Universidad Zaragoza

Almacenes de datos

Grado en Ingeniería en Informática y programa conjunto MAT-INF

Universidad Zaragoza

Curso 2023-2022

Fernando Tricas García (ftricas@unizar.es)

Raquel Trillo Lado (raqueltl@unizar.es)

Carlos Tellería Orriols (telleria@unizar.es)

Dpto. Informática e Ingeniería de Sistemas

Universidad Zaragoza
1542

Guion

- **Introducción**
  - Data Warehouses
  - Características: entornos OLTP y OLAP
- **Construcción de Data Warehouses**
  - Arquitecturas
  - Procesos ETL
  - Modelos multidimensionales
- **Referencias**

Universidad Zaragoza

Introducción

Problema

- Las organizaciones manejan enormes cantidades de datos…
  - … en distintos formatos.
  - … que residen en distintas bases de datos.
  - … organizados utilizando distintos tipos de gestores de bases de datos

Consecuencia

- Resulta difícil acceder y utilizar todos los datos en aplicaciones de análisis (las cuales requieren extraer, preparar e integrar los datos)

Universidad Zaragoza

Introducción

Data Warehouse (Def.):

1. Repositorio
2. Datos estructurados
3. A nivel de empresa
4. Datos históricos y actuales
5. Facilita la toma de decisiones

Business Intelligence

Universidad Zaragoza
1542

Introducción

Tipos de sistemas de información

Transaccionales (OLTP)	Analíticos (OLAP)
Datos operacionales	Datos consolidados (suelen provenir de distintas BD OLTP)
Muchas transacciones (INSERT, UPDATE, DELETE)	Pocas transacciones
Datos actuales	Datos actuales e históricos
Información detallada	Información detallada y resumida (integrada) (Consultas complejas – agregaciones □ Data mining)
Los datos cambian continuamente (volátiles)	Datos con mayor estabilidad y menos cambios (no volátiles)

Universidad Zaragoza
1542

Introducción

Características de los almacenes de datos

Orientados a un aspecto concreto

- La información en base a un tema de **interés para los directivos de la entidad** y no para facilitar la operatividad diaria:
  - Ej. Para un empresa dedicada al comercio en torno a las ventas, productos y proveedores

Integrados

- El almacén de datos suele contener, entre otros, todos los datos de los sistemas operacionales de la organización (empresa)
- Dichos datos deben ser consistentes
  - Ej. Agrega los datos de los sistemas de ventas, compras de productos, campañas de *marketing*, recursos humanos, etc.

Universidad Zaragoza
1542

Introducción

Características de los almacenes de datos

No volátiles

- Una vez los **datos** han sido incorporados al sistema (registrados) estos **no se borran ni actualizan**. Además están pensados para un horizonte de tiempo mucho mayor que los datos operacionales
  - Inserciones/borrados/actualizaciones constantes vs. lectura/agregación de datos
  - La misma consulta sobre el mismo periodo temporal siempre produce el mismo resultado

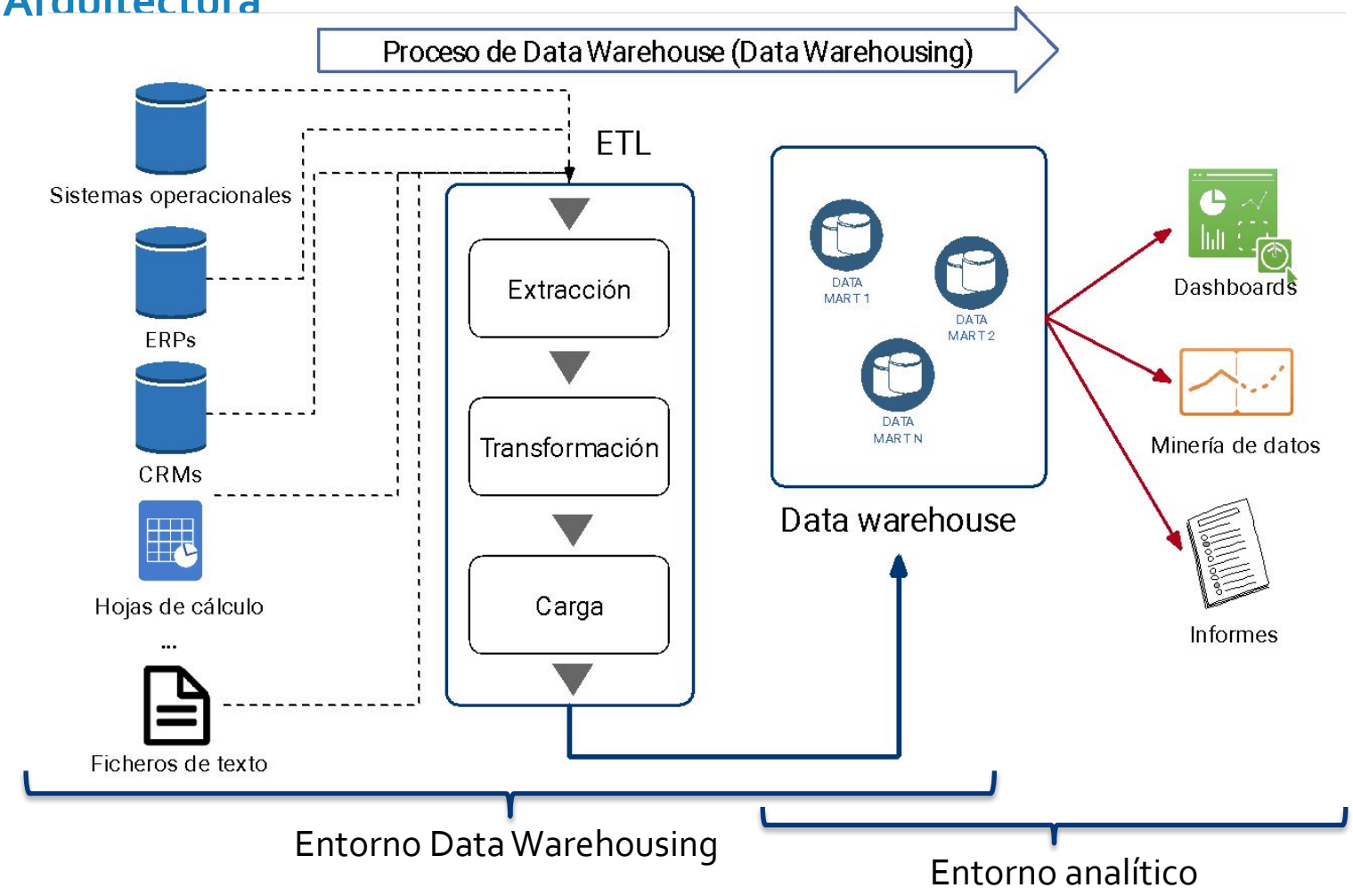
Universidad Zaragoza
1542

Guion

- **Introducción**
  - Data Warehouses
  - Características: entornos OLTP y OLAP
- **Construcción de Data Warehouses**
  - Arquitecturas
  - Procesos ETL
  - Modelos multidimensionales
- **Referencias**

Construcción de Data Warehouses

Arquitectura



Universidad Zaragoza
1542

Arquitectura

Procesos ETL

Extracción

- Heterogeneidad en las fuentes de datos
  - Normalmente: BD relacionales (OLTP), ERP, CRM, incluso ficheros de texto plano
  - Datos: estructurados, semi-estructurados o no estructurados
- La extracción puede llevarse a cabo:
  - Para realizar una imagen inicial
  - Para actualizar una imagen ya existente
- Muy costoso en tiempo, puede afectar al rendimiento de los sistemas fuentes de datos

Universidad Zaragoza

Arquitectura

Procesos ETL

Transformación

Arquitectura

Procesos ETL

Carga

- Se cargan los datos de la fase anterior (*Transformación*)
- Dos métodos
  - Carga completa
    - Primera carga (imagen inicial)
  - Carga incremental
    - Carga en intervalos de tiempo regulares y planificados
      - *Streaming* (volúmenes pequeños de datos)
      - Por lotes (grandes volúmenes de datos)
    - Mantenimiento de históricos (se puede hacer el seguimiento temporal de un dato)

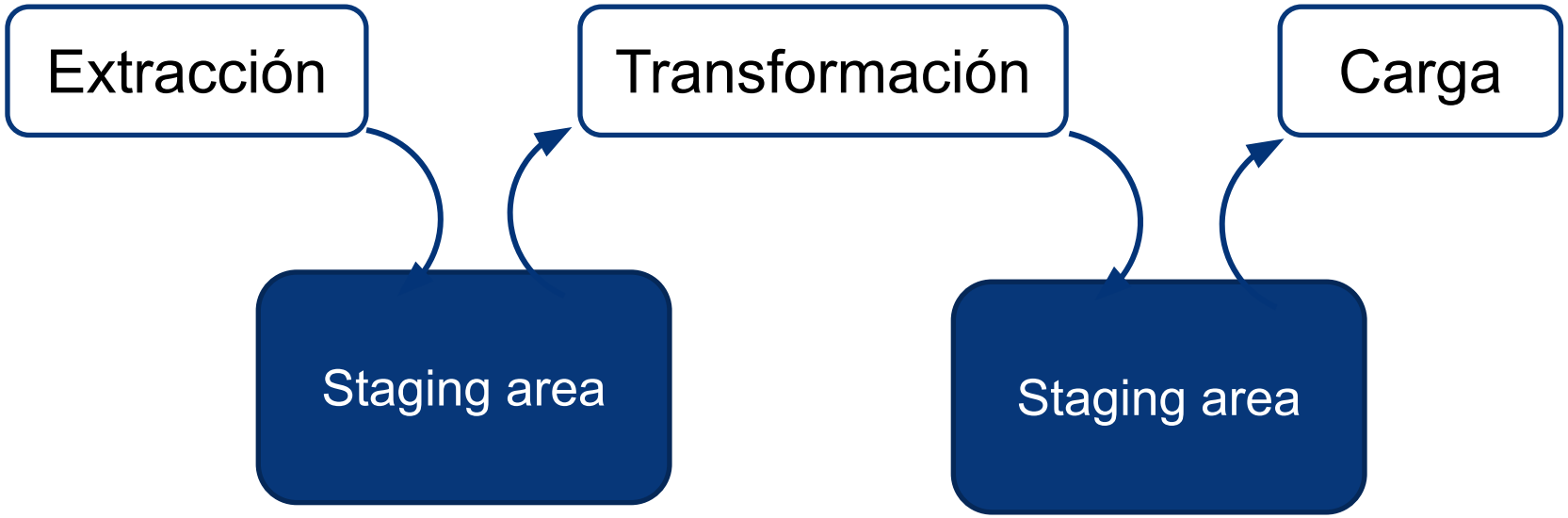
Universidad Zaragoza

Arquitectura

Procesos ETL

Staging Area

- Facilita los procesos de extracción y transformación de los datos antes de ser incluidos en el Data Warehouse



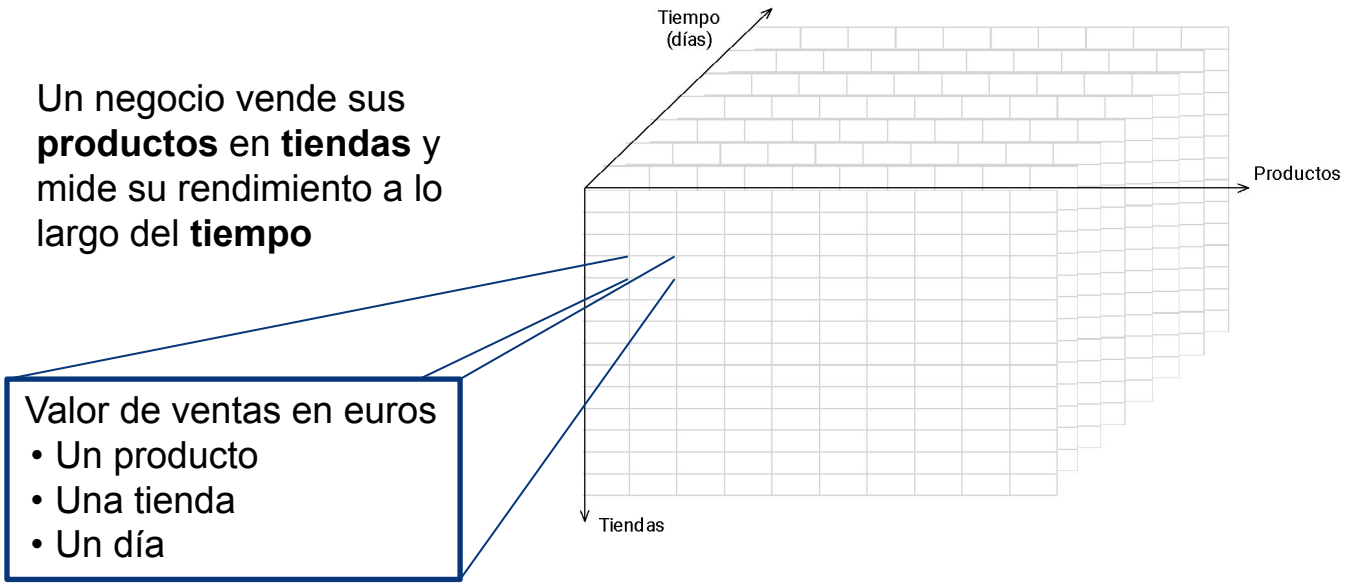
Universidad Zaragoza

Arquitectura

Modelos multidimensionales

- O cómo organizar los datos en un DW
- Cubo (n-dimensionales, hipercubo)
  - Estructura que se emplea para organizar los datos en el Data Mart. Tiene múltiples dimensiones.
  - Ej. 3 dimensiones

Un negocio vende sus **productos** en **tiendas** y mide su rendimiento a lo largo del **tiempo**

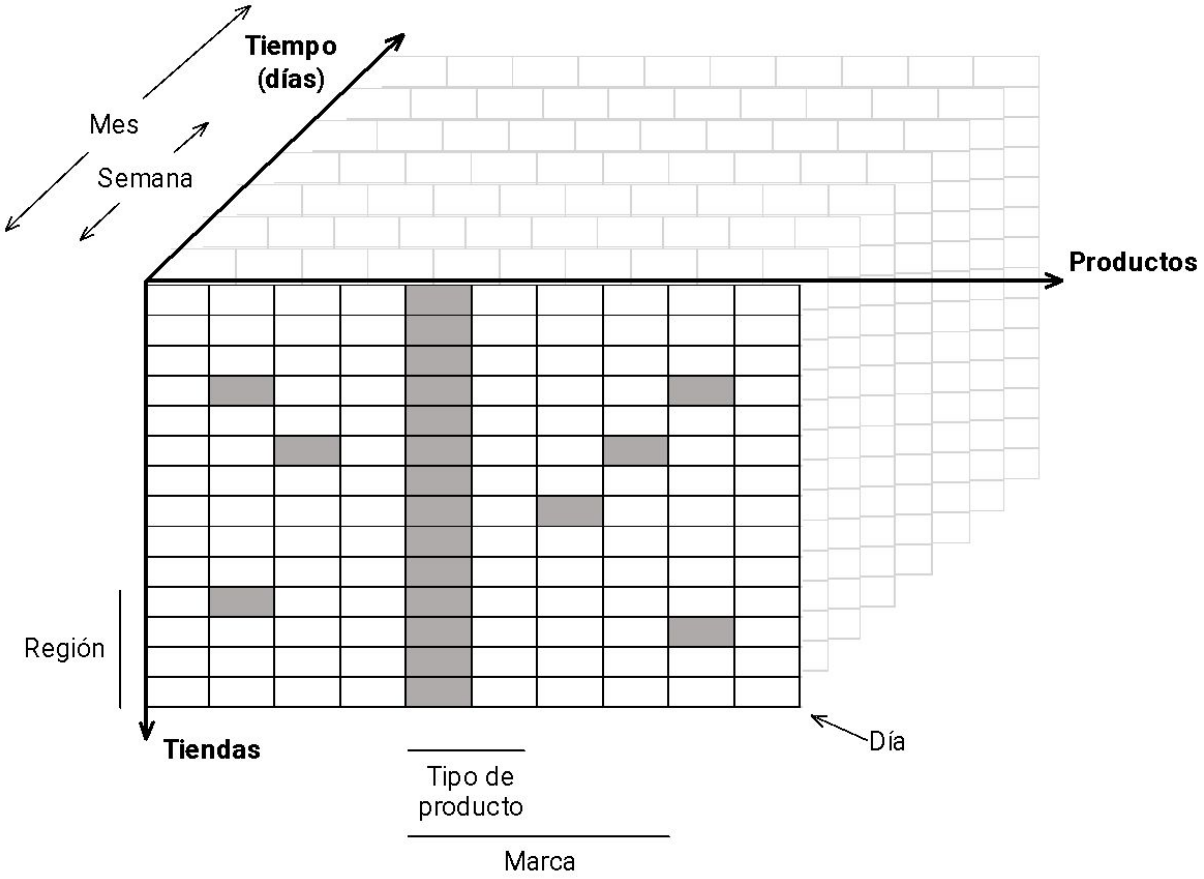


Universidad Zaragoza

Arquitectura

Modelos multidimensionales

- Diferente nivel de detalle en cada una de las dimesiones



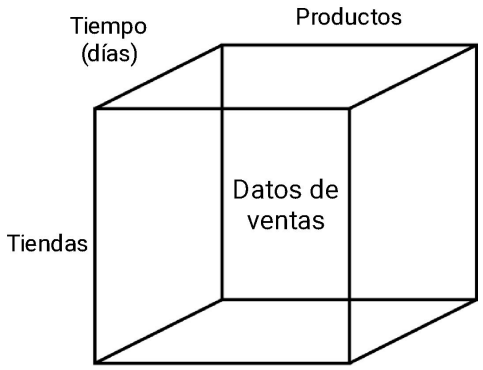
Universidad Zaragoza

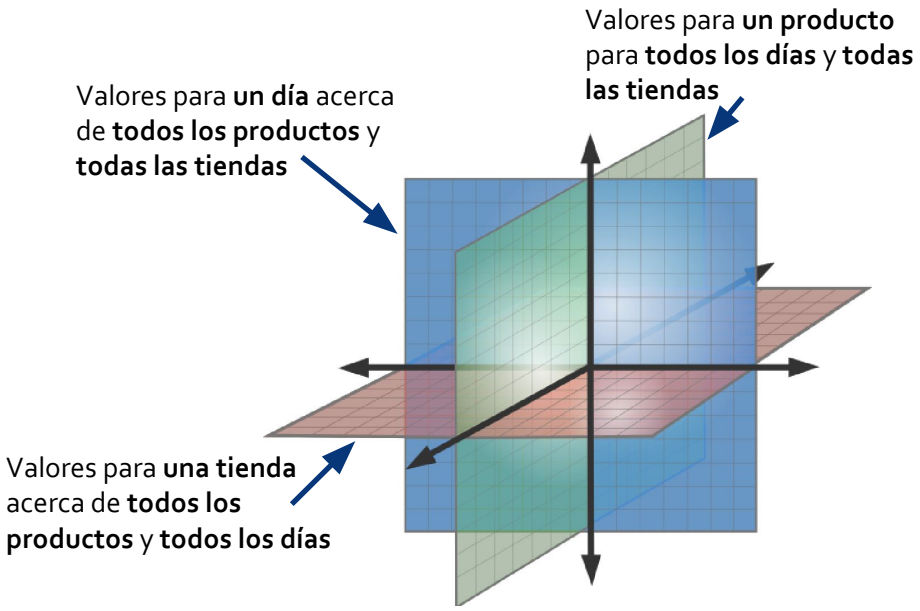
Arquitectura

Modelos multidimensionales

Slice (loncha)

- El subconjunto de datos multidimensionales definidos por **seleccionar valores específicos** para cada uno de los atributos que definen las dimensiones





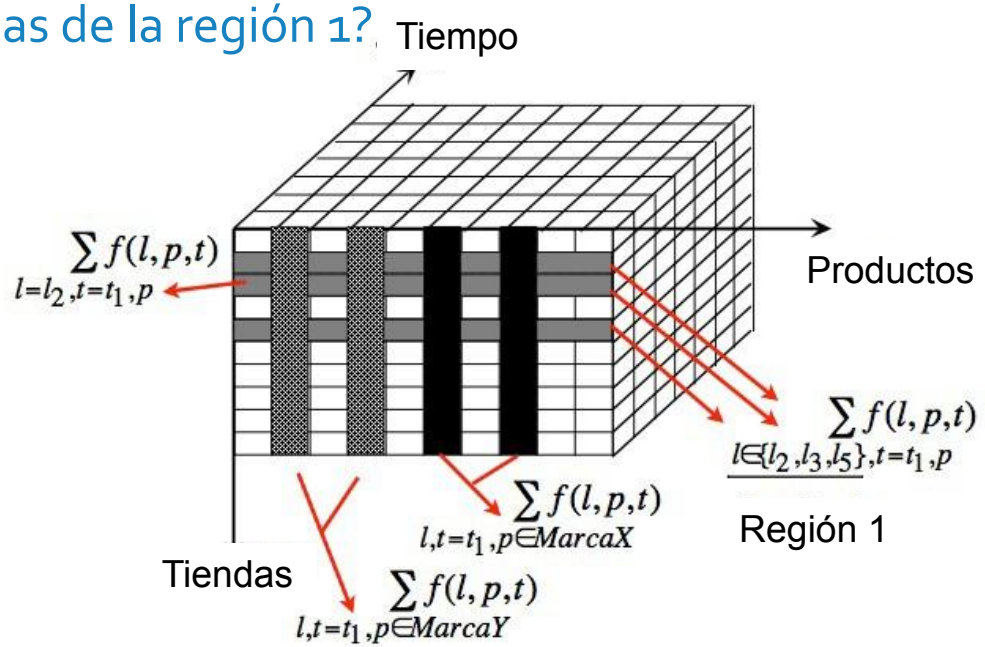
Universidad Zaragoza

Arquitectura

Modelos multidimensionales

Operación básica: la agregación

- ¿Qué cantidad de productos de la marca X se han vendido durante el mes actual en las diferentes tiendas?
- ¿Cuántas ventas de los diferentes productos se han realizado en las tiendas de la región 1?



Arquitectura

Modelos multidimensionales

Implementación de los cubos

Virtual

- Opción más simple:
  - **Una sola tabla** con múltiples **columnas** que representan o bien las **dimensiones** que se consideran o bien los datos de interés para el análisis (en nuestro ejemplo, el número de ventas de un producto)
  - Sigue un esquema en **estrella**

Física

- Bases de datos multidimensionales
  - Matriz n-dimensional almacenando los valores

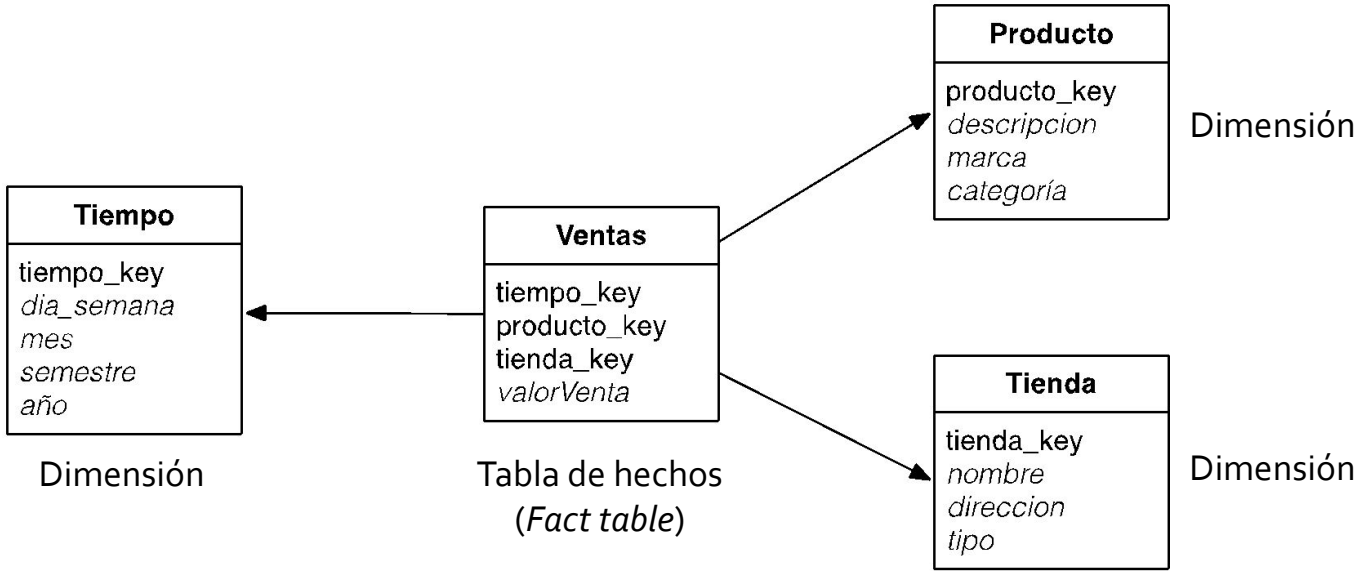
Universidad Zaragoza

Arquitectura

Modelos multidimensionales

Arquitectura en estrella

- Una **tabla central** que contiene la información de los **hechos** que se desea analizar (p. ej. las ventas) **conectada a diferentes tablas** que representan las diferentes **dimensiones**

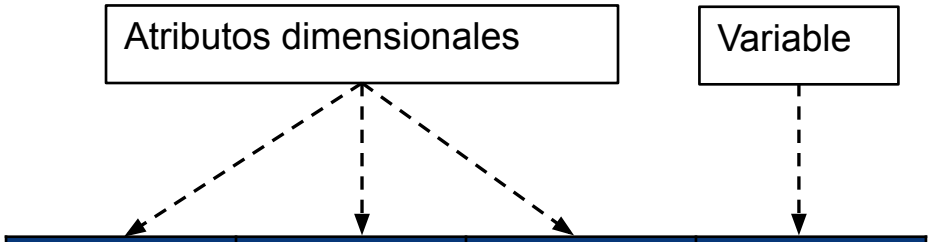


Universidad Zaragoza

Arquitectura

Modelos multidimensionales

- En general:
  - Implementación → Arquitectura de estrella
  - Vista analítica → Arquitectura de una sola tabla



Tiempo	Producto	Tienda	Ventas
2	10	12	225 €
2	11	12	152 €
2	12	12	89 €

Universidad Zaragoza

Arquitectura

Modelos multidimensionales: entorno analítico

Generación de informes

- Son configurables
  - Información a mostrar / Periodicidad
- **Operadores sobre los informes**
  - *Drill down*
    - Detallar los resultados obtenidos añadiendo un campo. Por ejemplo, el periodo temporal
  - *Roll up*
    - Agregar los resultados obtenidos eliminando un campo. Por ejemplo, agregando por marca todos los productos

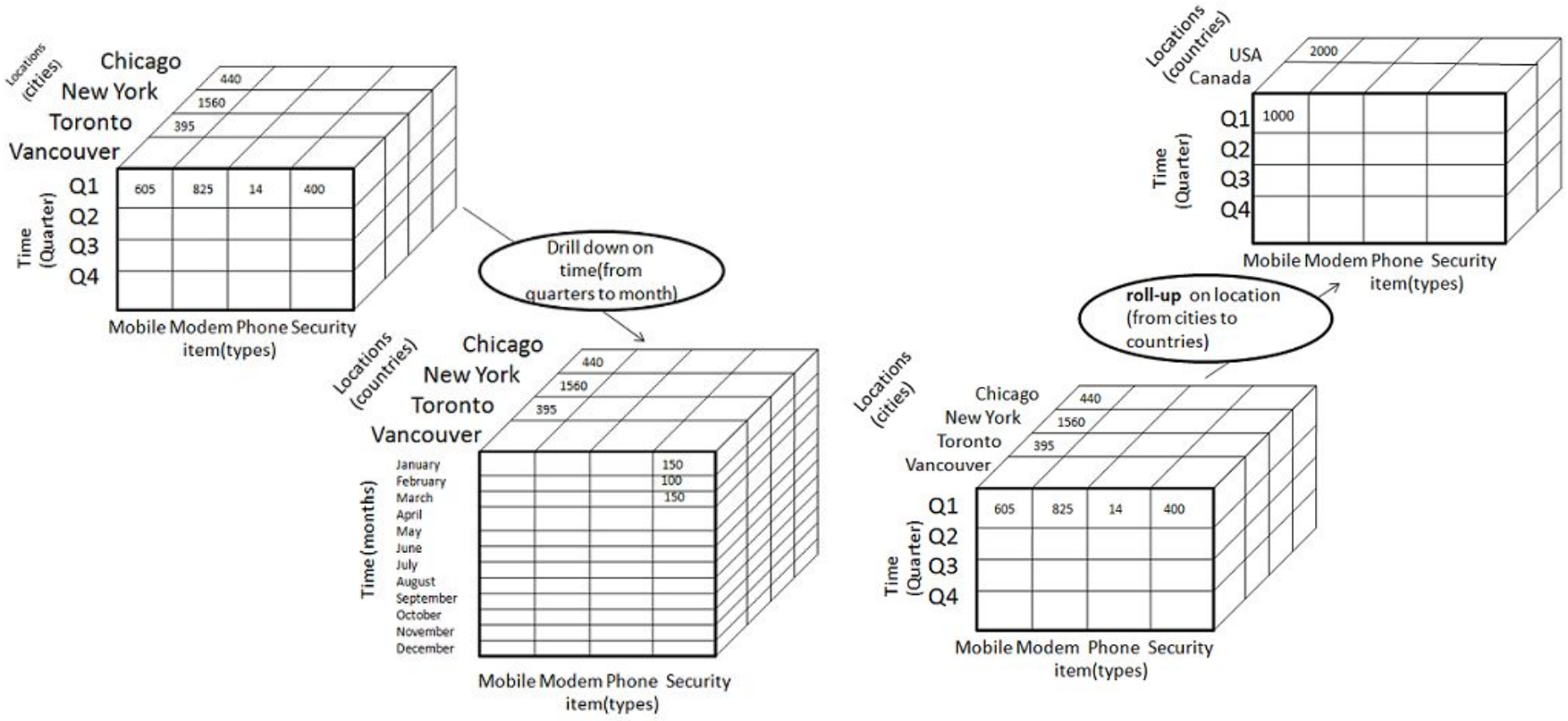
Universidad Zaragoza

Arquitectura

Modelos multidimensionales: entorno analítico

Drill down

Roll up



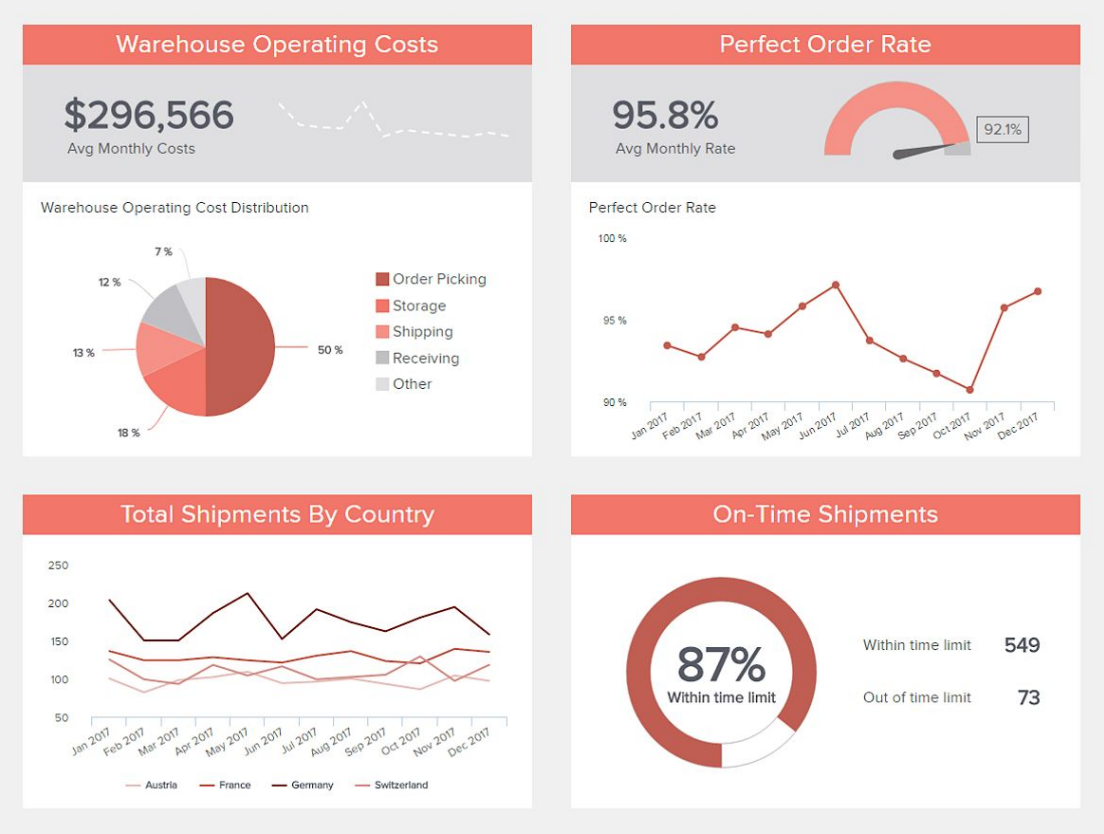
Fuente: https://www.tutorialspoint.com/dwh/dwh_olap.htm

Universidad Zaragoza

Arquitectura

Modelos multidimensionales: entorno analítico

Dashboard (Tablero de mandos)



Universidad Zaragoza
1542

Arquitectura

Factores de éxito

- **Integrar** datos **externos** con los datos de producción **internos** y gestionar **historiales**
- **Considerar** información útil, centrada en los objetivos de la empresa
- Emplear **datos de calidad** (coherentes, actualizados y documentados)
- **Arquitectura flexible** para garantizar escalabilidad (tanto a nivel hardware como a nivel software). Considerar también más usuarios, herramientas, volumen de negocio, etc.

Arquitectura

Errores comunes

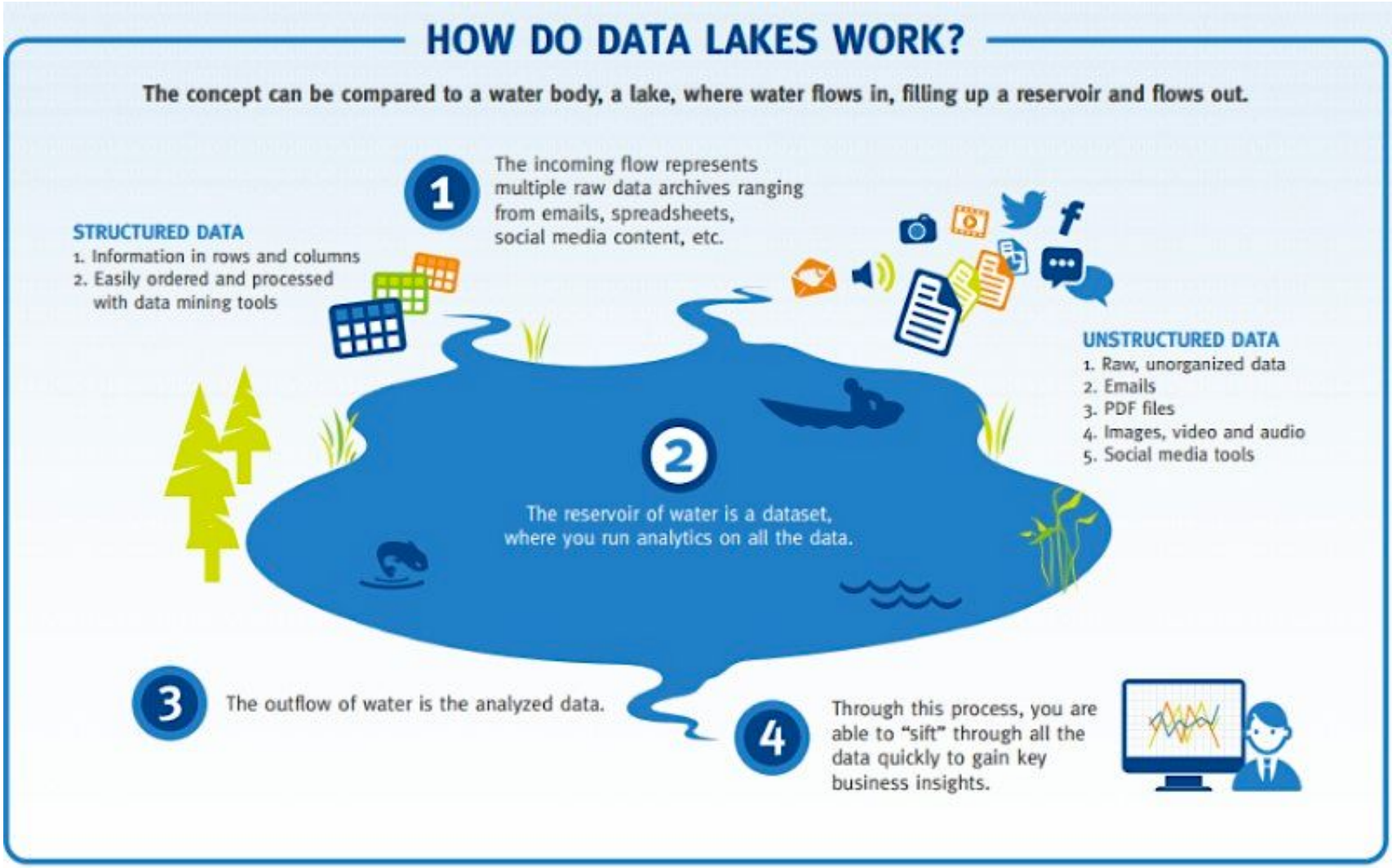
- Incluir datos solamente porque están disponibles (podrían no ser útiles)
- Crear un esquema de **BD relacionales tradicional**
- Crear el Data Warehouse **pensando en la tecnología** que se va a usar para su implementación
- Creer que los Data Warehouses acaban su **ciclo de vida** una vez son cargados los datos e instalado el sistema (incluir herramientas para el diseño de informes)

Temas relacionados (*Data lakes*)

Data warehouse	vs	Data lake
estructurados, preprocesados	DATOS	estructurados, semi-estructurados , no estructurados
esquema al escribir	PROCESAMIENTO	esquema al leer
costoso para grandes volúmenes	ALMACENAMIENTO	Diseñado para bajo coste
menos ágil, configuración fija	AGILIDAD	muy ágil, configuración bajo demanda
madura	SEGURIDAD	en proceso
directivos	USUARIOS	analistas de datos (entre otros)

Universidad Zaragoza

Temas relacionados (*Data lakes*)



Universidad Zaragoza
1542

Ejercicio

Ejercicio de modelado de la inteligencia de negocio para una empresa dedicada al servicio de música por Internet

- Modelo de negocio
- Información importante para la toma de decisiones
- Fuentes de datos
- Diseño del almacén de datos
- …

Universidad Zaragoza
1542

Guion

- **Introducción**
  - Data warehouses
  - Características: entornos OLTP y OLAP
- **Construcción de Data Warehouses**
  - Arquitecturas
  - Procesos ETL
  - Modelos multidimensionales
- **Referencias**

Universidad Zaragoza
1542

Referencias

- Ralph Stair y George Reynolds, Information Systems, 10th edición, International edition
- Imhoff y otros, Mastering Data Warehouse Design: Relational and Dimensional Techniques, 2003, Wiley
- J. M. Franco, El data warehouse. El Data Mining, 1997, Eyrolles

Universidad Zaragoza

Sistemas de información

Grado en Ingeniería en Informática y programa conjunto MAT-INF

Universidad Zaragoza

Curso 2023-2024

Fernando Tricas García (ftricas@unizar.es)

Raquel Trillo Lado (raqueltl@unizar.es)

Carlos Tellería Orriols (telleria@unizar.es)

Dpto. Informática e Ingeniería de Sistemas