Introducción a DW

Diseño Conceptual de DW

Alejandro Vaisman
Instituto Tecnológico de Buenos Aires, Argentina
avaisman@itba.edu.ar

Análisis de Requerimientos

- Previo al diseño, se deben relevar los requerimientos
- No está en los objetivos del curso, pero plantearemos un método simplificado, normalmente utilizado para definer KPIs
- En DW y OLAP los requerimientos son las consultas analíticas que deberán poder plantear los usuarios del Sistema
- Subject-oriented (def. de Immon): Plantear el objetivo principal (aumento de ventas, reducción de ausentismo, etc.)
- Para alcanzar un objetivo, el problema se divide en sub-objetivos
- Esta sub-división se debe realizar recursivamente, hasta que se puedan plantear consultas que ayuden a analizar cada sub-objetivo

Análisis de Requerimientos

- Cuanto más detallada la consulta, mejor será el diseño
- Las queries permitirán identificar facts, dimensiones, niveles de dimension, etc.

Análisis de Requerimientos Ejemplo

1. Contexto

La red pública de transporte en la actualidad incorpora el 60% del tráfico de personas de la Ciudad de Buenos Aires. A partir de fines del 2010 entró en vigencia el Sistema Único de Boleto Electrónico (SUBE). Concentrando entre otras funciones el pago de todos los medios de transporte para la ciudad. Cada usuario cuenta con una tarjeta SUBE a la cual le carga saldo y utiliza el mismo en los distintos medios de transporte.

Para el ministerio de transporte es fundamental tener visibilidad sobre la utilización de los subtes, colectivos, trenes, etc ya que a partir de los mismos puede tomar mejores decisiones en inversiones, subvenciones y concesiones. Un mayor uso, y más eficiente del transporte público permite descongestionar vías de tránsito y flujo de tránsito, brindar una mayor velocidad de transporte, y todos los beneficios que eso incluye: disminuir costos de combustibles, disminuir su contaminación por combustión y mejorar así la calidad de vida en la ciudad.

Análisis de Requerimientos Ejemplo

2. Planteo del problema y Objetivo Principal

Problema: El problema en la actualidad es que la red está subutilizada y organizada en una forma poco eficaz. Un ejemplo de esto es la superposición de recorridos, baja

frecuencia y colapso en horas pico producto de la escasa planificación de la red.

Esto incentiva el uso de transporte privado en detrimento del transporte público.

Objetivo Principal: Aumentar el uso del transporte público en la Ciudad de Buenos Aires en un 10%.

Análisis de Requerimientos Ejemplo

3. Sub-objetivos

1. Mejorar la red de puntos de carga en función del uso y operación.
2. Proponer políticas de reducción de tarifa eficientes en función de transbordos y frecuencia de uso.
3. Mejorar la frecuencia del transporte, en función de optimizar recorridos redundantes y distribución de nodos de alta concurrencia

Análisis de Requerimientos Ejemplo

4. Consultas

1. **Mejorar la red de puntos de carga en función del uso y operación.**
   1.1. Identificar los 1000 principales nodos de concentración de viajes en un rango de 1 hora (*aclaración: un nodo es un superficie previamente definida en el modelo de aprox 250mx250m. Todo el territorio se encuentra dividido en nodos, y cada parada pertenece a un nodo).*
   1.2. ¿Cuales son los 100 nodos con mayor cantidad de viajes realizados cuyo saldo remanente en las tarjetas sea menor al boleto mínimo, en un rango de 2 horas?
2. **Proponer políticas de reducción de tarifa eficientes en función de transbordos y frecuencia de uso.**
   2.1 ¿Qué cantidad de usuarios toman la misma línea más de 15 veces por mes?
   2.2. ¿Qué cantidad de personas realizan más de 4 viajes en un periodo de 8 horas?
   2.3. ¿Qué cantidad de personas toman más de un transporte en menos de 1 hora (transbordo) ?
3. **Mejorar frecuencias de transportes en función de optimizar recorridos redundantes y distribución de nodos de alta concurrencia**
   3.1. ¿Cuáles son las 10 líneas con menor cantidad de viajes diarios promedio?
   3.2. ¿Existen líneas que comparten más del 75% de los nodos que recorren y con un promedio de viajes diario menor al promedio global por línea?
   3.3. ¿Cuáles son las 10 líneas con menor proporción de viajes por vehículo sobre viajes por línea en un período de 2 horas?

Caso de Uso 1 - Tarjetas de Crédito

1. Contexto

- Las tarjetas de crédito XXX cuentan con un programa de bonificación durante los primeros 3 meses a partir de su emisión. Por otro lado, mientras el cliente tenga un consumo mensual mayor o igual a $100.000 para las tarjetas Platinum o $60.000 para las tarjetas Signature, se le continuará bonificando el costo del producto.
- Se percibe que, pasado el periodo inicial, los clientes solicitan la baja del producto.
- En este caso, se busca accionar únicamente con los tarjetahabientes Access Now que además cuentan con paquete de cuentas XXX.

Caso de Uso 1 - Tarjetas de Crédito

2. Objetivos

Principal

Aumentar la retención de clientes en un 20% anual, identificando y segmentando los distintos tipos de clientes para incluirlos en distintas campañas comerciales.

Secundarios

Bonificar a los clientes que pueden pagar el consumo mínimo, desde el alta al primer vencimiento después de la bonificación, sobre el total de la cartera activa

Encontrar todos los clientes que no tienen capacidad para llegar al mínimo de consumo para bonificación

Encontrar oportunidades de fidelización en base a los beneficios del producto

Caso de Uso 1 - Tarjetas de Crédito

3. Consultas

1. Bonificar a los clientes que pueden pagar el consumo mínimo, desde el alta al primer vencimiento después de la bonificación, sobre el total de la cartera activa:

- ¿Cuántos clientes del total de la cartera llegan con montos de saldos vista mensuales superiores a 15k? (y a 25k para Black)
- ¿Cuántos clientes realizaron transferencias mensuales a cuentas propias en otros bancos por montos suficientes como para bonificar el consumo?
- ¿Cuántos clientes tuvieron consumos superiores a 15k/25k durante los 3 meses de bonificación?
- ¿Cuántos clientes con tarjeta Black llegan con consumos mensuales entre 15k y 25k? (posiblesdowngrades)

Caso de Uso 1 - Tarjetas de Crédito

3. Consultas

2. Encontrar todos los clientes que no tienen capacidad para llegar al mínimo de consumo para bonificación:

- ¿Cuántos clientes no consumieron los $15k/$25k para la bonificación del costo mensual durante los primeros 3 meses?
- De los clientes que consumieron menos de 15k/25k para bonificación durante los 3 meses de bonificación ¿Cuántos pagaron el mínimo del resumen?

Caso de Uso 1 - Tarjetas de Crédito

3. Consultas

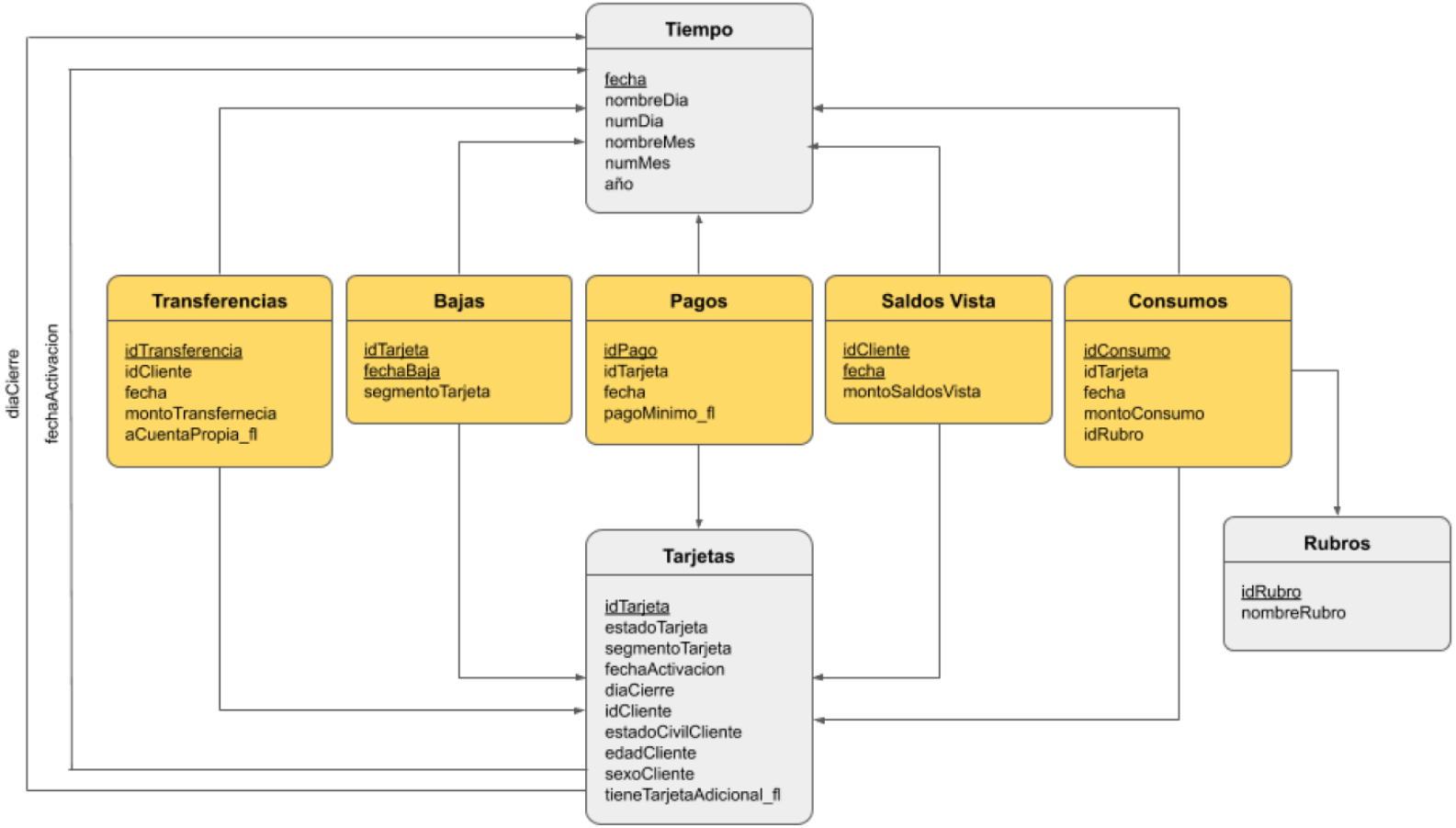
3. Encontrar oportunidades de fidelización en base a los beneficios del producto:

- ¿Cuántos clientes no están realizando consumos en tiendas online en las cuales tienen descuentos?
  - De este universo, ¿qué porcentaje corresponde a cada rango etario?
  - De este universo, ¿cuántos son hombres y cuántos mujeres?

- ¿Cuántos titulares de cuenta con estado civil “casado” (o similar) no cuentan con tarjetas adicionales?

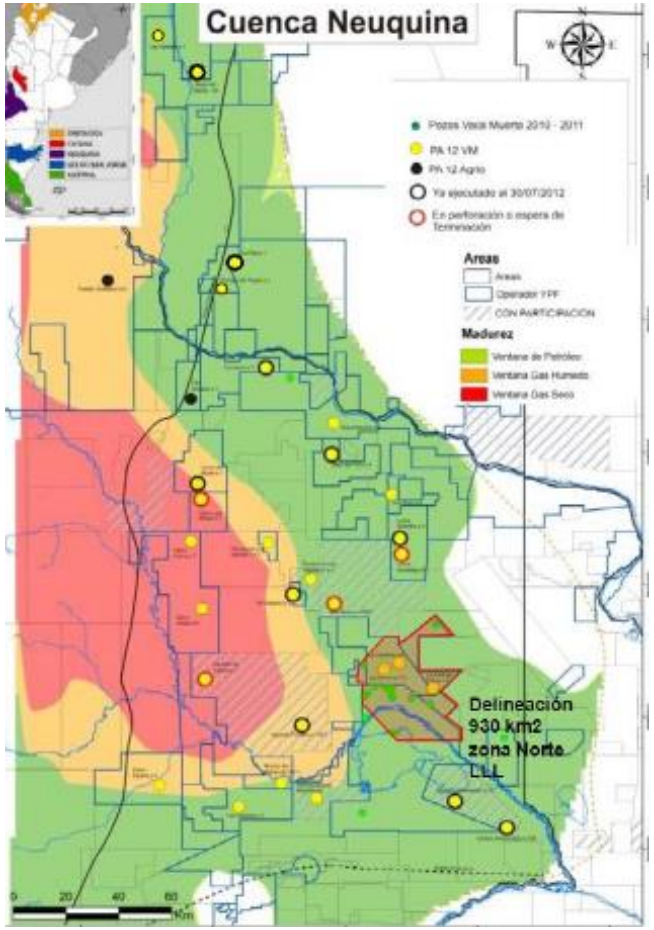
Caso de Uso 2 - Tarjetas de Crédito

4. Modelo



Caso de Uso 2 - Petróleo no convencional

1. Contexto



- Productividad de pozos
  - Difícil de predecir
  - Variabilidad geológica
  - Petróleo (verde), gas seco (rojo)
- Cada compañía tiene un bloque de desarrollo asignado

Caso de Uso 2 - Petróleo no convencional

2. Objetivos

- **Objetivo principal**

Mejorar indicadores de productividad de petróleo y gas (i.e., producción acumulada de petróleo o gas a los 3, 6 y 12 meses) en la formación Vaca Muerta mediante análisis del desempeño de los pozos perforados por diferentes operadores en diferentes bloques (campos) en el área.

Sub-objetivos:

1. Analizar si la diferencia en la productividad de los pozos es variable con el tiempo (por ejemplo, indicador de productividad a los 6 meses versus indicador de productividad a los 12 meses).
2. Analizar si la diferencia en la productividad de los pozos se debe a variaciones geológicas en la cuenca (por ejemplo, cambio en el tipo / calidad de fluido).
3. Analizar si la diferencia en la productividad de los pozos se debe a prácticas distintas de completamiento de pozos (por ejemplo, mayor longitud de la sección horizontal o mayor número de etapas de fractura).
4. Analizar el efecto de la colocación y tipo de sistema de levantamiento artificial en el desempeño de los pozos.
5. Definir curvas tipo de producción por área y tipo de fluido

Caso de Uso 2 - Petróleo no convencional

3. Consultas

Cada sub-objetivo está asociado a un conjunto de consultas

1. **Analizar si la diferencia en la productividad de los pozos es variable con el tiempo (por ejemplo, indicador de productividad a los 6 meses versus indicador de productividad a los 12 meses).**
   i. Evolución del número de pozos perforados por campaña (semestre) desde 2010 hasta Q2-2019, agrupados por tipo de pozo (Vertical y Horizontal).
   ii. Evolución del número de pozos perforados por trimestre desde 2010 hasta Q2-2019, agrupados por tipo de fluido producido (Petróleo y Gas).
   iii. Evolución del porcentaje de actividad de cada operador durante los últimos 10 años.
   iv. Evolución del número de pozos perforados por trimestre desde 2010 hasta Q2-2019, agrupados por área.
   v. Número total de pozos perforados por tipo de pozo (Vertical y Horizontal) desde 2010 hasta Q2-2019. Especificar porcentaje sobre el total.
   vi. Número total de pozos perforados por tipo de fluido (Black Oil, Volatile Oil, Wet Gas, Dry Gas) desde 2010 hasta Q2-2019. Especificar porcentaje sobre el total.
   vii. …..

Caso de Uso 2 - Petróleo no convencional

3. Consultas

Cada sub-objetivo está asociado a un conjunto de consultas

2. Analizar si la diferencia en la productividad de los pozos se debe a prácticas distintas de completamiento de pozos (p.ej., mayor longitud de la sección horizontal o mayor número de etapas de fractura).

i. Evolución trimestral de la producción acumulada de Petróleo y Gas desde 2010 hasta Q2-2019.
ii. Evolución anual de la producción inicial promedio de Petróleo desde 2010 hasta Q2-2019 para los pozos horizontales.
iii. Evolución trimestral de la producción acumulada de Petróleo y Gas desde 2010 hasta Q2-2019, discriminada por tipo de pozo (Vertical y Horizontal).
iv. Porcentaje de la producción acumulada de Petróleo y Gas a la fecha, discriminada por tipo de pozo (Vertical y Horizontal).
v. Evolución mensual de la tasa de producción promedio de Petróleo (bbl/d) y Gas (MMscf/d) de los pozos horizontales, agrupada por año o campaña.
vi. Evolución mensual de la producción acumulada promedio de Petróleo y Gas de los pozos horizontales, agrupada por año o campaña.
vii. ….

Caso de Uso 2 - Petróleo no convencional

3. Consultas

Cada sub-objetivo está asociado a un conjunto de consultas

3. Analizar si la diferencia en la productividad de los pozos se debe a variaciones geológicas en la cuenca (por ejemplo, cambio en el tipo / calidad de fluido).

i. Evolución de la producción acumulada de Petróleo (bbl) y Gas (MMscf) desde 2010 hasta Q2-2019, discriminada por área.
ii. Evolución de la producción acumulada de Petróleo (bbl) y Gas (MMscf) desde 2010 hasta Q2-2019, discriminada por tipo de fluido (Black Oil, Volatile Oil, Wet Gas, Dry Gas).
iii. Evolución de la producción diaria de Petróleo (bbl/d) y Gas (MMscf/d) desde 2010 hasta Q2-2019, discriminada por área.
iv. Evolución de la producción diaria de Petróleo (bbl/d) y Gas (MMscf/d) desde 2010 hasta Q2-2019, discriminada por tipo de fluido (Black Oil, Volatile Oil, Wet Gas, Dry Gas).
v. Porcentaje por área de la producción total acumulada de Petróleo y Gas
vi. Evolución de la producción promedio de Petróleo y Gas, discriminado por tipo de pozo (Vertical y Horizontal).
vii. Lista (ranking) de las áreas con mayor producción acumulada de Petróleo y Gas, normalizada por número de pozos.
viii. …..

Caso de Uso 2 - Petróleo no convencional

3. Consultas

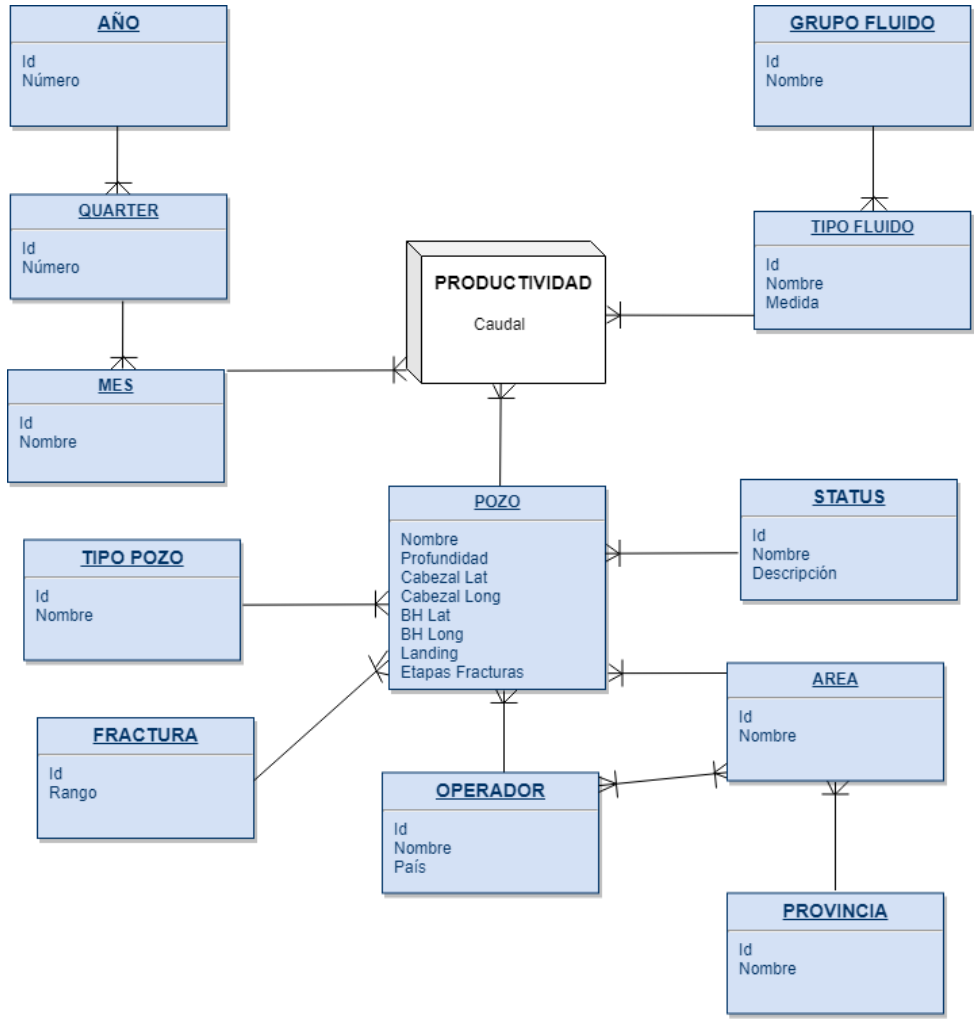
Cada sub-objetivo está asociado a un conjunto de consultas

4. Analizar el efecto de la colocación y tipo de sistema de levantamiento artificial en el desempeño de los pozos.

i. Evolución trimestral de la producción acumulada de Petróleo y Gas de pozos horizontales, discriminada por área y operador.
ii. Evolución anual (o por campaña) de la longitud lateral de los pozos horizontales discriminada por área y operador.
iii. Evolución anual (o por campaña) del número de etapas por fractura discriminada por área y operador.
iv. Evolución trimestral de la producción acumulada de Petróleo y Gas normalizada por longitud horizontal, discriminada por área y operador.
v. Evolución trimestral de la producción acumulada de Petróleo y Gas normalizada por longitud horizontal, discriminada por tipo de fluido (Black Oil, Volatile Oil, Wet Gas, Dry Gas).
vi. Evolución mensual de la producción de Petróleo y Gas por campaña de perforación, discriminado por tipo de fluido (Black Oil, Volatile Oil, Wet Gas, Dry Gas).
vii. …

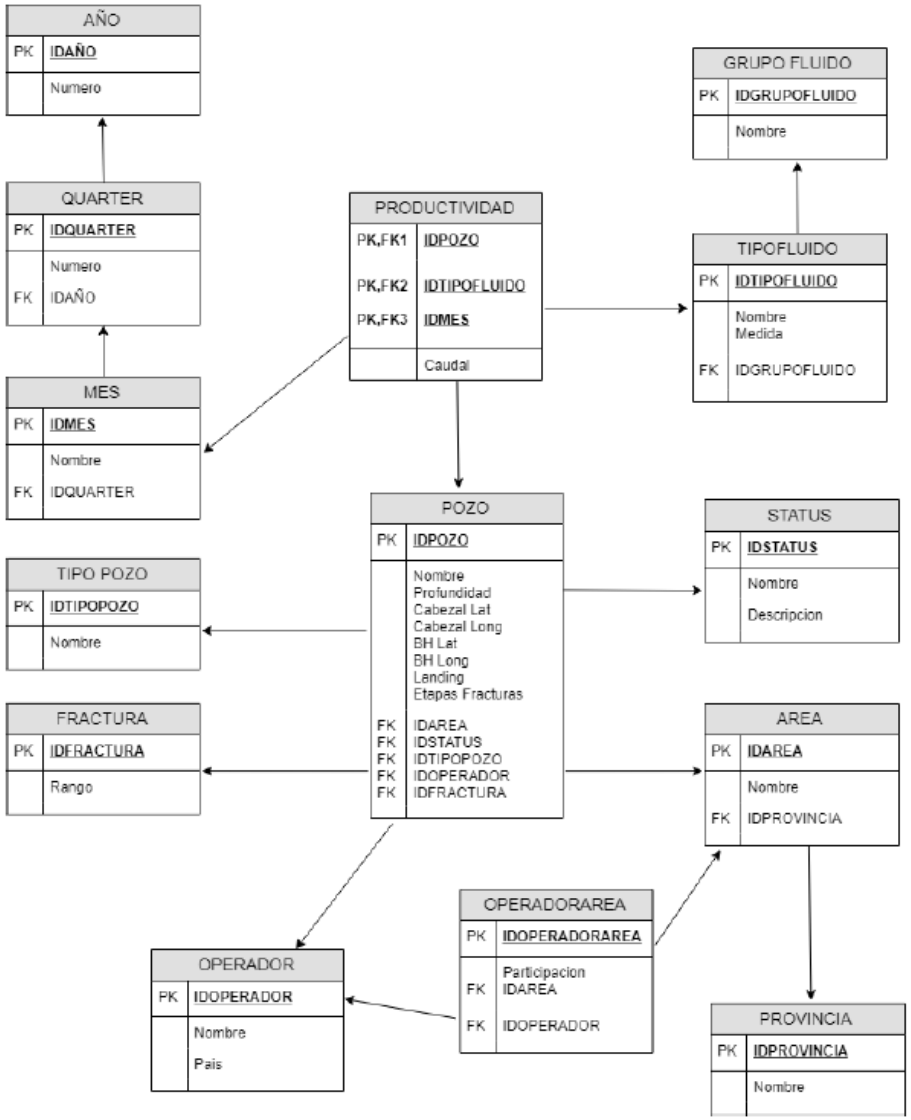
Análisis de datos en petróleo no convencional

Modelo Conceptual



Análisis de datos en petróleo no convencional

Modelo Tabular
-Snowflake-



Modelos Conceptuales

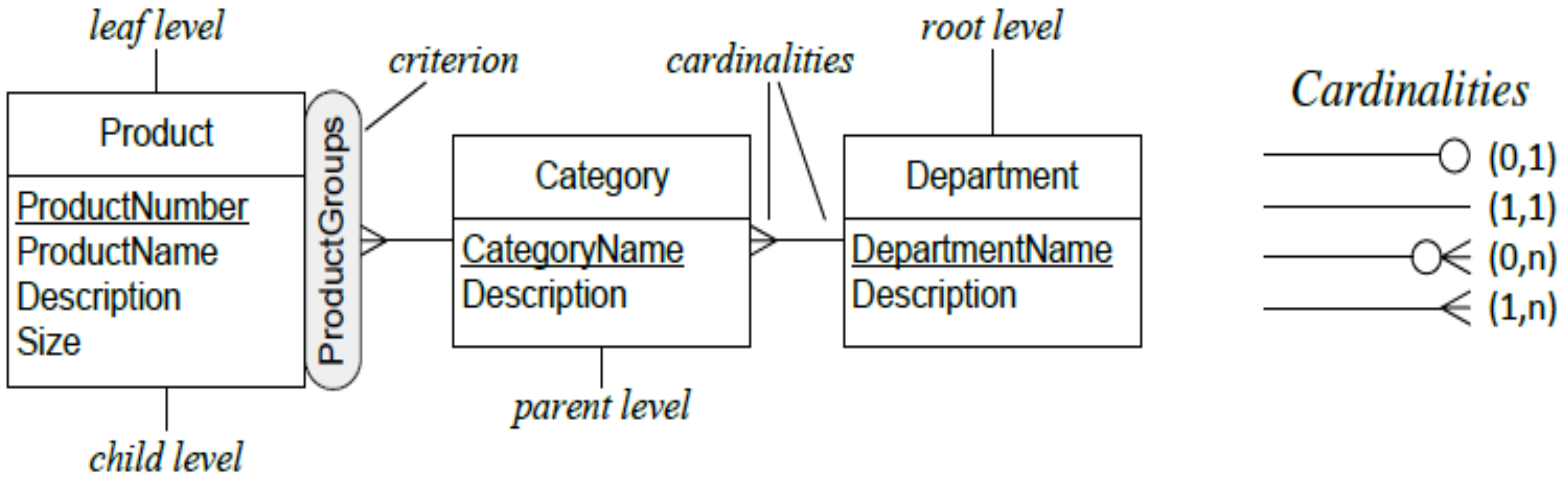
- Permiten una major comunicación entre diseñadores y usuarios para entender los requerimientos
- Son más estables que los modelos orientados a implementación (solo cambian si cambia el contexto)
- Problema: No hay una standard establecido en DW, como el DER o UML en bases de datos OLTP
- Actualmente los DWs se diseñan desde el modelo lógico
  - Dificulta la expression de requerimientos
  - Los usuarios ssólo pueden definer elementos que son parte del modelo (básicamente, tablas)
  - Ej: Sólo se pueden expresar jerarquías simples

El Modelo Multidim

- Basado en el DER
- Permite representar los conceptos de
  - Dimensiones
  - Hechos
  - Jerarquías
  - Métricas

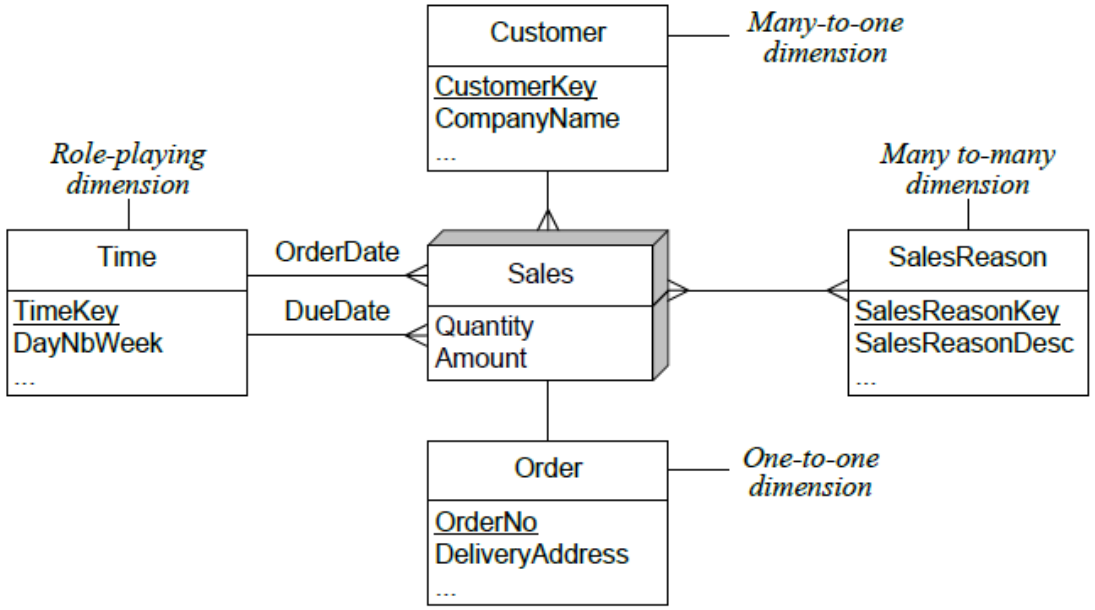
El Modelo Multidim Notación

- Dimensión: Un nivel, o una o más jerarquías
- Jerarquía: Conjunto de niveles relacionados (padre-hijo)
- Nivel: Un tipo de entidad
- Miembro: Una instancia en un nivel
- Cardinalidad: Máximo/mínimo numero de miembros en un nivel, relacionados con miembros de otro nivel (igual que en el DER)
- Criterio de agregación
- Atributos clave: Identifican unívocamente a un miembro de un nivel
- Atributos descriptivos: Decriben a un miembro de un nivel

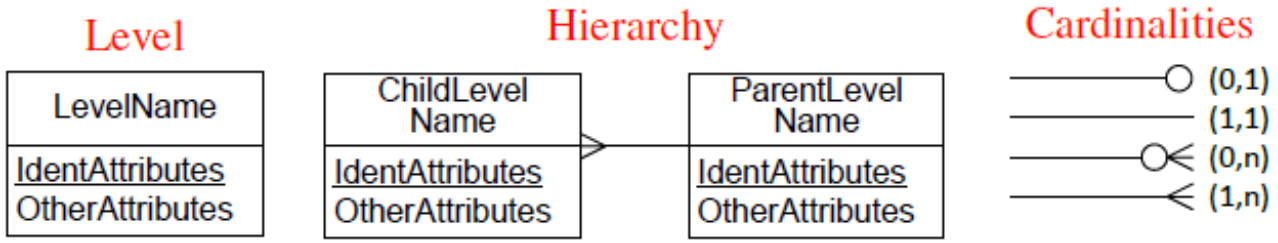


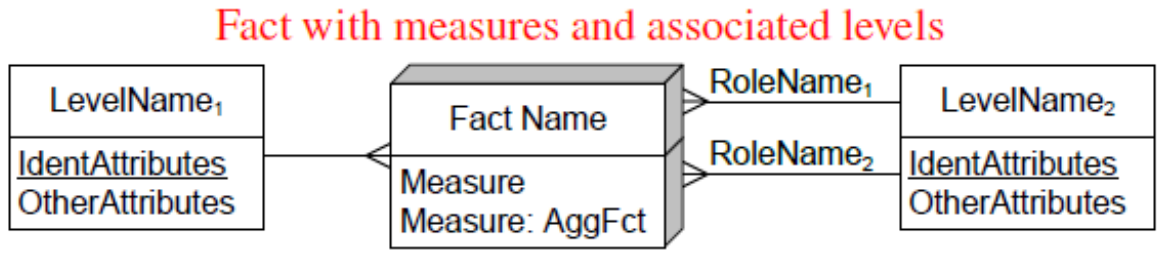
El Modelo Multidim Notación

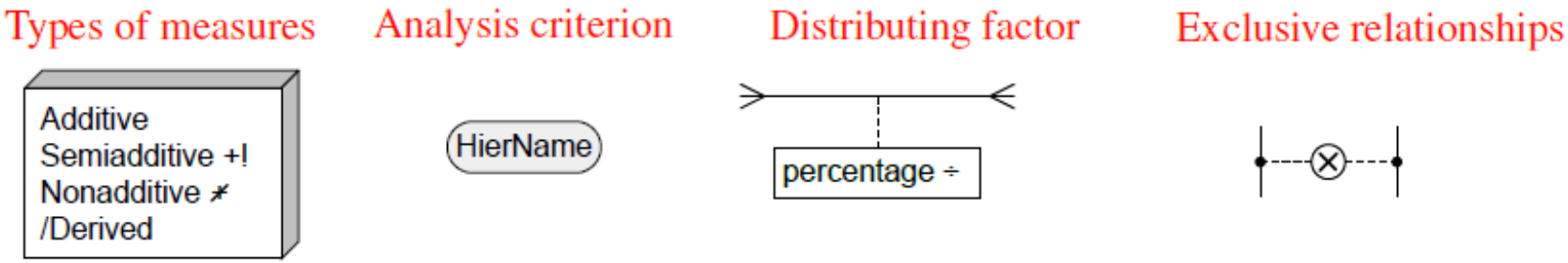
- Hecho (fact): Relaciona métricas con los niveles hoja (leaf) de las dimensiones involuctadas
- Representa una relación de muchos a muchos entre las dimensiones
- Las dimensiones de relacionan con los facts, con distintas cardinalidades
- Una misma dimension puede jugar más de un rol, como en el ejemplo siguiente => role-playing dimensions



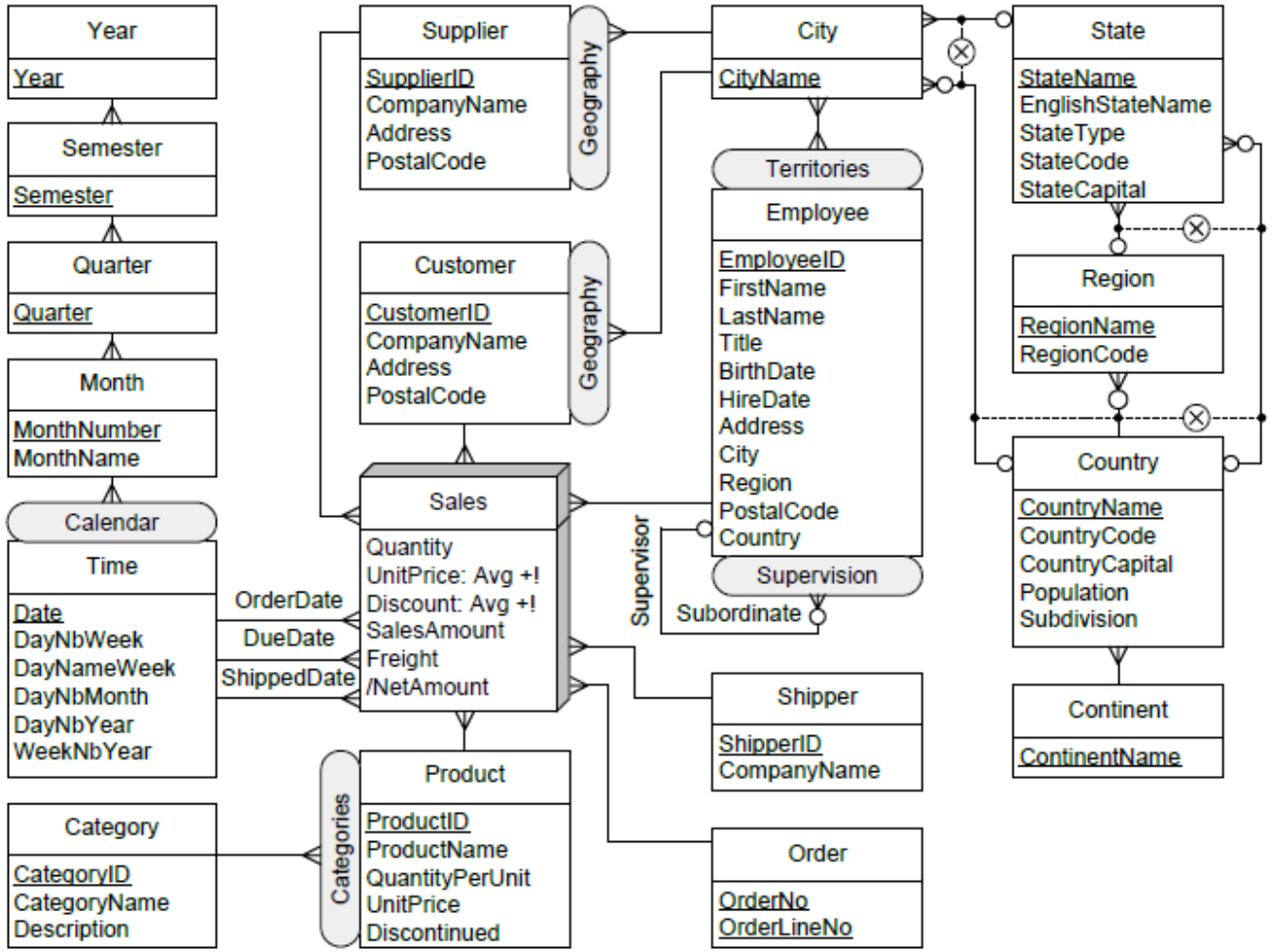
El Modelo Multidim Notación completa







Modelo Conceptual del Northwind DW



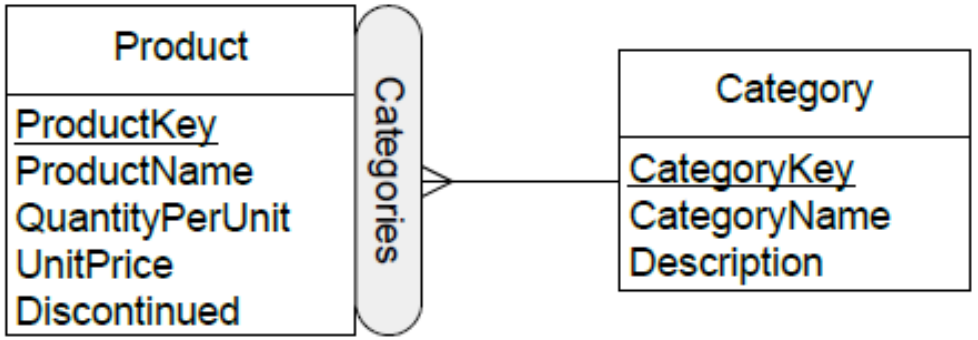
El Modelo Multidim Jerarquías Estrictas

- Sólo incluyen relaciones M:1 !!

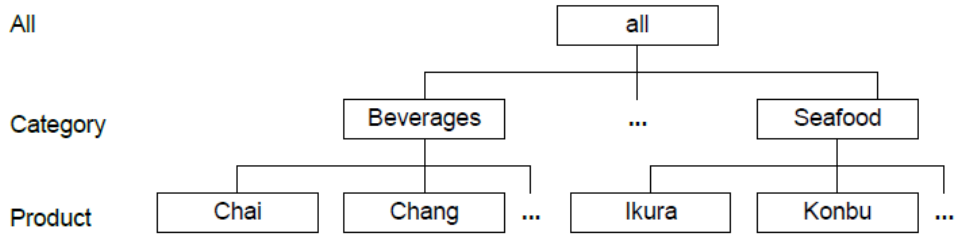
Jerarquías Balanceadas

- Esquema: Un único camino, todas las relaciones pade-hijo son m:1, obligatorias
- Instancia: Un árbol balanceado
- Garantizan la sumarizabilidad: Cada nivel define una partición del nivel inferior, no hay miembros sin padre o sin hijos, excepto el superior y los inferiores

Esquema



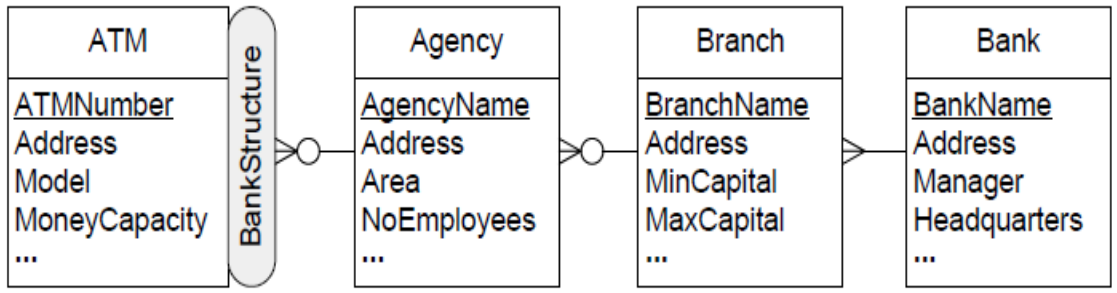
Instancia



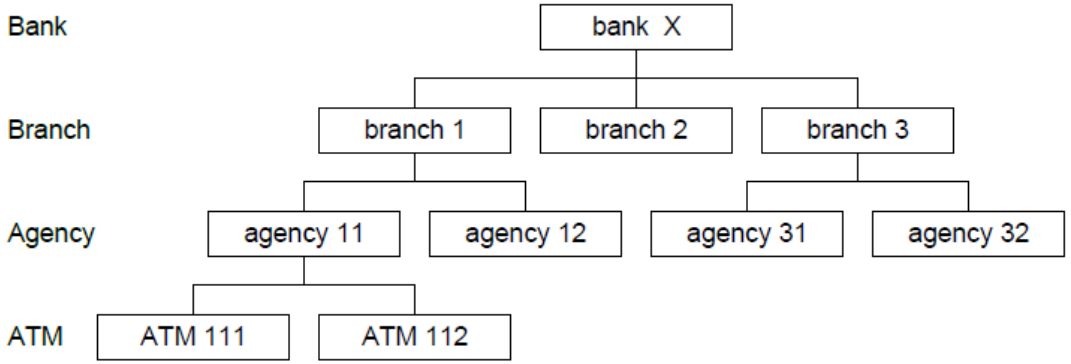
Jerarquías No-Balanceadas

- Esquema: Un único camino, todas las relaciones pade-hijo son m:1, opcionales
- Instancia: Un árbol desbalanceado
- Sólo garantizan la sumarizabilidad si no hay facts asociados a los miembros no-hoja
- Pueden ocurrir cuando:
  - los facts vienen a distintos niveles de granularidad (explicado luego)
  - La misma jerarquía se utiliza en diversos facts a distintos niveles de granularidad

Esquema



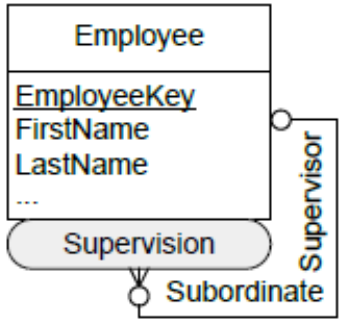
Instancia



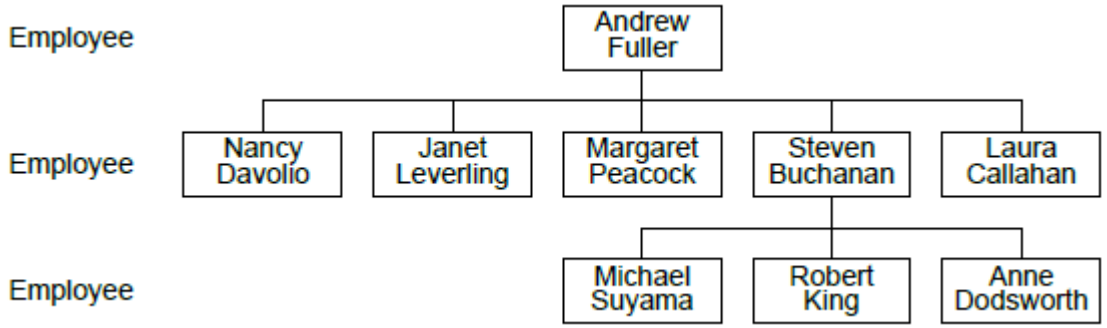
Jerarquías Recursivas

- Caso especial de no-balanceadas
- El mismo nivel está vinculado por los roles de las relaciones padre-hijo
- Las características de los niveles padre-hijo son similares
- Pueden ser reemplazadas por no-balanceadas tradcionales, a costa de expresividad

Esquema



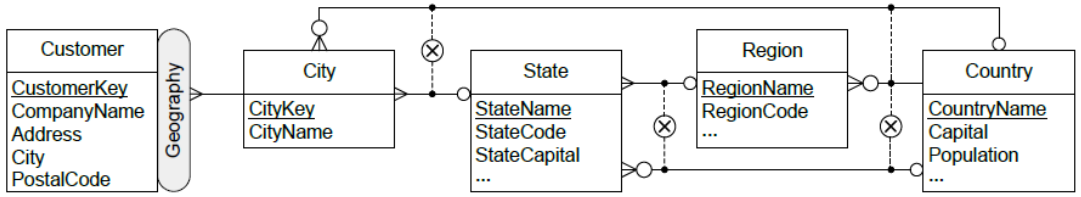
Instancia



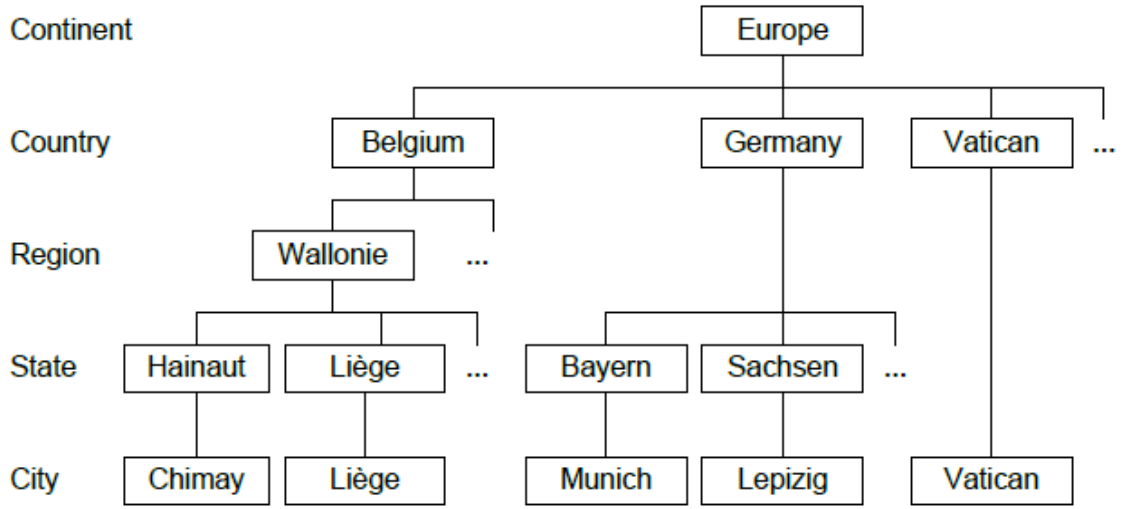
Jerarquías Ragged

- Caso particular de jerarquía generalizada
- A nivel esquema: caminos alternativos excluyentes, que pueden saltear niveles
- Muy útiles para representar jerarquías geográficas

Esquema



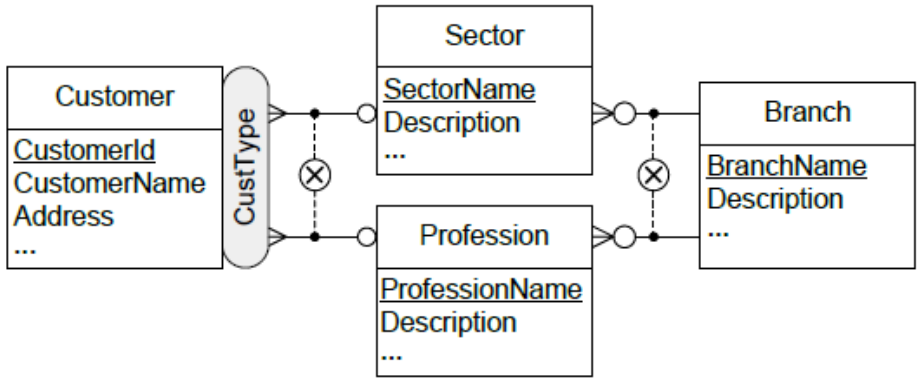
Instancia



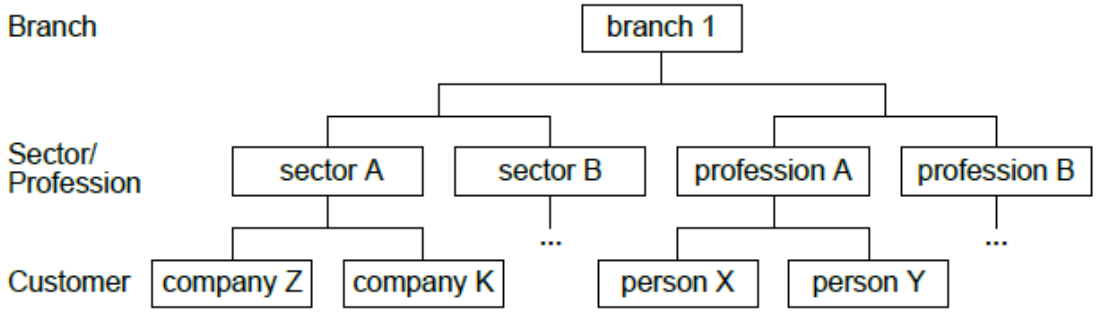
Jerarquías Generalizadas

- Esquema: Múltiples caminos, comparten al menos el nivel hoja
- Instancia: Cada miembro pertenece a un subárbol
- Precaución con sub-agregados: Para agregar una métrica (p.ej., ventas) por Branch, hay que considerar la unión de las ventas por sector (corporativas) y por profesión (a personas físicas)

Esquema



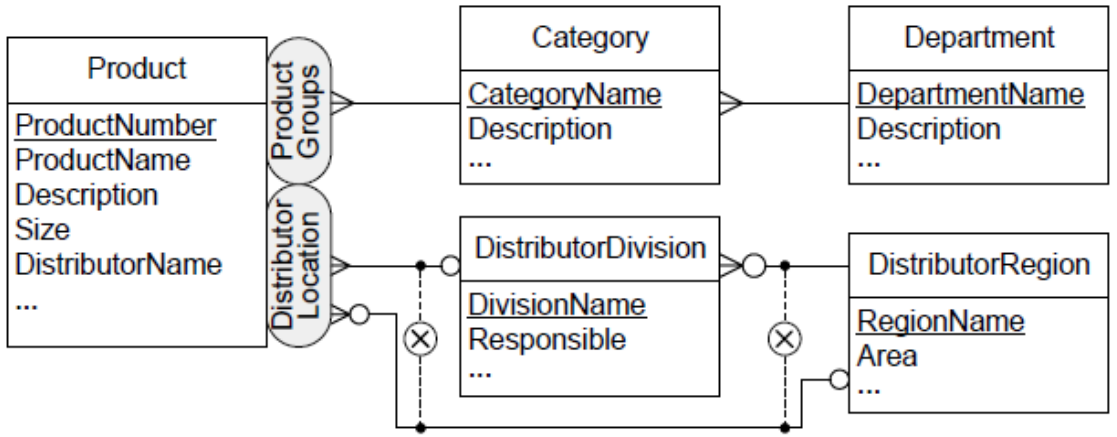
Instancia



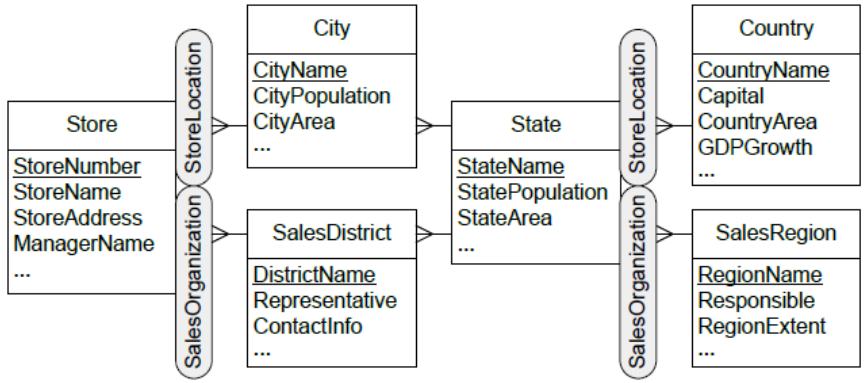
Jerarquías Paralelas

- Más de una jerarquía por dimension, que pueden ser independientes o dependientes
- Independientes: solo comparten el nivel hoja
- Dependientes: Comparten más que el nivel hoja

Independientes



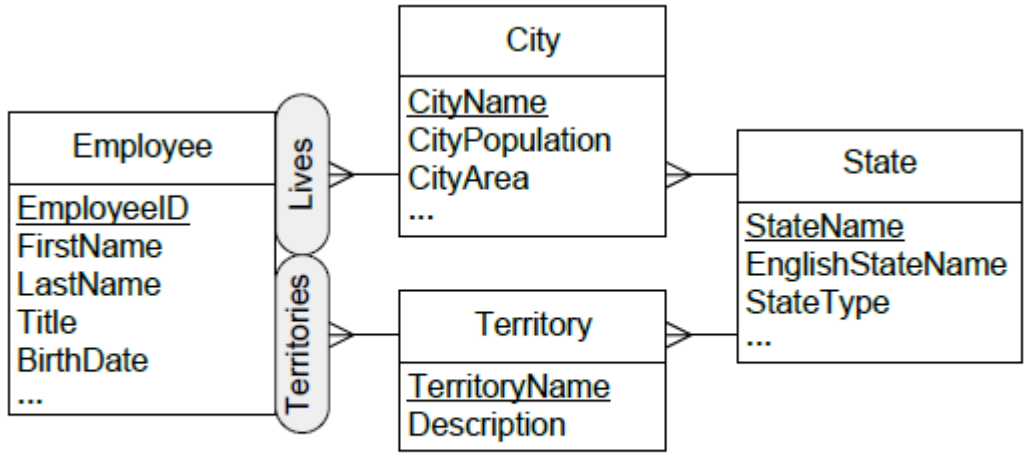
Dependientes



Jerarquías Paralelas

- Un caso típico: Empleados que viven en una ciudad y trabajan en otra
- Actividad: Qué problema presenta este esquema?
- Dibujar una posible instancia

Esquema



Instancia
(resolver como actividad)

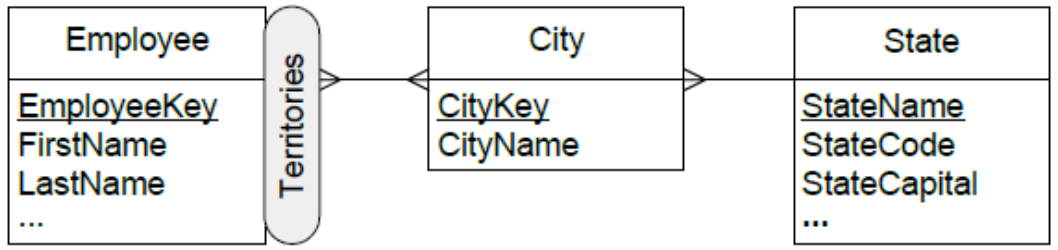
El Modelo Multidim
Jerarquías No-Estrictas

- Incluyen al menos una relación M:N
- Problemas de sumarizabilidad!!!

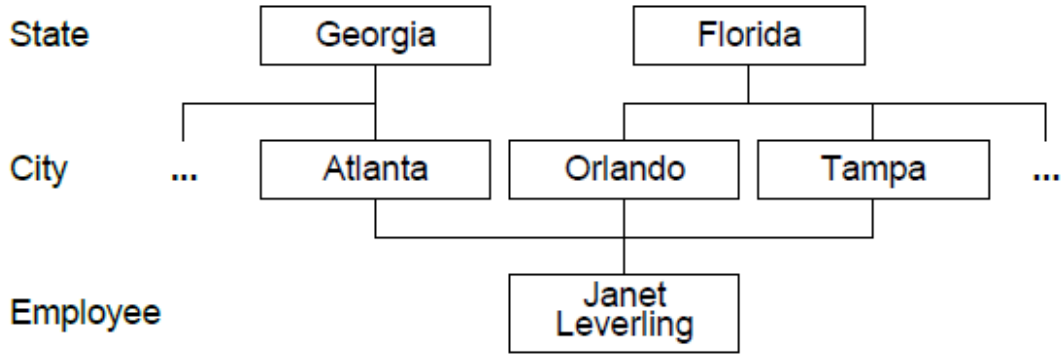
Jerarquías No-Estrictas

- Esquema: Al menos una relación M:N
- Instancia: Un GRAFO

Esquema



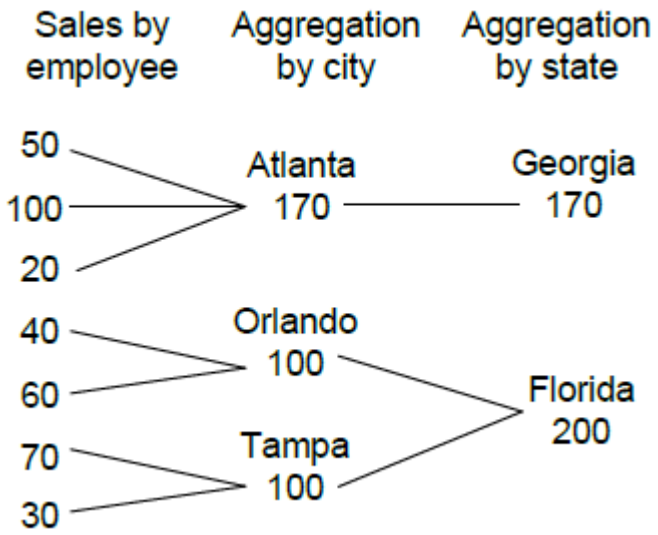
Instancia



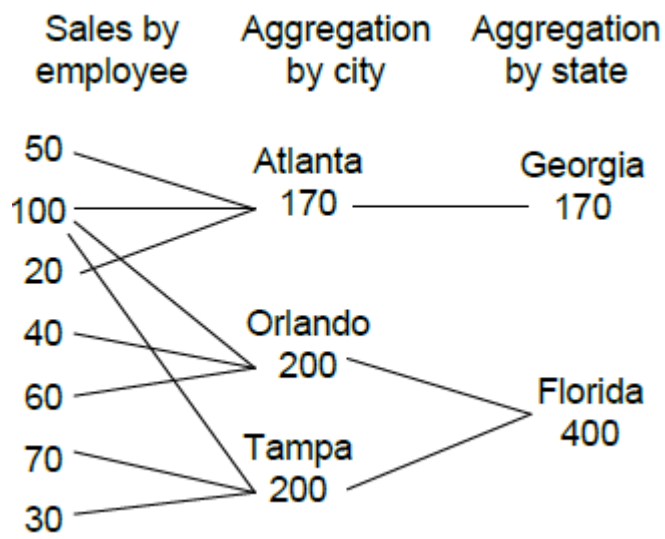
Jerarquías No-Estrictas

- PROBLEMA: **DOBLE CONTEO!**
- En el ejemplo, a la izquierda, asumimos que Janet L. vendió 100 unidades, y trabaja sólo en Atlanta
- A la derecha, las 100 unidades se asignan a las tres ciudades a las que está asignada Janet

Jerarquía Estricta



Jerarquía No-estricta



El Modelo Multidim Jerarquías No-Estrictas Soluciones

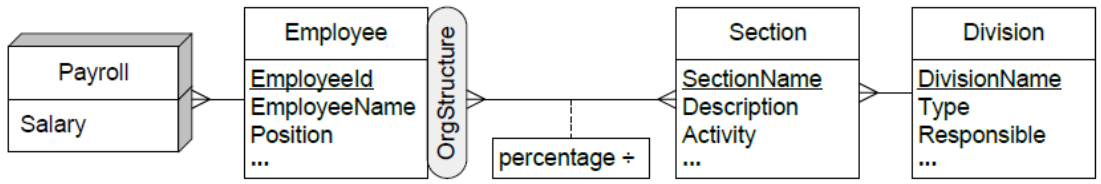
- Incluir el factor de distribución
- Elegir un miembro padre como primario, para asignarle la métrica, e ignorar el resto
- Formar grupos (ej., Orlando-Tampa-Georgia)
- Modificar el esquema (siguiente slide)

Jerarquías No-Estrictas

- Factor de distribución (izquierda)
- Reemplazar la relación m:n por un fact

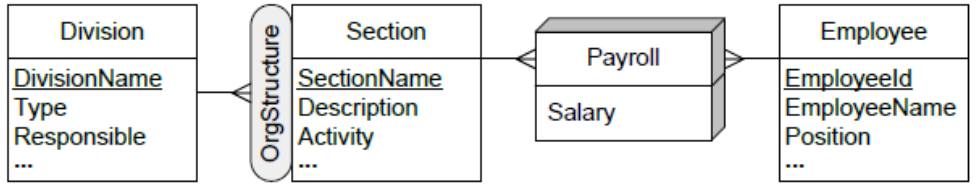
Factor de Distribución

- Asumimos que se conoce el salario por cada sector asignado al empleado
- No siempre es posible



Dividir la jerarquía

- La relación M:N se da a través del fact
- El doble problema no se elimina, se traslada
- **Actividad: Pensar dónde está el problema de doble conteo**

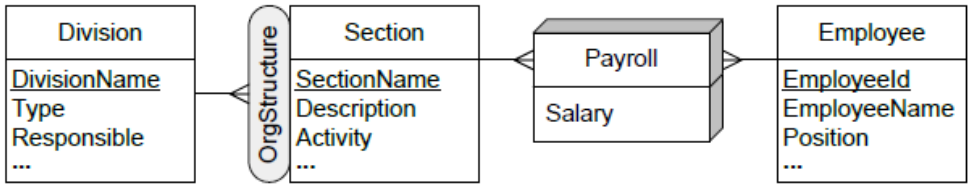


Jerarquías No-Estrictas

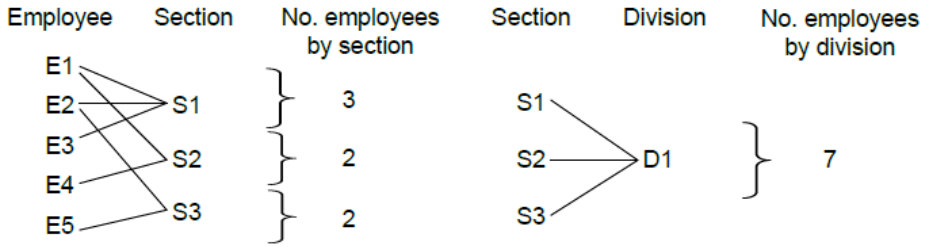
- Factor de distribución (izquierda)
- Reemplazar la relación m:n por un fact

Dividir la jerarquía

- La relación M:N se da a través del fact
- El doble problema no se elimina, se traslada



Doble conteo en la division de la jerarquía

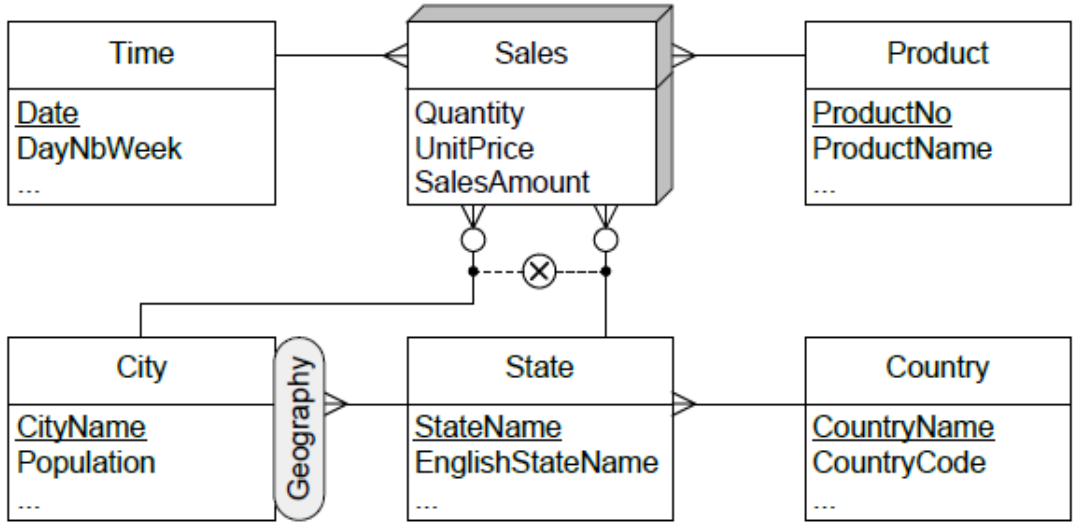


El Modelo Multidim
Casos especiales

- Facts con multiples granularidades
- Dimensiones M:N
- Relaciones entre Facts

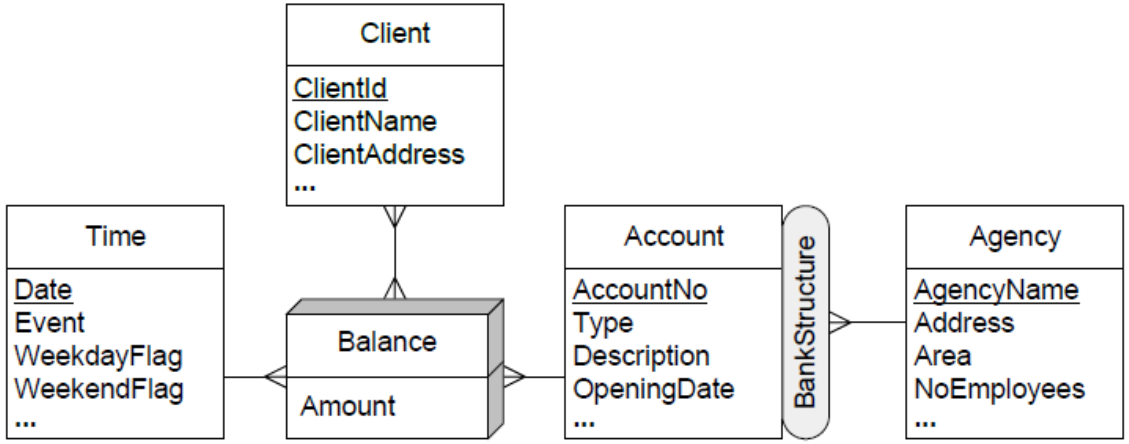
El Modelo Multidim
Hechos con más de una granularidad

- Situación muy común
- Ej: una empresa recibe datos de venta por ciudad, en ciudades grandes, o por provincial, para aquellas ciudades pequeñas
- Otro ejemplo: para algunos pacientes en un DW médico, se conoce la condición específica, para otros, una familia de condiciones (e.j, enfermedades respiratorias, y gripe, respectivamente)



El Modelo Multidim Dimensiones M:N

- Relación M:N entre fact y nivel hoja de una dimension (distinto de jerarquías no estrictas)
- Problema de doble conteo: el saldo de una cuenta se suma tantas veces como cotitulares tenga la cuenta (ver ejemplo parte inferior)



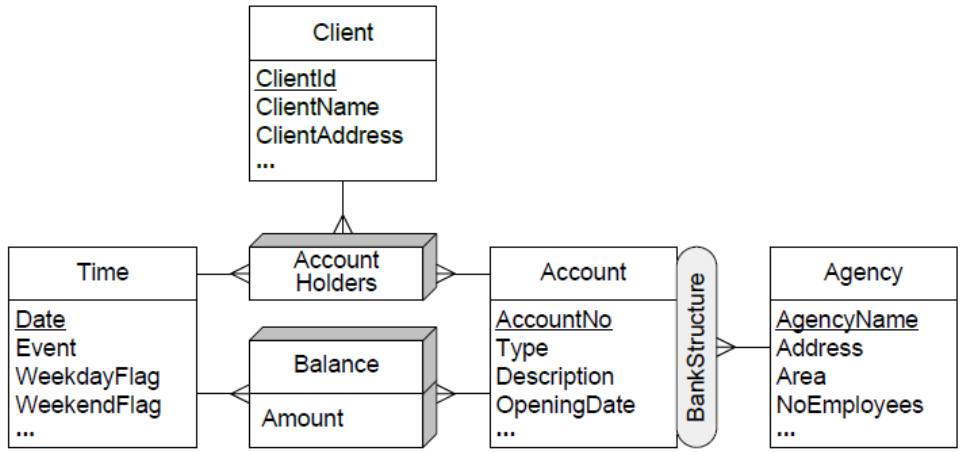
Time	Account	Client	Balance
T1	A1	C1	100
T1	A1	C2	100
T1	A1	C3	100
T1	A2	C1	500
T1	A2	C2	500

Jerarquías No-Estrictas

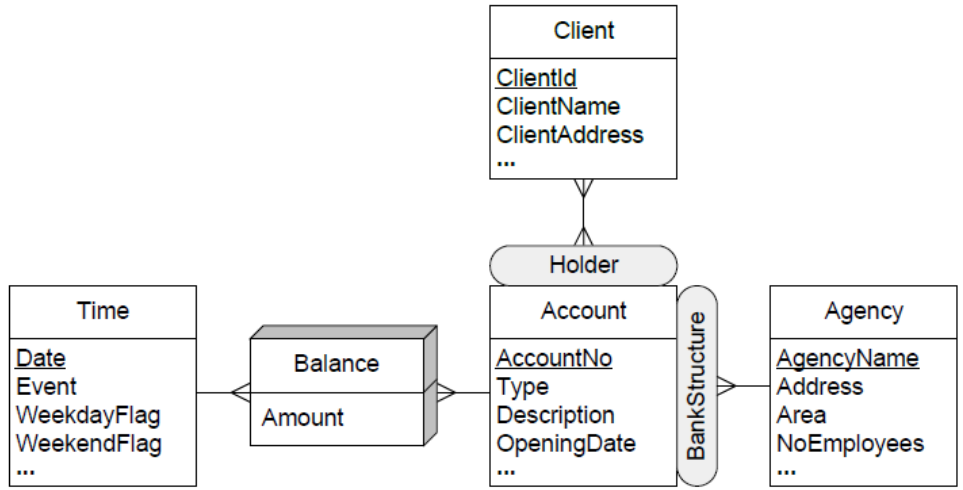
- Solución 1: crear dos facts
- Solución 2: crear una jerarquía no-estricta

- **ACTIVIDAD: Comparar ambas soluciones y decir si presentan problema de doble conteo**

Solución 1

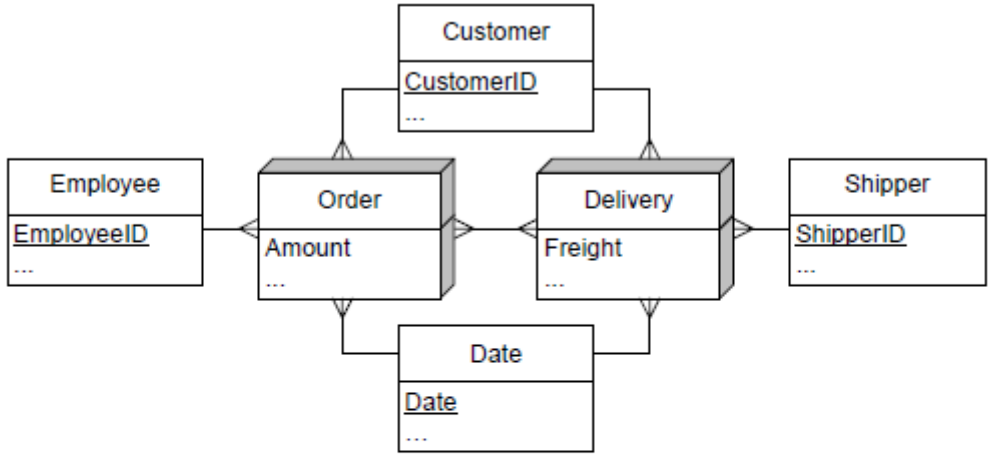


Solución 2

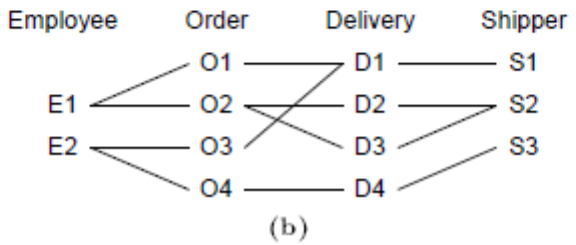


El Modelo Multidim
Relaciones entre Facts

- Relación M:N entre facts
- Varias órdenes pueden ser incluidas en el mismo envío, y una orden se puede divider en varios envíos
- Aunque los facts compartan dimensiones, el link entre ellos es necesario



(a)



(b)

Actividad

Actividad 1. Diseñe el DW para una empresa mayorista de venta de muebles. El DW debe permitir analizar las ventas de la empresa al menos con respecto a los muebles, los clientes y el tiempo. Además, la empresa necesita analizar las ventas de muebles con respecto a su tipo (silla, mesa, armario, armario ...), categoría (cocina, sala de estar, dormitorio, baño, oficina ...) y material (madera, mármol ...); compras de clientes con respecto a su ubicación geográfica, considerando al menos ciudades, regiones y estados. Se debe considerar que existen descuentos para ciertos muebles, en diferentes épocas del año. Se desea conocer la el monto de esos descuentos, y también los ingresos netos obtenidos de las ventas. Se pide dibujar el diagrama MultiDim para este escenario.

Actividad 2. Una compañía de teléfonos debe diseñar un data mart para analizar la eficiencia de los distintos planes ofrecidos a sus clientes. Luego del relevamiento realizado, se determinó que las consultas a responder, como mínimo, por el sistema, son:

a. Monto total facturado para cada plan, por año.
b. Duración total, por mes y por año, y por plan, de las llamadas iniciadas por clientes de cada provincia.
c. Número total de llamadas de fin de semana, por año, entre clientes de Buenos Aires y clientes de cada una de las capitales de las provincias.
d. Duración total, por año y por plan, de las llamadas internacionales iniciadas por los clientes de la empresa.
e. Facturación total, por plan, ciudad y provincia, en 2017, comparado contra los mismos valores obtenidos el año anterior.
f. Existen planes en los que la empresa permite al usuario definir un conjunto de números entre los cuales las llamadas son gratuitas. Computar el monto total de esas llamadas, por cliente, comparado contra el monto total por cliente de las llamadas que realiza.
g. Monto total por cliente de llamadas a líneas fijas, comparado contra las llamadas a teléfonos móviles.

Se pide dibujar el diagrama MultiDim para este escenario.