Modelo de datos Multidimensional Tema 3

Índice

[1. Ciclo de vida 2](#_Toc73115724)

[2. Tipos de arquitecturas OLAP 2](#_Toc73115725)

[1. OLAP Relacional (ROLAP) 2](#_Toc73115726)

[2. OLAP Multidimensional (MOLAP) 2](#_Toc73115727)

[3. OLAP Híbrido (HOLAP) 3](#_Toc73115728)

[3. Diseño conceptual 3](#_Toc73115729)

[Fases de diseño 3](#_Toc73115730)

[Tipos de mediciones 3](#_Toc73115731)

[4. Diseño lógico 4](#_Toc73115732)

[4.1. Dimensión degenerada 4](#_Toc73115733)

[4.2. Flags como dimensiones 5](#_Toc73115734)

[4.3. Dimensión “cajón de sastre” 5](#_Toc73115735)

[4.4. Dimensiones lentamente cambiantes 5](#_Toc73115736)

[4.5. Desdoblamiento de dimensiones 8](#_Toc73115737)

[4.6. Dimensión de mediciones 8](#_Toc73115738)

[5. Diseño físico 8](#_Toc73115739)

[5.1. Índices de mapa de bits y de join 9](#_Toc73115740)

# Ciclo de vida

Un ciclo de vida de datos constará de 5 etapas:

1. Ingestión: se ingresan las distintas fuentes de datos.
2. Identificación/limpieza/Enriquecimiento: se identifican los tipos de datos y nombres con que aparecen en las columnas. En esta etapa los datos también pueden enriquecerse.
3. Normalización: se transforman los datos a un modelo de datos neutro. Se establecen relaciones entre las diferentes entidades de datos y se introducen las reglas de negocio.
4. Presentación: Es el paso final.
5. Explotación de los datos: los datos los adaptamos a las necesidades del cliente.

Las dos primeras etapas son de alto volumen y bajo coste y esfuerzo. En un entorno tradicional, sin embargo, el almacenamiento de datos requiere un gran esfuerzo para garantizar la calidad de los datos.

# Tipos de arquitecturas OLAP

## OLAP Relacional (ROLAP)

Usa un esquema relacional para manejar la navegación. Incluye agregación y tiene una gran escalabilidad

## OLAP Multidimensional (MOLAP)

Almacenamiento con técnicas multidimensionales y tienen un acceso rápido a datos pre-calculados previamente

## OLAP Híbrido (HOLAP)

Bajo nivel MOLAP y Alto nivel ROLAP

# Diseño conceptual

Hay que prestar atención a la calidad de las instancias. Si se pueden enriquecer las dimensiones, mejor. Esto es en la etapa 2. El diseño conceptual se corresponde con los datos que se pueden extraer de la tabla de hechos. A la tabla de hechos se conectan las dimensiones (que están formadas por jerarquías). Todo esto es el modelo conceptual.

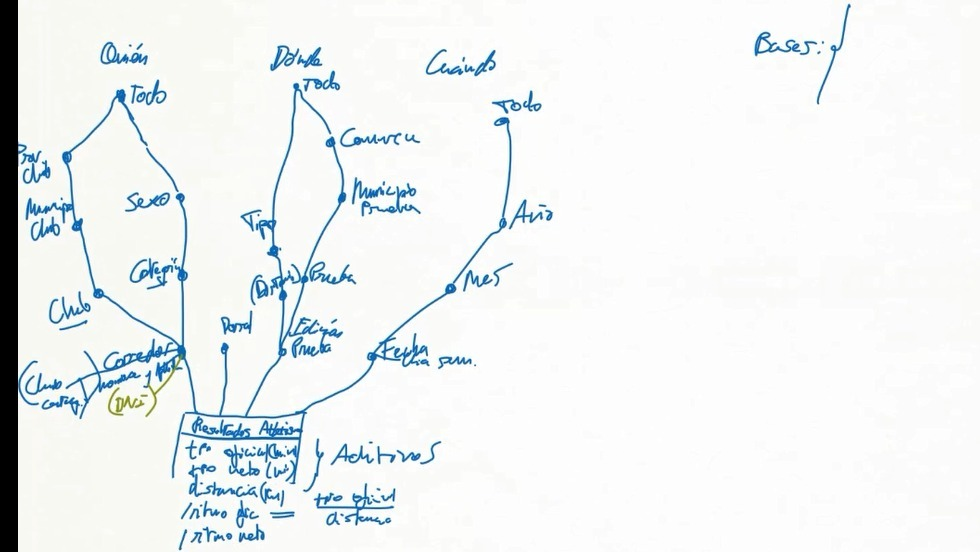
### Fases de diseño

1. Seleccionar la granularidad del proceso (cuánto de detalle). Puede ser un resumen, una línea, etc.
2. Diseñar las dimensiones (niveles y jerarquías). Para ello, debemos conectar lo que precede a cada cosa. Ejemplo: de una persona podemos conocer su género, pero de su género no podemos conocer la edad de la persona.
3. Seleccionar las mediciones (determinadas por las dimensiones)
4. Estimar el número de instancias (dimensiones y hechos)

### Tipos de mediciones

1. Medidas aditivas: pueden ser combinadas a lo largo de varias dimensiones (medir el total de ventas de un producto a lo largo del tiempo)
2. Medidas semiaditivas: no se permite la combinación a lo largo de una o varias dimensiones (sumar el stock de los productos presentes en un almacén)
3. Medidas no aditivas: no es posible combinarlas en ninguna dimensión.

Para ver de qué tipo es una medida, tenemos que preguntar lo siguiente:



Cuantos kilómetros en una prueba en una fecha ha corrido las mujeres? Tiene sentido

Cuantos kilómetros han corrido las mujeres en una edición en un año?

Tiene sentido

Cuantos kilómetros se han corrido en un año en una prueba los corredores?

Tambien tiene sentido.

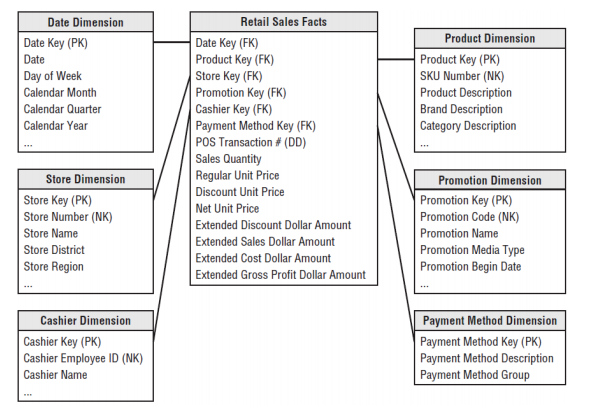
Entonces, es aditiva.

# Diseño lógico

Como se ha explicado anteriormente, se utilizará el modelo Estrella ROLAP como normal general. El diseño lógico es como un E-R. Se trata de representar las distintas dimensiones en tablas. Se usará el modelo Copo de nieve cuando se presenten las siguientes excepciones:

## Dimensión degenerada

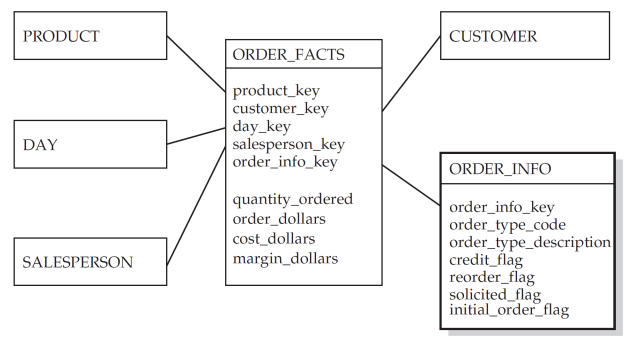
Existen atributos en la tabla de hechos que no son ni medidas ni llaves. Se puede formar una tabla con ellos. Son tablas que solamente tienen un solo atributo (como dorsal en el ejemplo de clase). Se puede poner el atributo directamente sobre el diseño lógico de la tabla de hechos, para ahorrar recursos.



## Flags como dimensiones

## Dimensión “cajón de sastre”

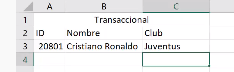
Se trata de una tabla de dimensiones que tiene atributos que no pertenecen a la tabla de hechos ni a ninguna de las tablas de dimensiones existentes. Suelen contener flags o indicadores.



## Dimensiones lentamente cambiantes

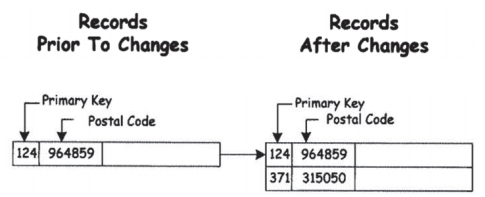
Hay dimensiones que cambian muy lentamente con el tiempo. Por ejemplo, los jugadores de un equipo permanecen por un largo periodo de tiempo. Existen varias soluciones para esto:

1. Sobrescribir (Tipo 1): el campo antiguo “se borra” y se escribe el nuevo dato. No se guarda ningún tipo de registro de este cambio. Este tipo se utiliza cuando este cambio no produce un gran impacto en el modelo de negocio.

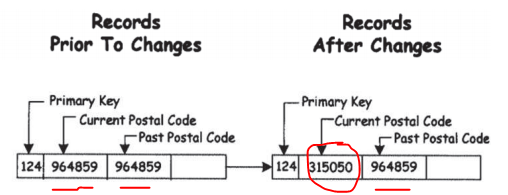


En este ejemplo, hemos actualizado el club del jugador por el antiguo, pisando el dato.

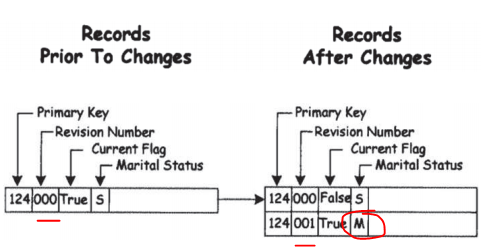
1. Añadir un nuevo registro (Tipo 2): se añade un nuevo registro a la tabla de la dimensión con el cambio realizado, manteniendo el registro antiguo intacto. Así podemos acceder a ambos registros si lo necesitamos.



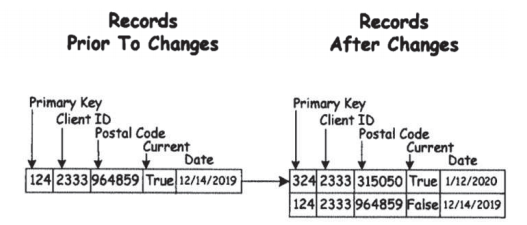
1. **Añadir un nuevo campo** (Tipo 3): se añade un campo en el registro. De esta forma, tenemos el registro con un campo “antiguo” y un campo “nuevo”. Lo más frecuente es tener dos.



1. **Añadir campo de versión de registro y añadir nuevo registro**(Tipo 4): es lo mismo que el tipo dos, pero añadiendo una revisión al registro. Se añade otro registro con los datos actualizados pero el atributo de revisión se actualiza.



1. Nuevo registro vinculado (Tipo 5): se añade un nuevo registro pero con una nueva llave generada. Hay un campo que nos permiten relacionar un registro con otro. Este es el que se suele adoptar usualmente.



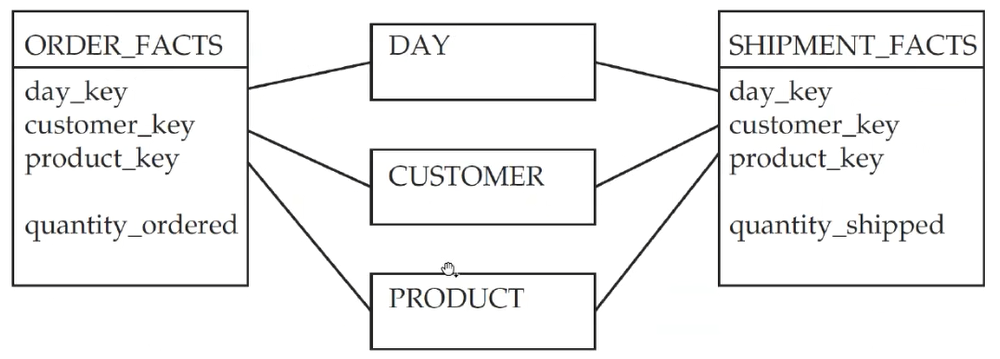
1. Tipo 1, 2 y 3 juntos (Tipo 6): Tenemos versiones de campos. Uno que se sobrescribe y otro que se añade un nuevo registro. Es un tipo 3 porque tiene versiones de campos, un tipo 2 porque añade un nuevo registro y un tipo 1 porque sobrescribe. Podremos analizar los hechos para la situación actual y los hechos para cada una de las situaciones en las que se ha estado. Nos permite analizar los datos actuales y pasados según la actualidad y analizar los datos pasados según los valores en cada momento.

## Desdoblamiento de dimensiones

Se desdoblan la tabla de hechos en varias tablas. Esto facilita las mediciones.

## Dimensión de mediciones

Se da el caso de que hay mediciones no definidas. En este caso, una solución es darse cuenta de que hay dos focos de atención. Se da el caso de que realmente son dos tablas de hechos que comparten dimensiones.

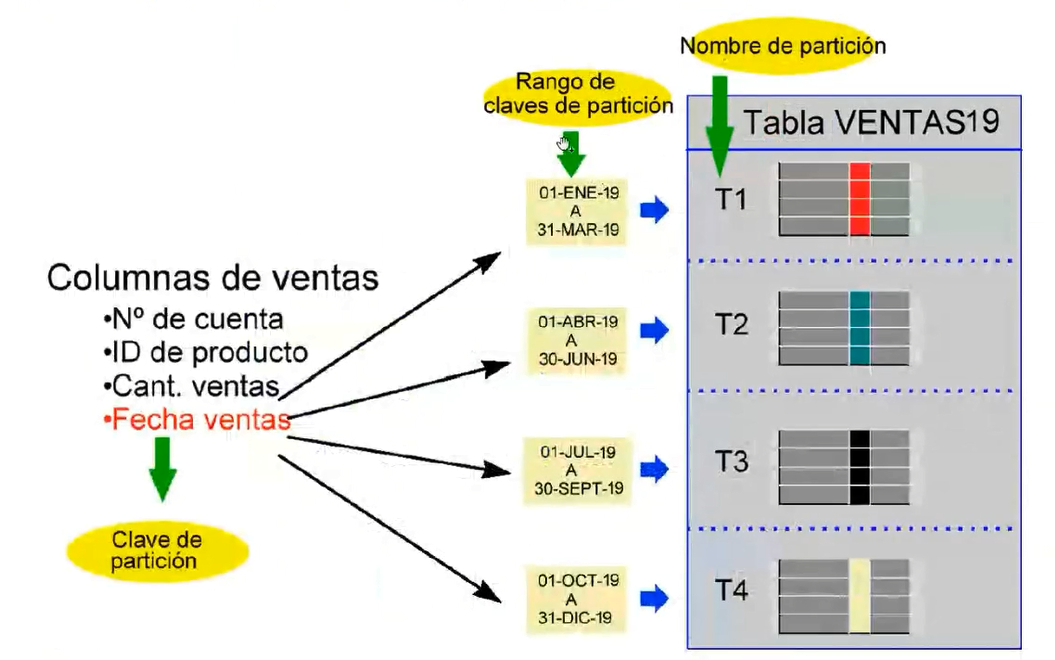


Sin embargo, si tuviésemos 20 focos de atención sería muy complicado. Una solución para esto es utilizar la **dimensión de mediciones**.

# Diseño físico

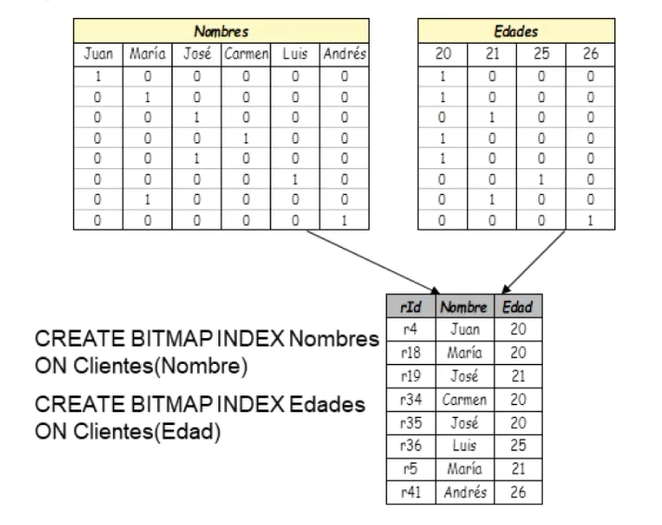
De todos los diseños lógicos nos quedamos con ROLAP. De las posibilidades de ROLAP nos quedamos con ROLAP en estrella. Sin embargo, a la regla general le metemos unas excepciones. Siempre nos interesa **coger mediciones aditivas**.

Particionar una tabla de hechos consiste en que de una tabla lógica hay varias físicas. Lo lleva el SGBD.



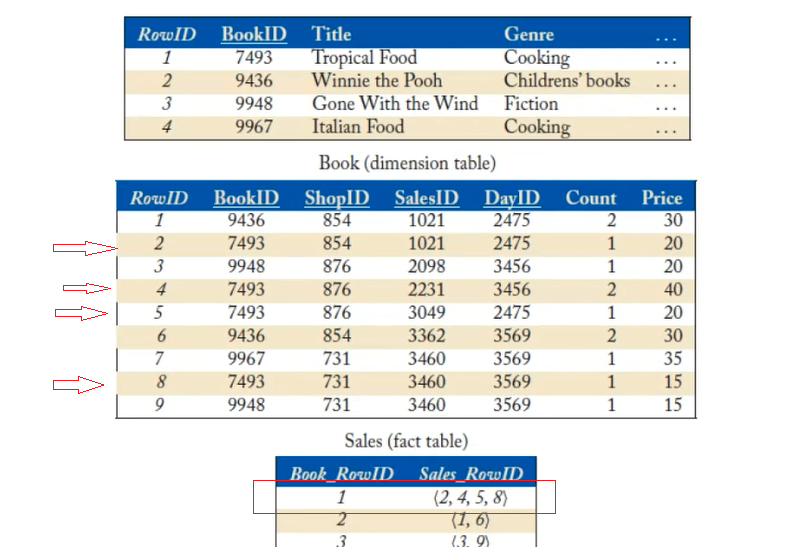
## Índices de mapa de bits y de join

El índice mapa de bits se puede usar para hacer join.



Para actualizaciones es más lento pero para consultas es más rápido.

Un **índice de Join**  es un join materializado en ese índice.



También podemos tener un join en mapa de bits. La ventaja de esto es , si queremos hacer un or o un and, con números tenemos que hacer un join de listas y es más lento.