Fundación iberCaja

KIMBALL Y EL MODELADO DIMENSIONAL

Fundamentos del modelado de datos moderno

Rubén Hermoso Díez

¿Se organizan los datos siempre de la misma forma?

SISTEMAS DE TRANSACCIONES (OLTP)

Sistemas diseñados para el manejo de datos, tanto para su almacenamiento como para su rápida recuperación.

Comúnmente asociados con los sistemas manejadores de base de datos. Estos sistemas suelen estar diseñados bajo la arquitectura cliente – servidor.

Modelo RELACIONAL

Operaciones CRUD

SISTEMAS DE ANÁLISIS (OLAP)

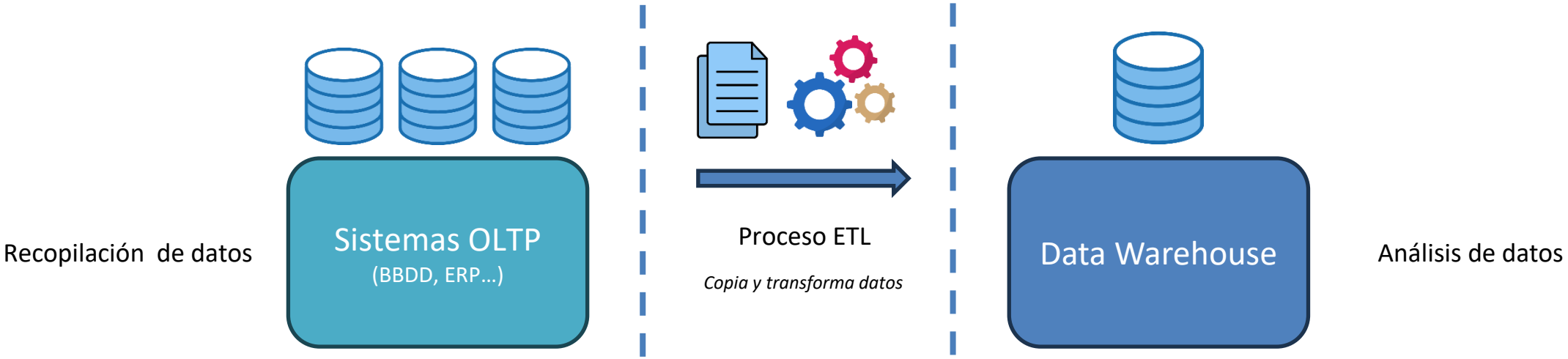
Es una categoría de tecnología de software, que permite a los analistas, gerentes y ejecutivos, obtener una mejor percepción de los datos, a través de un acceso rápido, consistente e interactivo, en una variedad de vistas posibles de la información que ha sido transformada, para reflejar la dimensionalidad real de la empresa de una manera que entienda el usuario.

Modelo DIMENSIONAL

! IMPORTANTE

Kimball propone organizar la información para que sea sencilla de analizar

Para conseguir que los datos de negocio estén bien definidos y consolidados, sean consistentes y se puedan acceder fácilmente, se estructuran y se almacenan en lo que se denomina un Data Warehouse, que son almacenes de datos corporativos que se cargan a través de un proceso de **extracción transformación y carga (ETL)**, para **posteriormente analizarlos** y obtener información relevante sobre los procesos mediante la realización de informes, estadísticas, modelos de machine learning…



! IMPORTANTE

Dentro del Datawarehouse… Se organiza la información

La idea es crear un lugar en el que cualquier usuario de negocio pueda utilizar la información del Datawarehouse para analizarla y crear sus propios informes corporativos. Muchas veces se necesitan realizar transformaciones. Esas transformaciones, se pueden realizar en fases, dando lugar a distintas capas del data warehouse.

Lo habitual es tener al menos dos capas:

- Capa bronce (una replica del transaccional original)
- Capa de consumo de datos (con las vistas de bbdd ya preparadas)

Datawarehouse

Una base de datos normal y corriente, a la que ponemos el nombre de DataWarehouse, y en la que vamos a organizar las tablas intentando seguir un enfoque de modelo de datos que NO es el relacional tradicional

! IMPORTANTE

Modelo dimensional: modelo de datos para el análisis

Dentro del DataWarehouse (una base de datos con fines analíticos), ya no organizaremos la información según el modelo relacional, usaremos el **modelo dimensional, mucho mas optimo para el análisis.** Se compone de:

Oro

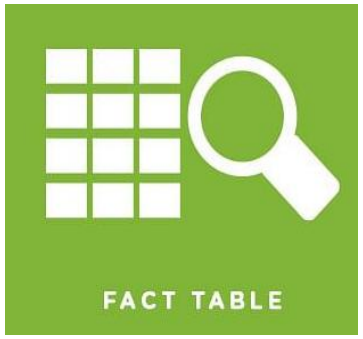
Agregaciones a nivel empresarial

Tablas de Hechos

"Lo que voy a medir"

Hace referencia a **eventos** que pueden **medirse.** Suelen tener millones de registros y pocas columnas.

Suele contener columnas de **IDENTIFICADOR** que servirán para unirse con las dimensiones (atributos del dato)

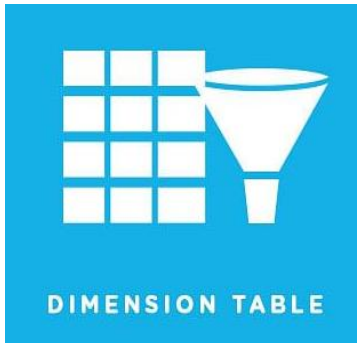


Tablas de dimensiones

"Por lo que voy a filtrar"

Se usan para **filtrar o segmentar** las tablas de hechos.

Suelen responder a las preguntas ¿Qué? ¿Qúién? ¿Cómo? ¿Cuándo? O ¿Dónde? Sucedió el evento medido. Hace referencia a datos poco cambiantes, maestros



! IMPORTANTE

Esquema Estrella

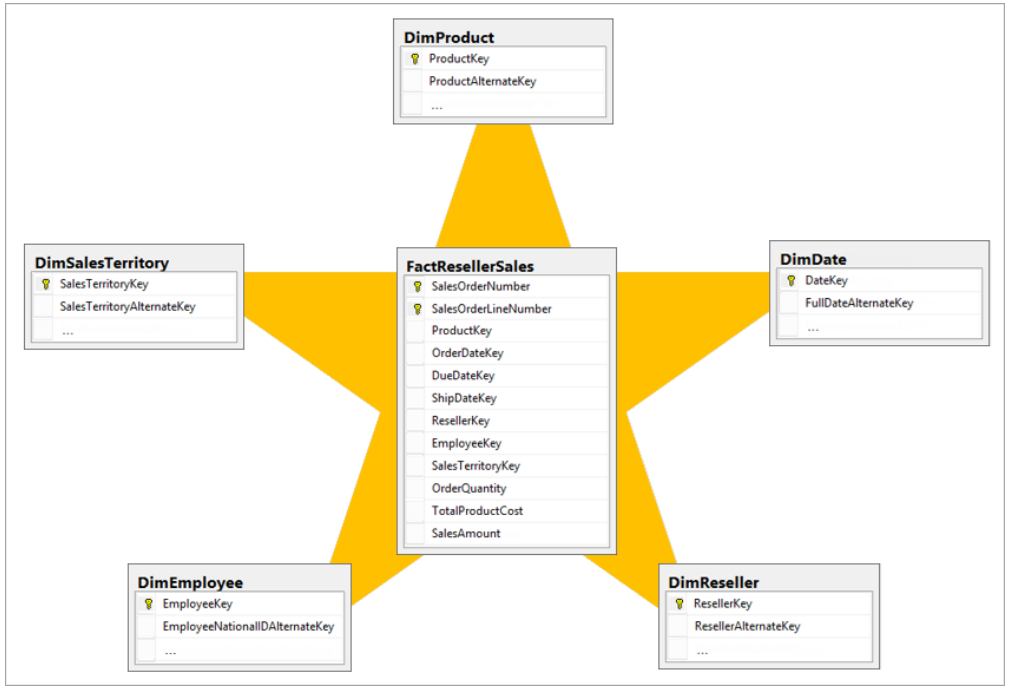
1. **Tabla de hechos central** con datos numéricos (ventas, unidades...) y claves hacia las dimensiones.

2. **Tablas de dimensiones** alrededor, con info descriptiva (producto, cliente, fecha...).

3. **Relaciones directas** entre hechos y dimensiones, como si fueran los puntos de una estrella 🌟.

4. **Diseño simple y orientado al análisis**, fácil de entender y muy usado en BI

5. **Puede haber redundancia** en las dimensiones, pero se prioriza la facilidad para el análisis.



Ejemplos: Caso de negocio de producción de piezas

Operarios

Registro de producción de piezas

Centro de trabajo

Máquinas

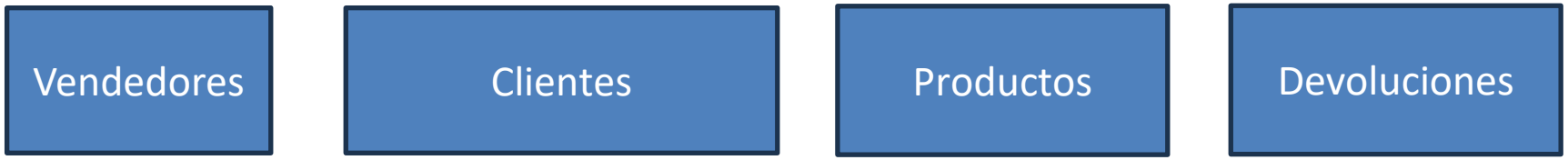
Operarios: datos de cada uno de los operarios (dni, ciudad, teléfono de contacto…)

Registro de producción de piezas: datos de cuántas piezas se van produciendo por las diferentes máquinas que manejan los operarios en los centros de trabajo

Centro de trabajo: listado de fábricas, con datos sobre la ubicación, el responsable, el teléfono…

Máquinas: listado de las diferentes máquinas que producen piezas (nº de referencia, marca….)

Ejemplos: Caso de negocio de registro de devoluciones



Vendedores: listado de vendedores con sus datos personales, teléfono de empresa, dni…

Clientes: registro de los clientes que han comprado en el establecimiento y están registrados en la web del supermercado

Productos: listado de productos, con información sobre la categoría, subcategoría a la que pertenecen y la marca

Devoluciones: listado de devoluciones que van haciendo los clientes a la empresa

¿Qué hemos hecho en los ejercicios?

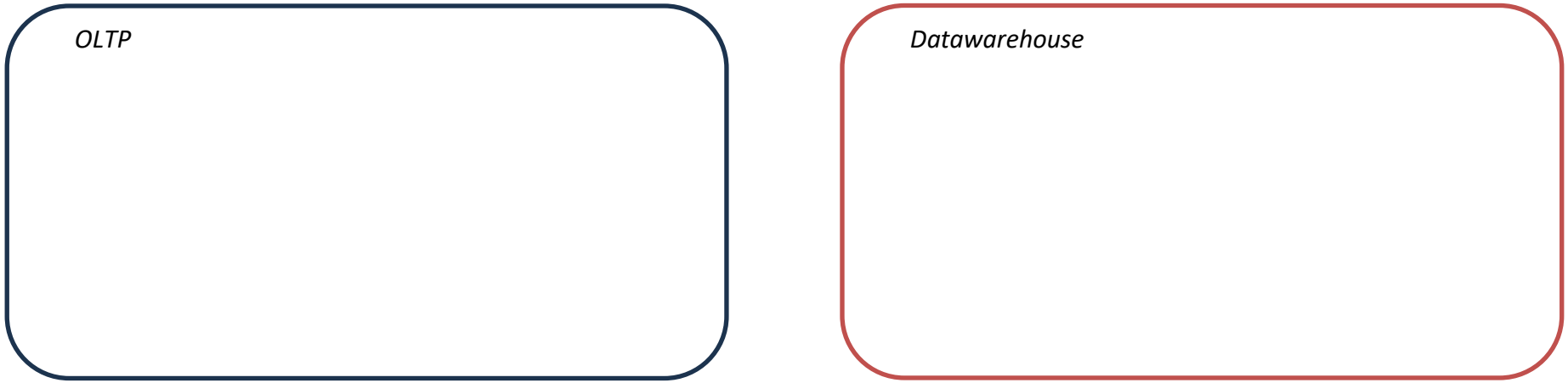
Con los ejercicios prácticos, se pretende que se tenga una idea a alto nivel del proceso de creación de un datawarehouse:

- Se identifica los datos que se necesitan del transaccional (modelo relacional original, OLTP)
- Se crea un datawarehouse si no existe
- Se replican las tablas que van a ser objeto de análisis al datawarehouse
- Se aplican transformaciones de datos necesarias

Ejercicio 2 – Creación de entornos

- Se crea un esquema OLTP: un esquema en el que vamos a crear un supuesto esquema relacional de tablas. El modelo de datos relacional no va a ser bueno para el análisis, y es por eso por lo que vamos a llevarnos la información al datawarehouse y a reorganizarla.

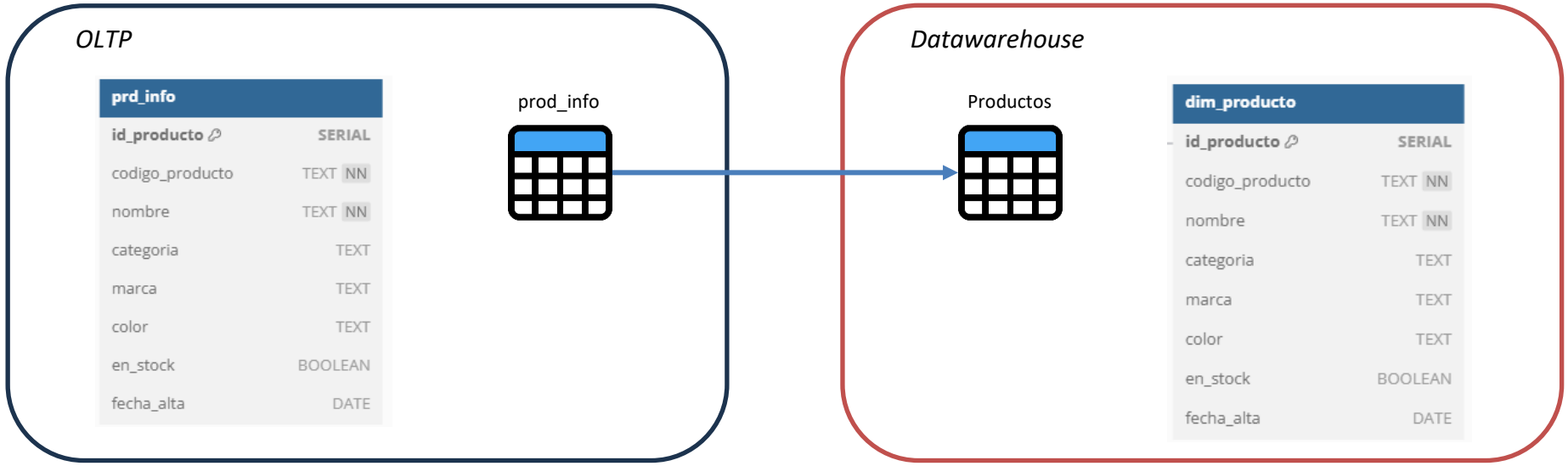
- Creamos un esquema DataWarehouse: idealmente sería un servidor de base de datos diferente, pero para nuestro caso, para simplificar, tendremos todo junto. En este esquema, es donde traeremos las tablas y crearemos un **modelo dimensional**



¿Qué hemos hecho en los ejercicios?

Ejercicio 3 – Tabla de dimensión de productos

- Directamente sobre el datawarehouse, creamos una tabla de dimensión de producto y la rellenamos con datos. Lo hacemos así por simplificar, pero realmente lo que ocurriría en un proceso empresarial real es:
  - Tenemos una tabla de producto en el OLTP
  - Nos la copiaríamos al datawarehouse para hacer nuestros análisis

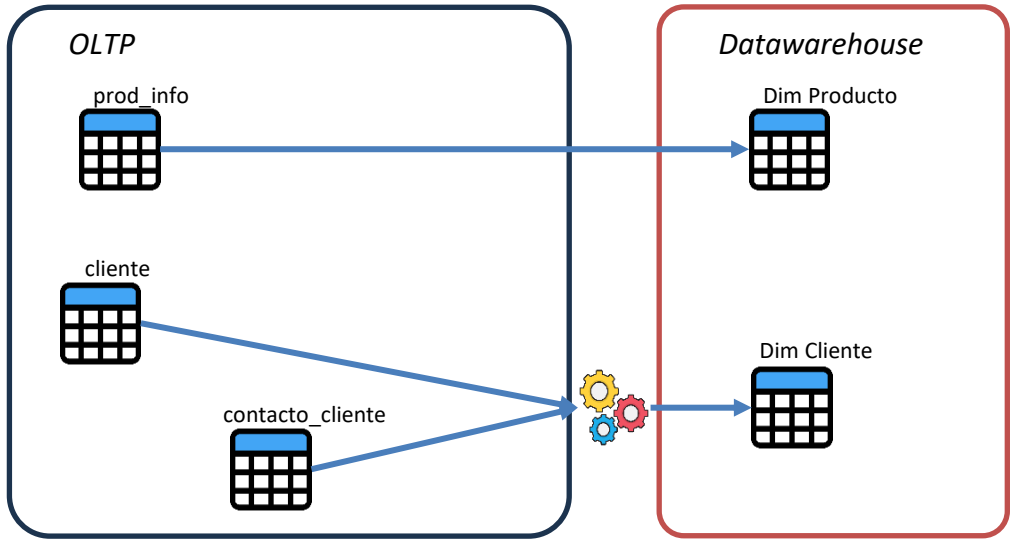


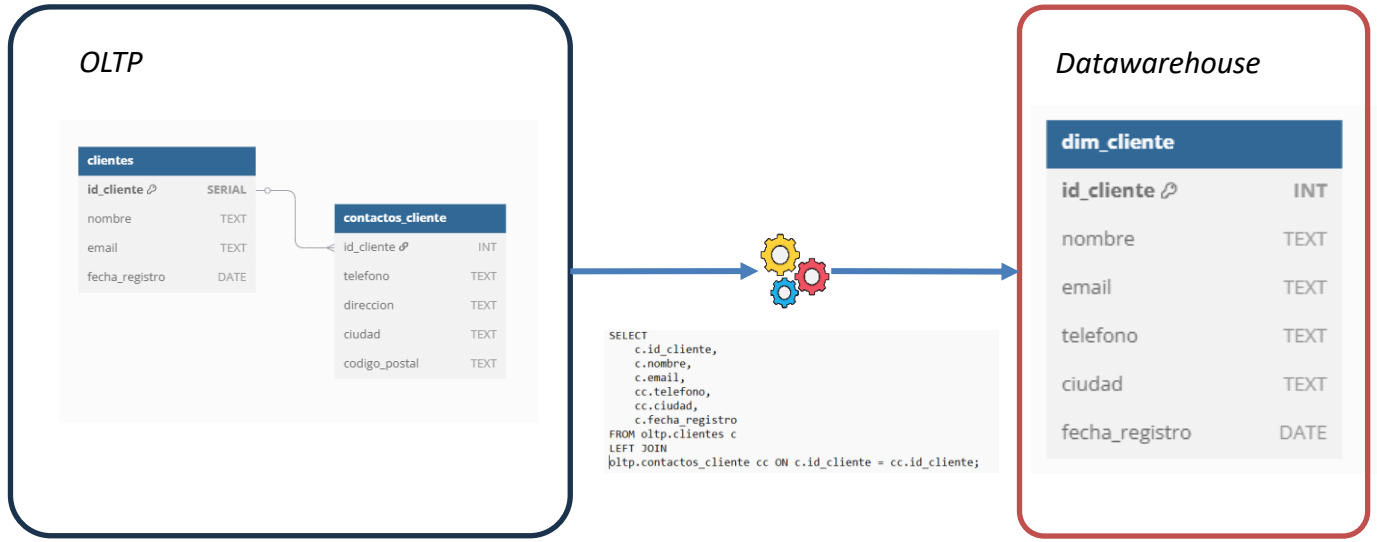
- Sobre esta tabla de dimensión, se puede analizar la información de productos y estuvimos haciendo algunas consultas sobre la tabla productos

¿Qué hemos hecho en los ejercicios?

Ejercicio 4 – Tabla de dimensión de cliente

- En el ejercicio vamos a afrontar un caso en el que no tenemos solo una tabla en el transaccional (OLTP) relacionada con los clientes. La información que necesitamos para mostrar en nuestros informes, sale de mas de una tabla de origen OLTP.
- Lo que hacemos es crear la dimensión cliente a partir de las dos tablas del relacional que vamos a usar en nuestros análisis.

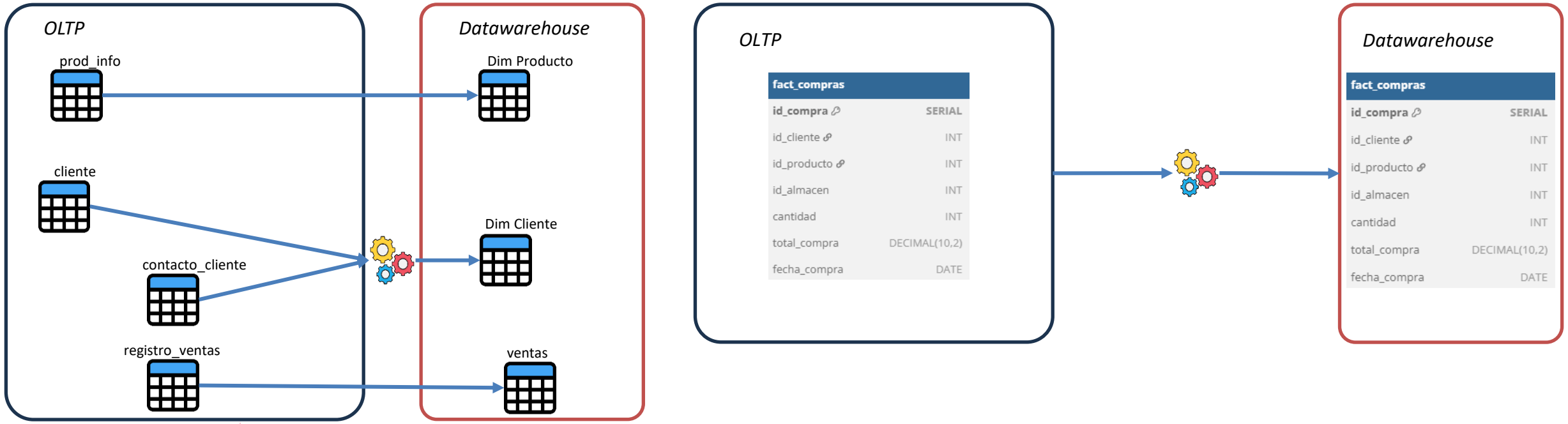




¿Qué hemos hecho en los ejercicios?

Ejercicio 7 – Añadimos una tabla de hechos

- Igual que en el primer ejercicio, por simplificar creamos la tabla de hechos directamente en el datawarehouse, pero la realidad, sería que saldría del OLTP, de las bases de datos relacionales una vez mas:

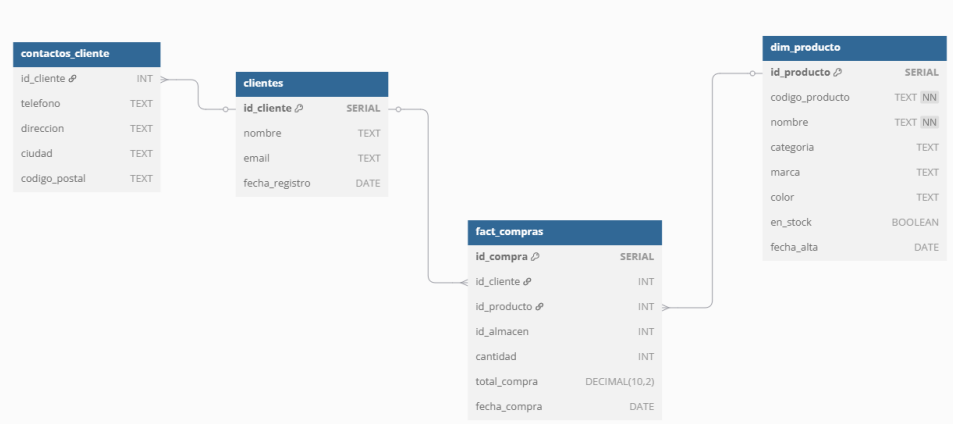


¿Qué hemos hecho en los ejercicios?

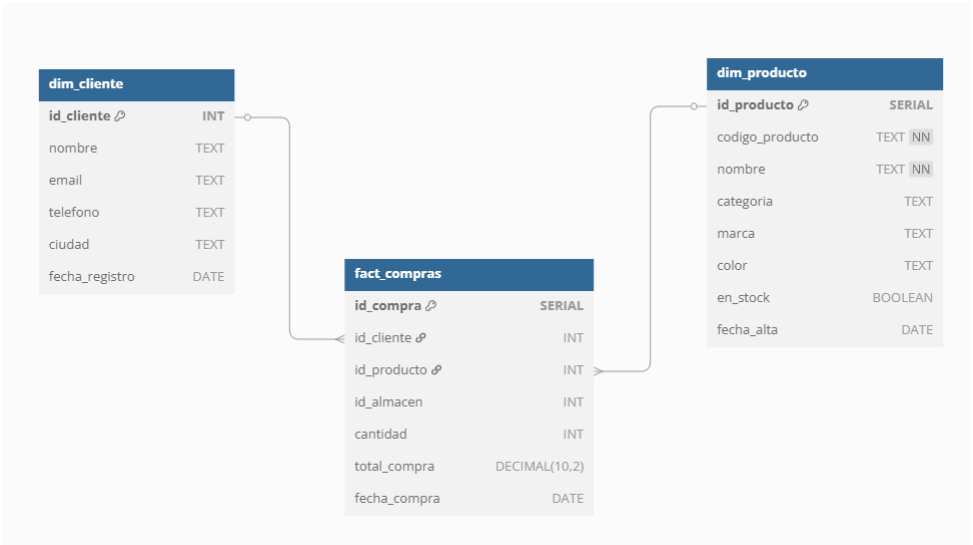
Ejercicio 7 – Añadimos una tabla de hechos

- Igual que en el primer ejercicio, por simplificar creamos la tabla de hechos directamente en el datawarehouse, pero la realidad, sería que saldría del OLTP, de las bases de datos relacionales una vez mas:

OLTP – Modelo relacional



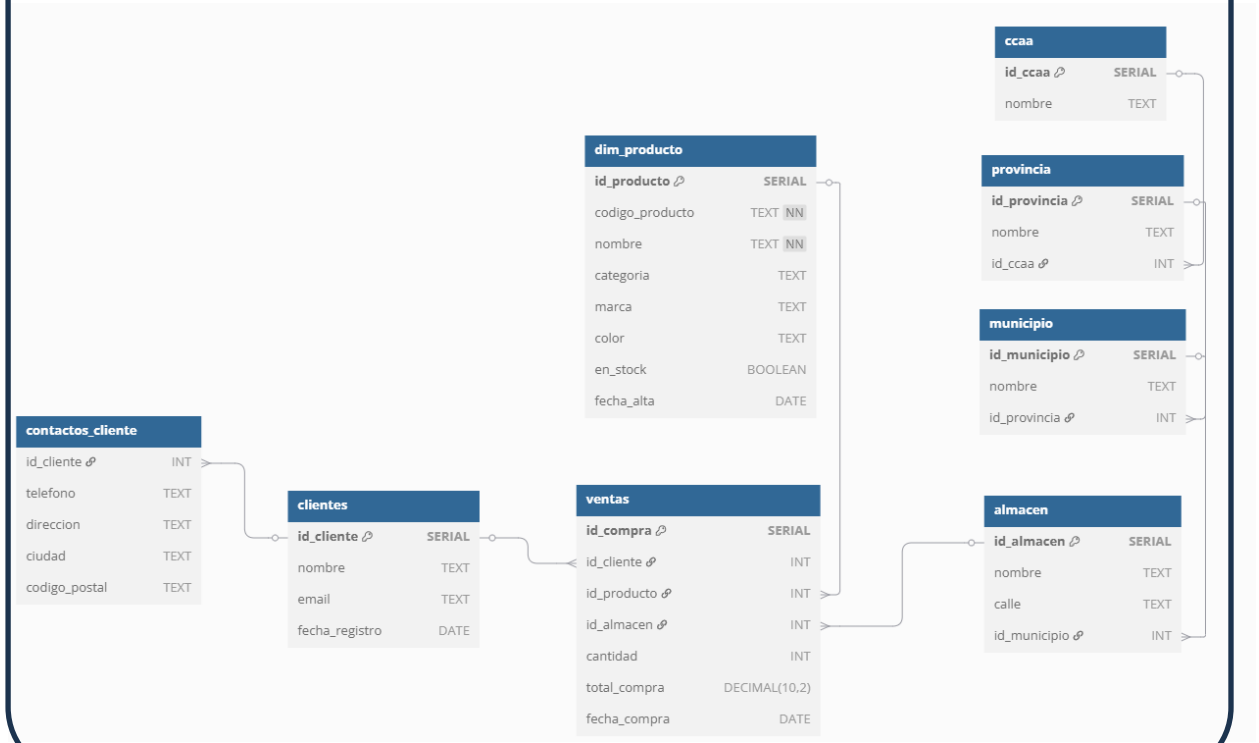
Datawarehouse – Modelo dimensional



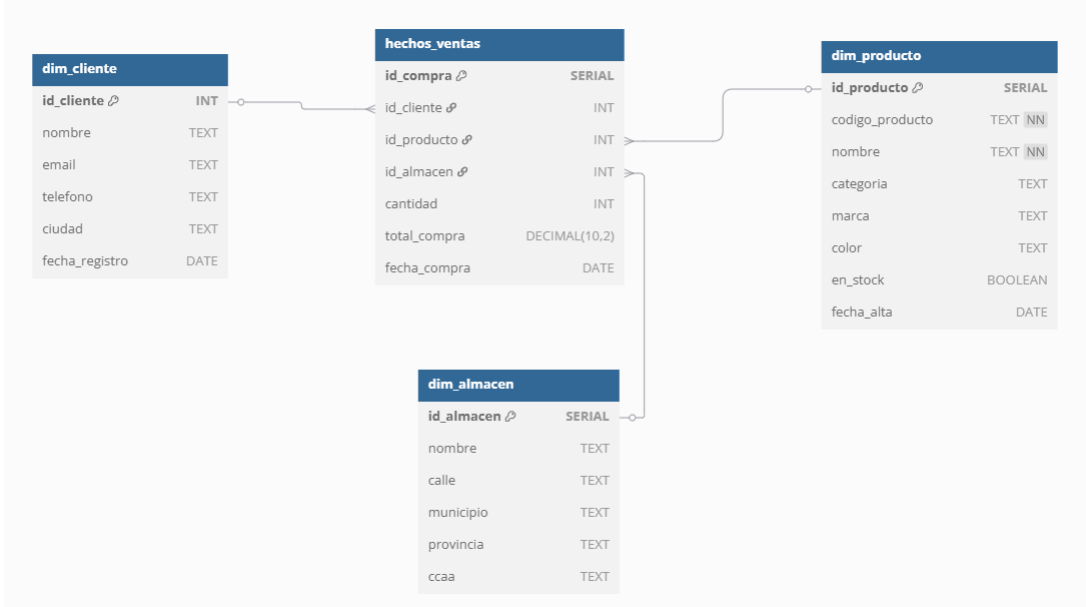
¿Qué hemos hecho en los ejercicios?

Supuesto en el que quisiésemos analizar también por almacén… Las cosas se pueden complicar si no uso un modelo dimensional desnormalizado

OLTP – Modelo relacional



Datawarehouse – Modelo dimensional



Fundación iberCaja

! IMPORTANTE

Granularidad

La **granularidad** es el nivel de detalle con el que se observa o registra algo. Cuanto mayor sea la granularidad, más detalles se capturan; cuanto menor sea, más resumida es la información.

Ejemplo: Ventas en una tienda

- **Granularidad alta (detallada)**: Si registramos las ventas por cada producto, cada cliente y cada transacción

ID Venta	Fecha	Producto	Cliente	Cantidad	Precio Unitario	Total
1	2025-05-01	Camisa	Juan	2	15€	30€
2	2025-05-01	Pantalón	María	1	30€	30€
3	2025-05-02	Zapatos	Pedro	1	50€	50€

- Granularidad baja: Si solo guardamos las ventas totales por mes o por día

Fecha	Total Ventas	Cantidad Total
2025-05-01	60€	3
2025-05-02	50€	1

! IMPORTANTE

Granularidad

La **granularidad** es el nivel de detalle con el que se observa o registra algo. Cuanto mayor sea la granularidad, más detalles se capturan; cuanto menor sea, más resumida es la información.

Ejemplo: Ventas en una tienda

- **Granularidad alta (detallada)**: Si registramos las ventas por cada producto, cada cliente y cada transacción

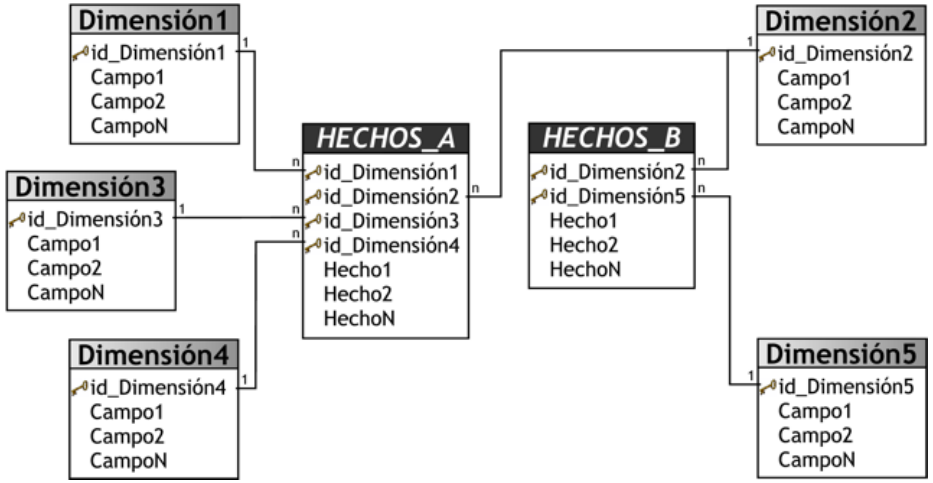
ID Pedido	Cliente	Producto	Fecha de Pedido	Cantidad	Precio Unitario	Total
5001	Jorge	Laptop	2025-05-01	1	700€	700€
5002	Carmen	Monitor	2025-05-02	2	150€	300€
5003	Pedro	Mouse	2025-05-02	3	20€	60€

- Granularidad baja: Si solo guardamos las ventas totales por mes o por día

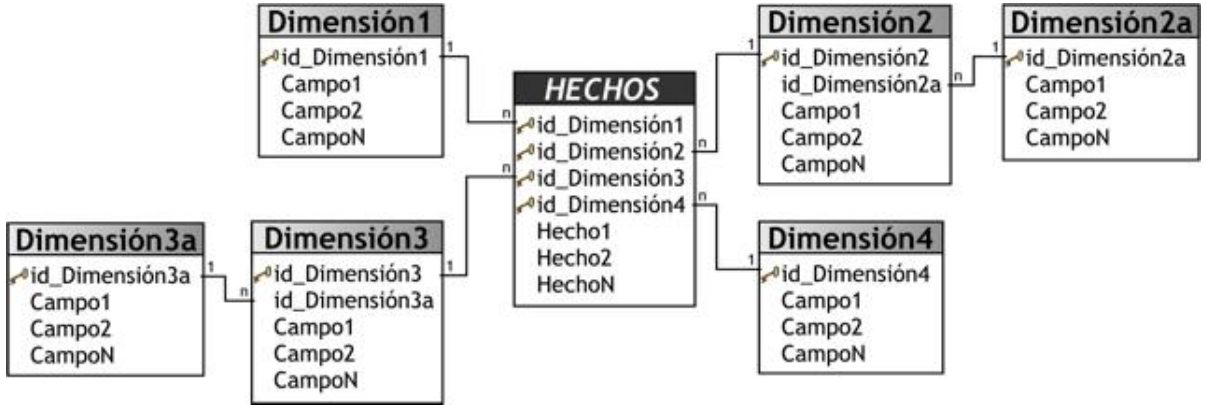
Cliente	Total Pedidos	Cantidad Total
Jorge	700€	1
Carmen	300€	2
Pedro	60€	1

Alternativas al modelo estrella

Esquema Constelación ✨



Esquema Copo de Nieve ❄️

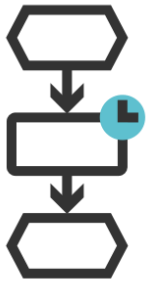


IMPORTANTE

Proceso de diseño de un Data Warehouse

Paso 1. Elegir los procesos o actividades de negocio a modelar o analizar

Proceso: actividad de la empresa soportada en una base de datos relacional (OLTP) de la cual se puede extraer información para crear un almacén de datos



Implica analizar las tablas y sus relaciones

Paso 2. Definir la granularidad (nivel de detalle para representar el proceso)

Granularidad: nivel de detalle de la información a almacenar sobre el proceso que vamos a modelar

Identificar lo que se desea medir:

- Ayuda a definir el nivel atómico de los datos en el almacén de datos
- Ayuda a determinar las dimensiones básicas del modelo de datos
- Ayuda a saber cómo se relacionarán los datos en la tabla de hechos

Proceso de diseño de un Data Warehouse

Paso 3. Pensar en las dimensiones que caracterizan el proceso

Dimensiones: las dimensiones que caracterizan la actividad al nivel de detalle (granularidad) que se ha elegido

DEFINIR LAS TABLAS DE DIMENSIONES

Paso 4. Definir información a almacenar sobre el proceso

Hechos: información (sobre la actividad) que se desea almacenar en cada tupla de la tabla de hechos y que será el objeto del análisis

DEFINIR LA TABLA DE HECHOS

IMPORTANTE

TIPS

- Claves primarias sin significado
- Evitar normalizar (no copo nieve)
- Dimensión tiempo

Fundación ibercaja