

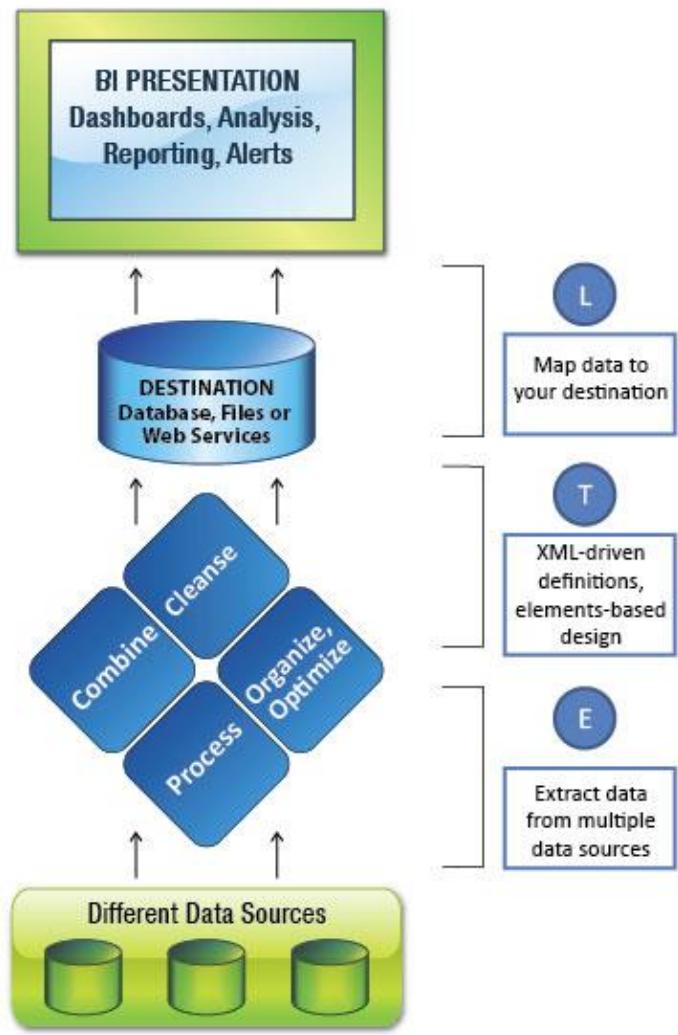


DATAWAREHOUSE Y MINERIA DE DATOS

**CLASE 3
CICLO 2-2020**



1 ETL





FUNCIONALIDAD DE UN ETL

- ◆ **Control de la extracción de los datos y su automatización;** disminuyendo el tiempo empleado en el descubrimiento de procesos no documentados, minimizando el margen de error y permitiendo mayor flexibilidad.
- ◆ **Acceso a diferentes tecnologías,** haciendo un uso efectivo del hardware, software, datos y recursos humanos existentes.



FUNCIONALIDAD DE UN ETL

- ◆ Proporcionar la gestión integrada del **Data Warehouse** y los **Data Marts existentes**, integrando la extracción, transformación y carga para la construcción del Data Warehouse corporativo y de los Data Marts.
- ◆ Uso de la arquitectura de **metadatos**, facilitando la definición de los objetos de negocio y las reglas de consolidación.
 - ◆ Acceso a una gran variedad de fuentes de datos diferentes.
 - ◆ Manejo de excepciones.



FUNCIONALIDAD DE UN ETL

- ◆ Planificación, logs, interfaces a schedulers (**planificadores**) de **terceros**, que nos permitirán llevar una gestión de la planificación de todos los procesos necesarios para la carga del DW.
- ◆ Interfaz independiente de hardware.
- ◆ Soporte en la explotación del Data Warehouse.



DEFINICIÓN



ETL son las siglas en inglés de **Extraer, Transformar y Cargar** (*Extract, Transform and Load*). Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra **base de datos, data mart, o data warehouse** para analizar, o en otro sistema operacional para apoyar un proceso de negocio.



Los procesos **ETL** también se pueden utilizar para la integración con **sistemas heredados** (aplicaciones antiguas existentes en las organizaciones que se han de integrar con los nuevos aplicativos, por ejemplo, ERP's. La tecnología utilizada en dichas aplicaciones puede hacer difícil la integración con los nuevos programas).

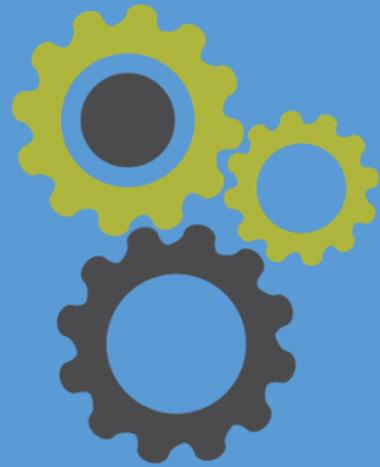
EXTRACCION



SQL SERVER ORACLE MySQL



TRANSFORMACION



CARGA





EXTRAER

- ◆ La primera parte del proceso ETL consiste en **extraer los datos** provenientes de **diferentes sistemas de origen**.
- ◆ La extracción convierte los datos a un **formato** preparado para iniciar el proceso de transformación.

EXTRACCION



EXTRAER

- ◆ Una parte intrínseca del proceso de extracción es la de analizar los datos extraídos (si los datos no cumplen la pauta o estructura esperada, estos datos podrían ser rechazados.)
- ◆ Un **requerimiento** importante que se debe exigir a la tarea de extracción es que ésta cause un impacto mínimo en el sistema origen.

EXTRACCION

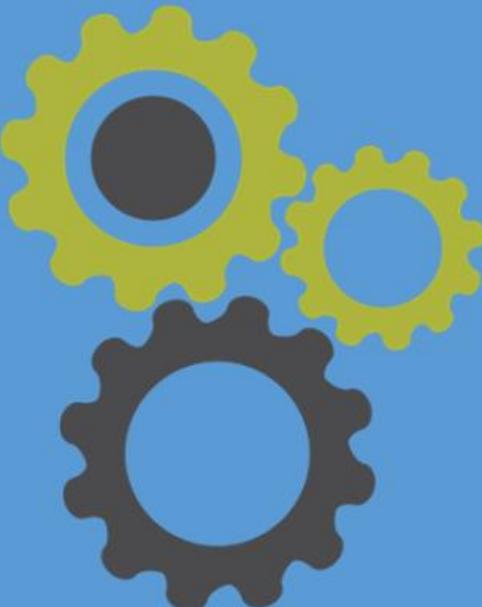




TRANSFORMAR

- ◆ La fase de transformación aplica una serie de **reglas de negocio** o funciones sobre los datos extraídos para convertirlos en datos que serán cargados.
- ◆ Algunas fuentes de datos requerirán alguna pequeña manipulación de los datos.

TRANSFORMACION





TRANSFORMAR

- ◆ No obstante en otros casos pueden ser necesarias aplicar algunas de las siguientes transformaciones:
 - ◆ Seleccionar sólo ciertas columnas para su carga (por ejemplo, que las columnas con valores nulos no se carguen).
 - ◆ Traducir códigos (por ejemplo, si la fuente almacena una "H" para Hombre y "M" para Mujer pero el destino tiene que guardar "1" para Hombre y "2" para Mujer).
 - ◆ Codificar valores libres (por ejemplo, convertir "Hombre" en "H" o "Sr" en "1").
 - ◆ Obtener nuevos valores calculados (por ejemplo, total_venta = cantidad * precio).
 - ◆ Unir datos de múltiples fuentes (por ejemplo, búsquedas, combinaciones, etc.).



TRANSFORMAR

- ◊ Calcular totales de múltiples filas de datos (por ejemplo, ventas totales de cada región).
- ◊ Generación de campos clave en el destino.
- ◊ Transponer o pivotar (girando múltiples columnas en filas o viceversa).
- ◊ Dividir una columna en varias (por ejemplo, columna "Nombre: García, Miguel"; pasar a dos columnas "Nombre: Miguel" y "Apellido: García").
- ◊ La aplicación de cualquier forma, simple o compleja, de validación de datos



CARGAR

- ◆ La fase de carga es el momento en el cual los datos de la fase anterior (**transformación**) son cargados en el sistema de destino.
- ◆ Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes.
- ◆ En algunas bases de datos se sobrescribe la información antigua con nuevos datos. Los **data warehouse** mantienen un historial de los registros de manera que se pueda hacer una auditoría de los mismos.

CARGA





DESAFÍOS

- ◆ **Los procesos ETL pueden ser muy complejos.** Un sistema ETL mal diseñado puede provocar importantes problemas operativos.
- ◆ **En un sistema operacional el rango de valores de los datos o la calidad de éstos pueden no coincidir con las expectativas** de los diseñadores a la hora de especificarse las reglas de validación o transformación.



DESAFÍOS

- ◆ El proceso ETL es clave para lograr que los datos extraídos asíncronamente de orígenes heterogéneos se integren finalmente en un entorno homogéneo.
- ◆ La **escalabilidad** de un sistema de ETL durante su **vida útil** tiene que ser establecida durante el análisis.
- ◆ El tiempo disponible para realizar la extracción de los sistemas de origen podría cambiar, lo que implicaría que la misma cantidad de datos tendría que ser procesada en menos tiempo.



HERRAMIENTAS

Algunas Herramientas ETL

- ◆ Ab Initio
- ◆ Benetl
- ◆ BITool – ETL Software
- ◆ CloverETL
- ◆ Cognos Decisionstream (IBM)
- ◆ Data Integrator (herramienta de Sap Business Objects)
- ◆ ETI*Extract (ahora llamada Eti Solution)
- ◆ IBM Websphere DataStage (antes Ascential DataStage)
- ◆ Microsoft Integration Services
- ◆ Oracle Warehouse Builder
- ◆ WebFocus-iWay DataMigrator Server
- ◆ Pervasive
- ◆ Informática PowerCenter

Oxio Data Intelligence ETL full.web

- ◆ SmartDB Workbench
- ◆ Sunopsis (Oracle)
- ◆ SAS Dataflux
- ◆ Sybase
- ◆ Syncsort: DMExpress.
- ◆ Opentext (antes Genio, Hummingbird).

Libres

- ◆ Kettle (ahora llamado Pentaho Data Integration).
- ◆ Scriptella Open Source ETL Tool.
- ◆ Talend Open Studio.
- ◆ Jitterbit.



PROPÓSITO DE UN ETL

- ◆ Las herramientas ETL no se tienen porque utilizar solo en entornos de construcción de un DW, sino que pueden ser útiles para multitud de propósitos, como por ejemplo:
 - ◊ **Tareas de Bases de datos:** También se utilizan para consolidar, migrar y sincronizar bases de datos operativas.
 - ◊ **Migración de datos entre diferentes aplicaciones** por cambios de versión o cambio de aplicaciones.
 - ◊ **Sincronización entre diferentes sistemas operacionales** (por ejemplo, nuestro entorno ERP y la Web de ventas).



PROPÓSITO DE UN ETL

- ◆ **Consolidación de datos:** sistemas con grandes volúmenes de datos que son consolidados en sistemas paralelos para mantener históricos o para procesos de borrado en los sistemas originales.
- ◆ **Interfaces de datos con sistemas externos:** envío de información a clientes, proveedores. Recepción, proceso e integración de la información recibida.
- ◆ **Interfaces con sistemas Frontoffice:** interfaces de subida/bajada con sistemas de venta.
- ◆ **Otros cometidos:** Actualización de usuarios a sistemas paralelos, preparación de procesos masivos (mailings, newsletter), etc.

2

MODELADO DE DATOS MULTIDIMENSIONAL



BASE DE DATOS MULTIDIMENSIONAL

- ◆ Las bases de datos multidimensionales (BDMB) son un tipo de base de datos optimizada para Data Warehouse que se utilizan principalmente para crear **aplicaciones OLAP**, una tecnología asociada al acceso y análisis de datos en línea.
- ◆ Las Bases de Datos Multidimensionales facilitan el manejo de grandes cantidades de datos dentro de empresas, dándole a esto una amplia aplicación dentro de varias áreas y diferentes campos del conocimiento humano.



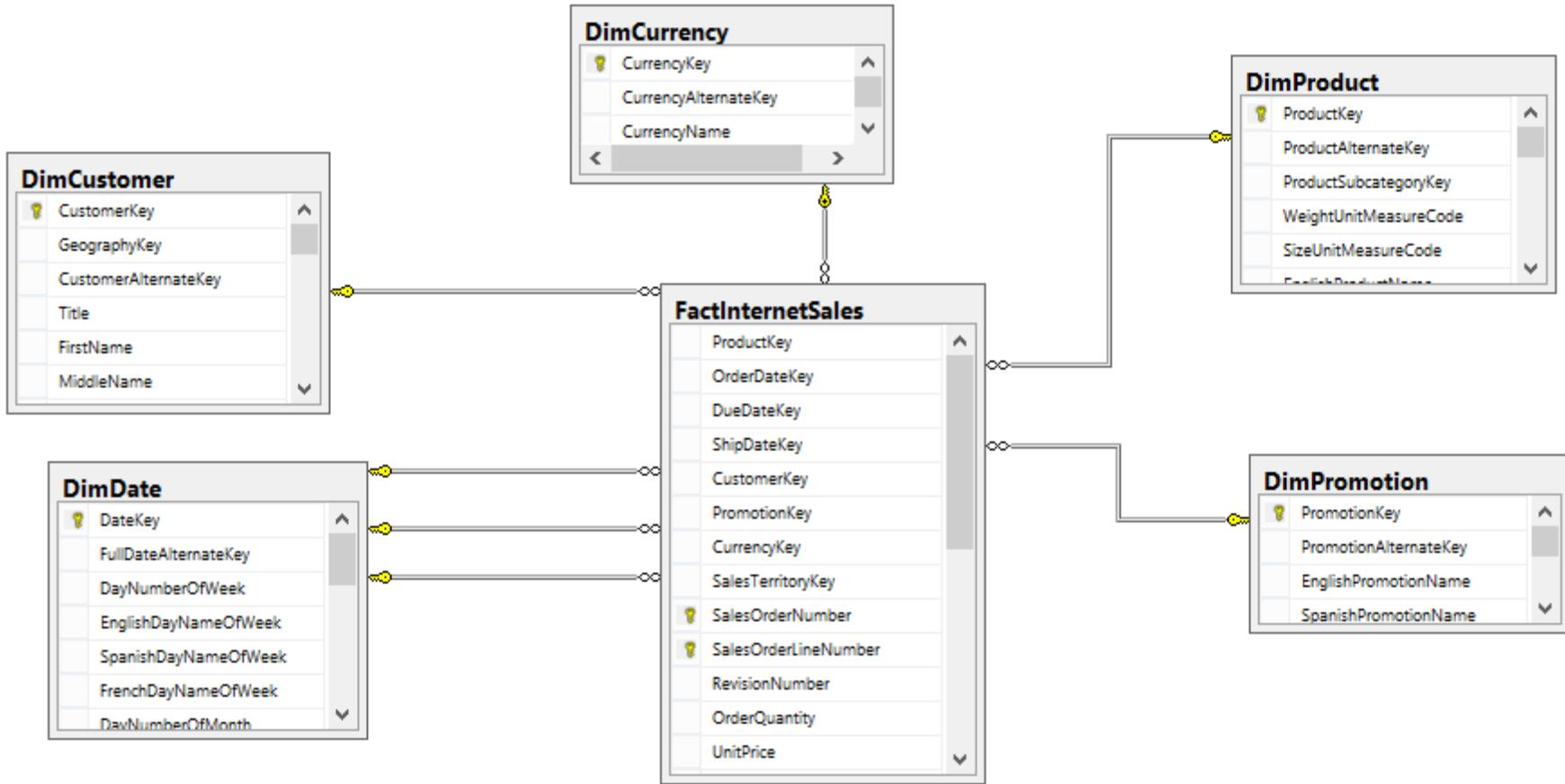
BASE DE DATOS MULTIDIMENSIONAL

- ◆ Adecuado para resumir y organizar datos.
- ◆ Enfocado para trabajar sobre datos de tipo numérico.
- ◆ Más simple: Más fácil de visualizar y entender.
- ◆ El modelamiento dimensional es una técnica para modelar bases de datos simples y entendibles al usuario final, la idea fundamental es que el usuario visualice fácilmente la relación que existe entre los distintos componentes del modelo.



BASE DE DATOS MULTIDIMENSIONAL

- ◆ Su información se almacena en forma multidimensional, es decir, a través de **tablas de hechos y tablas de dimensiones**.
- ◆ Cada dimensión contiene una llave primaria que se “conecta” a la tabla de hechos manteniendo una relación de 1 a muchos.





TABLAS DIMENSIONES Y HECHOS

- ◆ Para la construcción de un **modelo dimensional**, debemos tener en cuenta un conjunto de técnicas y conceptos para diseñar nuestros almacenes de datos.
- ◆ Una parte fundamental de estos, son los tipos de tablas donde guardamos la información, destacamos **las tablas de hechos** (aquellos que queremos medir o analizar) y **las tablas de dimensiones** (cómo lo queremos medir).



TABLA DIMENSIONES

- ◆ Las tablas de dimensiones definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio.
- ◆ Contienen datos cualitativos.
- ◆ Representan los aspectos de interés mediante los cuales el usuario puede filtrar y manipular la información almacenada en la tabla hechos



TABLA DIMENSIONES - CARACTERÍSTICAS

- ◆ Las tablas de dimensión (del inglés dimension table) son:
 - ◊ Tablas simples desnormalizadas
 - ◊ Se unen a las tablas de hechos a través de un campo clave
 - ◊ Los atributos de la tabla de dimensión ofrecen información característica de las tablas de hechos
 - ◊ No hay límite de tablas de dimensión
 - ◊ Las dimensiones pueden contener una o varias relaciones jerárquicas
 - ◊ Normalmente tiene pocos (miles) registros
- ◆ Por ejemplo: **clientes, productos, almacenes, proveedores, calendario...**



TABLA DIMENSIONES - CARACTERÍSTICAS

- ◆ Mas detalladamente, cada tabla de dimensión podrá contener los siguientes campos:
 - ◆ Clave principal o identificador único
 - ◆ Clave foránea
 - ◆ Datos de referencia primarios (ejemplo nombre de un cliente)
 - ◆ Datos de referencia secundarios (datos que complementan al cliente)



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

Como se puede observar cada tabla posee un identificador único y al menos un campo o dato de análisis relevantes para la organización, son por lo general de tipo texto.



JERARQUÍAS

- ◆ Una jerarquía representa una relación lógica entre dos o más atributos dentro de una misma dimensión.
- ◆ Las jerarquías poseen las siguientes características:
 - ◊ Pueden existir varias en una misma dimensión.
 - ◊ Están compuestas por dos o más niveles.
 - ◊ Se tiene una relación de n-1 (padre e hijo)

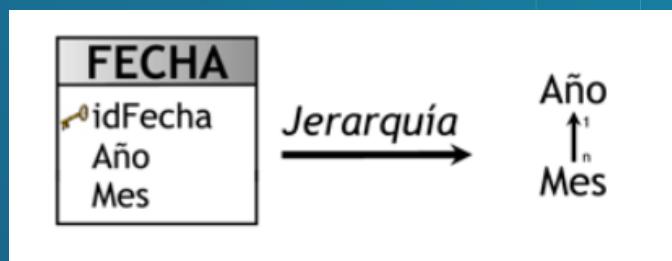




TABLA DE DIMENSIÓN TIEMPO

- En un DW, la creación y el mantenimiento de una tabla de **dimensión Tiempo es obligatoria**, y la definición de granularidad y estructuración de la misma depende de la dinámica del negocio que se este analizando.
- Toda la información dentro de la base de datos, posee su propio sello de tiempo que determina la ocurrencia de un hecho específico, representando de esta manera diferentes versiones de una misma situación.

FECHAS	
➔	id_Fecha
	Año
	Trimestre
	Mes
	Día





TABLA DE DIMENSIÓN TIEMPO

- ◆ Es importante tener en cuenta que la dimensión tiempo no es sola una secuencia cronológica representada de forma numérica, sino que mantiene niveles jerárquicos especiales que inciden notablemente en las actividades de la organización.
- ◆ Esto se debe a que los usuarios podrán por ejemplo analizar las ventas realizadas teniendo en cuenta el día de la semana en que se produjeron, quincena, mes, trimestre, semestre, año, estación, etc.





TABLA HECHOS

- ◆ La Tabla de hechos contiene precisamente los hechos que serán utilizados por los analistas del negocio para apoyar la toma de decisiones.
- ◆ Contiene datos cuantitativos.
- ◆ Los hechos son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de las dimensiones.

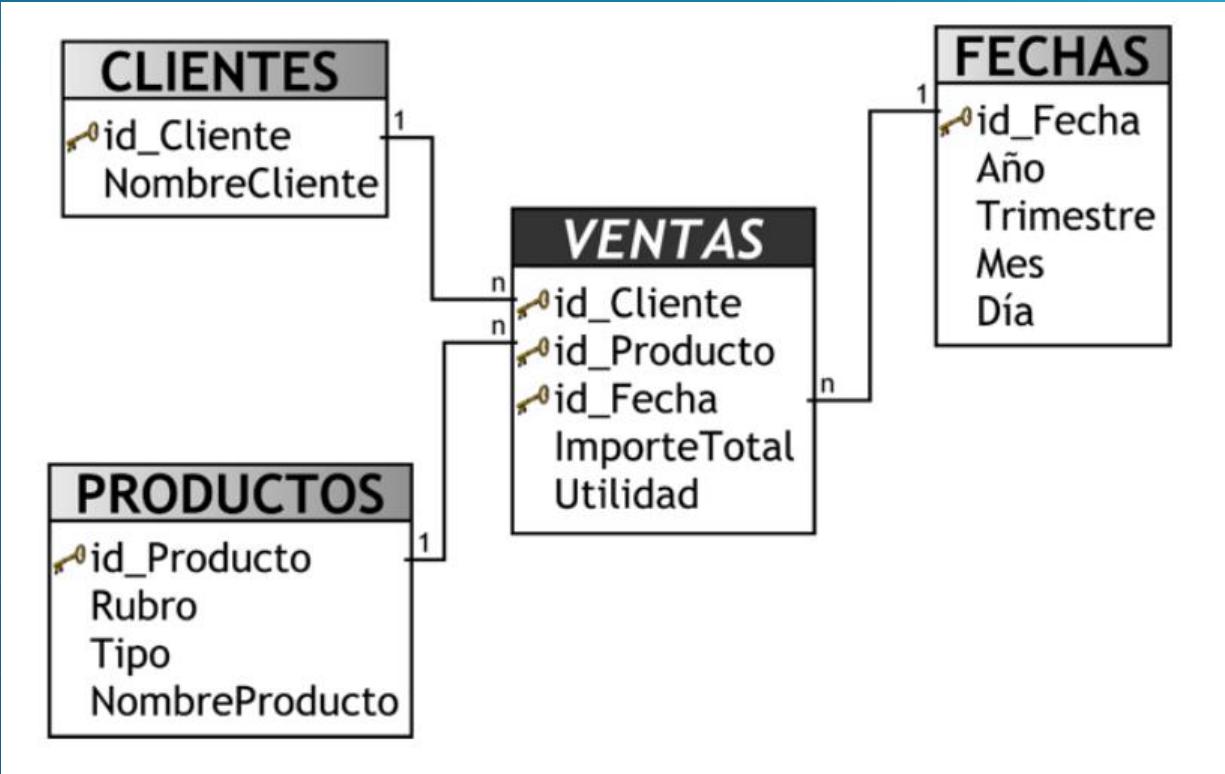




TABLA HECHOS - CARACTERÍSTICAS

- ◆ Las tablas de hechos (del inglés fact tables) son:
 - ◆ La tabla principal del modelo dimensional
 - ◆ Contienen campos claves que se unen a las tablas de dimensión
 - ◆ Contiene métricas o también llamadas medidas y es aquello que queremos medir o analizar. Generalmente son valores numéricos que se suelen agregar
 - ◆ Evitan la redundancia de atributos
 - ◆ Normalmente tienen muchos (millones) registros
- ◆ **Por ejemplo: ventas, compras, movimientos de contabilidad**







HECHOS

- ◆ Los hechos son aquellos datos que residen en una tabla de hechos y que son utilizados para crear indicadores, a través de summarizaciones preestablecidas al momento de crear un cubo multidimensional, Business Model, etc.
- ◆ Debido a que una tabla de hechos se encuentra interrelacionada con sus respectivas tablas de dimensiones, permite que los hechos puedan ser accedidos, filtrados y explorados por los valores de los campos de estas tablas de dimensiones, obteniendo de este modo una gran capacidad analítica.





TIPOS DE HECHOS

HECHOS	
id_Dimensión1	PK
id_Dimensión2	PK
id_DimensiónN	PK
precio	
cantidad	
total	

- ◆ **Hechos básicos:** son los que se encuentras representados por un campo en la tabla hechos, los campos "precio" y "cantidad" de la tabla son hechos básicos.
- ◆ **Hechos derivados:** se forma al combinar uno o mas hechos, a través de consultas SQL sencillas.
Ejemplo $\text{total} = \text{precio} * \text{cantidad}$

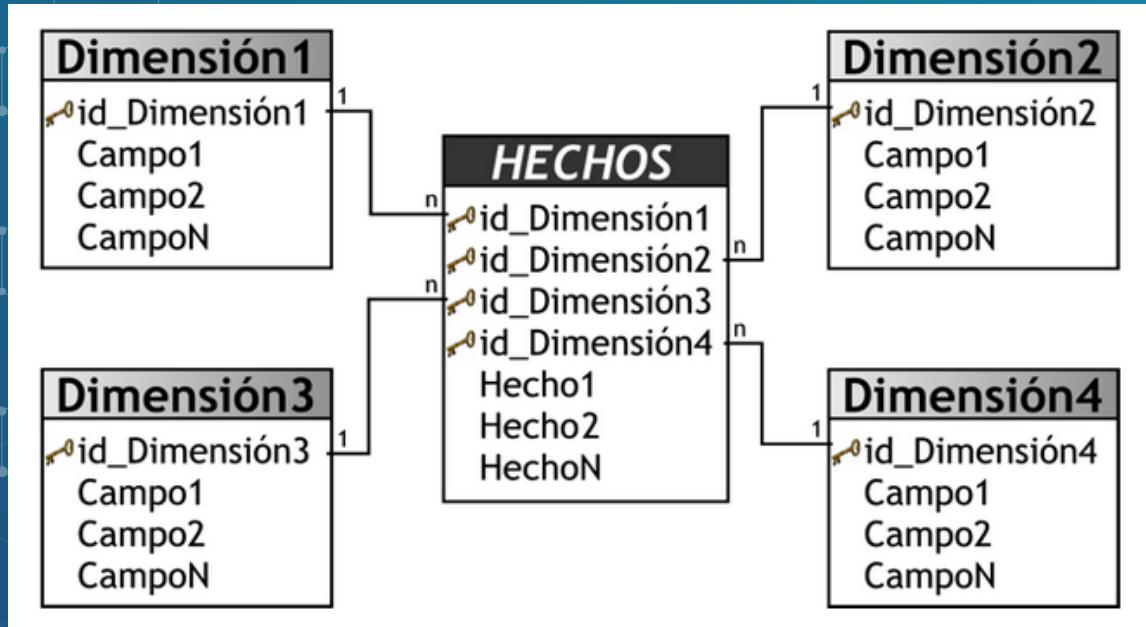


TIPOS DE MODELOS

- ◆ Las bases de datos multidimensionales implican tres variantes posibles de modelamiento, que permiten realizar consultas de soporte de decisión:
 - ◆ **Esquema en estrella.**
 - ◆ **Esquema copo de nieve.**
 - ◆ **Esquema constelación.**

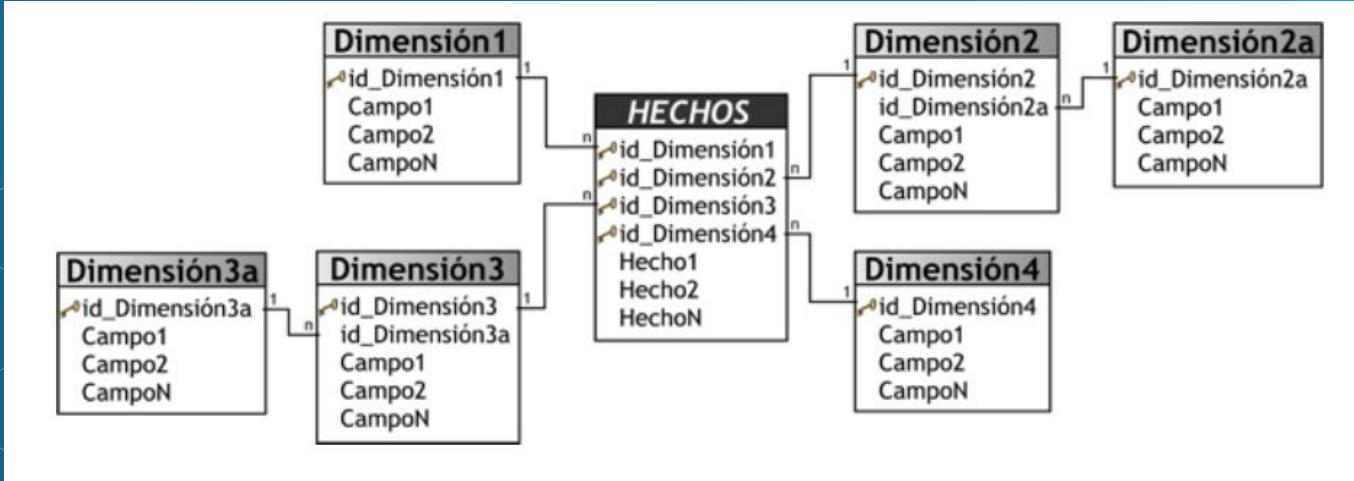


ESQUEMA DE ESTRELLA



El modelo o esquema de estrella es el más sencillo en estructura. Consta de una tabla central de "Hechos" y varias "dimensiones", incluida una dimensión de "Tiempo". Lo característico de la arquitectura de estrella es que sólo existe una tabla de dimensiones para cada dimensión

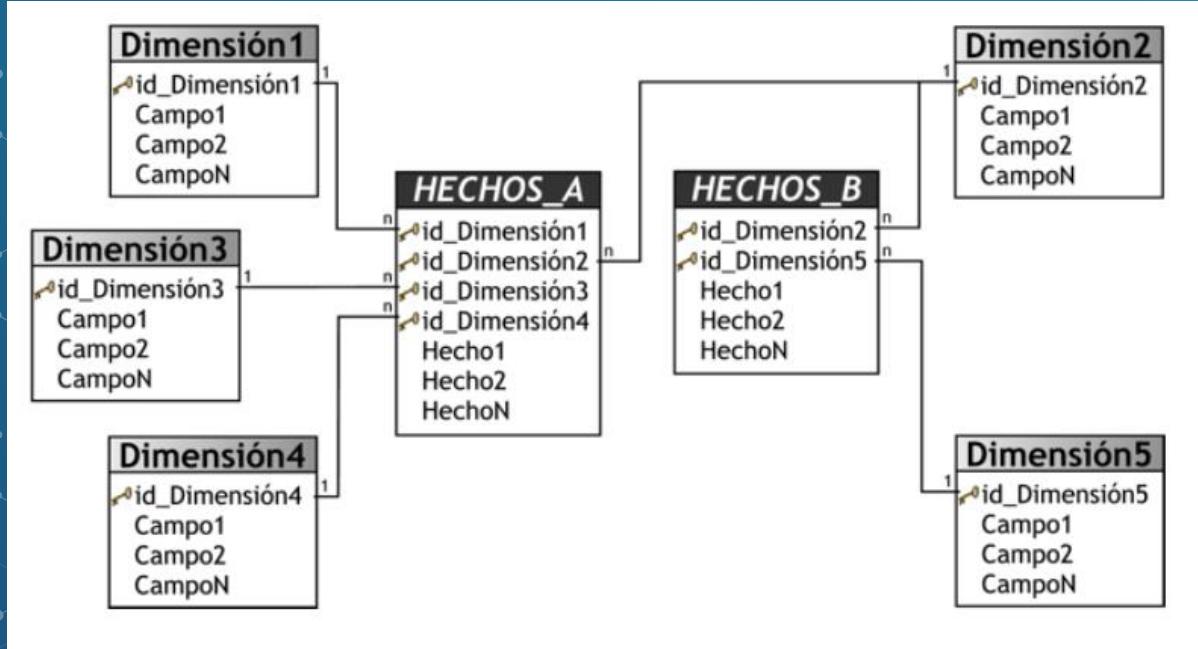
ESQUEMA DE COPO DE NIEVE



El modelo o esquema copo de nieve es una variación o derivación del modelo estrella. En este modelo la tabla de hechos deja de ser la única relacionada con otras tablas ya que existen otras tablas que se relacionan con las dimensiones y que no tienen relación directa con la tabla de hechos.

El modelo fue concebido para facilitar el mantenimiento de las dimensiones, sin embargo esto hace que se vinculen más tablas a las secuencias SQL, haciendo la extracción de datos más difícil así como vuelve compleja la tarea de mantener el modelo.

ESQUEMA DE CONSTELACIÓN



Este esquema puede estar formado por varios modelos en estrella definiéndose más de una tabla de hechos en la parte central del esquema relacionadas por sus respectivas tablas de dimensiones.

Es ampliamente utilizado y esquema más complejo que el esquema en estrella y el esquema de copo de nieve.

MUCHAS GRACIAS!

¿PREGUNTAS?

