Introducción a la Minería de Datos

Instituto de Computación - CPAP

Antecedentes

Grandes cantidades de información son almacenadas en la actualidad:

- Web Data
- E-comercio
- Tiendas de autoservicio
- Bancos

Grandes computadoras bajan precios.

La presión competitiva es mas fuerte:

- Provee mejores servicios personalizados para el cliente.

Antecedentes

Esto dificulta la realización de análisis de aspectos relevantes.

Frecuentemente hay información "**oculta**" en los datos que realmente no es evidentemente.

A los analistas humanos les puede tomar semanas **descubrir** información que sea útil.

Mucha de esta información no es **analizada** del todo.

La extracción de información se vuelve un tema relevante.

Antecedentes

La búsqueda tradicional de datos se realiza mediante análisis estadísticos.

A finales de los 80’s la estadística se amplió a técnicas como lógica difusa, razonamiento heurístico y redes neuronales.

Actualmente, las técnicas anteriores se aprovechan para generar conocimiento.

Definición

La **Minería de Datos** es la extracción automática de información predictiva escondida desde bases de datos.

La **Minería de Datos** estudia métodos y algoritmos que permiten la extracción automática de información sintetizada que permite caracterizar las relaciones escondidas.

Definición

En las aplicaciones de la Minería de Datos se hace sobre datos previamente recolectados.

Los datos no cambian mientras están siendo analizados.

Por lo que los datos generados son confiables y consistentes para éstos datos.

MD vs BS

La Minería de Datos y las Bases de Datos comerciales están disponibles para resolver problemas de decisión de negocios.

La Minería de Datos es una tecnología que ayuda a enfocarse en la información más importante en los almacenes de datos.

MD vs BS

Minería de Datos:

- No es una solución a negocios.
- Es sólo tecnología.
- Encuentra las "gemas pérdidas" en montañas de información.

Bases de Datos Comerciales:

- Involucra decisiones de información.
- Da decisiones de negocios.

¿Qué es y no es MD?

Que no es Minería de Datos:

- Localizar un número telefónico en el directorio.
- Consultar en un buscador información acerca de un tópico en particular.

Que es Minería de Datos:

- Grupos de documentos/usuarios similares.
- Preferencias de compras de los usuarios – ecommerce.
- Zonas de mayor criminalidad - prevención.
- Predicciones meteorológicas - agro.

Herramientas

Las Herramientas de la Minería de Datos:

- Predicen tendencias futuras y comportamientos.
- Pueden responder a preguntas que consumarían demasiado tiempo para resolverlas.

La automatización, provee herramientas típicas de soporte de decisión.

Técnicas

Las Técnicas de la Minería de Datos son el resultado de un largo proceso de investigación y desarrollo de productos.

La Minería de Datos esta soportada por tres tecnologías que son lo suficientemente maduras:

- Colección masiva de datos.
- Computadoras con multiprocesamiento.
- Algoritmos de minería de datos.

Evolución

En la siguiente tabla se muestra la evolución del tipo de consultas.

Evolución	Preguntas de Negocios	Tecnologías permitidas	Características
Colección de Datos (1960)	Cuales fueron los ingresos en los últimos 5 años?	Computadoras, cintas y discos	Liberación de datos estáticos retrospectiva.
Acceso a Datos (1980)	Que rebajas se tuvieron en Nueva Inglaterra en marzo?	Bases de datos relacionales y lenguajes de consulta estructurados (SQL)	Nivel de registro en liberación de datos dinámicos retrospectiva.
Almacén de Datos y Soporte de Decisión (1990)	Que rebajas se tuvieron en Nueva Inglaterra en marzo? Repetir para Boston.	Procesamiento analítico en línea, bases de datos multidimensionales y almacenes de datos.	Niveles múltiples en liberación de datos dinámicos retrospectiva.
Minería de Datos (1995)	Que es lo más probable que pase con las rebajas en Boston el próximo mes?	Algoritmos avanzados, computadoras con multiprocesador y bases de datos masivas	Liberación de información proactiva prospectiva.

Evolución

El componente principal en la Tecnología de la Minería de Datos ha sido desarrollado en:

- Estadística
- Inteligencia Artificial
- Máquinas de Aprendizaje

Actualmente, existe gran relevancia en:

- Ambientes de negocios
- Las descripciones básicas de las arquitecturas de almacenes de datos.

Funcionalidad

Predictiva:

- En base a una clasificación: por ejemplo si el cliente pagará o no pagará, o el tipo de dolencia que puede tener un paciente.

- En base a una regresión: por ejemplo calcular el tiempo previsible que se empleará en corregir los errores de un desarrollo de software.

Funcionalidad

Descriptiva:

- Agrupamiento (clustering): clasificar individuos en grupos en base a sus características. Por ejemplo, clasificar pacientes del hospital.
- Reglas de asociación: conocer cómo se relacionan los datos o campos. Por ejemplo conocer en el hipermercado que un cliente que compra leche muy probablemente comprará también pan.
- Secuenciación: intentar predecir el valor de una variable en función del tiempo. Por ejemplo la demanda de energía eléctrica.

Relación con otras disciplinas

Las bases de datos. Uso de almacenes de datos y/o OLAP (On-Line Analytical Processing).

- OLAP Este tipo de procesamiento en tiempo real maneja operaciones únicamente de consulta sobre grandes cantidades de información con la finalidad de realizar informes y resúmenes → toma de decisiones.

La recuperación de información. Obtener información desde datos textuales.

Relación con otras disciplinas

La estadística. Son necesarios cálculos para obtener: la media, la varianza, las distribuciones, el análisis univariante y multivariante, la regresión lineal y no lineal, la teoría del muestreo, la validación cruzada, la modelación paramétrica y no paramétrica, técnicas bayesianas.

El aprendizaje automático. La máquina usa algunos ejemplos para aprender un modelo y los utiliza para resolver el problema.

Relación con otras disciplinas

Los sistemas para la toma de decisión. El análisis ROC (Receiver Operating Characteristic) y los árboles de decisión.

La visualización de datos. Uso de diagramas de barras, graficas de dispersión, histogramas, coloreado de imágenes.

La computación paralela y distribuida. Distribuir las tareas más complejas entre diferentes procesadores o nodos.

Relación con otras disciplinas

Procesamiento del lenguaje natural. Es una disciplina encargada de producir sistemas informáticos que ayuden en la comunicación, por medio de la voz o del texto.

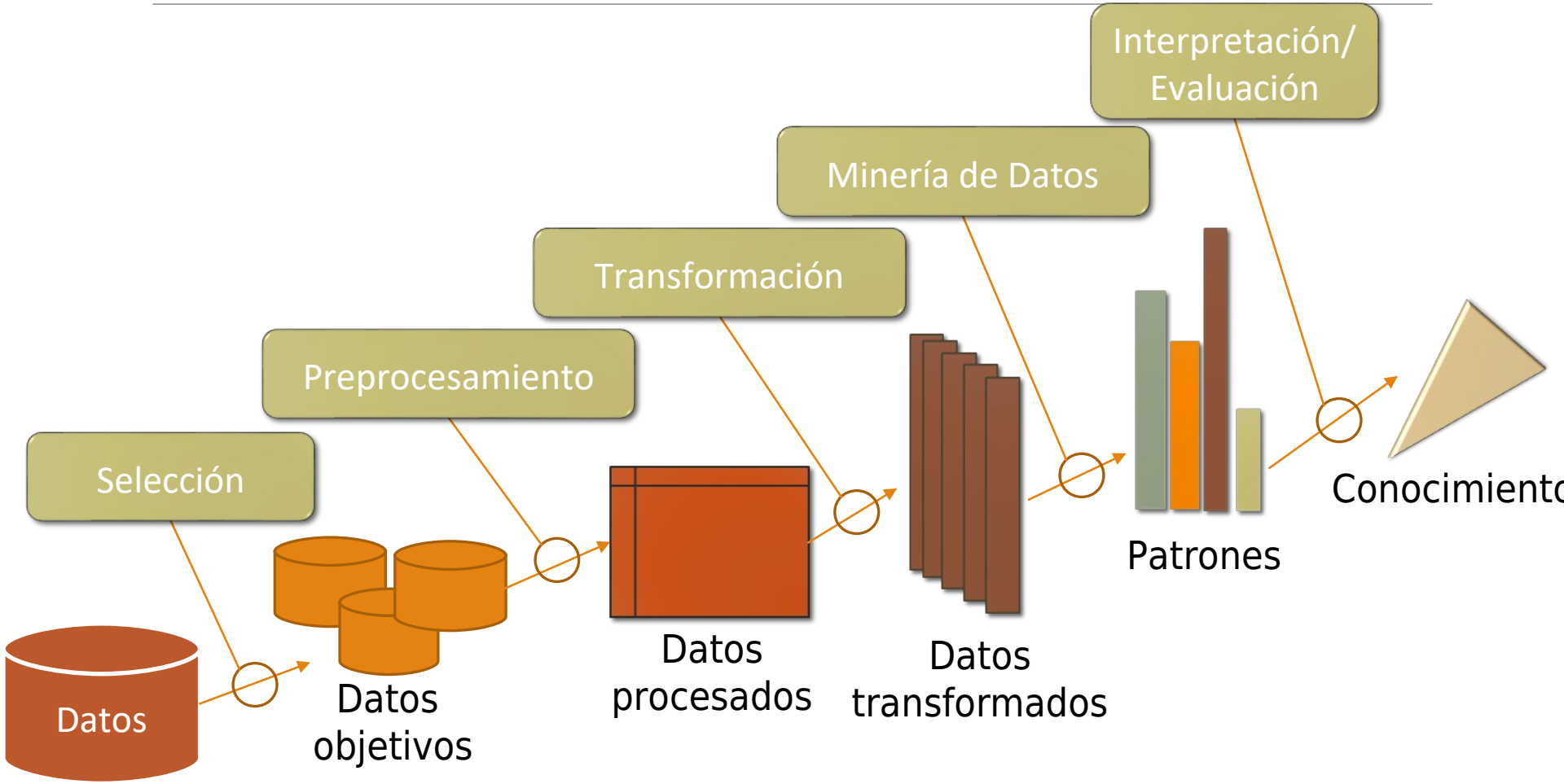
El proceso de KDD

KDD

El **KDD** (Knowledge Discovery from Databases) es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos. (Fayyad)

El **objetivo** fundamental del KDD (Knowledge Discovery from Databases), es encontrar conocimiento útil, válido, relevante y nuevo sobre una determinada actividad mediante algoritmos, dadas las crecientes órdenes de magnitud en los datos

Etapas de KDD



Etapas de KDD

Selección de datos. Consiste en buscar el objetivo y las herramientas del proceso de minería, identificando los datos que han ser extraídos, buscando los atributos apropiados de entrada y la información de salida para representar la tarea.

Esto quiere decir, primero se debe tener en cuenta lo que se saber lo que se quiere obtener y cuáles son los datos que nos facilitarán esa información para poder llegar a nuestra meta, antes de comenzar el proceso en tal.

Etapas de KDD

Limpieza de datos. En este paso se limpian los datos sucios, incluyendo los datos incompletos (donde hay atributos o valores de atributos perdidos), el ruido (valores incorrectos o inesperados) y datos inconsistentes (conteniendo valores y atributos con nombres diferentes).

Los datos sucios en algunos casos deben ser eliminados ya que pueden contribuir a un análisis inexacto y resultados incorrectos.

Etapas de KDD

Integración de datos. Combina datos de múltiples procedencias incluyendo múltiples bases de datos, que podrían tener diferentes contenidos y formatos.

Transformación de datos. Consisten en modificaciones sintácticas llevadas a cabo sobre datos sin que supongan un cambio para la técnica de minería aplicada. Las transformaciones discretas de los datos tienen la ventaja de que mejoran la comprensión de las reglas descubiertas al transformar los datos de bajo nivel en datos de alto y también reduce el tiempo de ejecución del algoritmo de búsqueda.

Etapas de KDD

Su principal desventaja es que se puede reducir la exactitud del conocimiento descubierto, debido a que puede causar la perdida de alguna información.

Reducción de datos. Reducir el tamaño de los datos, encontrando las características Más significativas dependiendo del objetivo del proceso.

Se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas, o para encontrar otras representaciones de los datos.

Etapas de KDD

- reducción de dimensiones (la extracción irrelevante y débil de atributo), compresión de datos (reemplazando valores de datos con datos alternativos codificados),
- reducción de tamaño (reemplazando valores de datos con representación alternativa más pequeña),
- una generalización de datos (reemplazando valores de datos de niveles conceptuales bajos con niveles conceptuales más altos), etc.

Etapas de KDD

Minería de Datos. Consiste en la búsqueda de los patrones de interés que pueden expresarse como un modelo o simplemente que expresen dependencia de los datos.

Se tiene que especificar un criterio de preferencia para seleccionar un modelo de un conjunto de posibles modelos. También se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está determinado en el algoritmo de minería).

Herramientas

Las Herramientas obtienen de las bases de datos patrones escondidos.

Las Técnicas de la Minería de Datos pueden ser implementadas rápidamente en software y en las plataformas de hardware existente.

Las Herramientas de Minería de Datos pueden ser implementadas en plataformas cliente-servidor o computadoras de procesamiento paralelo.

Etapas de KDD

Evaluación de los patrones. Se identifican verdaderamente patrones interesantes que representan conocimiento usando diferentes técnicas incluyendo análisis estadísticos y lenguajes de consultas.

Interpretación de resultados. Consiste en entender los resultados del análisis y sus implicaciones y puede llevar a regresar a algunos de los pasos anteriores

Almacén de Datos

Definición

Colección de datos orientada a un **dominio**, **integrado**, no volátil, y que varía poco en el tiempo.

Ayuda a la toma de decisiones de la empresa u organización.

Es sobre todo, un expediente de una empresa más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficientes de datos (especialmente **OLAP**).

Características

El almacenamiento de los datos no debe usarse con datos de uso actual.

Los almacenes de los datos contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas que son conocidas como los centros comerciales dependientes de los datos.

Características

Generalmente, dos ideas básicas dirigen la creación de un almacén de los datos:

- Integración de los datos, que facilita una descripción global y un análisis comprensivo en el almacén de los datos.
- Separación de los datos usados en operaciones diarias de los datos usados del almacén para los propósitos de la divulgación.

Arquitectura

Nivel operacional: Contiene datos primitivos (operacionales) que están siendo permanentemente actualizados, usados por los sistemas operacionales tradicionales que realizan operaciones transaccionales.

Almacén de datos: Contiene datos primitivos correspondientes a sucesivas cargas del Almacén de Datos y algunos datos derivados. Los datos derivados son datos generados a partir de los datos primitivos al aplicarles algún tipo de procesamiento (resúmenes).

Función

Un **almacén de datos** debe entregar la información correcta a la gente indicada en el momento adecuado en el formato correcto.

El almacén de datos da respuesta a las necesidades de usuarios conocedores, utilizando sistemas de ayuda en la decisión (DSS), Sistemas de Información Ejecutiva (EIS) o herramientas para hacer consulta o informes.

Los usuarios finales fácilmente pueden hacer consultas sobre sus almacenes de datos sin tocar o afectar la operación del sistema.

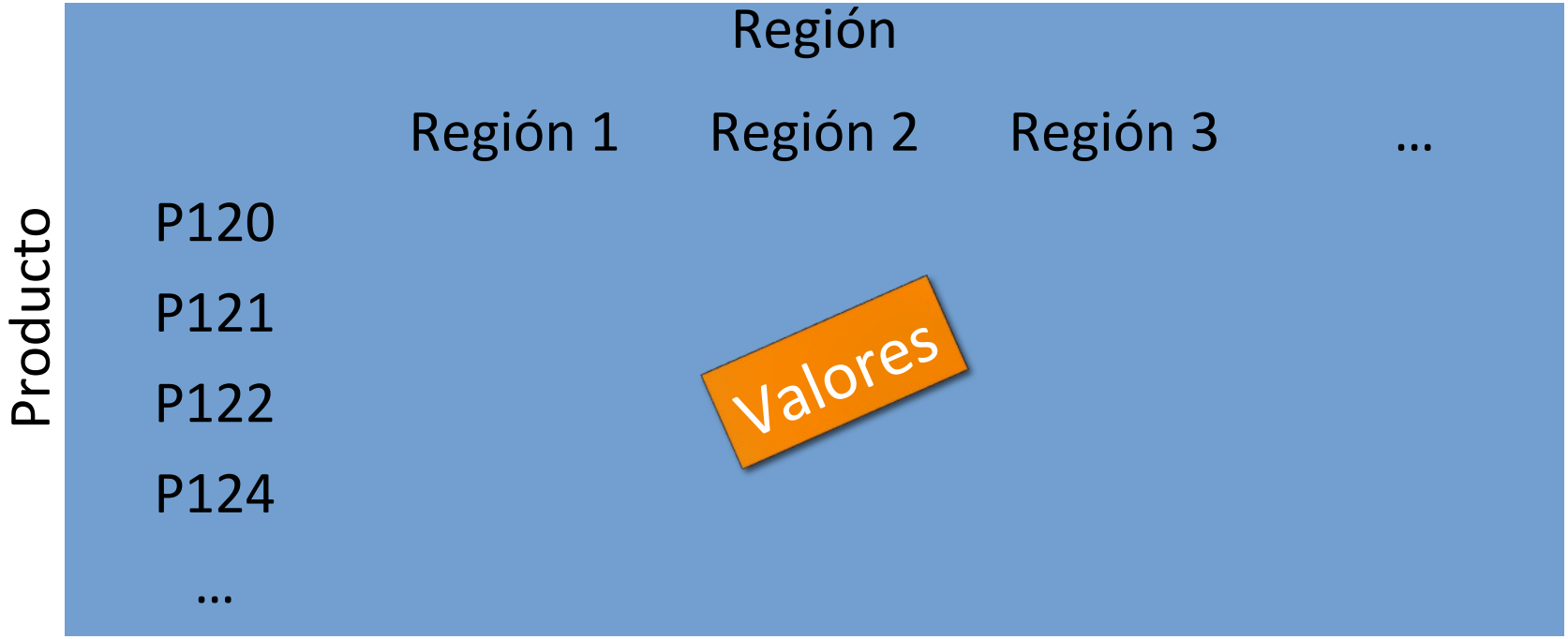
Arquitectura

Nivel departamental (Data Mart): Contiene casi exclusivamente datos derivados. Va a ser el blanco de salida sobre el cual los datos en el almacén son organizados y almacenados para las consultas directas por los usuarios finales, los desarrolladores de reportes y otras aplicaciones.

Nivel individual: Contiene pocos datos, resultado de aplicar heurísticas, procesos estadísticos, etc., a los datos contenidos en el nivel anterior. El nivel individual es el objetivo final de un Almacén de Datos.

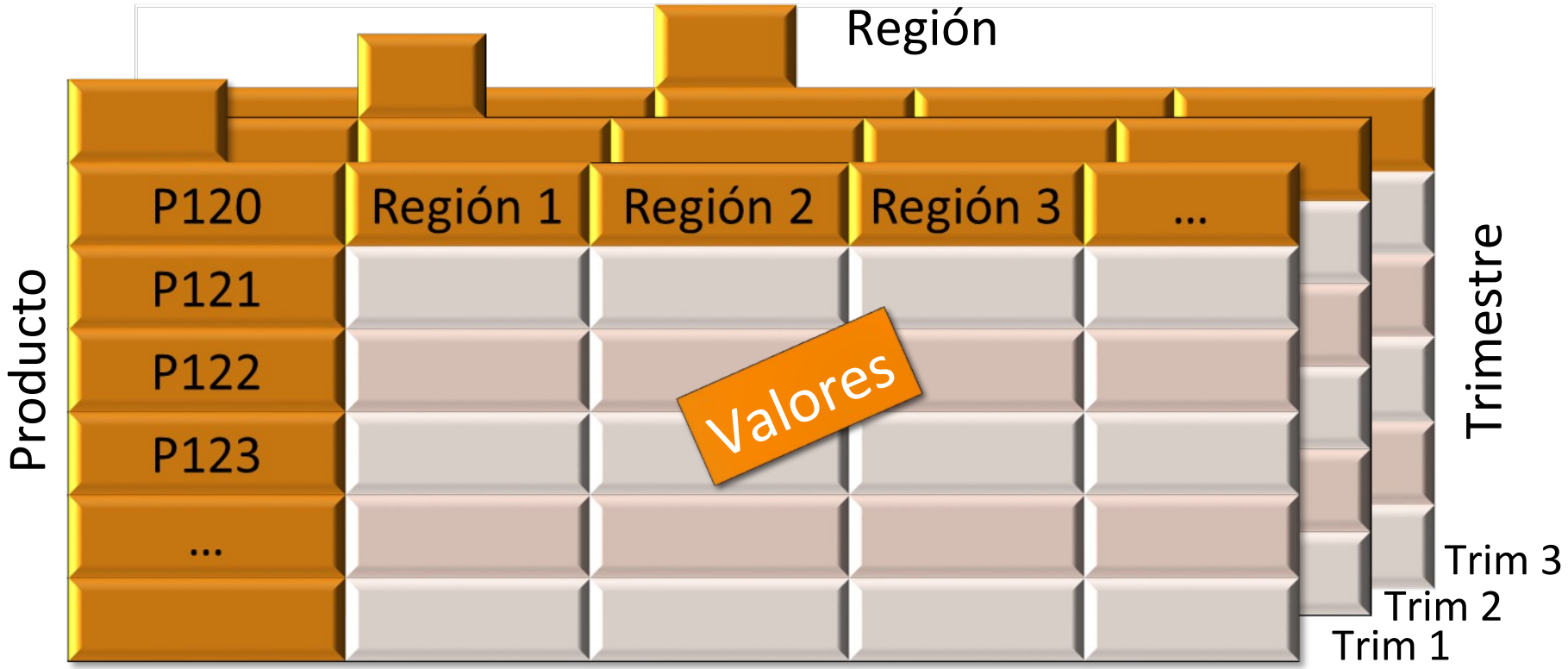
Modelado de datos

Una hoja de cálculo constituye una matriz.



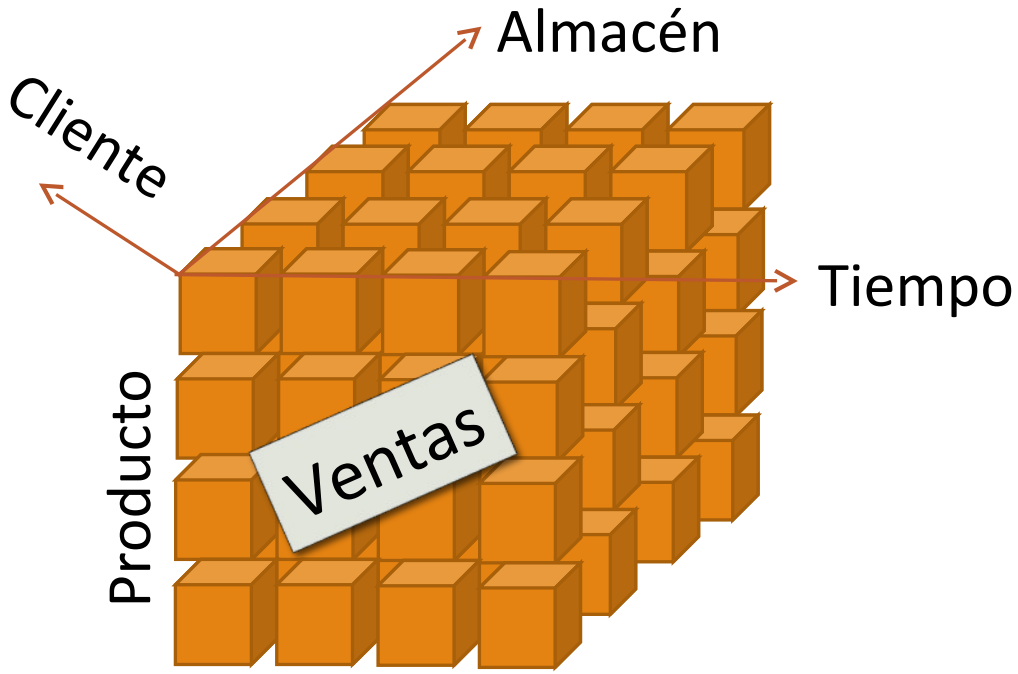
Modelado de datos

Añadiendo una dimensión se tendría una matriz tridimensional.



Modelado de datos

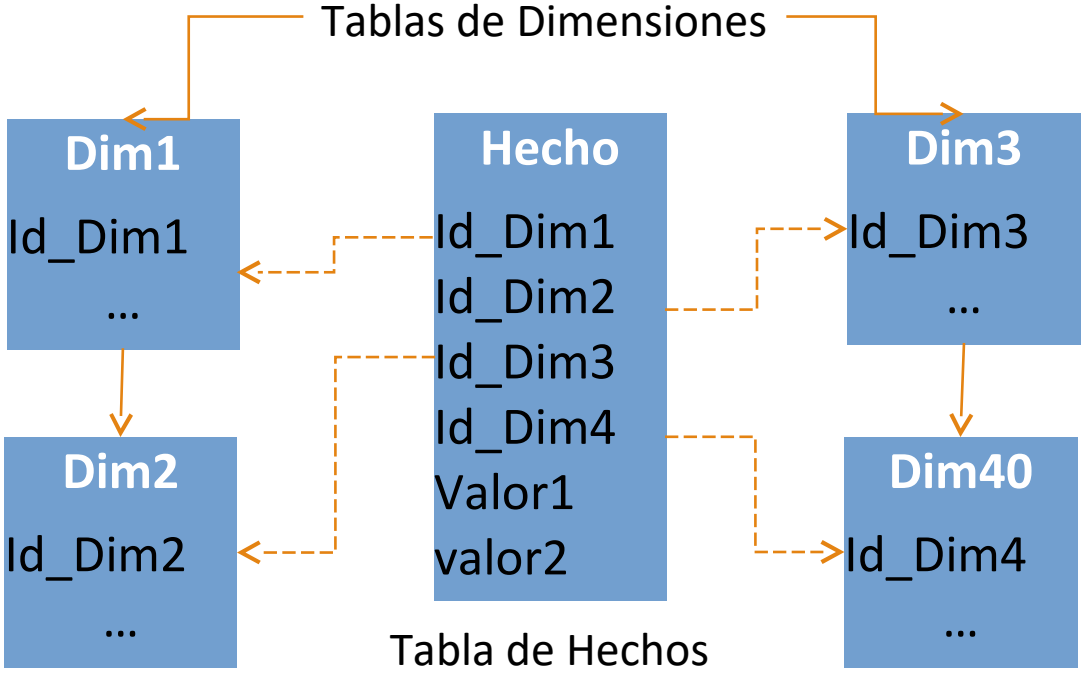
Las herramientas de explotación OLAP de los almacenes de datos han adoptado un modelo multidimensional de datos.



Modelado de datos

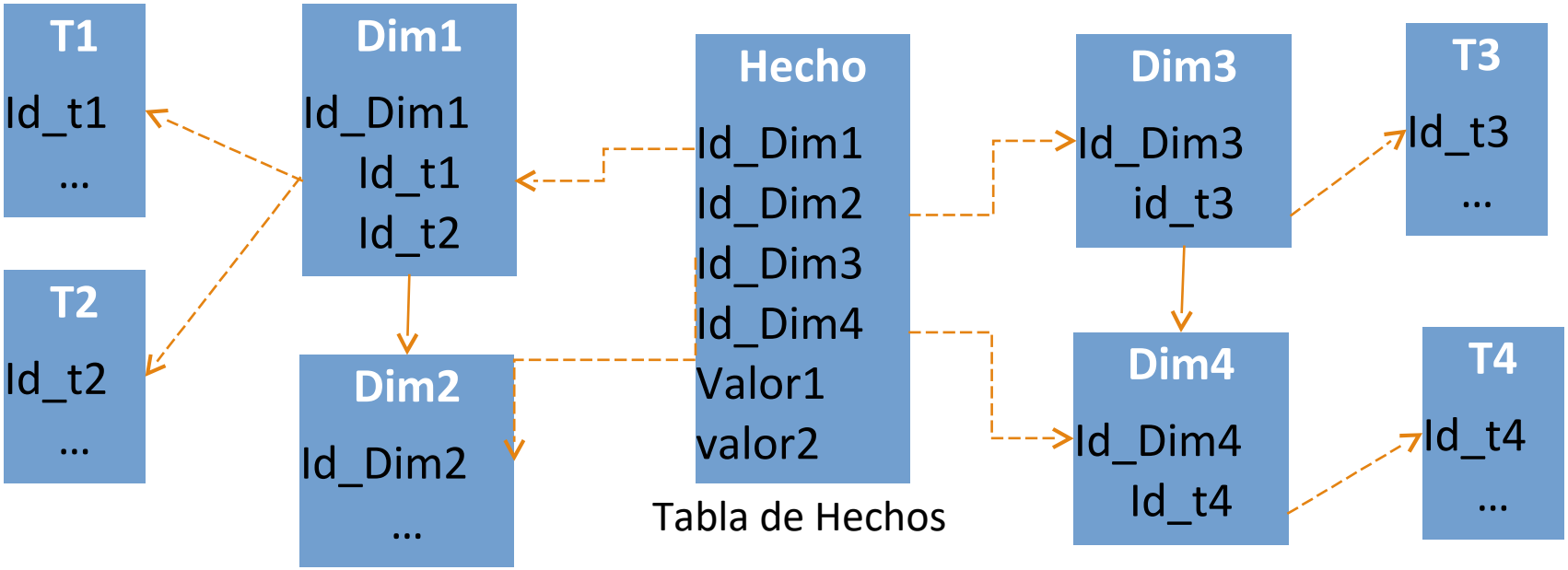
Tres son los esquemas multidimensionales comunes:

- Esquema en estrella: formado por una tabla de hechos con una única tabla para cada dimensión.



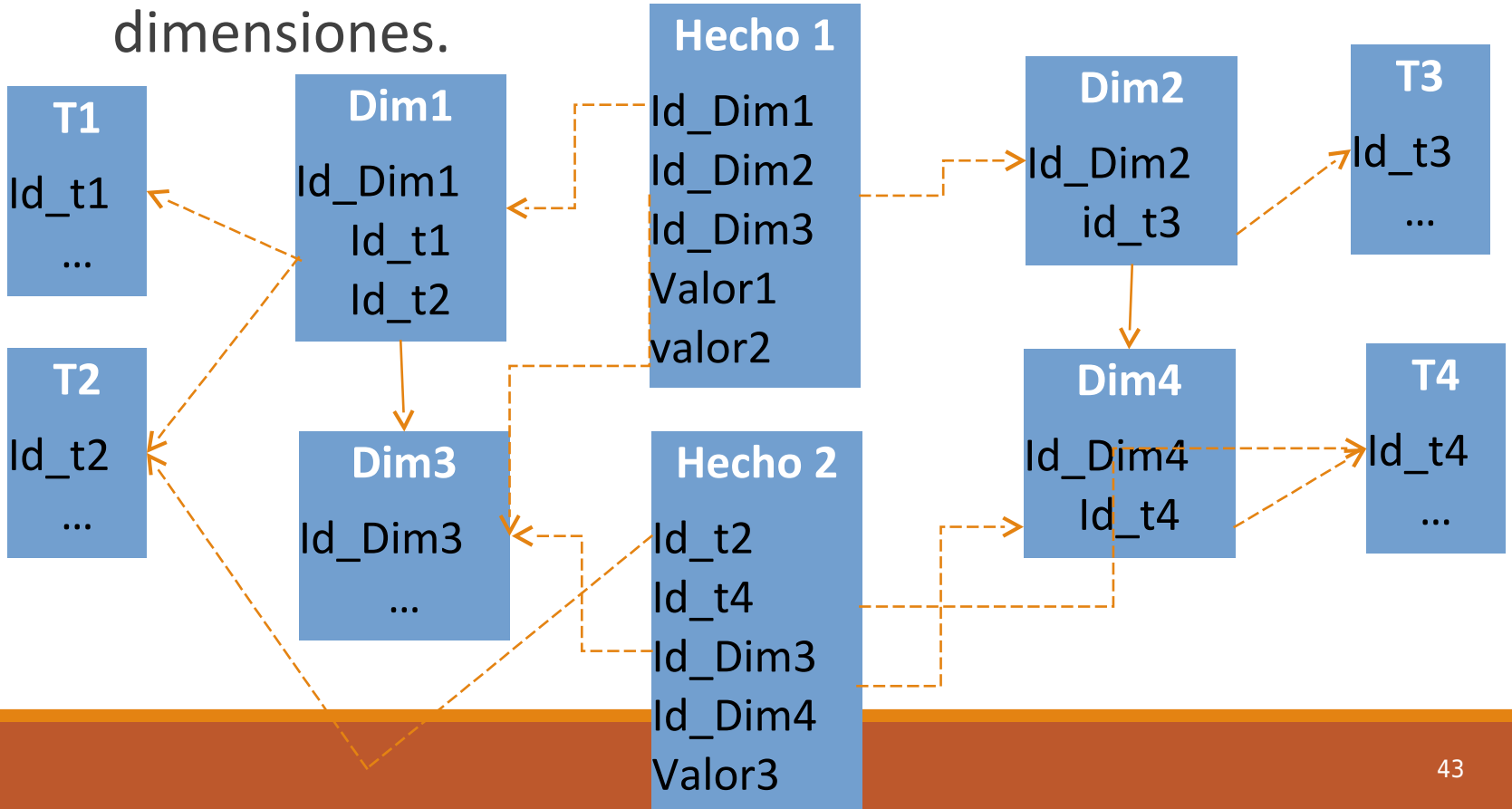
Modelado de datos

- Esquema en copos: es una variante del esquema de estrella en el que las tablas dimensionales de este último se organizan jerárquicamente mediante su normalización.

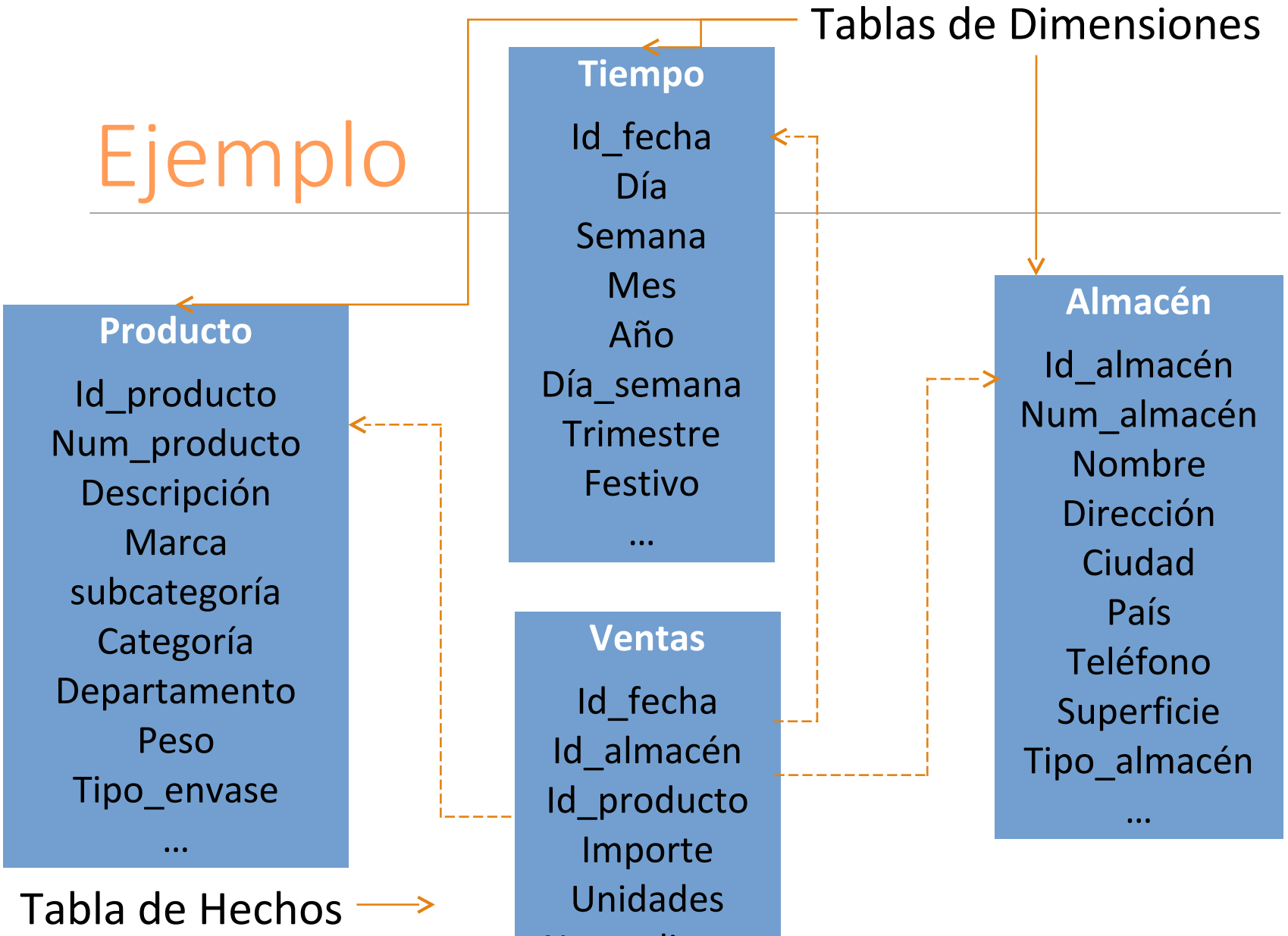


Modelado de datos

- Constelación de hechos: es un conjunto de tablas de hechos que comparten algunas tablas de dimensiones.



Ejemplo


Tablas de Dimensiones
Tiempo
Id_fecha
Día
Semana
Mes
Año
Día_semana
Trimestre
Festivo
…
Producto
Id_producto
Num_producto
Descripción
Marca
subcategoría
Categoría
Departamento
Peso
Tipo_envase
…
Ventas
Id_fecha
Id_almacén
Id_producto
Importe
Unidades
Num_cliente
Almacén
Id_almacén
Num_almacén
Nombre
Dirección
Ciudad
País
Teléfono
Superficie
Tipo_almacén
…
Tabla de Hechos