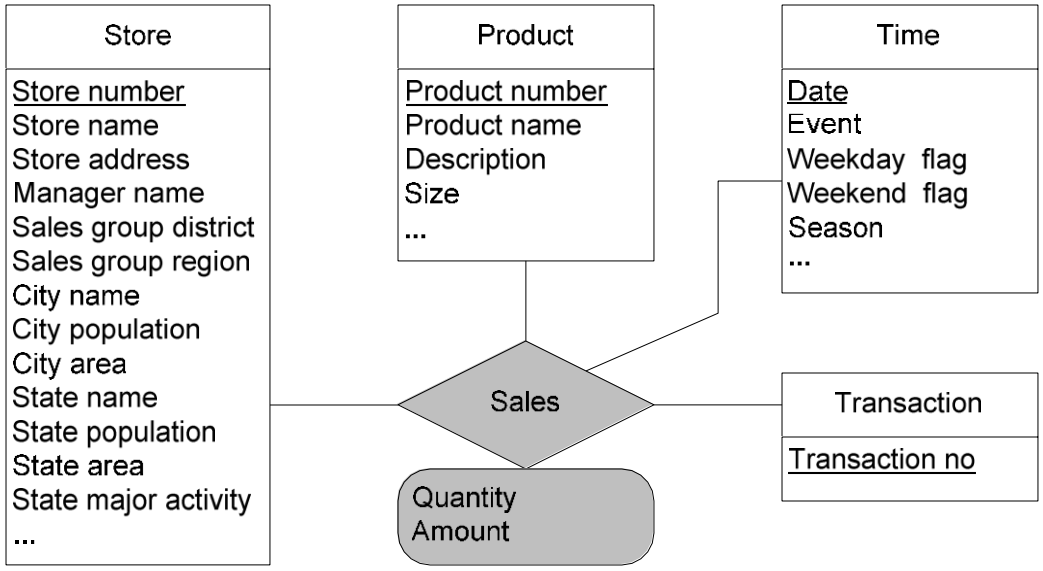
- Hay muchas situaciones reales que requieren incluir dimensiones de juego de roles en el esquema DW, por ejemplo para modelar:
  - El estado de entrega en inventario DW
  - La información de viajero frecuente
  - Las telecomunicaciones

Dimensiones degeneradas o de hecho

- Muchos de los esquemas multidimensionales giran en torno a algún tipo de control como un pedido, una factura o un ticket.
- Por lo general, estos documentos de control son una especie de contenedor con uno o más artículos en su interior.
- Un grano natural para una relación de hechos en estos casos es una línea de pedido por artículo

Dimensiones degeneradas o de hecho

- Un ejemplo típico, donde los usuarios necesitan analizar las ventas de productos con la granularidad más baja, que está representada por la línea de transacción (o pedido).
  - El número de transacción es útil, ya que sirve como clave de agrupación para reunir todos los productos solicitados en una orden de compra



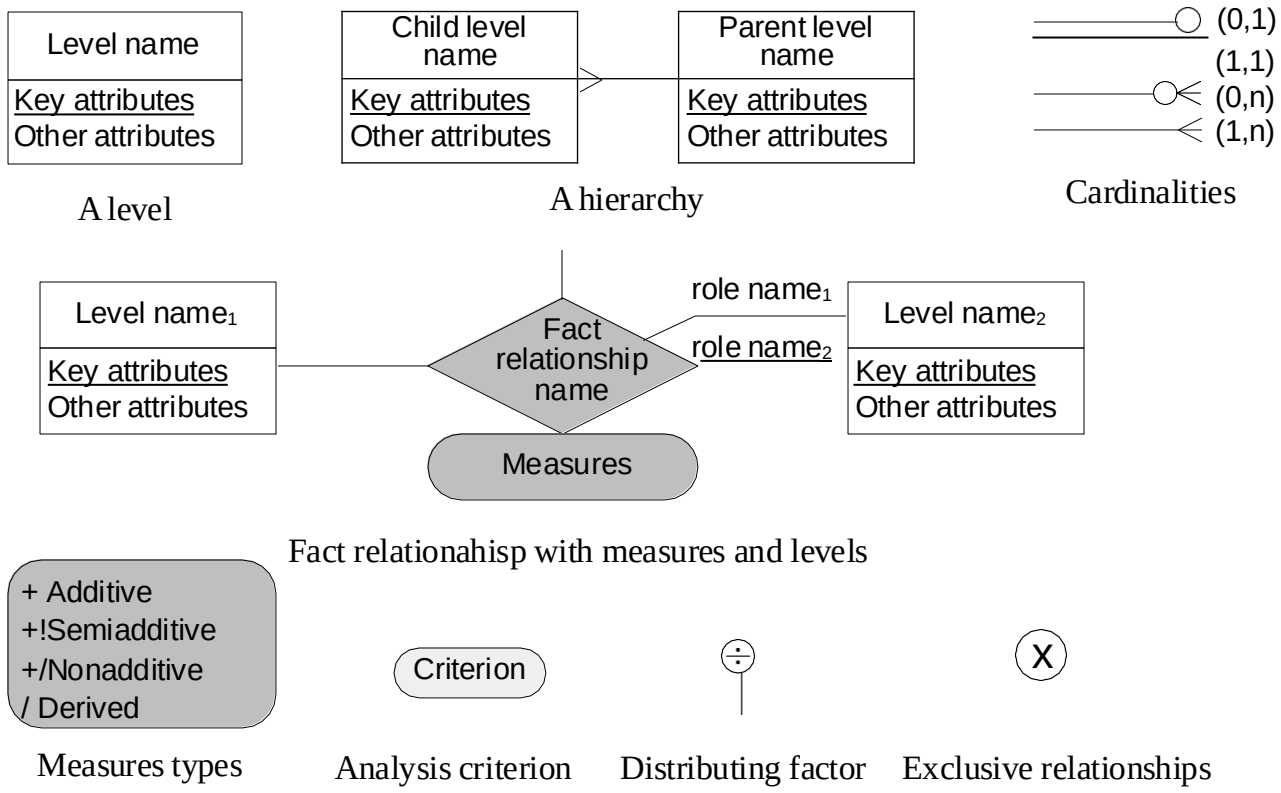
Dimensiones degeneradas o de hecho

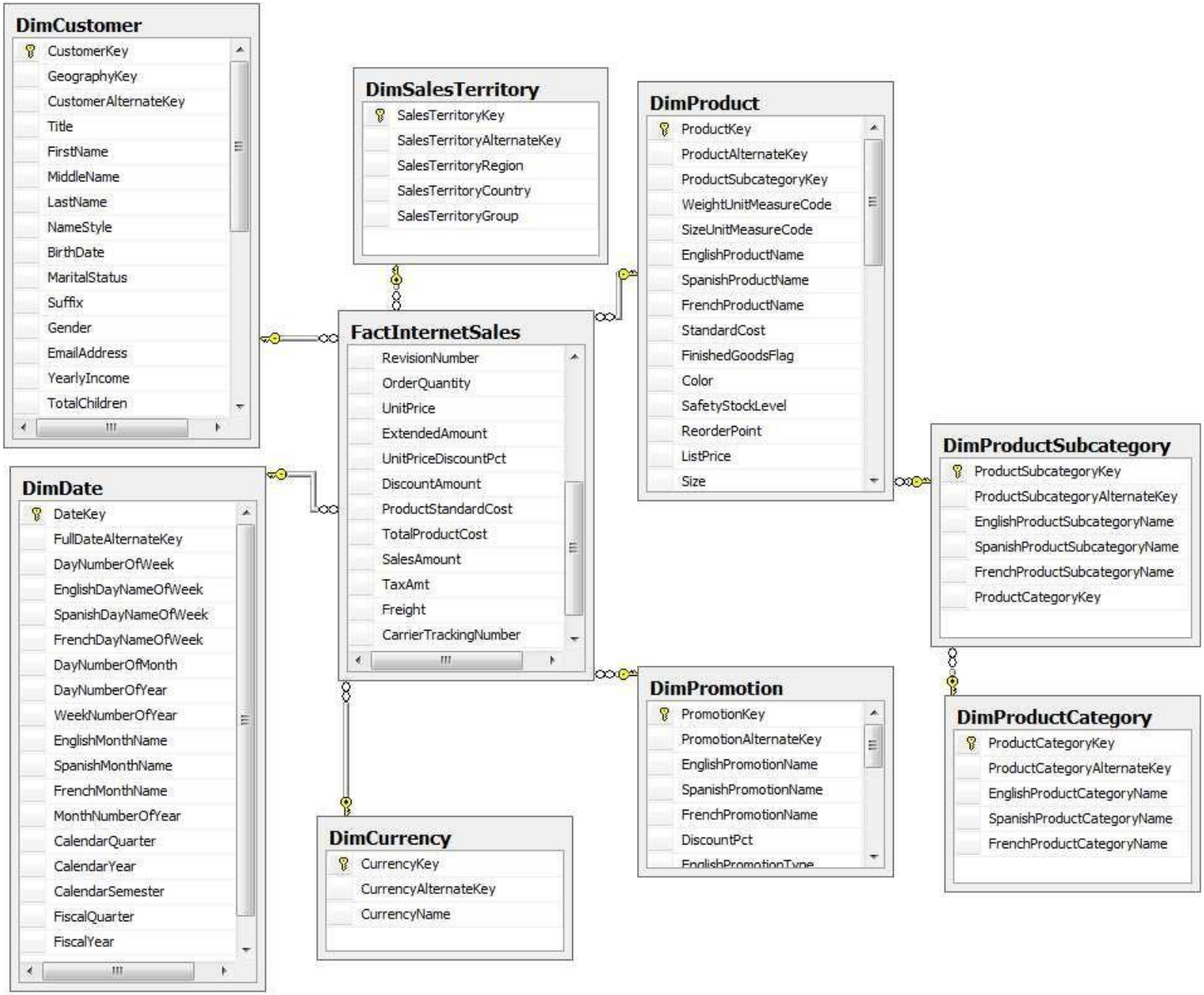
- La dimensión Transacción incluye solo el número de transacción, ya que otros datos, como la fecha de la transacción o el número de tienda, ya están incluidos en las otras dimensiones
- En las representaciones lógicas de este esquema, esta dimensión normalmente se convierte en un atributo de la tabla de hechos
- En consecuencia, esta dimensión se llama
  - Una dimensión degenerada, ya que no tiene otros atributos que la describan
  - O una dimensión de hecho, ya que se incluye como parte de la relación de hecho mientras desempeña el papel de una dimensión, es decir, se utiliza con fines de agrupación

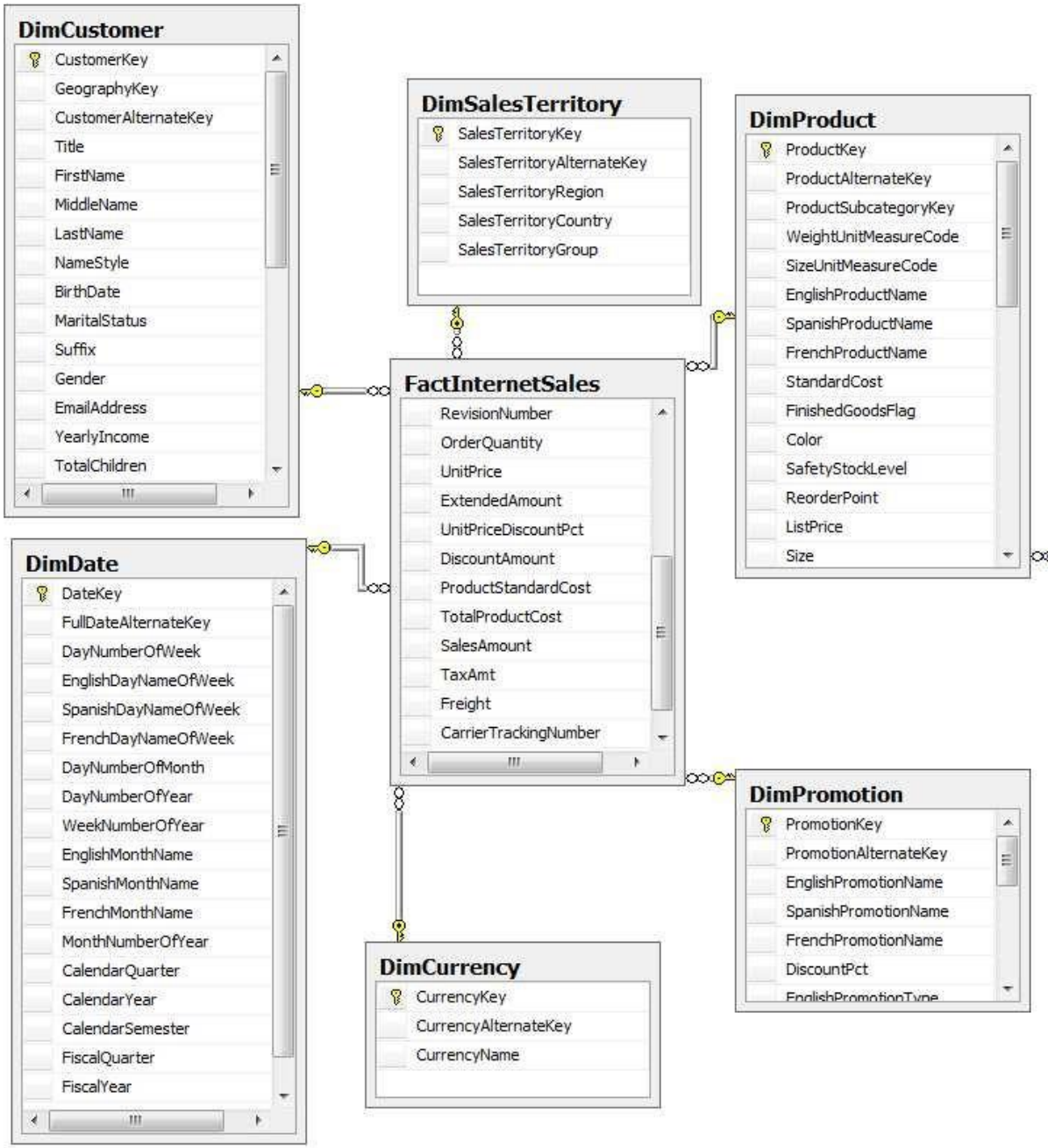
Dimensiones degeneradas o de hecho

- En nuestra técnica de modelado:
  - El esquema conceptual debe mantener esta dimensión para indicar que los usuarios pueden requerir agruparse según los miembros de la dimensión, es decir, según el número de transacción
  - La transformación al nivel lógico o de implementación puede incluir esta dimensión como un atributo de una tabla de hechos

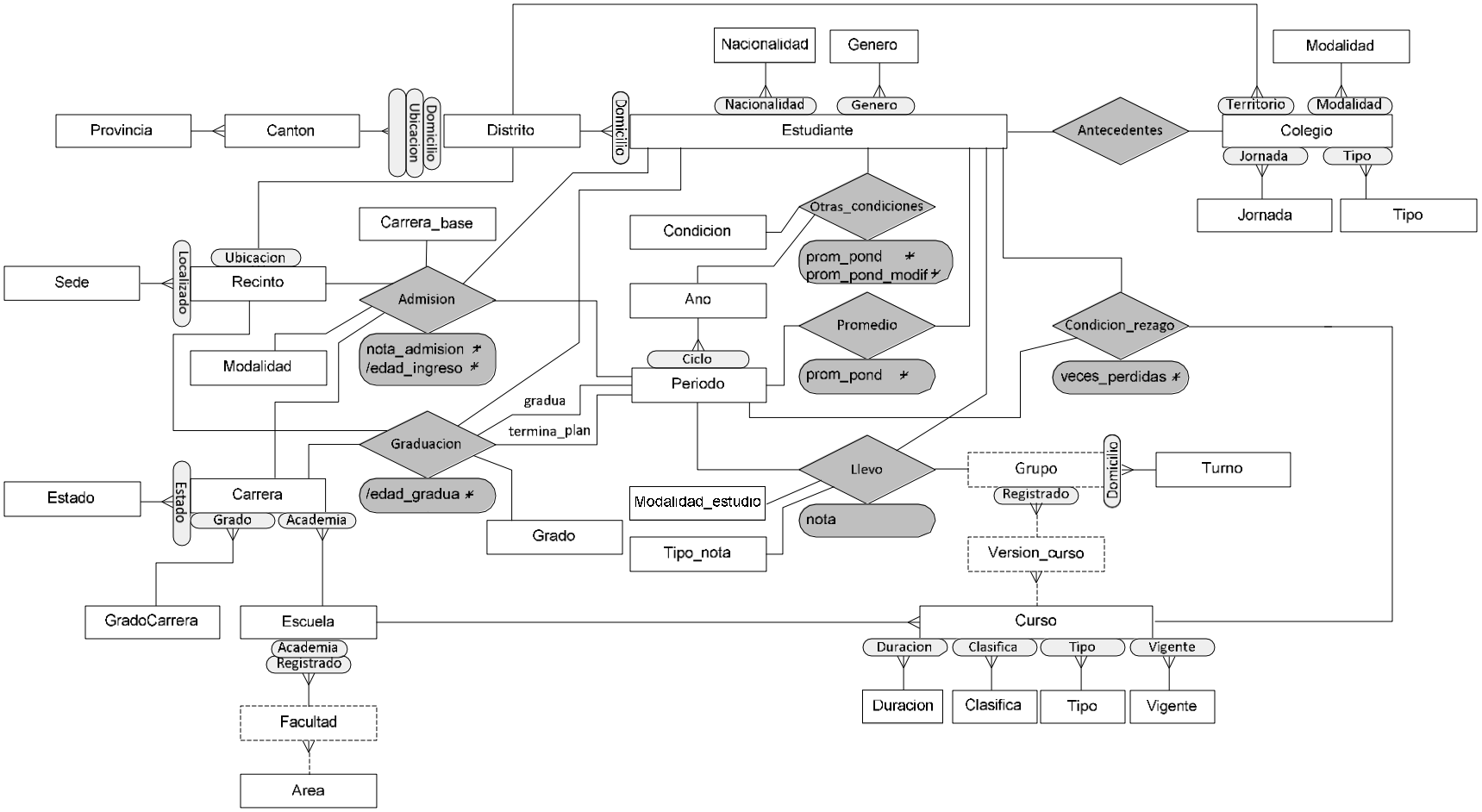
Resumen de la notación del modelo MultiDim




DimCustomer
CustomerKey
GeographyKey
CustomerAlternateKey
Title
FirstName
MiddleName
LastName
NameStyle
BirthDate
MaritalStatus
Suffix
Gender
EmailAddress
YearlyIncome
TotalChildren
DimSalesTerritory
SalesTerritoryKey
SalesTerritoryAlternateKey
SalesTerritoryRegion
SalesTerritoryCountry
SalesTerritoryGroup
DimProduct
ProductKey
ProductAlternateKey
ProductSubcategoryKey
WeightUnitMeasureCode
SizeUnitMeasureCode
EnglishProductName
SpanishProductName
FrenchProductName
StandardCost
FinishedGoodsFlag
Color
SafetyStockLevel
ReorderPoint
ListPrice
Size
FactInternetSales
RevisionNumber
OrderQuantity
UnitPrice
ExtendedAmount
UnitPriceDiscountPct
DiscountAmount
ProductStandardCost
TotalProductCost
SalesAmount
TaxAmt
Freight
CarrierTrackingNumber
DimProductSubcategory
ProductSubcategoryKey
ProductSubcategoryAlternateKey
EnglishProductSubcategoryName
SpanishProductSubcategoryName
FrenchProductSubcategoryName
ProductCategoryKey
DimDate
DateKey
FullDateAlternateKey
DayNumberOfWeek
EnglishDayNameOfWeek
SpanishDayNameOfWeek
FrenchDayNameOfWeek
DayNumberOfMonth
DayNumberOfYear
WeekNumberOfYear
EnglishMonthName
SpanishMonthName
FrenchMonthName
MonthNumberOfYear
CalendarQuarter
CalendarYear
CalendarSemester
FiscalQuarter
FiscalYear
DimPromotion
PromotionKey
PromotionAlternateKey
EnglishPromotionName
SpanishPromotionName
FrenchPromotionName
DiscountPct
DimProductCategory
ProductCategoryKey
ProductCategoryAlternateKey
EnglishProductCategoryName
SpanishProductCategoryName
FrenchProductCategoryName
DimCurrency
CurrencyKey
CurrencyAlternateKey
CurrencyName



Medida	Función de aggregación
Order quantity	Sum
Extended amount	
Sales amount	
Discount amount	
Product standard cost	
Total product cost	
Tax amount	
Freight	
Unit price	No aditiva
Unit price discount	Max


Nacionalidad
Genero
Modalidad
Nacionalidad
Genero
Provincia
Canton
Ubicacion
Domicilio
Distrito
Domicilio
Estudiante
Antecedentes
Territorio
Modalidad
Colegio
Jornada
Tipo
Jornada
Tipo
Otras_condiciones
Carrera_base
Condicion
prom_pond *
prom_pond_modif *
Ubicacion
Localizado
Sede
Recinto
Admision
Ano
Promedio
Condicion_rezago
nota_admision *
/edad_ingreso *
Modalidad
Ciclo
prom_pond *
veces_perdidas *
Periodo
gradua
termina_plan
Graduacion
Llevo
Grupo
Domicilio
Turno
Estado
Estado
Carrera
/edad_gradua *
Modalidad_estudio
Registrado
Grado
Academia
Grado
nota
Tipo_nota
Version_curso
GradoCarrera
Escuela
Curso
Academia
Registrado
Duracion
Clasifica
Tipo
Vigente
Duracion
Clasifica
Tipo
Vigente
Facultad
Area