

Clase 3

Modelo Dimensional
Profundizando en el tema





De la clase pasada: El modelo dimensional. Una metodología de diseño

Según Kimball (capítulo 3 del libro), el proceso de diseño dimensional está formado por 4 pasos:

1. Seleccionar el Proceso de negocio: identificar las actividades operativas de la organización que interesa focalizar. Genera eventos o captura métricas para que se trasladan a hechos.
2. Declarar la Granularidad: aspecto fundamental que especifica el nivel de detalle que se toman los datos.
3. Identificar las Dimensiones: proveen el quien, que, cuando, donde, por que y como del evento del proceso de negocio.
4. Identificar los Hechos: evento físico que resulta del proceso de negocio.



1. Seleccionar el Proceso de Negocio

¿Qué actividades del negocio nos interesa focalizar? Los modelos de negocio son simplificaciones de la realidad que nos sirven para comprender qué está sucediendo. Si está bien definido nos permitirá responder preguntas claves de la organización.

Los **indicadores** claves del negocio (KPI) sirven para evaluar si se han alcanzado los objetivos.

Están relacionados con los objetivos y con las actividades de la organización. Deben ser medibles, comparables y cuantificables.

Por ejemplo, en un sistema de ventas podemos definir KPI sobre las unidades, el período de tiempo, el objetivo de ventas a conseguir.



1. Seleccionar el Proceso de Negocio

KPI

Pensemos un portal de venta de pasajes por Internet. Las ventas se han estancado pero el departamento de ventas no presenta dificultades para conseguir nuevos clientes. Esto significa que el cliente no repite su compra entonces es necesario analizar la razón de esta situación y mejorarla.

Un **indicador** podría ser definido de la siguiente manera:

Nombre: Repetición de venta de clientes.

Definición: venta a clientes que ya eran clientes en el trimestre anterior.

Calculo: $\text{importe de las ventas a clientes que teníamos el trimestre anterior} / \text{total de ventas}$.

Objetivo: mejorar un 3% trimestral, partiendo del 48% del período actual.

1. Seleccionar el Proceso de Negocio

Una vez definido el modelo de negocios, es necesario analizar **qué información** disponemos para analizarlo. Para esto, sigamos con el ejemplo de ventas en un supermercado, el caso más típico.

Tomamos la base de datos relacional donde se almacena información de los tickets de ventas.

BUSINESS INTELLIGENCE: COMPETIR CON INFORMACIÓN. Xavier Mendonça
de ESADE Business School

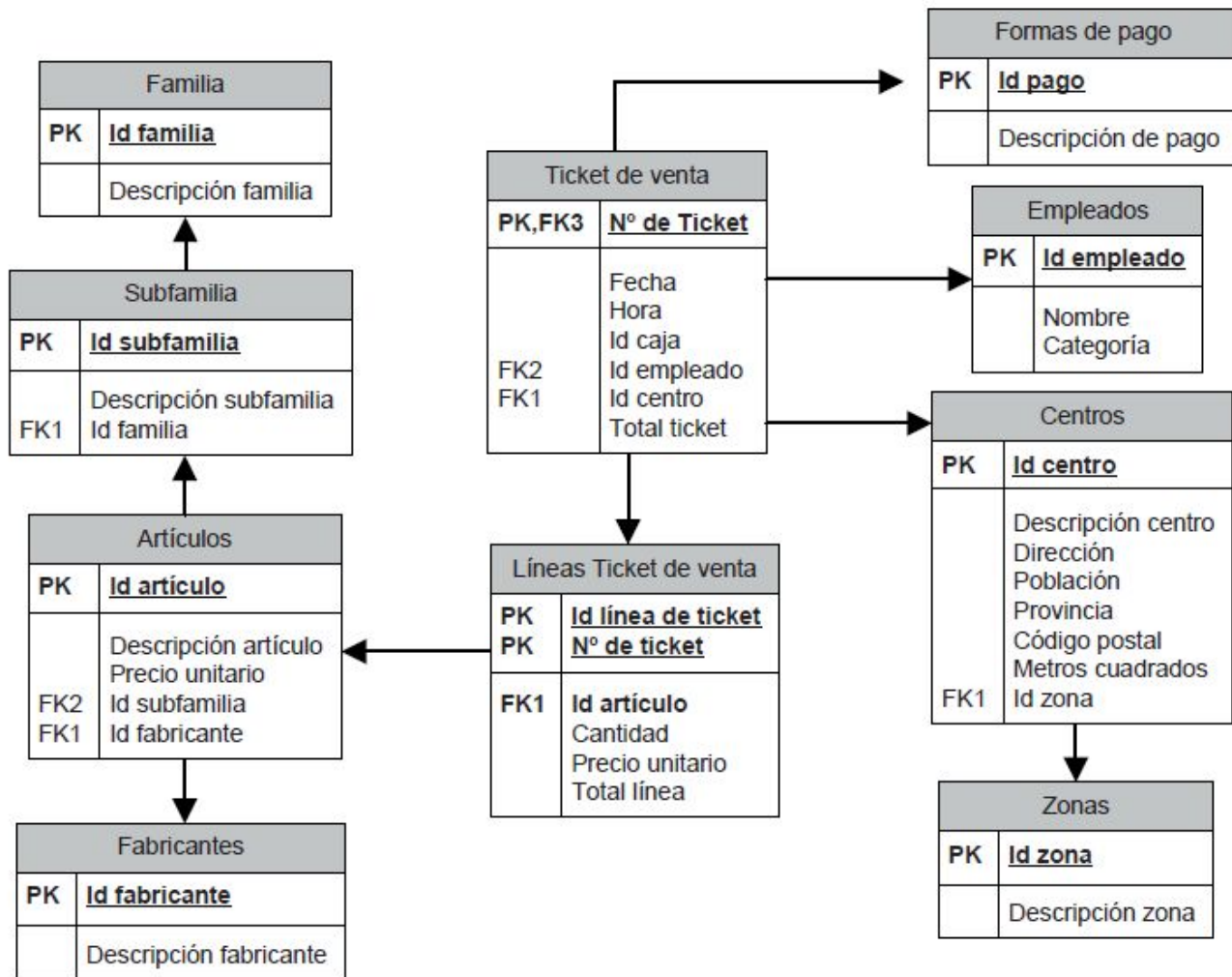
Ticket Ejemplo	
Dirección de Ejemplo Línea 1	
Dirección de Ejemplo Línea 2	
Teléfono de Ejemplo	
Fecha 11/08/14	Hora 16:19:01
001072 ACEITUNAS RELLE	10,00
2,000 x 17,50	
549225 AGUA SABORIZADA	35,00
000160 AMERICANO GANCI	57,00
TOTAL	102,00
Muchas gracias por su compra	



Ejemplo:
Ventas de un
supermercado.

Modelo Relacional

¿Qué nos interesa
analizar? ¿Los tickets
o los artículos más
vendidos? ¿O la
evolución del ticket
medio?





1. Seleccionar el proceso de negocio



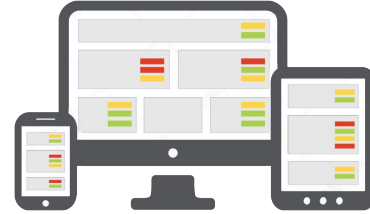
Contando con la **información operativa** de los tickets podríamos analizar:

1. Total de ventas por día
2. Nro de ticket por día, hora o fracción de tiempo
3. Análisis por cajero.
4. Análisis por artículo
5. Importe por distintas formas de pago: efectivo, débito, crédito.
6. Importe por ticket medio
7. Nro medio de tickets por día, hora, cajero

Los datos se transforman en información.

1. Seleccionar el proceso de negocio

Analizar la información



A este nivel nos permite tomar decisiones acerca de:

1. Definir artículos a reponer
2. Definir el esquema de turnos de los cajeros
3. ¿Cuales son los productos más vendidos?
4. ¿Cuál es el medio de pago más utilizado?

Si se produce una disminución en las ventas, y se cuenta con el ticket medio y el nro de tickets se puede determinar la o las causas.

¿Tenemos un problema de afluencia al supermercado? o ¿la gente compra menos?

A partir de identificar el problema la solución es distinta.

1. Seleccionar el proceso de negocio

Analizar la información

Cada modelo de datos permite responder un número limitado de preguntas y permite responder preguntas distintas.

En el modelo de ticket medio podemos responder cuestiones como el centro que más vende, en qué horario, en que período y comparar trimestres o años, por ejemplo. No hay referencia a productos.

El modelo de líneas de tickets de ventas podemos responder a preguntas como el artículo más vendido en un centro en un período, o la familia de productos más vendida.

La terminología utilizada y las definiciones deben ser las mismas. Se recomienda el uso de Metadatos o Diccionarios de datos con los atributos de las tablas, fuentes, campos recalculados o transformados.

Ticket Ejemplo	
Dirección de Ejemplo Línea 1	
Dirección de Ejemplo Línea 2	
Teléfono de Ejemplo	
Fecha 11/08/14	Hora 16:19:01
001072 ACEITUNAS RELLE	10,00
2,000 x 17,50	
549225 AGUA SABORIZADA	35,00
000160 AMERICANO GANCI	57,00
TOTAL	102,00
Muchas gracias por su compra	

1. Seleccionar el Proceso de negocio

Analizar la información disponible



Para conocer la evolución del **ticket medio** tenemos que definir el período de análisis: anual, trimestral, por cada día de la semana, diferenciar días festivos. En general, estos días o períodos no están discriminados en el sistema de gestión.

Otro análisis posible sería si hay diferencias en el ticket medio de los clientes que pagan al contado de los que usan tarjeta de crédito o débito. De esto puede surgir lanzar una tarjeta propia.

¿El ticket medio de todos los centros son iguales? Utilizando información sobre los centros, como los m2, nos permite analizar si el tamaño permite ofrecer mayor cantidad de productos y se encuentran tickets medios superiores. Esto permitirá seleccionar los nuevos centros.



1. Seleccionar el proceso de negocio

Analizar la información



También se puede analizar la población. Si está cerca de un mercado o colegio, está info no está en el sistema de gestión entonces es necesario incluirla en nuestro modelo de análisis.



2. Granularidad

Identificar la granularidad.

Es una de las actividades más importantes a nivel de diseño del DW, es la clave de la **reusabilidad**.

Alto nivel de detalle -> Baja granularidad

Bajo nivel de detalle -> Alta granularidad

El nivel de granularidad afecta a las preguntas que se pueden hacer y el espacio que ocupa la información.

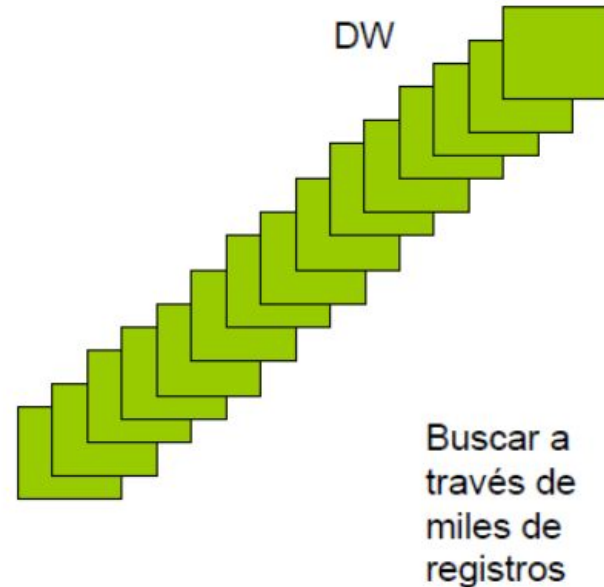
Esto implica también la gestión de índices adecuada.



En el ejemplo del supermercado, ¿cuál sería el mejor nivel de granularidad?

2. Granularidad

- ¿Cuántas llamadas se realizaron a Mar del Plata en el último mes?





2. Granularidad

Si se define un grano más grueso de granularidad (como las ventas por mes) no podrán luego consultarse las ventas por día. Entonces es recomendable utilizar un mayor nivel de detalle.

En general es mejor almacenar la información a nivel transaccional, por eso las tablas de hecho se denominan transaccional fact tables y garantizan el mayor nivel de detalle, permite construir modelos más robustos, proveen mayor flexibilidad.

Permite también realizar minería de datos y análisis por clientes.



2. Granularidad

Especificar cada fila de la tabla de hechos, indicando el nivel de detalle asociado a las medidas que contiene.

Datos Atómicos	→	Mayor Flexibilidad y mayor procesamiento
Información Agregada	→	Menor Flexibilidad y peligro de sumalizaciones no válidas.



2. Granularidad

Tipos de Granularidad

Los tipos de granularidad más usados de las tablas de hechos son:

Transacciones individuales

Un registro en la tabla de hechos por cada transacción, con una única medida. Permite analizar el comportamiento a nivel de detalle extremo. Es útil para identificar comportamientos, análisis de rangos horarios o de canasta, aplicaciones de data mining.

No es útil para análisis del estado actual del negocio, por ejemplo para saber los ingresos no siempre hay que sumar las transacciones, porque hay de diferentes tipos y estas afectan los cálculos.



2. Granularidad

Tipos de Granularidad

Fotografías (snapshots) periódicas (diaria, semanal, mensual, etc.)

La tabla de hechos se carga al final del día, o del mes o pasado un intervalo de tiempo regular y predecible, con los eventos que ocurrieron al pasar este intervalo.

Pueden existir varias medidas de la actividad ocurrida en el intervalo, por ejemplo importe total, cantidad de ventas, cantidad de clientes que compraron, etc.

Permite medir rápidamente el estado de la organización.



2. Granularidad

Tipos de Granularidad

Snapshot acumulativas

Representan un lapso de tiempo indeterminado que permite realizar un seguimiento de una transacción o un producto completo.

Es útil para identificar el estado actual de una orden y la velocidad en que un producto va superando las diferentes etapas.

Cada hecho es un registro de control. Tienen múltiples fechas representando los hitos que atraviesan.

Pueden existir diferentes medidas y en general existe una dimensión estado para guardar el valor actual de cada ítem. Se actualizan en forma frecuente.

En general se usa en conjunción con las fotografías periódicas.

2. Granularidad. Tipos de Granularidad

Comparación

Característica	Granularidad transaccional	Granularidad de fotografías periódicas	Granularidad de fotografías acumulativas.
Período de tiempo representado	Punto en el tiempo	Intervalos regulares y predecibles	Lapso de tiempo Indeterminado, típicamente de corta vida
Granularidad	Una fila por evento transaccional	Una fila por período	Una fila por “vida” (al nivel de detalle que se haga el seguimiento)
Tipo de cargas en tablas de hechos	Inserciones	Inserciones	Inserciones y actualizaciones
Actualizaciones en tablas de hechos	No	No	Se actualiza el registro cuando hay actividad.
Dimensión fecha	Fecha de la transacción	Fecha del fin del período	Múltiples fechas para hitos estándares
Métricas	Actividad transaccional	Rendimiento para intervalos de tiempo predefinidos	Rendimiento sobre el tiempo de vida finito

Dimensiones





3. Definición de Dimensiones

- Surgen de la definición del proceso de negocio, recogen los puntos de análisis. Serán mostradas al usuario para la navegación.
- Se listan todos los atributos que proveerán el detalle de la dimensión y que aplican a la tabla de hechos.
 - Descripción de los atributos: entendibles, prolijas y significativas.
 - Claves subrogadas, en general son un solo campo numérico y secuencial.

3. Definición de Dimensiones

Sub dimensiones

- Dentro de una dimensión se puede identificar una entidad independiente.
- Suelen modelarse en forma separada y facilita el reuso.
- De acuerdo a la herramienta usada para modelar, se puede visualizar como una sub dimensión o como una dimensión más.

Cliente
Tipo de cliente
Cliente
Localidad

Localidad
Identificación
Nombre
Provincia



3. Definición de Dimensiones Jerarquías Múltiples

- Los atributos de una dimensión se modelan como dimensiones separadas.
- Por ejemplo para una dimensión producto, ítem, línea, categoría, rubro y también marca y color.
- Si la herramienta lo permite se generarán múltiples jerarquías sino se mostrarán como jerarquías separadas.



3. Definición de Dimensiones

Dimensiones con valores variables

slowly changing dimension

Cambios en los atributos de la dimensión por corrección de errores o por cambios en el sistema fuente.

- Sobre escribir el valor anterior.
- Crear un nuevo registro con el valor.
- Crear un campo dimensión anterior
- No reflejar nunca estos cambios.



3. Definición de Dimensiones

Dimensiones “grandes” o Monster

- Contiene gran cantidad de elementos, por ejemplo alumnos.
- La descripción es la clave en el sistema operacional.
- Los códigos no se regeneran sino que se cargan en la tabla de hechos.
- Es necesario aplicar técnicas para controlarlas.
- Es fundamental conocer la información primordial a mantener dentro del DW



3. Definición de Dimensiones Dimensiones Degeneradas

- Se encuentran como atributos en la tabla de hechos, por ejemplo el género.
- Información de baja cardinalidad, dicotómica.



3. Definición de Dimensiones

Dimensiones Conformadas

- Permiten compartir información entre las dimensiones y poder realizar consultas comunes.
- Por ejemplo, en el contexto académico, el calendario podría ser períodos lectivos y turnos de exámenes.



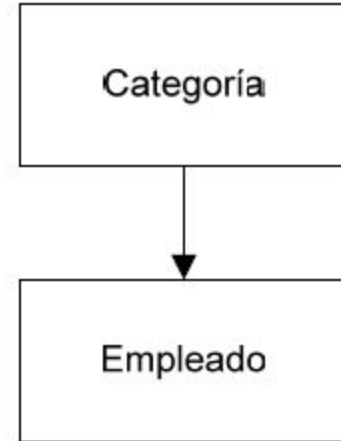
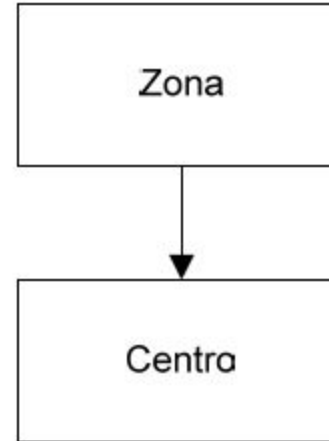
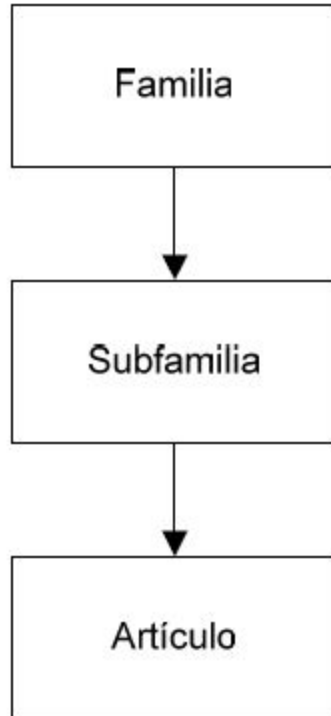
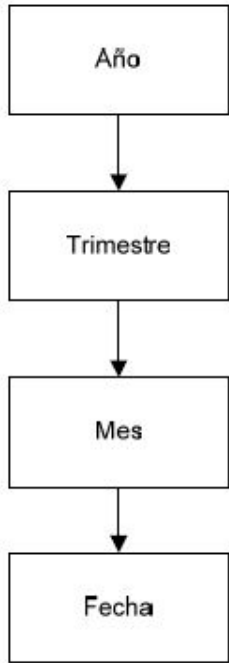
3. Definición de Dimensiones

Dimensiones Role-playing

Tienen asociado un significado, por ejemplo fecha y fecha de entrega pueden ser dos dimensiones separadas.

3. Definición de Dimensiones

Dimensiones - Jerarquías



MULTIDIMENSIONALIDAD:
analizar la información por distintas dimensiones a la vez:
analizamos las ventas por artículo de un mes y/o un centro.

Profundizar el análisis, agregar o desagregar la información

Tablas de Hechos



4. Tablas de Hechos

Representación en el data warehouse de los procesos de negocio de la organización.

Permiten almacenar:

- Medidas del proceso/actividad/flujo de trabajo que se desea modelizar.
- Claves foráneas a las dimensiones.



4. Tablas de Hechos. Tipos

Transaction Fact Table

Análisis con el máximo nivel de detalle.

Representan eventos que suceden en un determinado espacio-tiempo

Periodic Snapshot Fact Tables

Recoger información acumulada a intervalos de tiempo regulares, en forma periódica.

Factless Fact Table/Coverage

Representan el hecho de que el evento suceda.

No tiene medidas asociadas.

Accumulating Snapshot Fact Tables

Representan el ciclo de vida de una actividad o proceso en forma completa. Múltiples dimensiones relacionadas con un proceso.



Tablas de Hechos sin medidas

- *Factless fact tables*: no incluye medidas en sus campos. Son útiles para seguimiento de eventos y de cobertura.
- Seguimiento de eventos
 - Se tiene una fila por cada evento, por ejemplo una inscripción a una materia en un período académico.
 - Compuesta por claves de todas las dimensiones.
- Cobertura: cuando la tabla de hechos es “rala”. Permite dar respuesta a preguntas como el grado de uso de un aula en un edificio a partir del registro de cada bloque horario por aula.

Métricas



Identificar las Medidas

- Métricas del rendimiento del proceso de negocio a analizar.
- Todas las medidas tienen que tener el mismo nivel de granularidad.
- En general son sumalizaciones efectuadas sobre los hechos o expresiones que forman parte de un cubo multidimensional.
- Las medidas típicas son numéricas o aditivas: sumar valores a través de todas las posibles dimensiones, por ejemplo métricas de actividad, como unidades o importes.
- Los datos en el nivel de granularidad más bajo suele ser aditivo.





Medidas de Intensidad

- También son numéricas, pero la dimensión tiempo no entra dentro de la suma. Por ejemplo un inventario o un balance.
- Se pueden combinar incluyendo la medida *promedio en el tiempo*, que se basa en el promedio de los hijos.

Fecha	
Año	
Mes	
Día	

Promedio Anual: la suma de los valores en los meses dividido la cantidad de meses.

Promedio Mensual: la suma de los valores en los días dividido la cantidad de días.



Otras Medidas

- Medidas no aditivas, como promedio o temperatura, que no afectan a todas las dimensiones pueden surgir al no estar correctamente definido el nivel de granularidad.
- Medida no numérica, modelar estos casos como dimensiones, y de llevar los textos libres a una lista de alternativas.



Medidas con múltiples unidades

- Conflictos con las medidas de cantidades.
- Por ejemplo se expresan en distintas unidades de medición: unidades de envío, unidades vendidas, unidades escaneadas, etc. O tienen distintas valuaciones económicas como valor de inventario, precio de lista, etc.
- Se deberán incluir identificar las medidas fundamentales, los factores de conversión y de valuación, e incorporar estos valores a la tabla de hechos.



Medidas Básicas y Derivadas

- Una medida es Derivada si necesita ser calculada.
 - Aditivas: se calcula a partir de las otras medidas:
 $\text{ganancia} = \text{ventas} - \text{costos}$
 - No aditivas: como porcentajes, es recomendable incluirlos en la tabla de hechos para simplificar los cálculos.
- Es importante considerar desde el diseño el cálculo de medidas derivadas.



Normalización de la tabla de Hechos

Permite analizar cuando se tienen diferentes medidas asociadas a un evento. Por ejemplo el precio bruto del producto, con descuento, con IVA, entre otras.

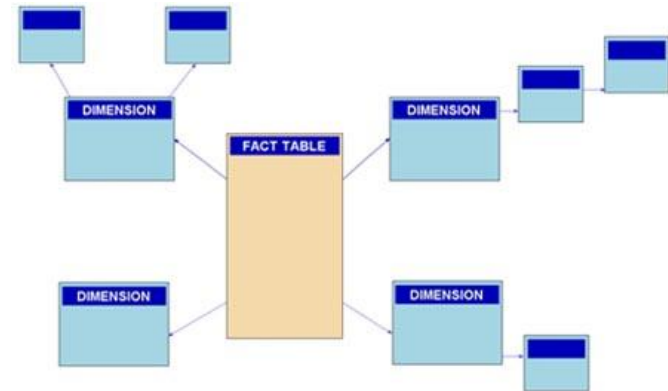
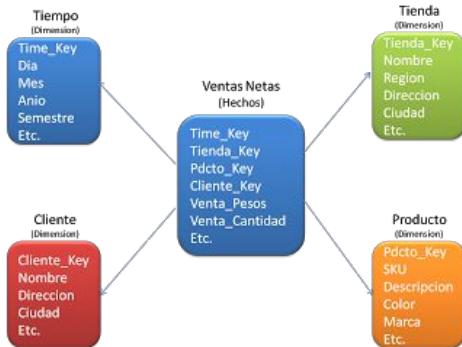
Lo más habituales es que cada una de ellas sea una columna separada en la tabla de hechos. tener las medidas como columnas de un mismo registro es en general mucho más útil.

También puede definirse una sola columna monto, junto con una dimensión adicional con el tipo de monto, cuando se tienen muchas medidas sin valor.

Esquemas para estructurar los datos

Existen principalmente dos tipos de esquemas para estructurar los datos:

- Esquema estrella
- Esquema copo de nieve.

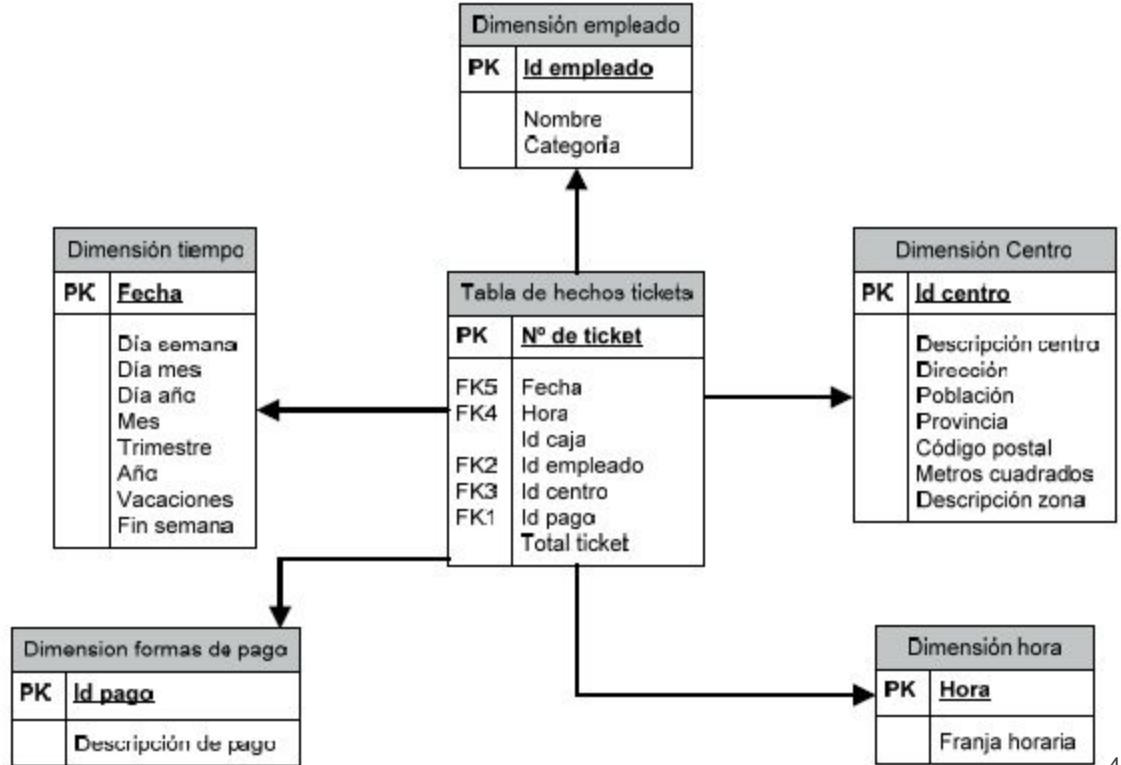


Esquema Estrella

Tabla de Hechos
(fact table): aquello
que queremos
medir o analizar.

Dimensiones: cómo
lo queremos medir.
Agrupa los hechos.

¿Total de
tickets de
venta de
una zona?



Esquema Estrella - Dimensiones

Dimensión tiempo	
PK	<u>Fecha</u>
	Día semana
	Día mes
	Día año
	Mes
	Trimestre
	Año
	Vacaciones
	Fin semana

Atributos Nuevos como Trimestre, Vacaciones, Fin semana. ¿Se vende igual en finde o en día laboral?

Dimensión hora	
PK	<u>Hora</u>
	Franja horaria

Cuatro franjas horarias: de 9:00 a 11:59, de 12:00 a 14:59, de 15:00 a 17:59 y de 18:00 a 21:00
¿En que franja se vende más?

Dimension formas de pago	
PK	<u>Id pago</u>
	Descripción de pago

Formas de pago: tarjeta de crédito, efectivo
¿Cual es la principal forma de pago de mis clientes?

Dimensión empleado	
PK	<u>Id empleado</u>
	Nombre
	Categoría

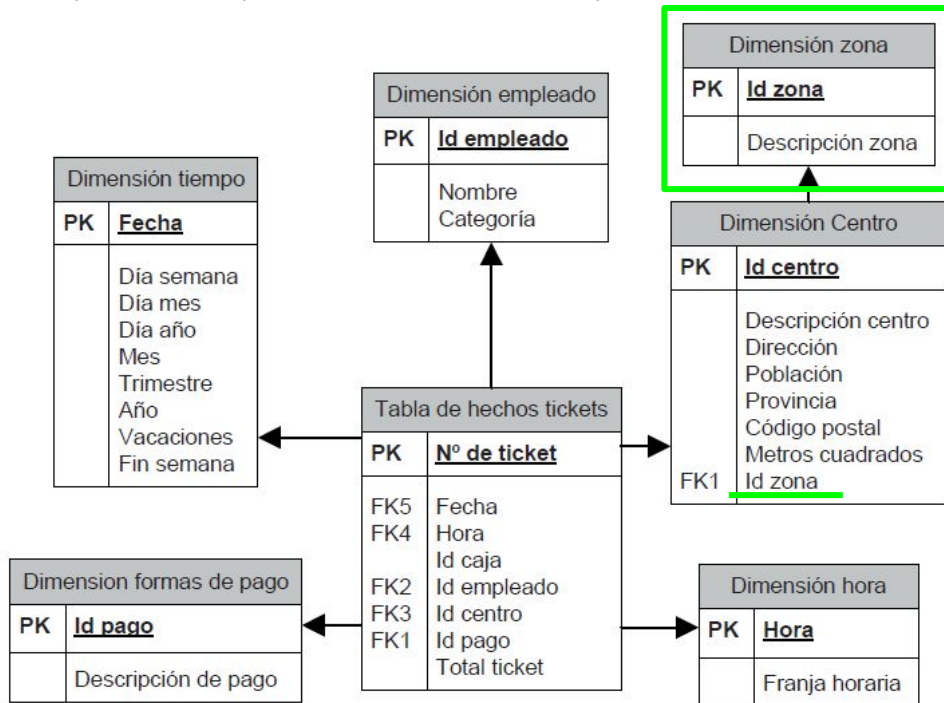
Empleado o grupo de Empleados

Dimensión Centro	
PK	<u>Id centro</u>
	Descripción centro
	Dirección
	Población
	Provincia
	Código postal
	Metros cuadrados
	Descripción zona

Ventas de una población o provincia, tamaño del centro y la zona poblacional

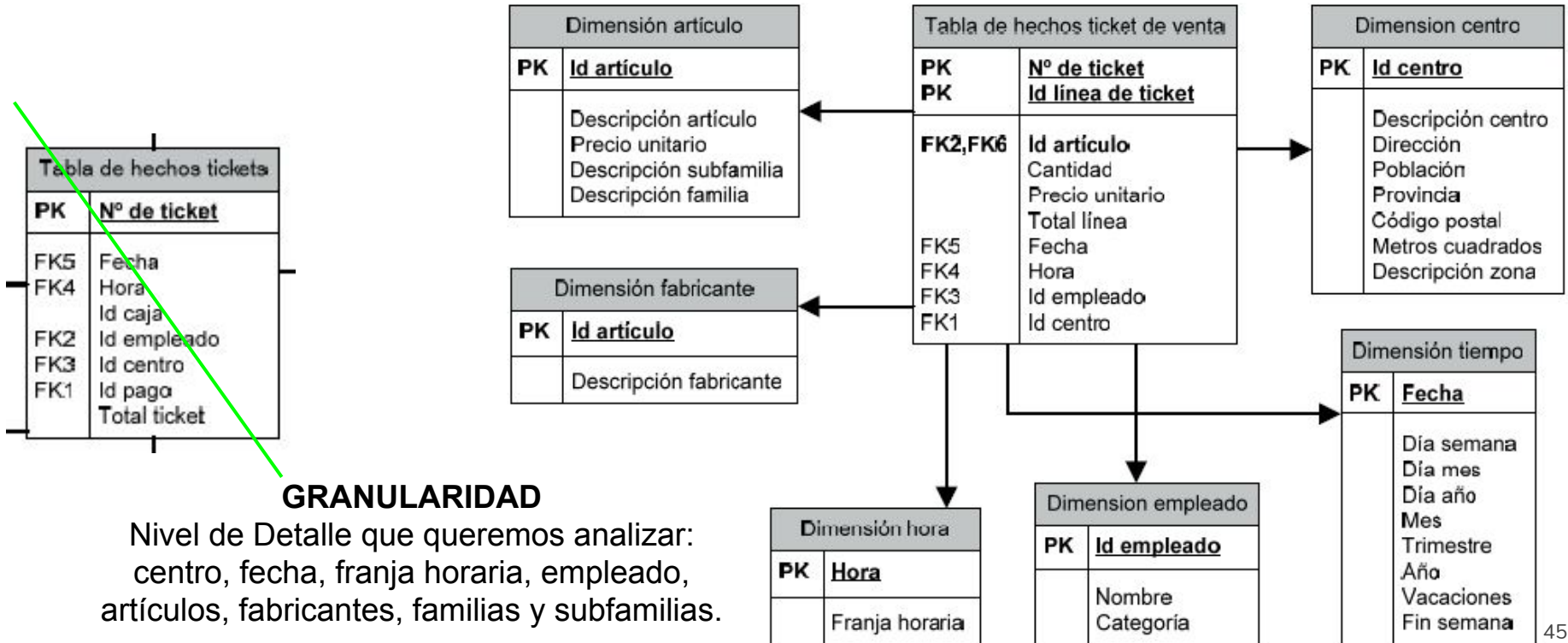
Esquema Copo de Nieve - Snowflake

El esquema estrella no está normalizado, no está e 3º forma normal. Tiene redundancia, por ejemplo “Descripción de la zona”: se repetirá tantas veces como centros existan en la misma.



Aparecen relaciones entre las dimensiones y se disminuye el almacenamiento, aunque a veces esta diferencia no es significativa.

¿Y si ahora analizamos los artículos vendidos en los tickets?





Resumiendo

¿Cuál es el modelo de negocio?

¿Qué queremos medir? Los hechos

¿Como los queremos analizar? Dimensiones de análisis: ¿Qué? ¿Quién? ¿Por qué?
¿Dónde? ¿Cuándo?

El nivel de detalle responderá a las necesidades que queramos responder con el mismo. Lo ideal es construirlo al mayor nivel de detalle.

Este ejemplo simple se puede extender a tickets más complejos, el procedimiento es el mismo.