

A high-angle, black and white photograph of a massive concrete dam. The dam's surface is composed of large, rectangular panels with visible vertical joints and some weathering. A narrow walkway with a metal railing runs along the top edge of the dam. A small figure of a person stands on this walkway, providing a sense of scale to the enormous structure. The sky is a uniform, dark grey.

PABLO FIGUEROA

---

# MINERÍA DE DATOS

# DISEÑO DE LA INVESTIGACIÓN

# CONOCIMIENTO CIENTÍFICO

- ▶ **Sistemático**, no puedo eliminar arbitrariamente pasos, se deben seguir de forma rigurosa.
- ▶ **Ordenado**.
- ▶ **Metódico**, debe seguir un camino.
- ▶ **Racional / Reflexivo**, implica una reflexión por parte del investigador y tiene que ver con una ruptura con el sentido común.
- ▶ **Crítico / Subversivo**, intenta producir conocimiento.

# PROBLEMA DE INVESTIGACIÓN

- ▶ Se hace necesaria la existencia de un problema, luego la toma de conciencia sobre el problema y, por último, la existencia de una solución posible.
- ▶ El problema de investigación yace en la discrepancia existente entre un modelo ideal y un modelo real.
- ▶ La discrepancia entre el modelo ideal y real debe ser significativa y se requiere la toma de conciencia de esa discrepancia.
- ▶ El trabajo se orienta a la solución del problema, si no tiene solución no se investiga.

# TÉCNICAS Y PASOS DE LA INVESTIGACIÓN

- ▶ Tema.
- ▶ Delimitación del tema.
- ▶ Formulación del problema.
- ▶ Reducción del problema a nivel empírico.
- ▶ Determinación de las unidades de análisis - Recolección de datos.
- ▶ Análisis de datos.
- ▶ Informe Final.



# DELIMITACIÓN DEL TEMA

- ▶ Contextualización.
  - ▶ Espacial, Temporal, Sociodemográfico y Sociocultural.
- ▶ Torbellino de ideas.
- ▶ Ayudas metodológicas.
- ▶ Observación de casos típicos y atípicos.
- ▶ Acercamiento al campo.

# FORMULACIÓN DEL PROBLEMA

- ▶ Formulación de objetivos.
- ▶ Marco teórico.
- ▶ Formular Hipótesis.
- ▶ Formular Interrogantes.
- ▶ Definición de variables.

# CLASIFICACIÓN DE LA HIPÓTESIS

- ▶ Hipótesis general, poseen un alto contenido de abstracción.
- ▶ Hipótesis intermedia, establece relación intermedia entre la teoría y el campo empírico.
- ▶ Hipótesis empírica, construída a partir de definiciones operacionales o indicadores, directamente contrastables y medibles.
- ▶ Hipótesis generalización, permite extender las conclusiones tomadas para las muestras al conjunto o población de sujetos o fenómenos.



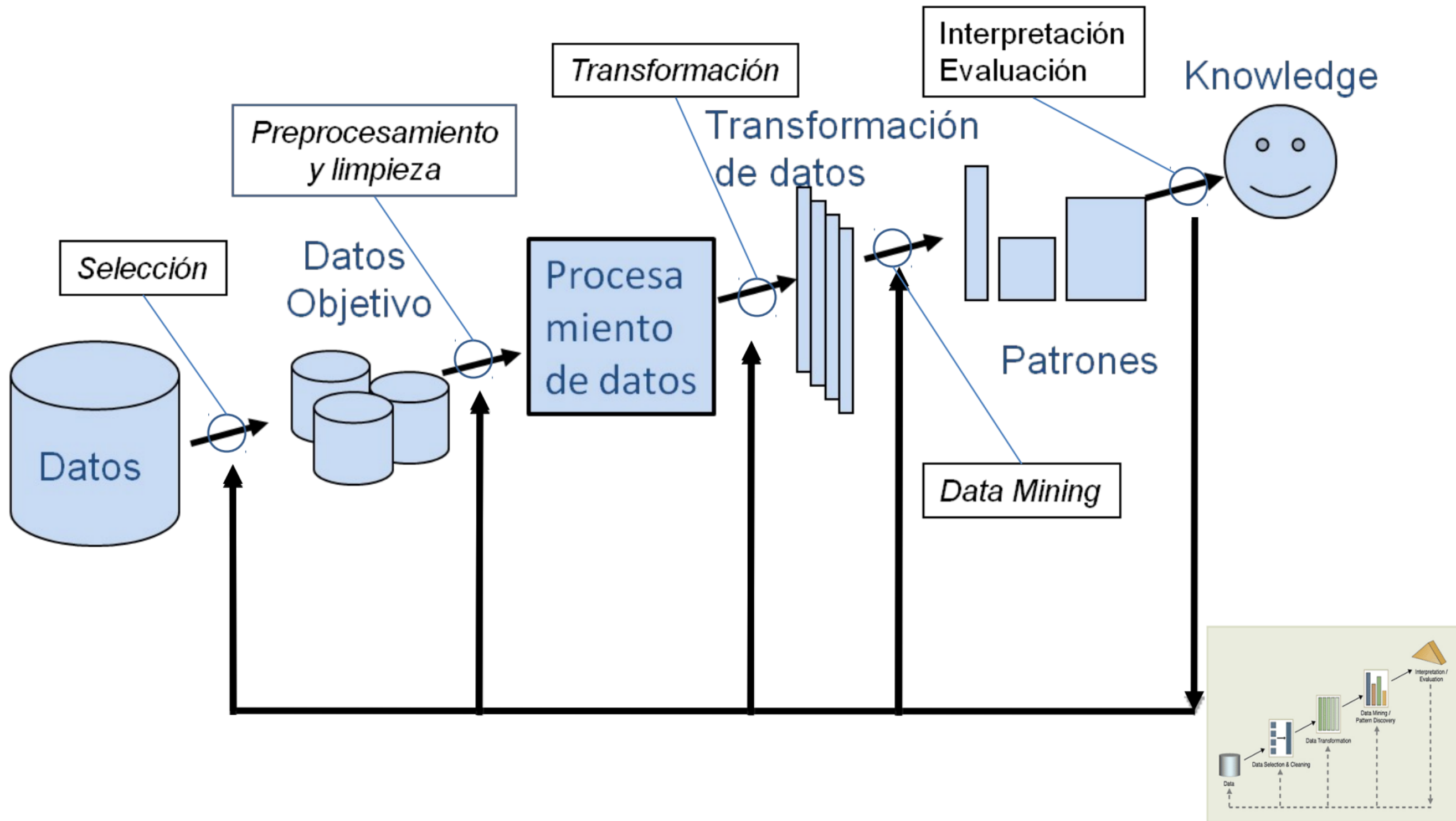
# CLASIFICACIÓN DE LA HIPÓTESIS – SEGÚN NEXO VARIABLES

- ▶ Hipótesis descriptiva: señala la frecuencia o característica de un fenómeno sin establecer relaciones causales entre sus variables.
  - ▶ Asociativas y Correlacionales.
- ▶ Hipótesis explicativas: dan cuenta del porque o causa de los fenómenos.
  - ▶ Causales o determinísticas, estocásticas o probabilísticas, contingentes, predictivas.

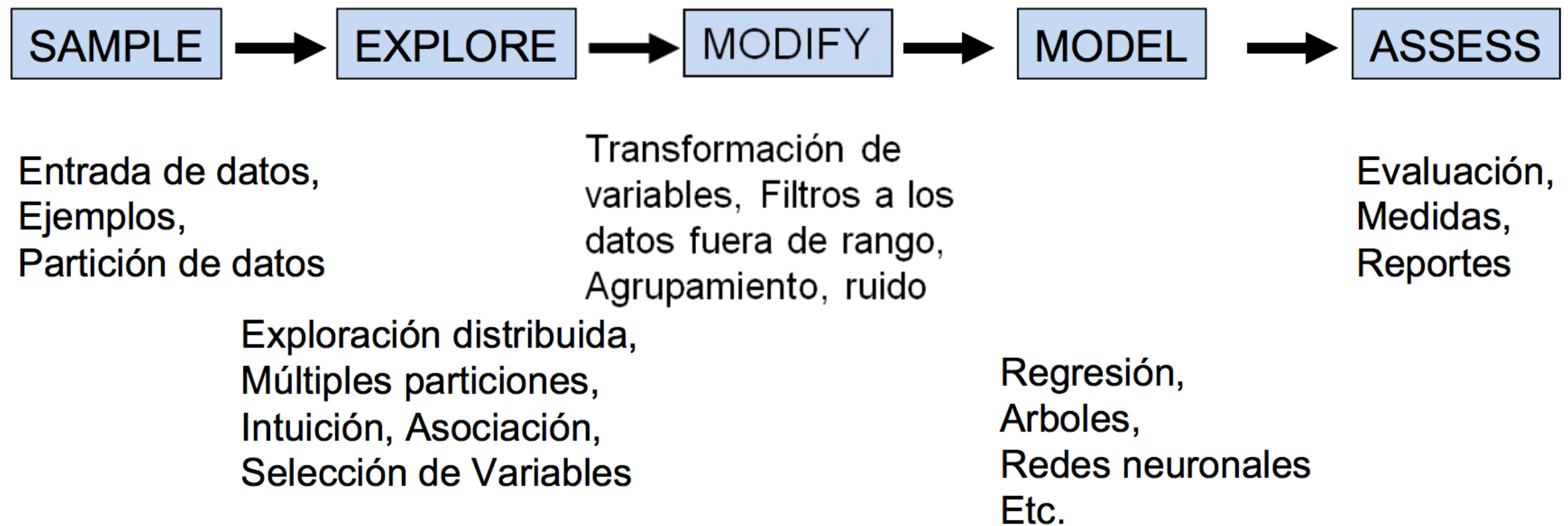
# METODOLOGÍAS MINERÍA DE DATOS – KDD

- ▶ Knowledge Discovery in Databases.
- ▶ Es una metodología propuesta por Fayyad en 1996, propone 5 fases:
  - ▶ Selección.
  - ▶ Preprocesamiento.
  - ▶ Transformación.
  - ▶ Minería de datos.
  - ▶ Evaluación e implantación.
- ▶ Es un proceso iterativo e interactivo.

# METODOLOGÍAS MINERÍA DE DATOS – KDD



# METODOLOGÍAS MINERÍA DE DATOS – SEMMA



## METODOLOGÍAS MINERÍA DE DATOS – CRISP-DM

### ► Cross-Industry Standard Process for Data Mining

#### 1. **Comprensión del negocio:**

- ✓ Entendimiento de los objetivos y requerimientos del proyecto.
- ✓ Definición del problema de Minería de Datos

#### 2. **Comprensión de los datos**

- ✓ Obtención conjunto inicial de datos.
- ✓ Exploración del conjunto de datos.
- ✓ Identificar las características de calidad de los datos
- ✓ Identificar los resultados iniciales obvios.

#### 3. **Preparación de Datos**

- ✓ Selección de datos
- ✓ Limpieza de datos

#### 4. **Modelamiento**

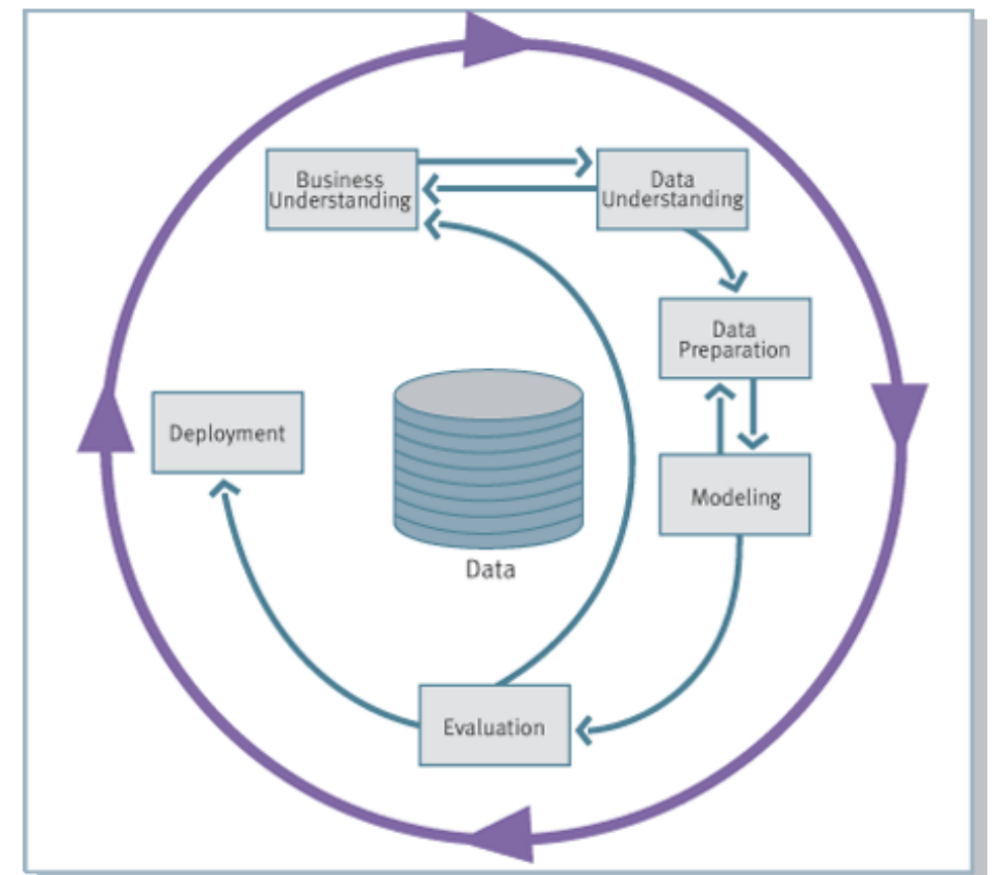
Implementación en herramientas de Minería de Datos

#### 5. **Evaluación**

- ✓ Determinar si los resultados coinciden con los objetivos del negocio
- ✓ Identificar las temas de negocio que deberían haberse abordado

#### 6. **Despliegue**

- ✓ Instalar los modelos resultantes en la práctica
- ✓ Configuración para minería de datos de forma repetida ó continua



# FUENTES DE DATOS



# Tipo de datos

- **Datos estructurados (Structured Data)**. Datos con formato o esquema fijo y que poseen campos fijos.
- **Datos semiestructurados (Semi-Structured Data)**. No tienen formatos fijos, pero contienen etiquetas y otros marcadores que permiten separar los elementos dato. Los registros *weblogs*.
- **Datos no estructurados (Unstructured Data)**. Son datos sin tipos predefinidos. Se almacenan como documentos u objetos sin una estructura uniforme.
- **Datos en tiempo real (Real-Time Data)**. A los anteriores se les añade la capacidad de visionarios en tiempo real, mientras están ocurriendo.

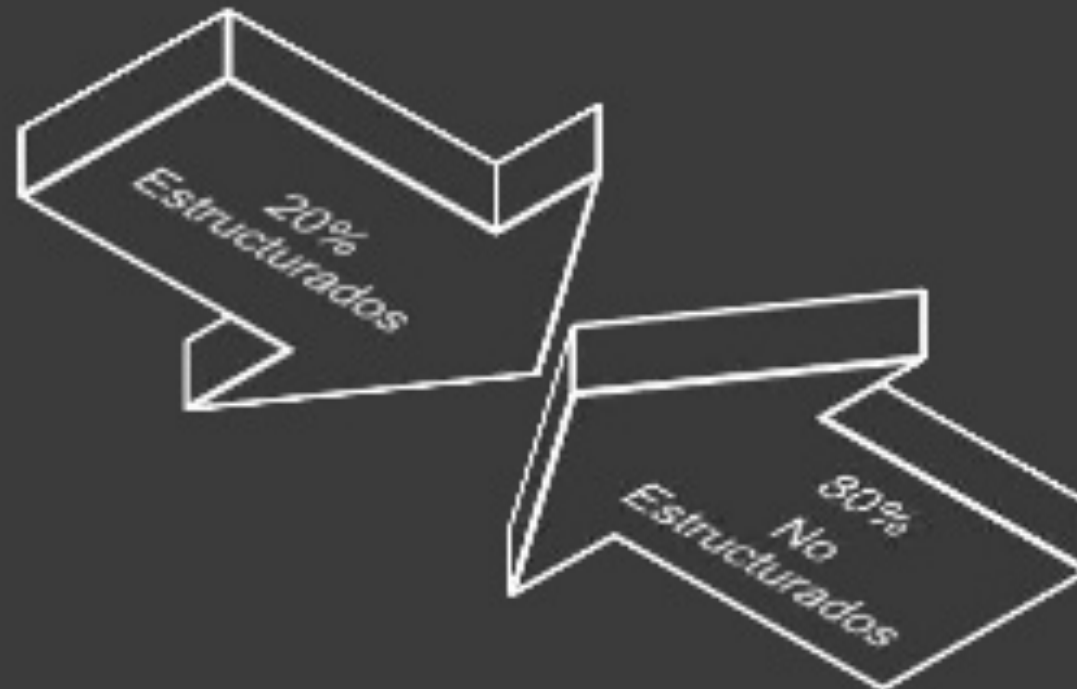


# TIPOS DE DATOS

## Tipos de Datos Big Data

### Datos Estructurados

**(Structured Data).** Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres.



### Datos No Estructurados (Unstructured Data).

Datos en el formato tal y como fueron recolectados, carecen de un formato específico.

# LOS CINCO TIPOS DE FUENTES DE DATOS



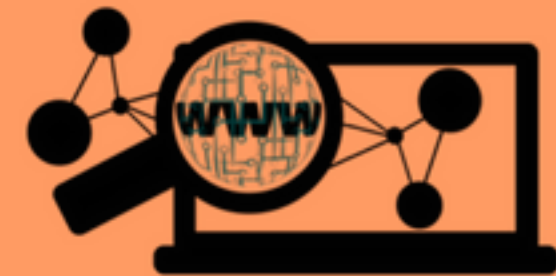
## BIOMÉTRICOS

Referidos a la identificación automática de una persona basada en sus características anatómicas o trazos personales..  
(Reconocimiento facial, Genética)



## MAQUINA A MAQUINA

Se trata de aquellas tecnologías que permiten la conexión de diferentes dispositivos entre sí (Internet de las Cosas). Un ejemplo son los GPS, pero también los denominados chips NFC (aquella tecnología que se sustenta en la comunicación inalámbrica y que permite la transmisión de datos de forma segura: integrada fundamentalmente en smarphone y tablets)



## WEB Y MEDIOS SOCIALES

Son los que se originan en la red y configuran, el trozo más grande del pastel llamado Big Data  
(Facebook, Twitter, Contenido Web)

## TRANSACCIONES

Los datos que se registran en los departamentos de facturación, centros de llamada, mensajería, reclamaciones, presentación y registro de documentos

## GENERADOS POR HUMANOS

Registros de voz de centros de llamada, correo electrónico, registros médicos electrónicos, faxes, documentos electrónicos, notas de voz...

# OLAP – OLPT

- ▶ Sobre estas mismas bases de datos ya se puede extraer conocimiento:
  - ▶ OLPT, On-Line Transactional Processing.
  - ▶ OLAP, On-Line Analytical Processing.
- ▶ Problemas:
  - ▶ Killer queries.
  - ▶ Diseñada para trabajo transaccional, no para el análisis de datos.

# DATA-WAREHOUSING

- ▶ Para poder operar eficientemente con esos datos y debido a que los costes de almacenamiento masivo y conectividad se han reducido drásticamente en los últimos años, parece razonable recoger (copiar) los datos en un sistema unificado.
- ▶ DATA-WAREHOUSES (Almacenes de Datos): Se separan de los datos a analizar con respecto a sus fuentes transaccionales (se copia/almacena toda la información histórica).
- ▶ Existe toda una tecnología creciente de cómo organizarlos y sobretodo de cómo tenerlos actualizados (cargas periódicas) respecto a los datos originales.

# DATA-WAREHOUSING

- ▶ Facilita el análisis de los datos en tiempo real (OLAP),
- ▶ No interviene el OLTP de las bases de datos originales.

	BASE DE DATOS TRANSACCIONAL	ALMACÉN DE DATOS
Propósito	Operaciones diarias. Soporte a las aplicaciones.	Recuperación de información, informes, análisis y minería de datos.
Tipo de datos	Datos de funcionamiento de la organización.	Datos útiles para el análisis, la sumariaización, etc.
Características de los datos	Datos de funcionamiento, cambiantes, internos, incompletos...	Datos históricos, datos internos y externos, datos descriptivos...
Modelo de datos	Datos normalizados.	Datos en estrella, en copo de nieve, parcialmente desnormalizados, multidimensionales...
Número y tipo de usuarios	Cientos/miles: aplicaciones, operarios, administrador de la base de datos.	Decenas: directores, ejecutivos, analistas (granjeros, mineros).
Acceso	SQL. Lectura y escritura.	SQL y herramientas propias ( <i>slice &amp; dice, drill, roll, pivot...</i> ). Lectura.

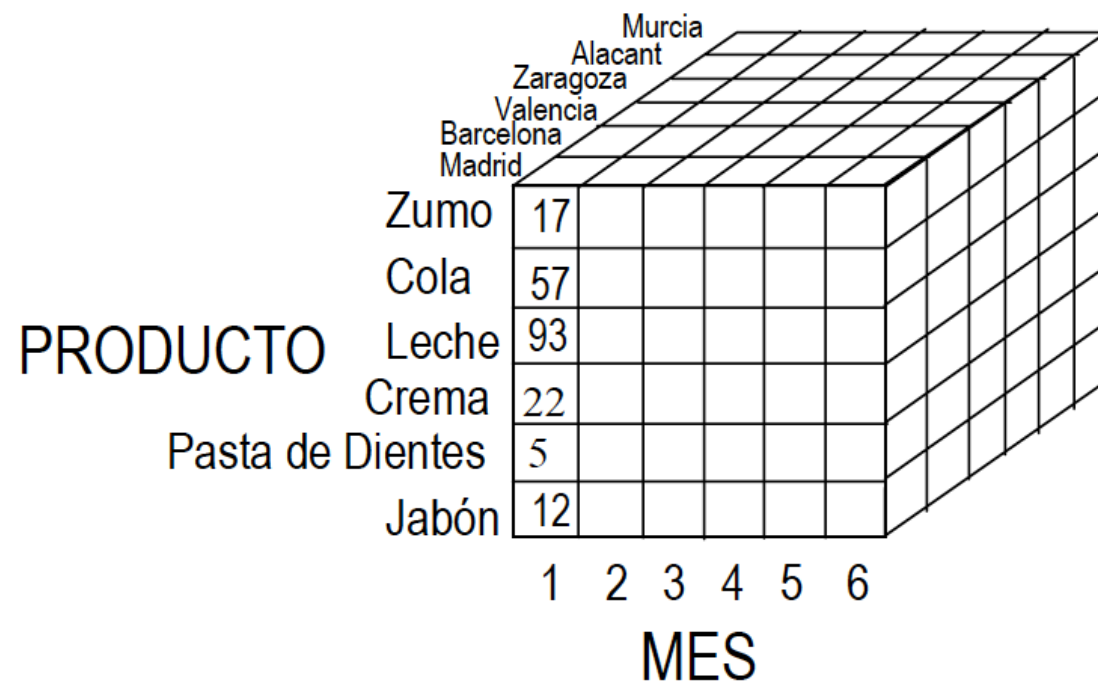


# DATA-WAREHOUSING

- ▶ Según la organización de la información copiada se distingue:
  - ▶ ROLAP (Relational OLAP): el almacén de datos es relacional.
  - ▶ MOLAP (Multidimensional OLAP): el almacén de datos es una matriz multidimensional.
- ▶ Aunque un MOLAP puede estar implementado sobre un sistema de gestión de base de datos relacional (RDBMS).

## MOLAP

- ▶ Cada atributo relevante se establece en una dimensión, que se puede agregar o desagregar. La base de datos está completamente desnormalizada.



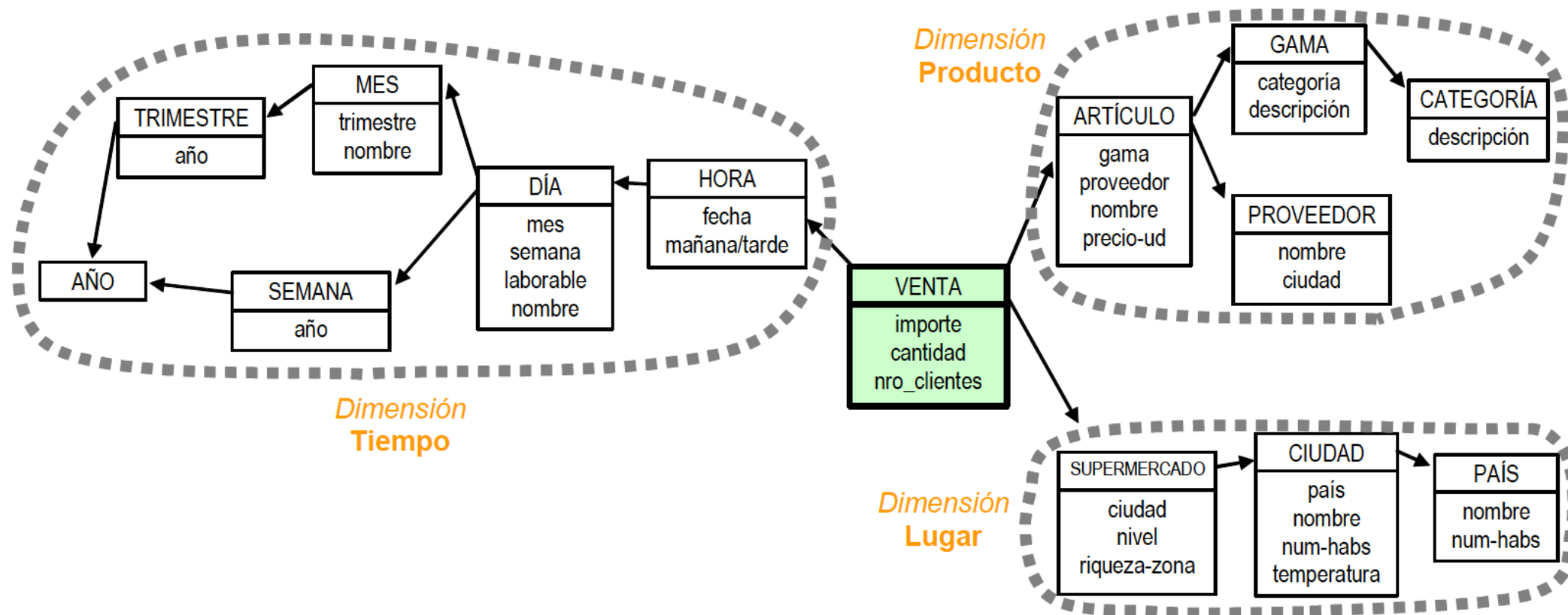
*Ventas en millones de Euros*

*Las dimensiones se agregan:*

<i>Industria</i>	<i>País</i>	<i>Año</i>	
<i>Categoría</i>	<i>Región</i>	<i>Cuatrimestre</i>	
		/	\
<i>Producto</i>	<i>Ciudad</i>	<i>Mes</i>	<i>Semana</i>
		\	/
	<i>Supermercado</i>	<i>Día</i>	

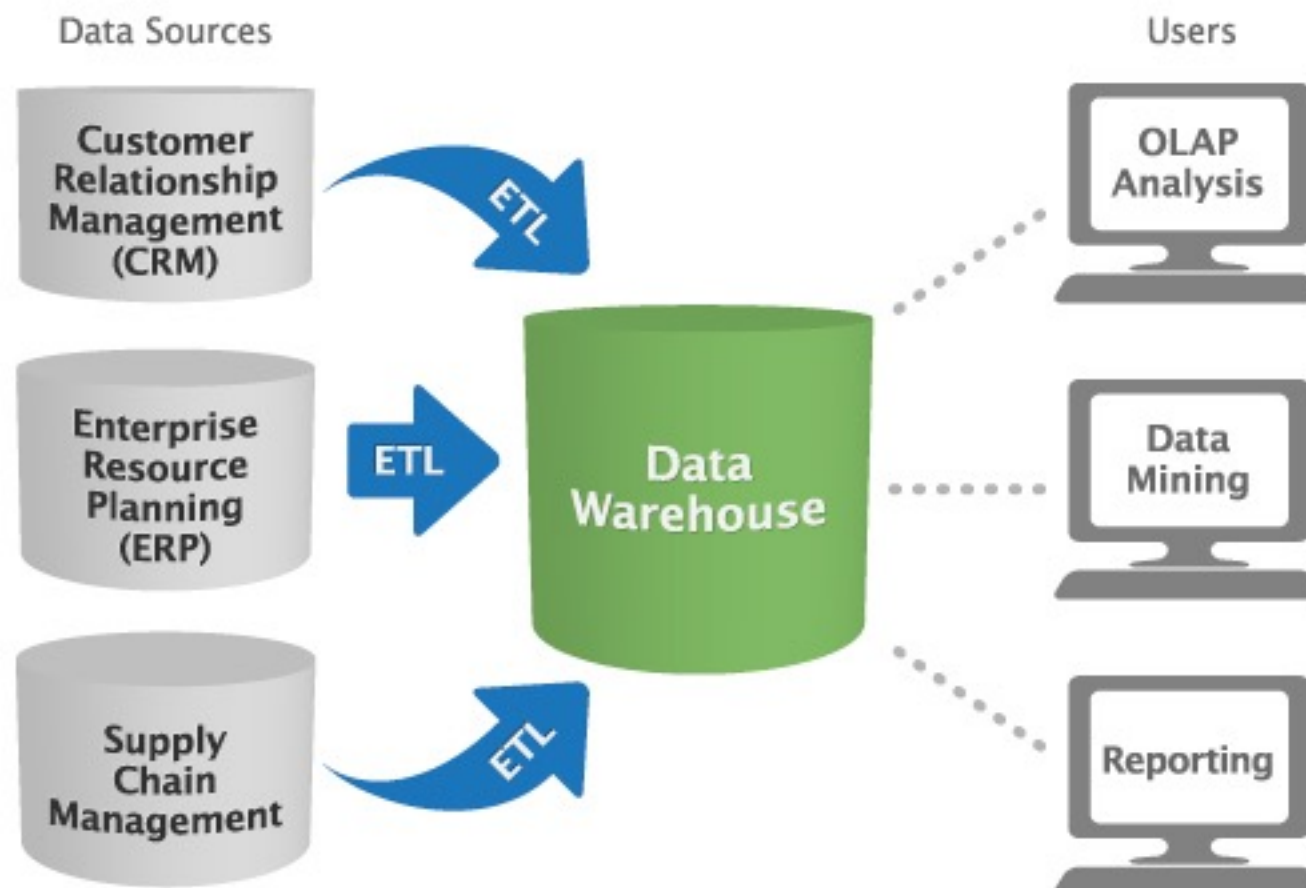
## ROLAP

- Las dimensiones, que se puede agregar o desagregar, siguiendo claves ajenas. Se conserva parte de la normalización.



## DATA-WAREHOUSING


- ▶ Esquemas de almacenes de datos más comunes:
  - ▶ estrella simple
  - ▶ estrella jerárquica (copo de nieve).
- ▶ Esta estructura permite la sumalización, la visualización y la navegación según las dimensiones de la estrella.



## DATA-WAREHOUSING \* SUMARIZACIÓN Y OPERADORES

- ▶ Estas estructuras permiten 'navegar' sumalizando (agregando) o desagregando.
- ▶ **Drill.** Se utiliza para desagregar dimensiones. Este operador permite entrar más al detalle en el informe.

CATEGORÍA	TRIMESTRE	IMPORTE
Refrescos	T1	150.323 euros
Refrescos	T2	233.992 euros
Refrescos	T3	410.497 euros
Refrescos	T4	203.400 euros
Congelados	T1	2.190.103 euros
Congelados	T2	1.640.239 euros
Congelados	T3	1.904.401 euros
Congelados	T4	2.534.031 euros

  
**drill**  
categoría= "refrescos"  
ciudad= {"Valencia", "León"}

CATEGORÍA	TRIMESTRE	CIUDAD	IMPORTE
Refrescos	T1	Valencia	13.267
Refrescos	T1	León	3.589
Refrescos	T2	Valencia	27.392
Refrescos	T2	León	4.278
Refrescos	T3	Valencia	73.042
Refrescos	T3	León	3.780
Refrescos	T4	Valencia	18.391
Refrescos	T4	León	3.629

# DATA-WAREHOUSING \* SUMARIZACIÓN Y OPERADORES

- ▶ Roll. Operador inverso a drill. Obtiene información más agregada.

CATEGORÍA	TRIMESTRE	IMPORTE
Refrescos	T1	150.323 euros
Refrescos	T2	233.992 euros
Refrescos	T3	410.497 euros
Refrescos	T4	203.400 euros
Congelados	T1	2.190.103 euros
Congelados	T2	1.640.239 euros
Congelados	T3	1.904.401 euros
Congelados	T4	2.534.031 euros



roll

un nivel por "tiempo"

CATEGORÍA	IMPORTE
Refrescos	998.212 euros
Congelados	10.458.877 euros



## DATA-WAREHOUSING \* SUMARIZACIÓN Y OPERADORES

- ▶ El operador **pivot** permite cambiar algunas filas por columnas.

CATEGORÍA	TRIMESTRE	Valencia	León
Refrescos	T1	13.267	3.589
Refrescos	T2	27.392	4.278
Refrescos	T3	73.042	3.780
Refrescos	T4	18.391	3.629
Congelados	T1	150.242	4.798
Congelados	T2	173.105	3.564
Congelados	T3	163.240	4.309
Congelados	T4	190.573	4.812



**pivot**

categoría × ciudad

CATEGORÍA	TRIMESTRE	Refrescos	Congelados
Valencia	T1	13.267	150.242
Valencia	T2	27.392	173.105
Valencia	T3	73.042	163.240
Valencia	T4	18.391	190.573
León	T1	3.589	4.798
León	T2	4.278	3.564
León	T3	3.780	4.309
León	T4	3.629	4.812

## DATA-WAREHOUSING \* SUMARIZACIÓN Y OPERADORES

- ▶ **Slice & dice.** Este operador permite escoger parte de la información mostrada, no por agregación sino por selección.

CATEGORÍA	TRIMESTRE	Valencia	León
Refrescos	T1	13.267	3.589
Refrescos	T2	27.392	4.278
Refrescos	T3	73.042	3.780
Refrescos	T4	18.391	3.629
Congelados	T1	150.242	4.798
Congelados	T2	173.105	3.564
Congelados	T3	163.240	4.309
Congelados	T4	190.573	4.812



**slice & dice**

trimestre = {T1, T4}  
ciudad = Valencia

CATEGORÍA	Trimestre	Valencia
Refrescos	T1	13.267
Refrescos	T4	18.391
Congelados	T1	150.242
Congelados	T4	190.573

# DATA-WAREHOUSING

- ▶ Necesidad de los Almacenes de Datos:
  - ▶ No son imprescindibles para hacer KDD pero sí convenientes.
- ▶ Especialmente indicada para las dos tipologías de usuarios:
  - ▶ ‘picapedreros’ (o ‘granjeros’): se dedican fundamentalmente a realizar informes periódicos, ver la evolución de determinados parámetros, controlar valores anómalos, etc.
  - ▶ ‘exploradores’: encargados de encontrar nuevos patrones significativos utilizando técnicas de minería de datos.

# DATA-WAREHOUSING

- ▶ Recogida de Información Externa:
  - ▶ Aparte de información interna de la organización, los almacenes de datos pueden recoger información externa:
    - ▶ Demografías (censo), páginas amarillas, psicografías, gráficos web, información de otras organizaciones.
    - ▶ Datos compartidos en una industria o área de negocio, organizaciones y colegios profesionales, catálogos, etc.
    - ▶ Datos resumidos de áreas geográficas, distribución de la competencia, evolución de la economía, información de calendarios y climatológicas, programaciones televisivas deportivas, catástrofes,...
    - ▶ Bases de datos externas compradas a otras compañías.

# PROYECTOS