

Semana 4 – Modelado dimensional

OLTP

- On-Line Transaction Processing.
- Sistemas operacionales que capturan transacciones y las almacenan en Base de Datos.
- Características:
 - Transacciones en tiempo real (con día a día)
 - Datos almacenados cambian continuamente.
 - Mantienen los datos (INSERT; DELETE; UPDATE)
 - Estructuras de datos optimizadas – normalizadas.
 - Basado en reglas.
 - Limitado para la toma de decisiones, las consultas históricas producen un impacto en la operación del sistema.
 - Usa Diagrama Entidad Relación (DER).

OLAP

- On-Line Analytical Processing.
- Respuesta rápida y flexible a consultas, orientada al análisis de datos.
- Características:
 - Optimizado para responder rápidamente a consultas.
 - Consulta interactiva de los usuarios.
 - Almacenan varios niveles de datos optimizadas para responden a consultas.
 - Proporciona una vista de datos multidimensional.
 - Se puede cambiar fácilmente filas, columnas, y páginas en informes de OLAP.

Ejemplos

- Operación en sitio Web:
 - Validar al cliente y autenticarlo en el sistema.
 - Tomar el pedido.
 - Controlar los topes de créditos.
 - Informar los valores parciales de la compra y acumulados.
 - Requerir confirmación del cliente antes de enviar el pedido.
 - Enviar el pedido.
 - Descontar del stock las cantidades vendidas.
 - Informar el número de venta y la fecha de entrega.
 - Saludar al cliente.
- Realizar una transferencia:
 - Verificar que está autorizado para realizarla.
 - Verificar que tiene saldo.
 - Inicializar la transferencia manejándola como una transacción.
 - Emitir comprobante.
 - Saludar al Cliente.
- Sistemas de Información para ejecutivos
 - Alertas.
 - Toma de decisiones.
- En la Actividad Financiera
 - Reportes analíticos.
 - Planeamiento.
 - Análisis.
- En el Marketing
 - Análisis de productos.
 - Análisis de Clientes.
 - Análisis de Facturación.
- Otros Usos
 - Análisis de la Producción.
 - Análisis de Servicios al cliente.
 - Evolución del Costo del producto.

OLTP

- Generalmente se usa el modelo relacional.
- Se busca eliminar redundancias.
- Dividimos la información en entidades discretas.
- Se diseñan buscando el satisfacer los requerimientos de un sistema de información.
- Está basado en tablas con distintos atributos o campos y las relaciones entre las tablas. Cada tabla tiene un Clave primaria ("Primary key" o PK en nuestro esquema) formada por uno o más atributos y las tablas se relacionan entre ellas mediante las Claves externas ("Foreign Key" o FK en nuestro esquema) que actúan como claves primarias en sus propias tablas.

OLAP

- Usamos el modelado dimensional
- También conocido como Modelo Estrella (Star join Schema)
- Técnica de diseño lógico.
- Representa diagramas orientados a temas.
- Busca presentar la información dentro de un marco intuitivo.
- Permite un acceso de alta performance, por su diseño es de alta performance en las consultas.
- Fácilmente accesible y entendible.

Cuando hay problemas de performance en una db, se tienen 2 soluciones

- Escalamiento vertical: crecer la capacidad del equipo
- Escalamiento horizontal: incrementar el número de equipos que ejecutan el sistema

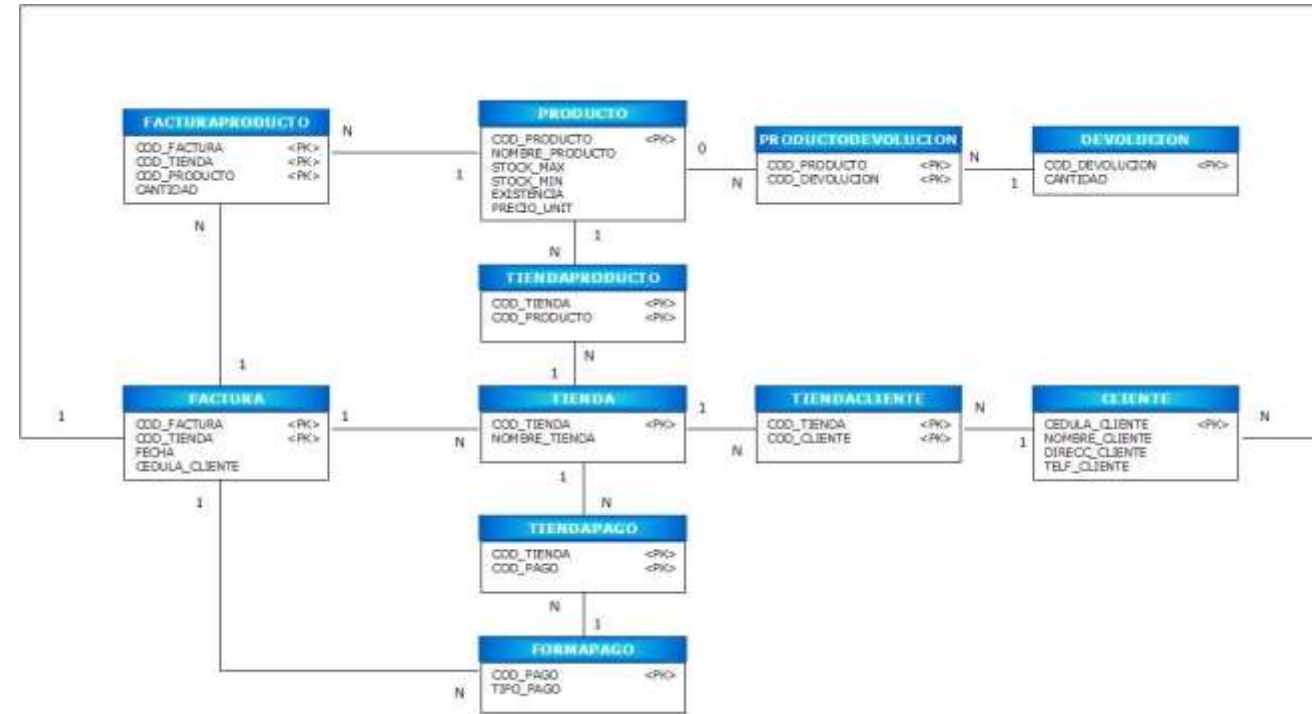
El escalamiento horizontal en las db es difícil de hacer ya que la data debe de estar sincronizada.

Modelo Relacional

Modelo de datos Conceptual = Base de Datos Relacional

- Utiliza un serie de símbolos y reglas para representar datos y relaciones.
- Representa de manera gráfica la estructura lógica de una Base de datos.
- Refleja tan solo la existencia de los datos, no lo que se hace con ellos.
- No tiene en cuenta espacio, almacenamiento, ni tiempo de ejecución.
- Está abierto a la evolución del sistema.
- Es el más utilizado.

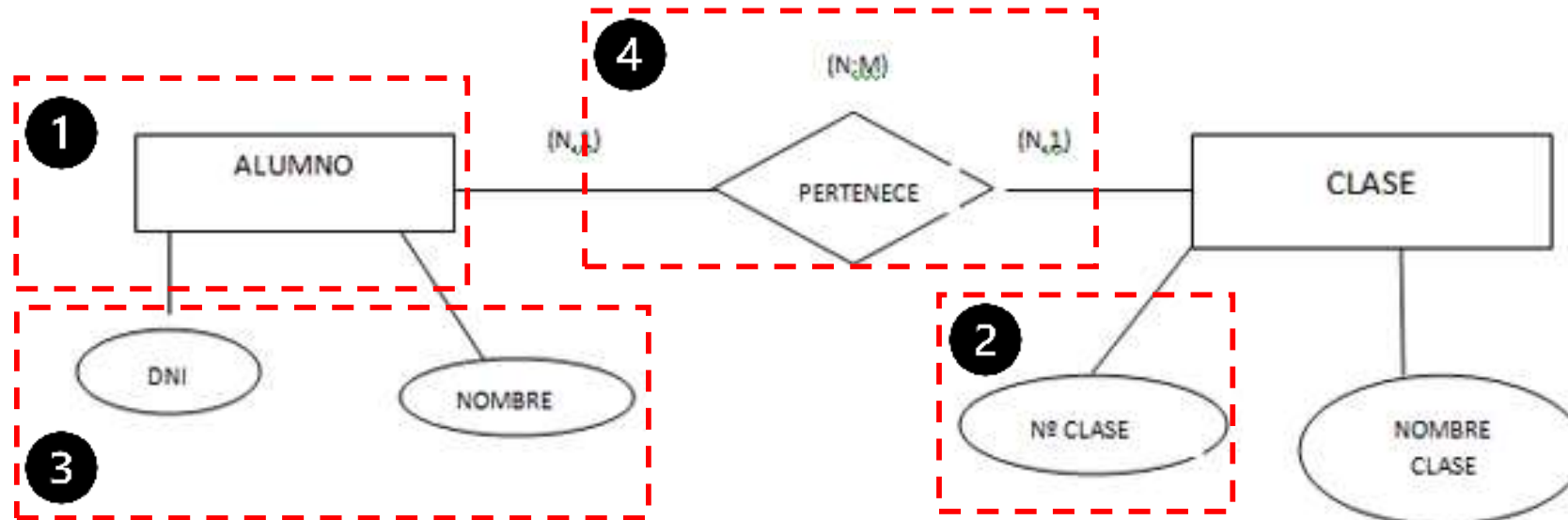
- Entidad
- Atributo
- Clave principal
- Relaciones



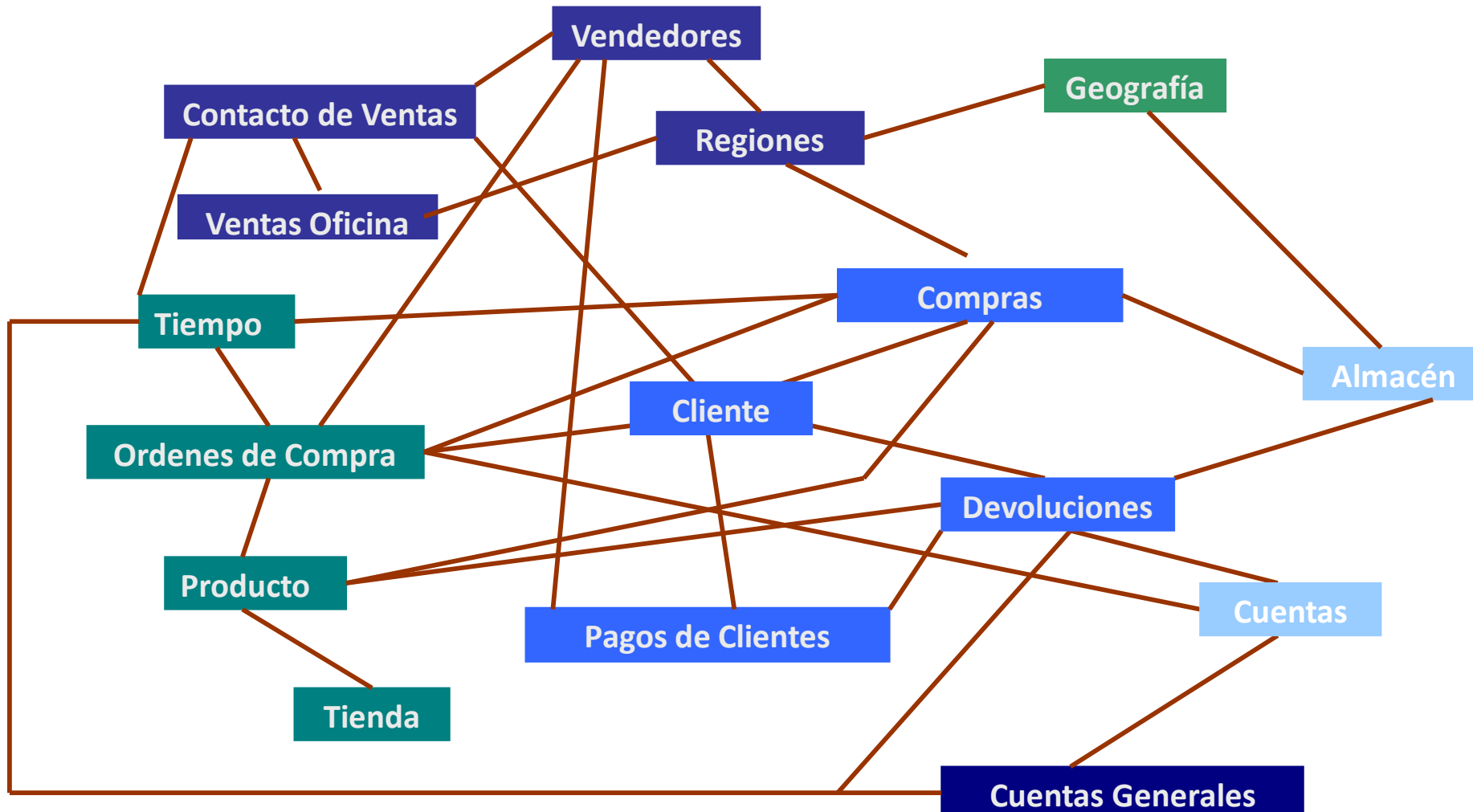
¿Cómo hacerlo?

1. Identificar las entidades
2. Determinar las claves primarias
3. Describir los atributos de las entidades
4. Establecer relaciones entre las entidades
5. Dibujar el modelo de datos
6. Realizar comprobaciones.

- 1 Las entidades representan cosas u objetos (ya sean reales o abstractos), que se diferencian claramente entre sí.
- 2 Es el atributo de una entidad, al que le aplicamos una restricción que lo distingue de los demás registros.
 - Clave Primaria
 - Clave Foránea
- 3 Los atributos definen o identifican las características de entidad (es el contenido de esta entidad). Cada entidad contiene distintos atributos, que dan información sobre esta entidad. Estos atributos pueden ser de distintos tipos (numéricos, texto, fecha...).
- 4 Es un vínculo que nos permite definir una dependencia entre varias entidades, es decir, nos permite exigir que varias entidades compartan ciertos atributos de forma indispensable.



Ejemplo



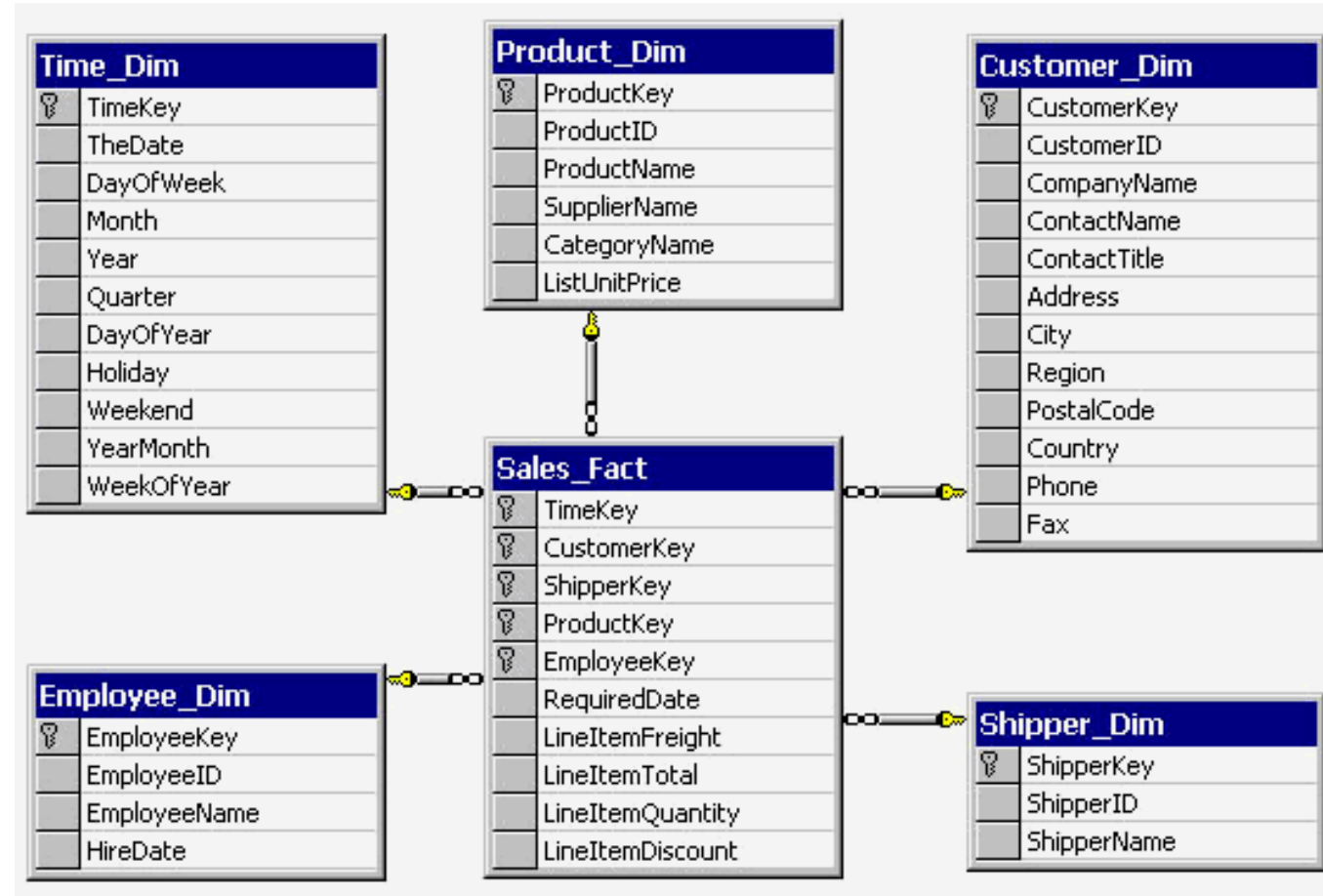
Modelo dimensional

Modelo de datos Lógico = Alto desempeño

- Una técnica para diseñar el modelo lógico de la bodega de datos.
- Permite alto rendimiento en el momento de acceder a los datos (orientado a consultas)
- Dimensional (orientado al negocio)
- Usa algunos conceptos del modelo entidad/relación
- Diferente del modelo relacional.

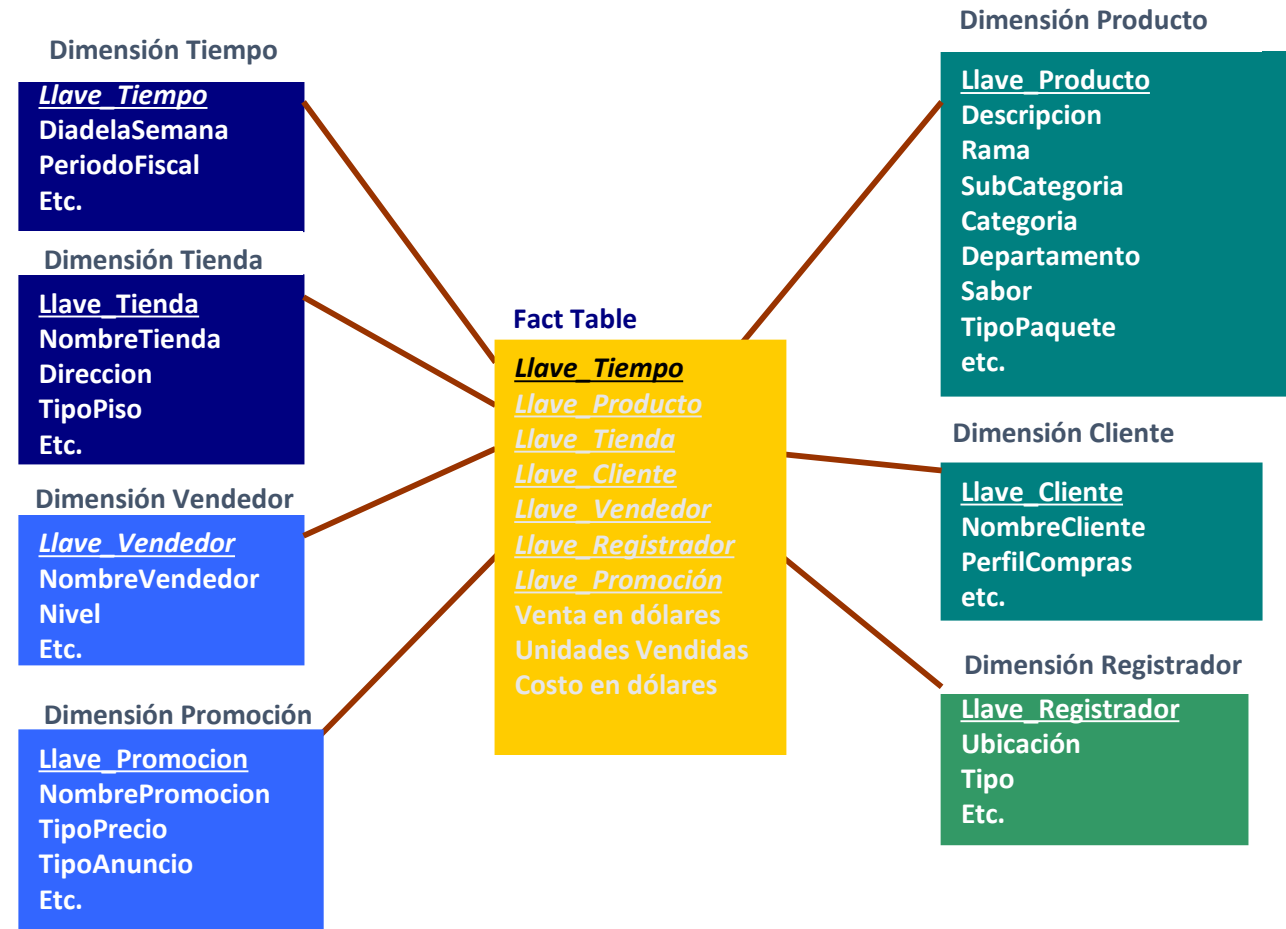
En el modelo dimensional se cuenta con

Hechos
Medidas
Dimensiones
Atributos
Relaciones



¿Cómo hacerlo?

1. Seleccionar el proceso de negocio a modelar
2. Definir el nivel de granularidad del proceso de negocio
3. Escoger las dimensiones que aplican en cada fila de la tabla de hechos.
4. Identificar los hechos numéricos que poblaran la tabla de hechos.

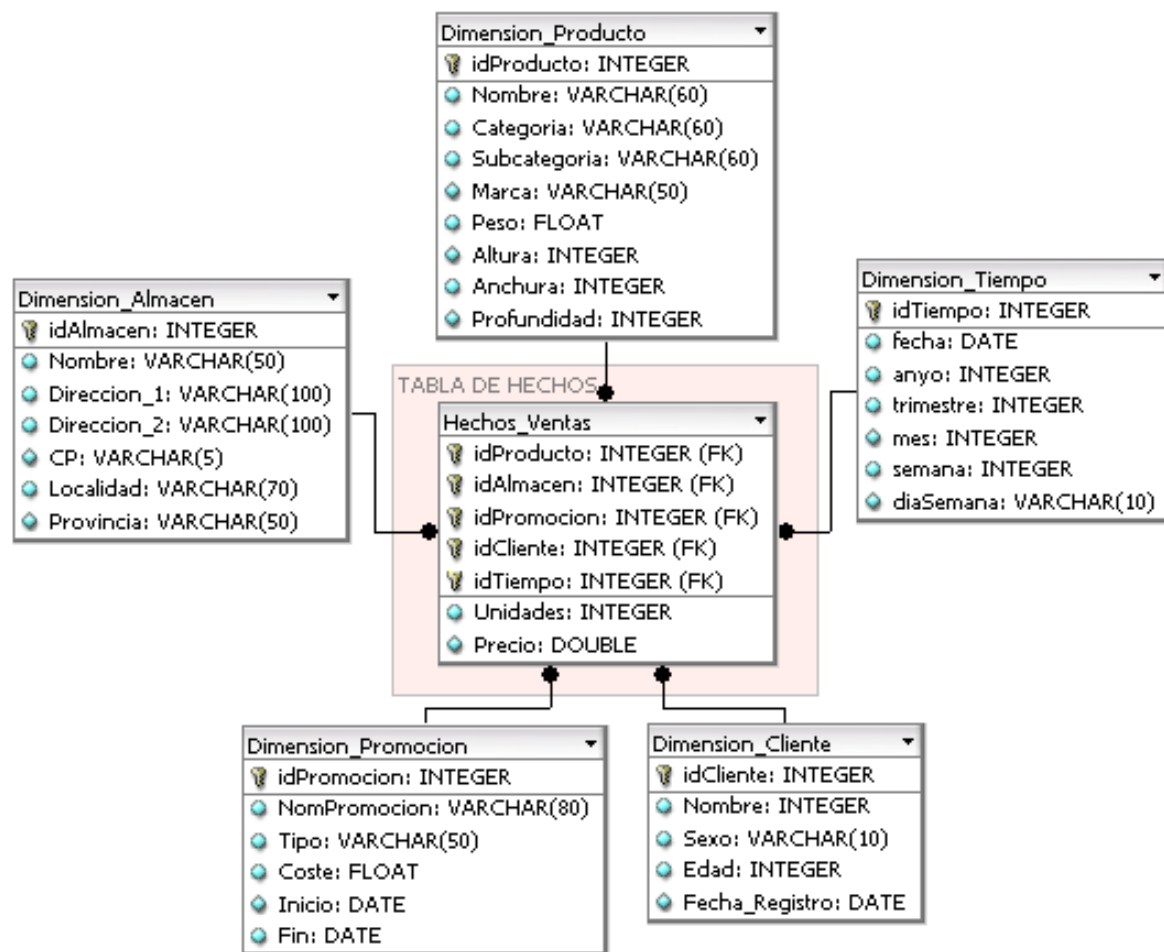


Diferencias

E-R	Dimensional
Enfocado a la actualización: Menos redundancia, coordinar actualizaciones y repetir el mismo tipo de operaciones muchas veces en el día.	Enfocado a consulta
Altamente normalizadas para soportar actualizaciones consistentes y mantenimiento de la integridad referencial	Altamente desnormalizada puesto que se requiere disminución de tiempos en la obtención de grandes cantidades de datos.
Tiempos de respuesta en segundos o inferior	Tiempos de respuesta aceptables pueden ser segundos, minutos, horas
Almacenan pocos datos derivados	Gran cantidad de datos derivados (redundancia)
Pocos datos agregados	Agregación: Varios niveles de datos precalculados

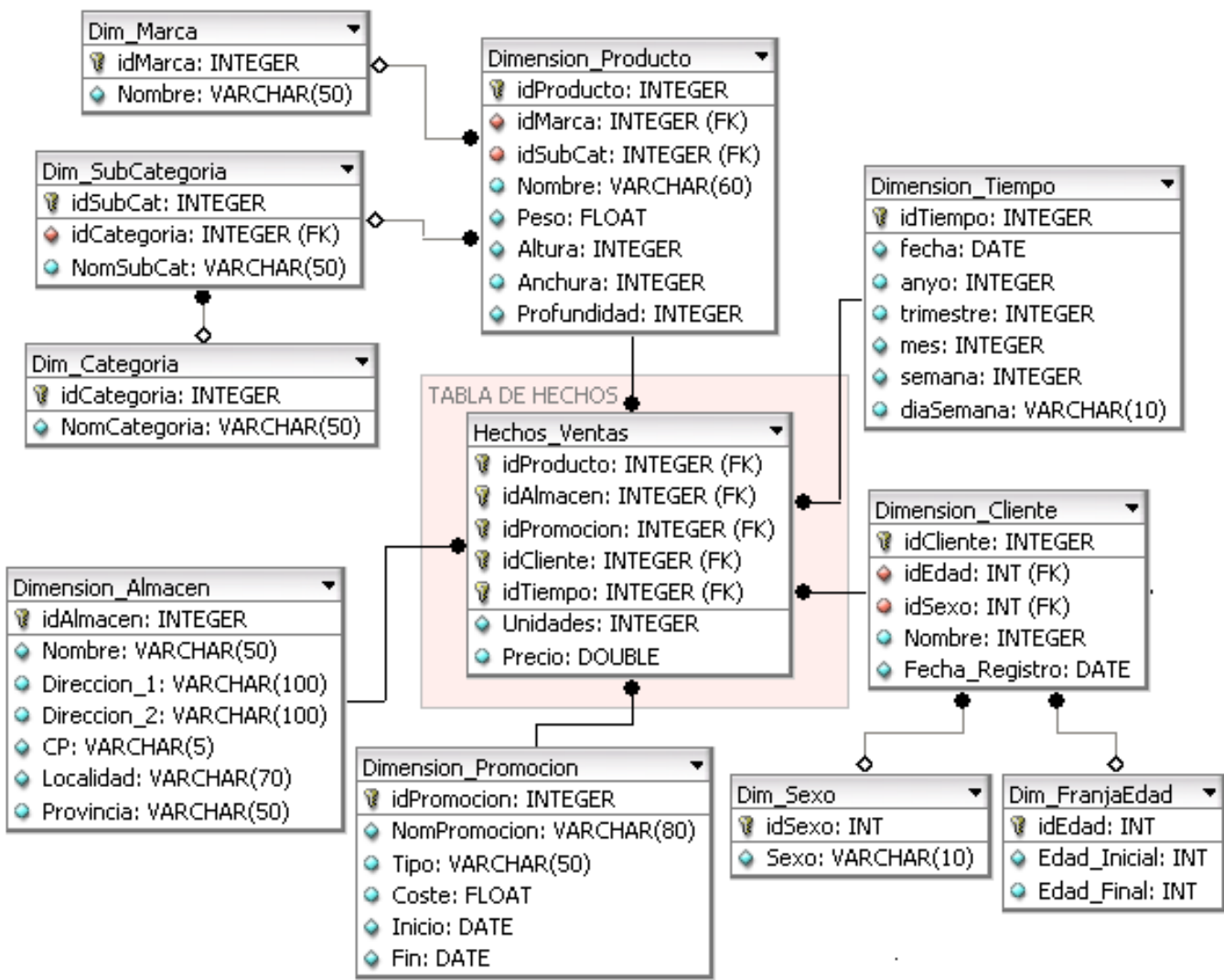
Esquema estrella

- Representación más importante del Modelo Dimensional.
- Todo objeto de análisis es un hecho.
- Los hechos son analizados a través de las dimensiones.
- Los hechos contienen columnas llamadas métricas y las dimensiones contienen columnas llamadas Niveles jerárquicos.
- Las tablas de dimensiones tendrán siempre una clave primaria simple, mientras que en la tabla de hechos, la clave principal estará compuesta por las claves principales de las tablas dimensionales.



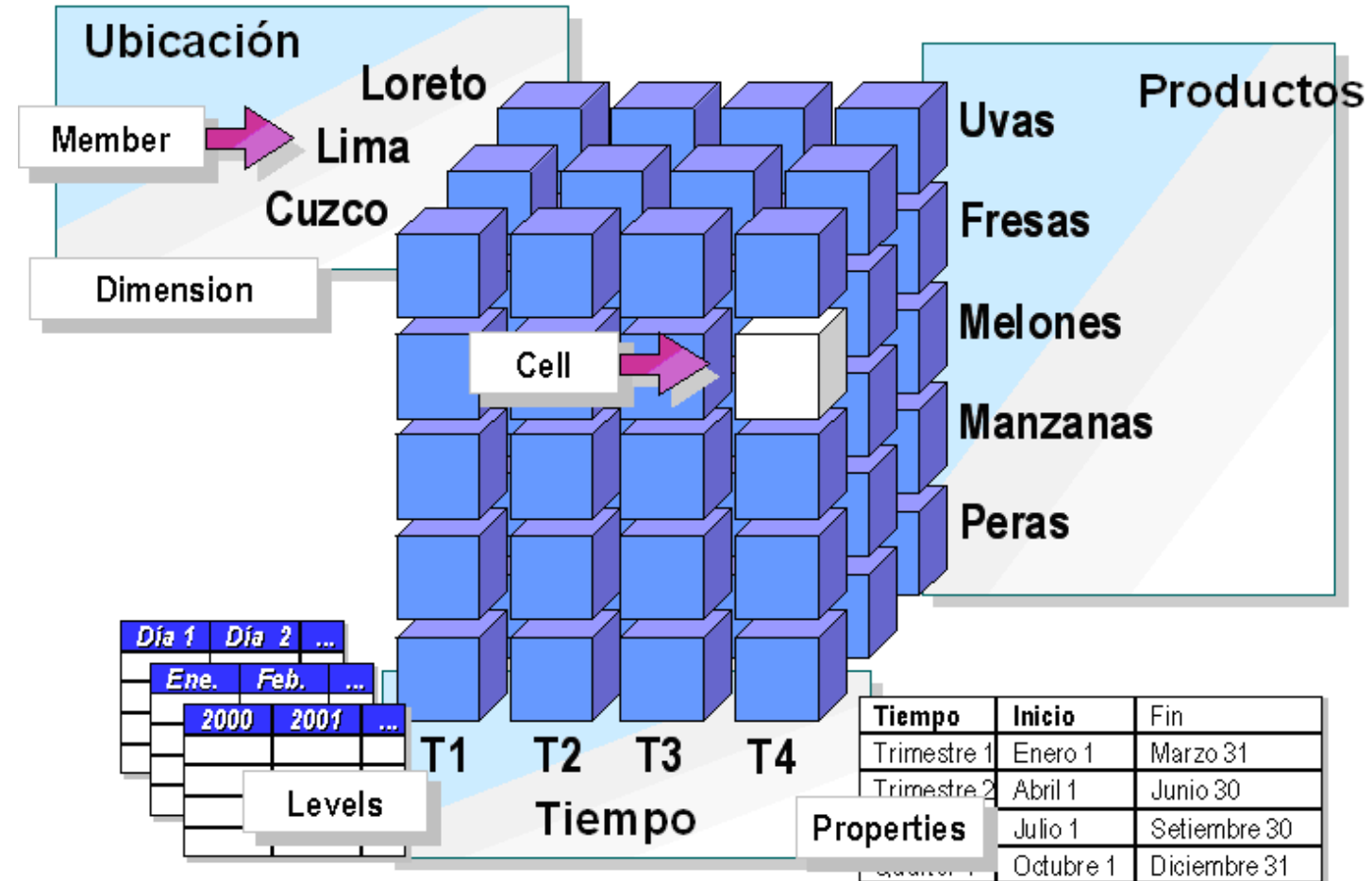
Esquema copo de nieve

- Esquema derivado del esquema estrella.
- Tablas de dimensión normalizadas en múltiples tablas, diferencia principal con el esquema estrella.
- Necesitamos un esquema estrella para obtener uno.
- Existen 2 tipos, el completo y el parcial.

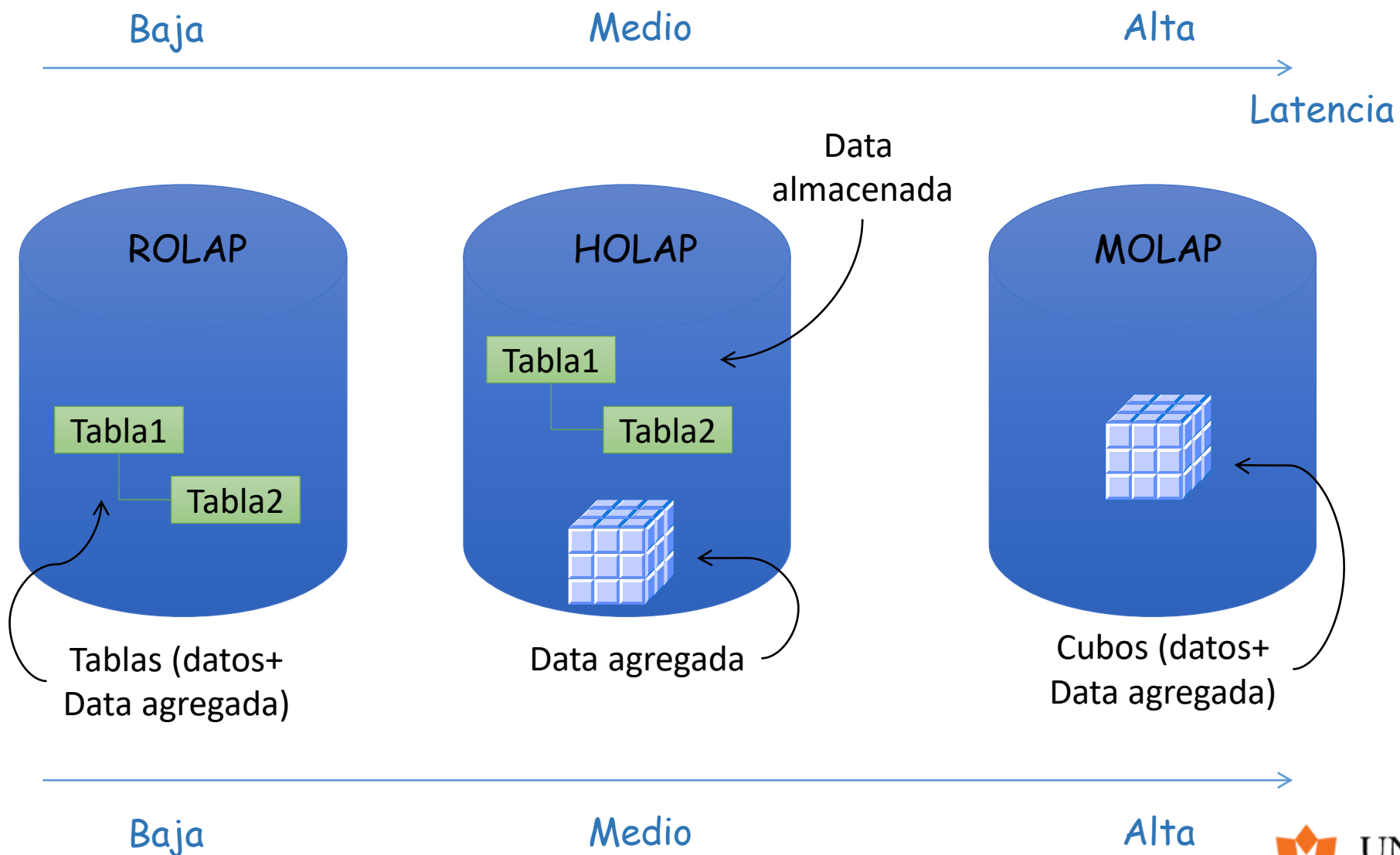


- Colección de dimensiones y métricas relacionadas. Contiene información en una estructura multidimensional
- Se puede tener 3 , 4, 10 o más dimensiones siempre se llamara "cubo"

Componentes de un Cubo



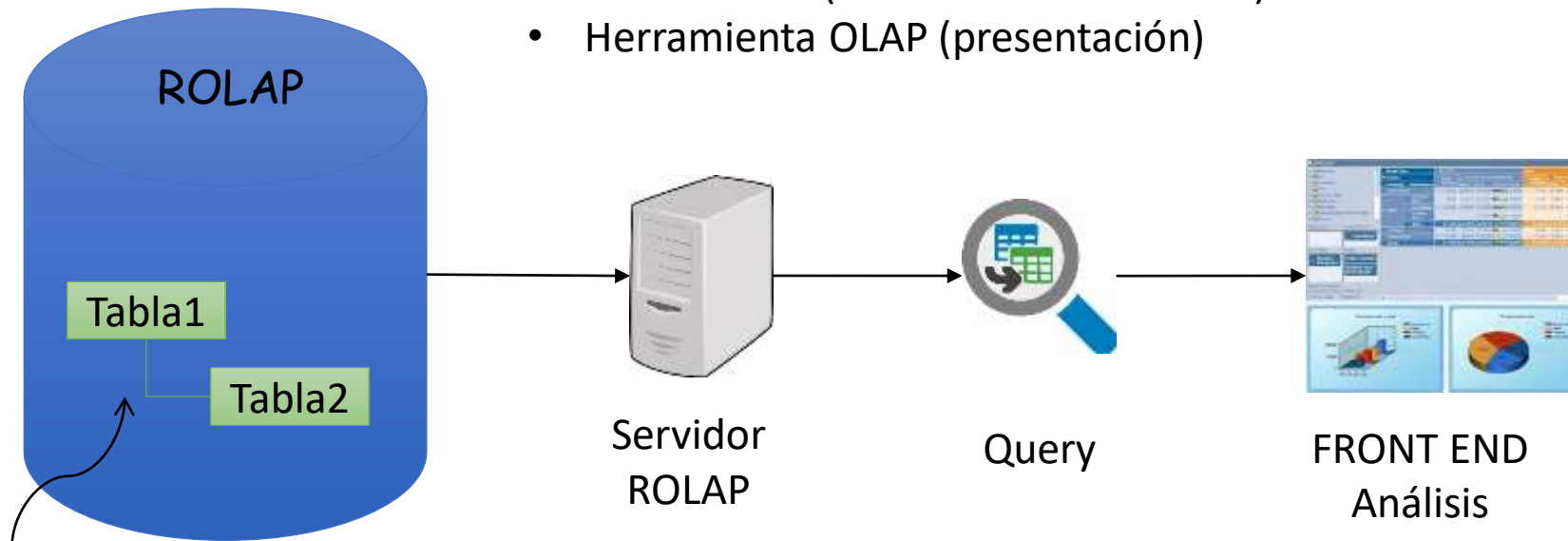
Tipos de OLAP



Tipos de OLAP

Arquitectura de tres niveles.

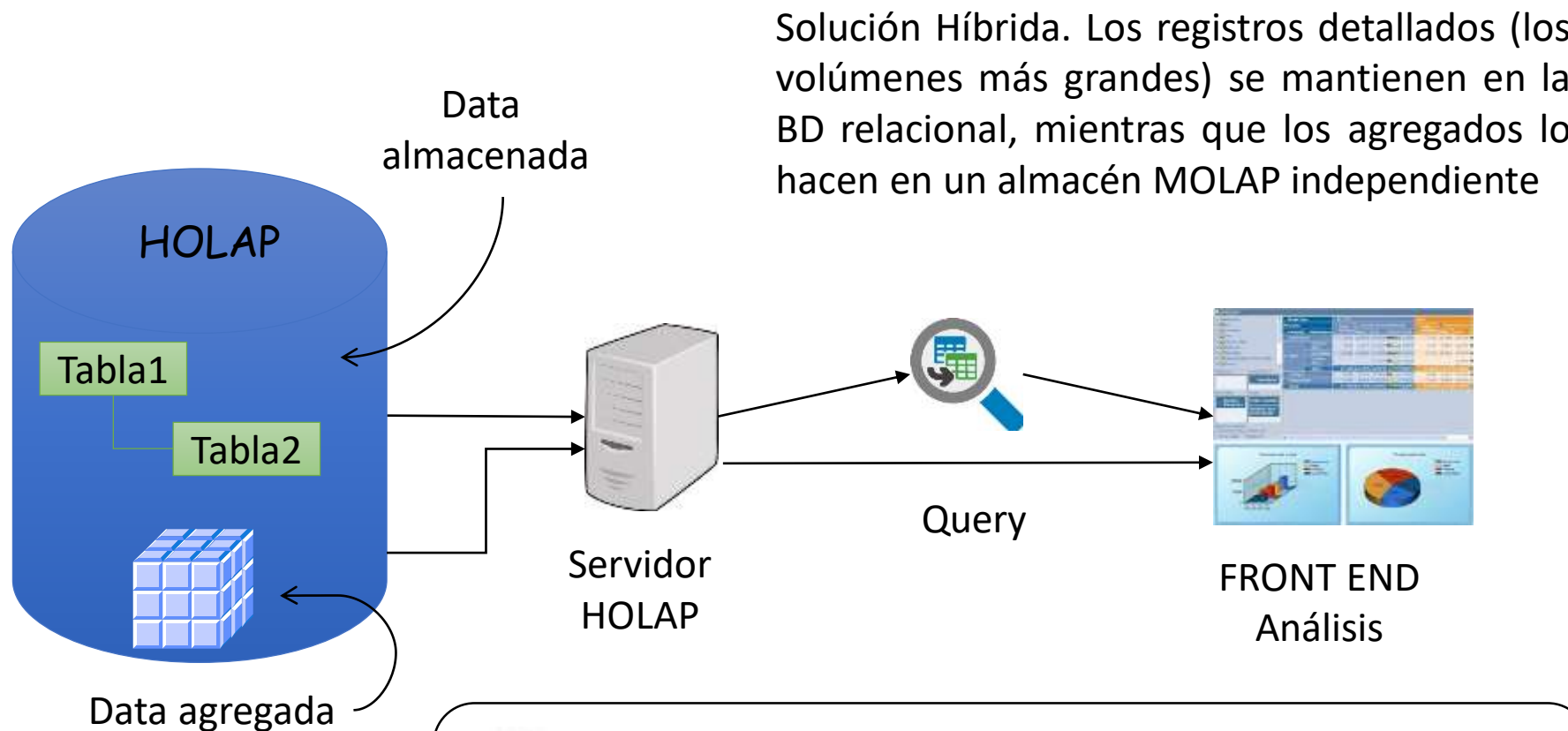
- BD relacional (almacenamiento de datos)
- Motor OLAP (funcionalidad analítica)
- Herramienta OLAP (presentación)



Es capaz de usar datos precalculados (si estos están disponibles), o de generar dinámicamente los resultados desde la información elemental (menos resumida).

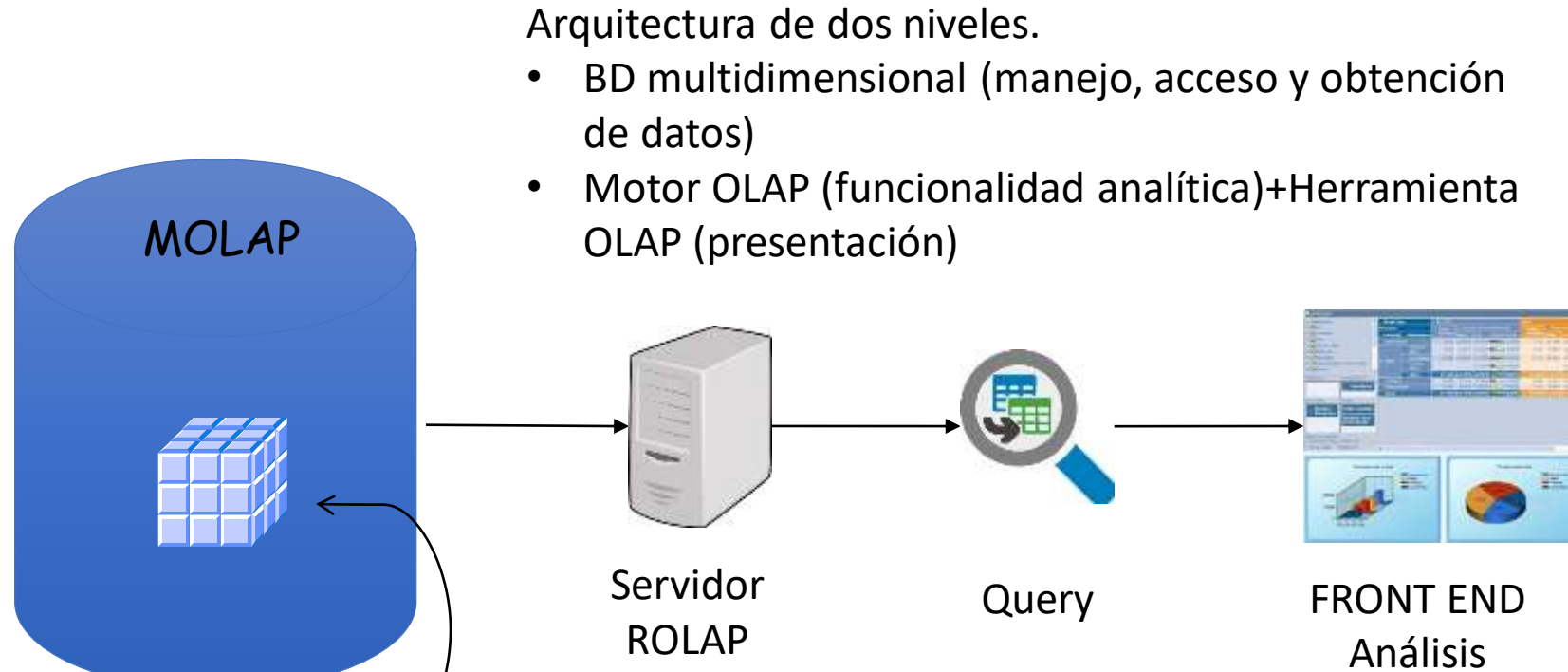


Tipos de OLAP



Resuelve el problema de dispersión, dejando los datos menos agregados en la BD relacional, pero almacena los agregados en un formato multidimensional, minimizando la presencia de celdas vacías.

Tipos de OLAP



Una de las características distintivas de MOLAP es la preconsolidación de los datos. En una BDMD, estos totales se calculan rápidamente usando operaciones sobre arreglos. Una vez calculados, los totales se pueden almacenar en estructuras de la misma BDMD.

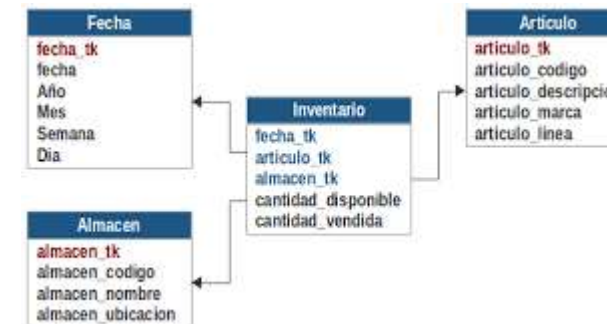
Comparando Estructuras de almacenamiento

Almacenamiento	MOLAP	HOLAP	ROLAP
Data Base	<i>Cubo</i>	<i>Tabla Relacional</i>	<i>Tabla Relacional</i>
Agregación	<i>Cubo</i>	<i>Cubo</i>	<i>Tabla Relacional</i>

Perspectiva del Cliente	MOLAP	HOLAP	ROLAP
Rendimiento de Consultas	<i>Rapidísimo</i>	<i>Más Rápido</i>	<i>Rápido</i>
Almacenamiento	<i>Alto</i>	<i>Medio</i>	<i>Bajo</i>
Mantenimiento del Cubo	<i>Alto</i>	<i>Medio</i>	<i>Bajo</i>

Fundamentos del Modelado Dimensional

- Es una adaptación del modelo relacional.
- Consiste de **tablas de hechos** que se caracterizan usando **dimensiones** y **medidas**.
- La información sobre un **hecho** (actividad) se representa mediante **indicadores** (medidas o atributos de hecho).
- La información de cada **dimensión** se representa por un conjunto de atributos (atributos de dimensión).
- Una **dimensión** en el contexto de un **hecho**, tienden a ser discretas y jerárquicas.
- Un **indicador** es una cantidad que describe el **hecho**, debe ser agregables.

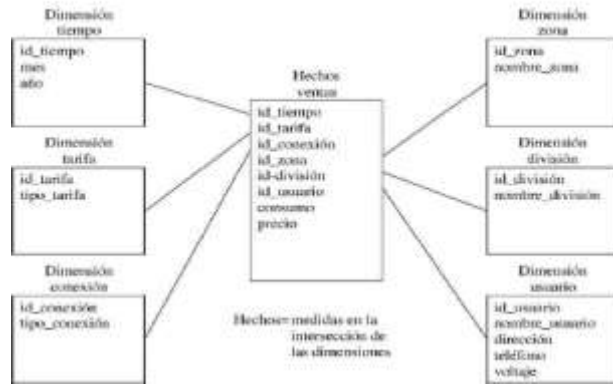


Conceptos Básicos

1. **Hecho**. Evento, actividad, ítem transacción del negocio.
2. **Medida**. métrica de hechos, métricas del negocio
3. **Dimensión**. Característica de un hecho.
4. **Jerarquía**. Relaciones padre-hijo dentro de una dimensión. Son las estructuras lógicas que utilizan niveles pedidos como los medios de ordenamiento de datos.
5. **Tabla de hechos**: Almacena eventos y las métricas. Estas son las tablas centrales en un esquema estrella de un modelo DW. Las tablas fact representan el conocimiento del negocio y sus datos generalmente son numéricos y/o añadidos para ser analizados.
6. **Tabla de dimensión**. También conocidas como Tablas de la Referencia, contienen los datos relativamente estáticos en el DW. Una dimensión es una estructura, integrada a menudo por unas o más jerarquías, que categoriza datos.

Hechos

Ejemplo de hechos y dimensiones en un modelo multidimensional.



- Representan un evento o actividad específica, tiene **dimensiones** y **medidas**.
- Representan un ítem de negocio, una transacción o un evento que tiene significancia para el negocio.
- Corresponden a una colección de ítems de datos y datos de contexto.
- Son aquellos datos que residen en una tabla de **hechos** y que son utilizados para crear **indicadores**, a través de sumalizaciones preestablecidas.
- Un **hecho** debe estar relacionado al menos con una dimensión: "El tiempo".

Medidas – Métricas

- Es un atributo numérico de un hecho que representa la performance o comportamiento del negocio relativo a la dimensión
- Ejemplos:
 - Ventas en \$\$
 - Cantidad de productos
 - Total de transacciones
 - Cantidad de pacientes admitidos
 - Llamadas efectuadas.
 - $\text{ImporteTotal} = \text{precioProducto} * \text{cantidadVendida}$
 - $\text{Rentabilidad} = \text{utilidad} / \text{PN}$
 - $\text{CantidadVentas} = \text{cantidad}$
 - $\text{PromedioGeneral} = \text{AVG}(\text{notasFinales})$



Medidas-Métricas

- Representan los valores que son analizados.
- Características de las medidas:
 - Deben ser numéricas. Porque estos valores son las bases de las cuales el usuario puede realizar cálculos.
 - Cruzan todas las dimensiones en todos los niveles.
- Si la medida es no numérica debemos codificarla a un valor numérico y cuando tengamos que exponerla decodificarla para mostrarla con el valor original.

$$\begin{aligned} \frac{1}{3} &= 0.3333333333333333\ldots \\ 1 &= 0.9999999999999999\ldots \\ \sqrt{2} &= 1.4142135623730950\ldots \\ \sqrt{3} &= 1.732050807568877293\ldots \\ e &= 2.718281828459045235\ldots \\ \pi &= 3.141592653589793238\ldots \end{aligned}$$

Medidas-Metricas

- Las medidas pueden clasificarse en:
Naturales.

- Son aquellas que se obtiene por agregación de los datos originales.
 - Suma: suma los valores de las columnas
 - Cuenta: conteo de los valores
 - Mínima: valor mínimo
 - Máxima: valor máximo
 - Cuenta de Distintos: valores diferentes

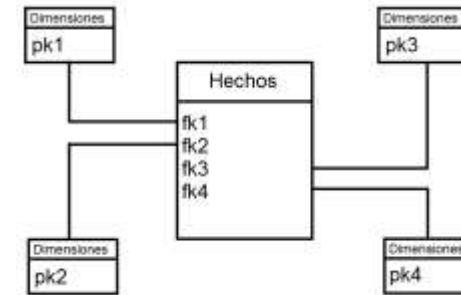
Calculadas

- Si se derivan de una medida natural.
 - Cálculos Matemáticos
 - Expresiones condicionales
 - Alertas



Dimensiones

- Es una característica de un **hecho** que permite su análisis posterior, en el proceso de toma de decisiones.
- Determina el contexto del **hecho** (quién participó, cuándo y donde pasó y su tipo).
- Es una entidad de negocios respecto de la cual se deben calcular las **métricas** (clientes, productos, tiempo)
- Tienden a ser discretas y jerárquicas <país, región, departamento, provincia, distrito>.
- Es una colección de miembros o unidades o individuos del mismo tipo que permite categorizar un **hecho**.



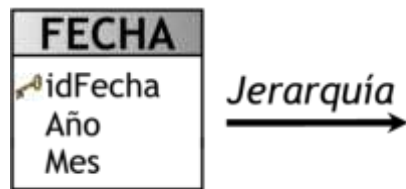
Dimensiones

- Se utilizan como parámetros para los análisis OLAP
- Las dimensiones habituales son:

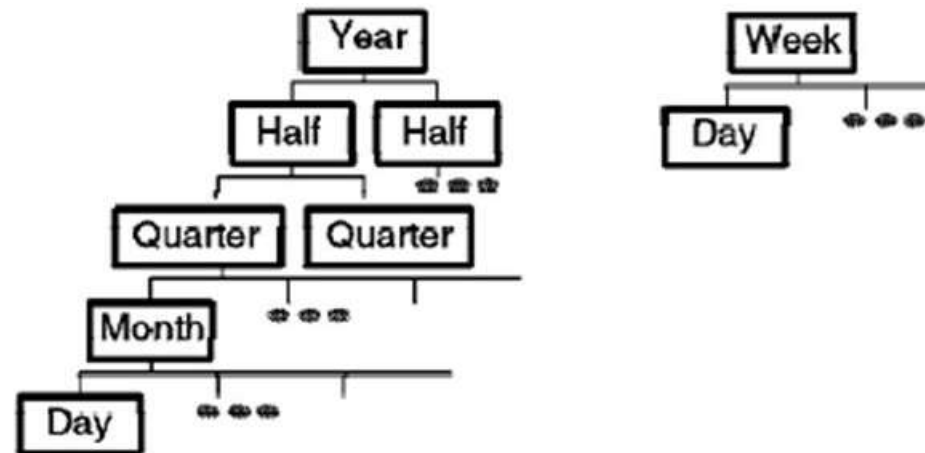
Dimensión	Miembro
Tiempo	Meses, Trimestre, Años
Geografía	País, Región, Ciudad
Cliente	Id Cliente
Vendedor	Id Vendedor

Jerarquía de las dimensiones

- Una **jerarquía** representa una relación lógica entre los datos de una dimensión.
- Estos datos poseen una relación “padre-hijo”.

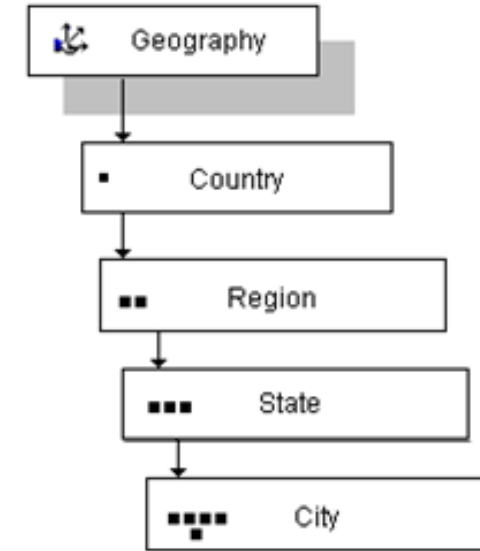


Año
↑
Mes



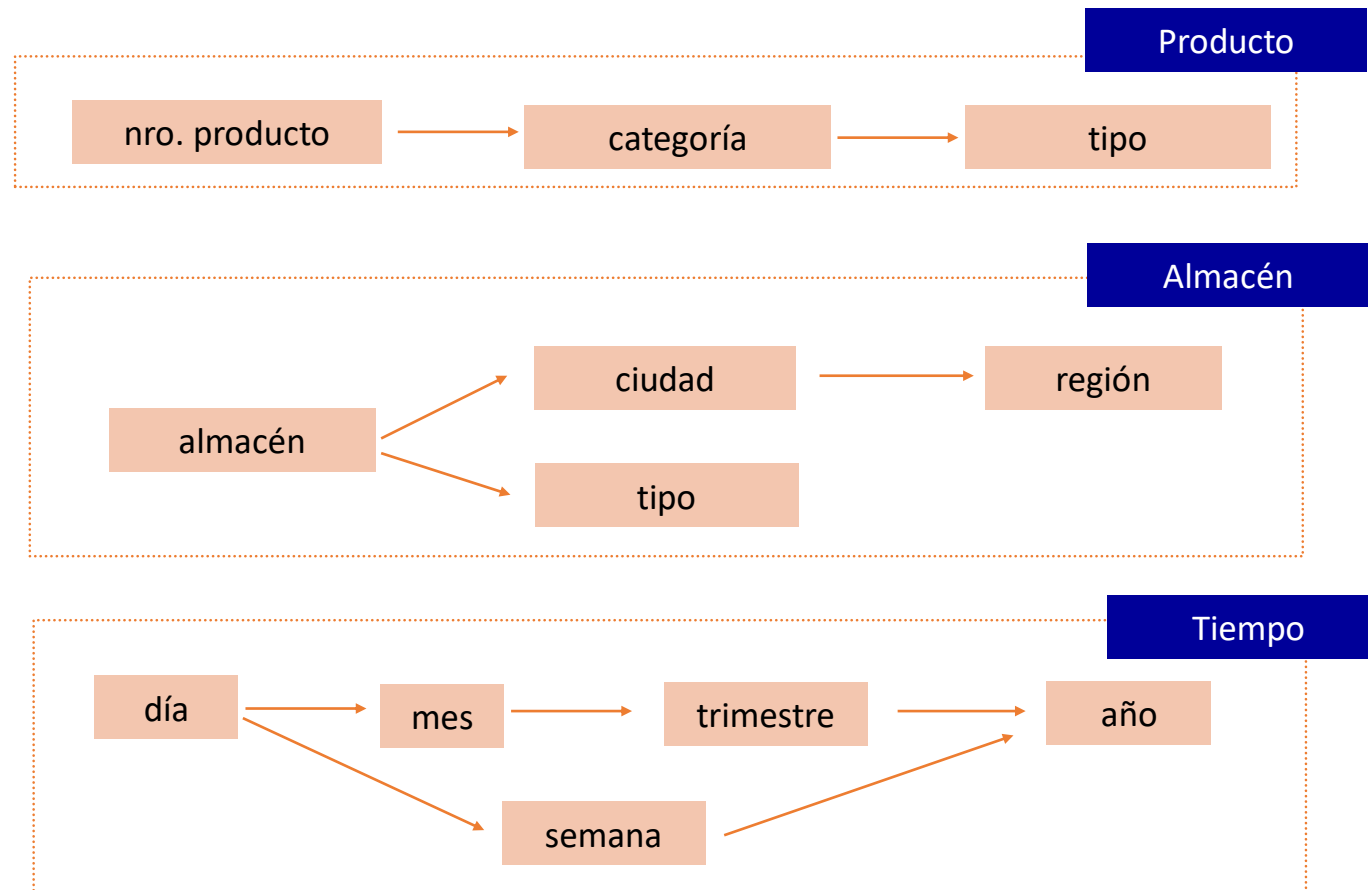
Jerarquía de las dimensiones

- Tienen las siguientes características:
 - Se presentan al interior de una dimensión.
 - Pueden existir varios niveles (dos o más)
 - Relación “1-n” o “padre-hijo” entre atributos consecutivos de un nivel superior y uno inferior.
- Se pueden identificar cuando existen relaciones “1-n” o “padre-hijo” en la dimensión.



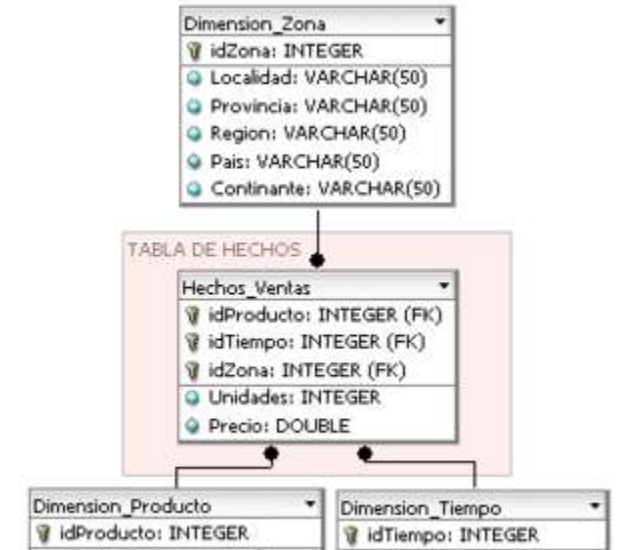
Origen de las Jerarquías

- Entre los atributos de una dimensión se definen **jerarquías**.



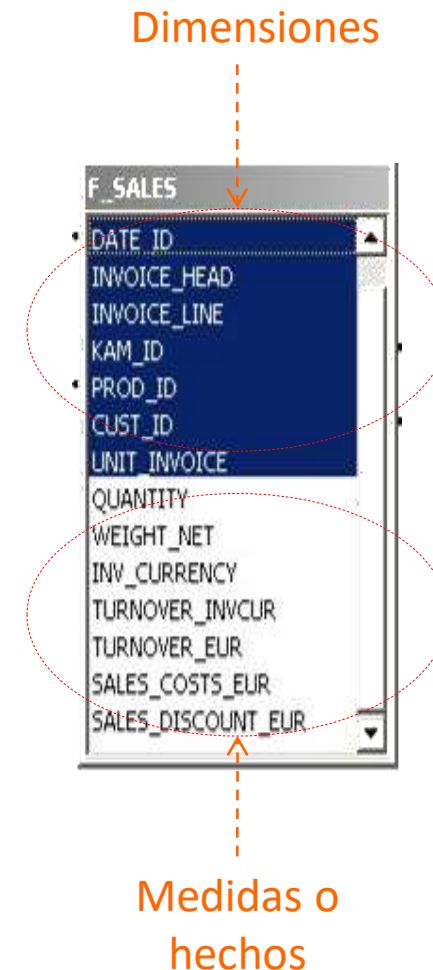
Granularidad

- La granularidad es el nivel de detalle en que se almacena la información.
- Por ejemplo:
 - Datos de ventas o compras de una empresa, pueden registrarse día a día
 - Datos pertinentes a pagos de sueldos o cuotas de socios, podrán almacenarse a nivel de mes.
- A mayor nivel de detalle, mayor posibilidad analítica, ya que los mismos podrán ser resumidos o sumariados.
- Los datos con granularidad fina (nivel de detalle) podrán ser resumidos hasta obtener una granularidad media o gruesa. No sucede lo mismo en sentido contrario.







Tablas de Hechos

- Las tablas de hechos contienen las dimensiones y las medidas de los hechos.
- Los hechos o medidas son los valores de datos que se analizan (son métricos).
- La tabla de hechos tiene una clave primaria compuesta por las claves primarias de las tablas de dimensiones relacionadas a este.



Tablas de Dimensiones

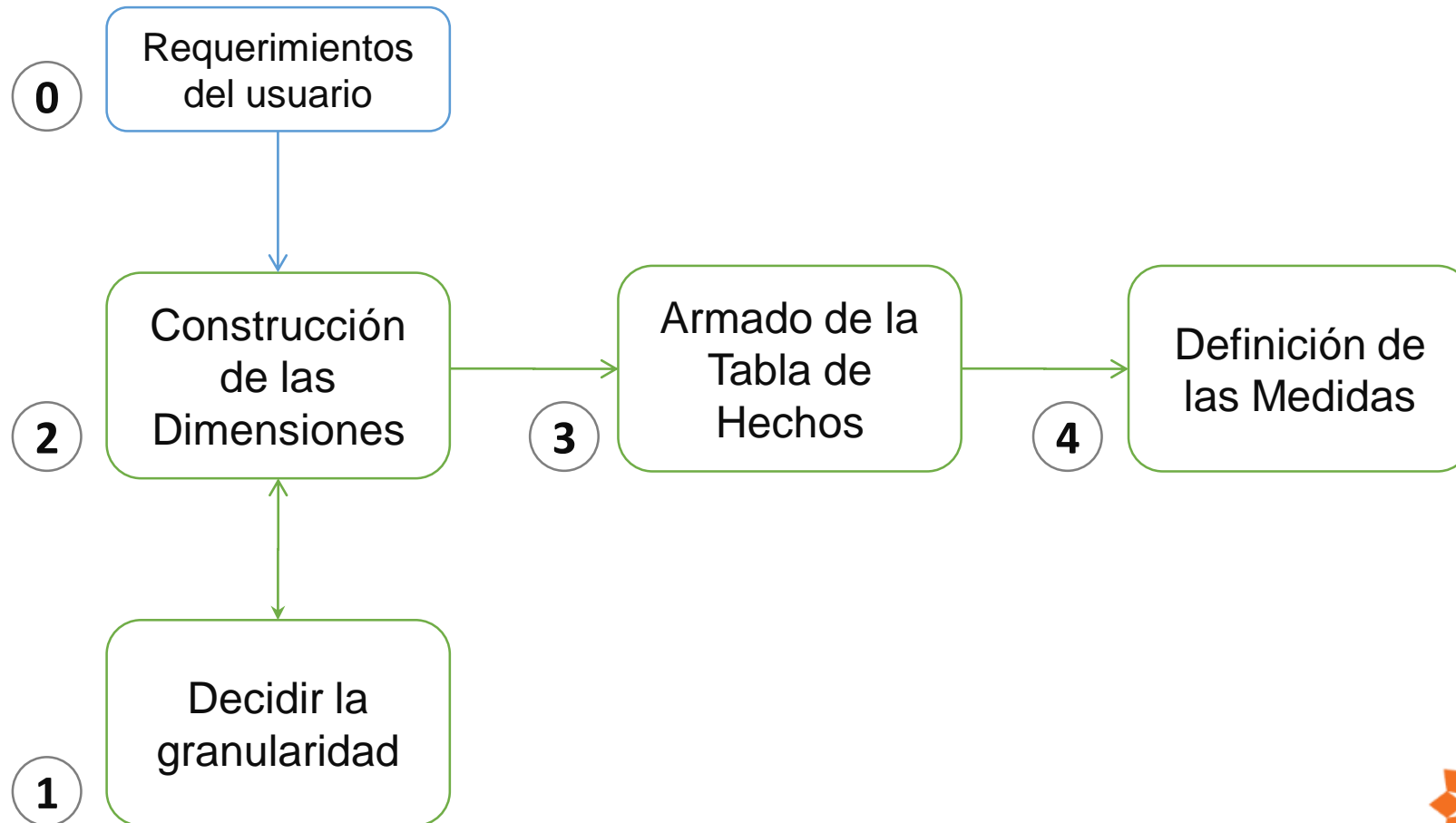
- Definen la organización lógica de los datos.

GEOGRAFIA	PRODUCTOS	CLIENTES	FECHAS
 id_Geografía País Provincia Ciudad Barrio	 id_Producto Rubro Tipo NombreProducto	 id_Cliente NombreCliente	 id_Fecha Año Trimestre Mes Día

- Tiene una PK (única) y columnas de referencia:
 - Clave principal (PK) o identificador único.
 - Clave foráneas.
 - Datos de referencia primarios (identifican la dimensión)
 - Datos de referencia secundarios (complementan la descripción).
- No siempre la PK del OLTP, corresponde con la PK de la tabla de dimensión relacionada.

Ejercicio

- Etapas en la construcción de un modelo dimensional:



Requerimientos del usuario

	Dimensiones				
Medidas	Tiempo	Sucursal	Vendedor	Cliente	Producto
Ventas_Importe	X	X	X	X	X
Ventas_Costo	X	X	X	X	X
Ventas_Unidades	X	X	X	X	X
Ventas_ImporteTotal	X	X	X	X	X
Ventas_Ganancia	X	X	X	X	X
Ventas_Promedio	X	X	X	X	X

Decidir la granularidad

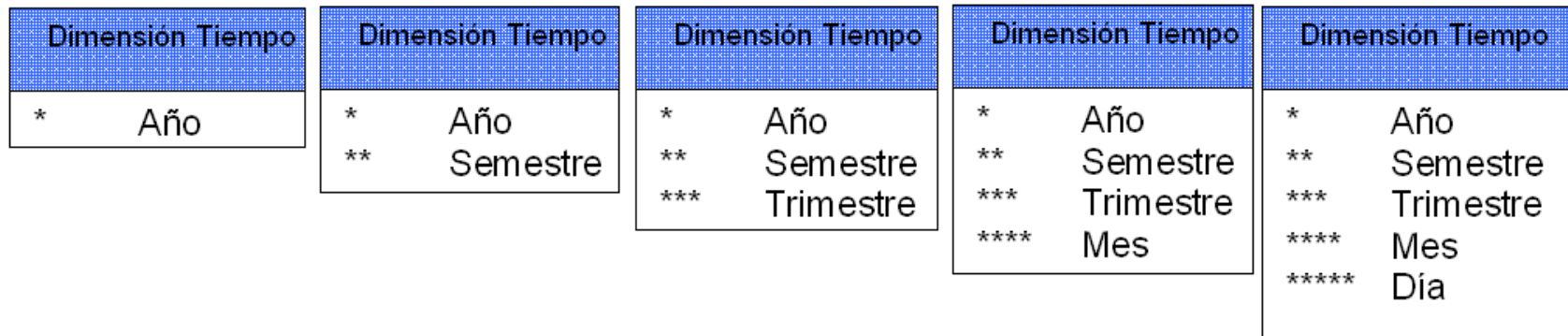
- La granularidad:
 - Es el nivel de detalle al que se desea almacenar información sobre la actividad a modelar.
 - Define el nivel atómico de datos en el almacén de datos.
 - Determina el significado de las tuplas de la tabla de hechos.
 - Determina las dimensiones básicas del esquema.
- Por ejemplo en la dimensión Sucursal:

Dimensión Sucursal	Dimensión Sucursal	Dimensión Sucursal	Dimensión Sucursal
* ** Sucursal Tipo Sucursal	* ** *** Sucursal Tipo Sucursal País	* ** *** **** Sucursal Tipo Sucursal País Provincia	* ** *** **** ***** Sucursal Tipo Sucursal País Provincia Ciudad

Decidir la granularidad

- Ejemplo de la dimensión fecha. Se desea los datos por:
 - Información anual
 - Información semestral
 - Información trimestral
 - Información mensual.
 - Información semanal
 - Información diaria
 - Transacción en el OLTP

+ granularidad
+ detalle



Construcción de las dimensiones

- Identificar las dimensiones que caracterizan el proceso al nivel de detalle (gránulo) que se ha elegido.
- De cada dimensión se debe decidir los atributos (propiedades) relevantes para el análisis de la actividad.
- Entre los atributos de una dimensión existen jerarquías naturales que deben ser identificadas (día-mes-año)
 - Tiempo. Cuándo se produce la actividad
 - Sucursal. Donde está ubicado el almacén
 - Vendedor. Quién ha vendido
 - Cliente. Quién es el destinatario de la actividad
 - Producto. Cuál es el objeto de la actividad

Dimensión Tiempo	Dimensión Sucursal	Dimensión Vendedor	Dimensión Cliente
* Año ** Semestre *** Trimestre **** Mes ***** Día	* Sucursal ** Tipo Sucursal *** País **** Provincia ***** Ciudad	* Sucursal ** Sección *** Vendedor	* País ** Provincia *** Ciudad **** Razón Social

	Dimensiones				
Medidas	Tiempo	Sucursal	Vendedor	Cliente	Producto
Ventas_Importe	X	X	X	X	X
Ventas_Costo	X	X	X	X	X
Ventas_Unidades	X	X	X	X	X
Ventas_ImporteTotal	X	X	X	X	X
Ventas_Ganancia	X	X	X	X	X
Ventas_Promedio	X	X	X	X	X



Fact_Ventas
ID_Tiempo ID_Producto ID_Cliente ID_Vendedor ID_Sucursal



	Dimensiones				
Medidas	Tiempo	Sucursal	Vendedor	Cliente	Producto
Ventas_Importe	X	X	X	X	X
Ventas_Costo	X	X	X	X	X
Ventas_Unidades	X	X	X	X	X
Ventas_ImporteTotal	X	X	X	X	X
Ventas_Ganancia	X	X	X	X	X
Ventas_Promedio	X	X	X	X	X



Definición de las medidas

