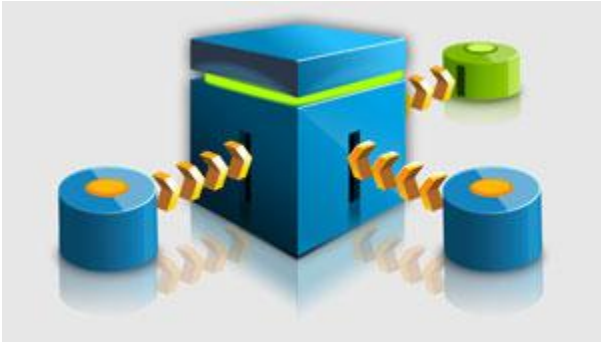
Integración de datos (ETL) Y Almacenes de datos

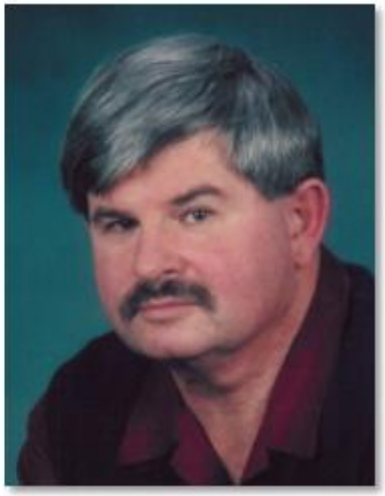


Bases de Datos
Otoño 2012
Maestría en Ingeniería de Software
L.I Yessica Sugeidy Morales Mateo

Antecedentes

A principios de la década de los sesenta, el software de acceso a datos consistía en aplicaciones independientes, basadas en ficheros maestros almacenados en ***cinta magnética***; lo que significaba un acceso secuencial a los datos. La aparición de los discos magnéticos en la década de los setenta representó un cambio cualitativo, éstos permitían el acceso directo a los datos (***DASD, del inglés Direct Access Strorage Device***), favoreciendo el desarrollo de nuevas organizaciones de ficheros. A partir de ese momento se produjo una acelerada evolución en la tecnología de acceso a datos que no ha parado hasta nuestros días.

La primera generación de sistemas de almacenes de datos fue construida sobre ciertos principios establecidos por líderes de la industria. Se reconoce a dos grandes pioneros en el área de Almacenes de Datos: ***Bill Inmon y Ralph Kimball***. Estos dos científicos, han proporcionado las definiciones y los principios de diseño que la mayoría de los profesionales utilizan hoy en día. Aunque sus guías no sean seguidas exactamente, es común hacer referencia a la definición de almacén de datos de Inmon y a las reglas de diseño de Kimball.



Concepto

Bases de Datos es un conjunto de datos persistentes que es utilizado por los sistemas de aplicación de alguna ***empresa*** dada.

Un almacén de datos (data warehouse) es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.

Introducción

El cada vez mayor poder de procesamiento y sofisticación de las herramientas y técnicas analíticas ha dado como resultado la creación de los almacenes de datos.

Proporcionan *almacenamiento, funcionalidad y receptividad* a las consultas que van más allá de las posibilidades de las bases de datos destinadas a transacciones.

Las bases de datos tradicionales equilibran el requisito de acceso a datos con la necesidad de asegurar la integridad de los mismos

Los almacenes de datos difieren de las bases de datos tradicionales en su estructura, funcionamiento, rendimiento y propósito.

W.H. Inmon definió un almacén de datos como:

“un conjunto de datos orientado a temas, integrado, no volátil, variante en el tiempo, como soporte para la toma de decisiones”

Los almacenes de datos proporcionan acceso a datos para análisis complejos, revelación de conocimientos y toma de decisiones.

Dan respuesta a las demandas de alto rendimiento de datos e información de una organización. Soportan varios tipos de aplicaciones, como OLAP, DSS y aplicaciones de minería de datos.

Definiciones

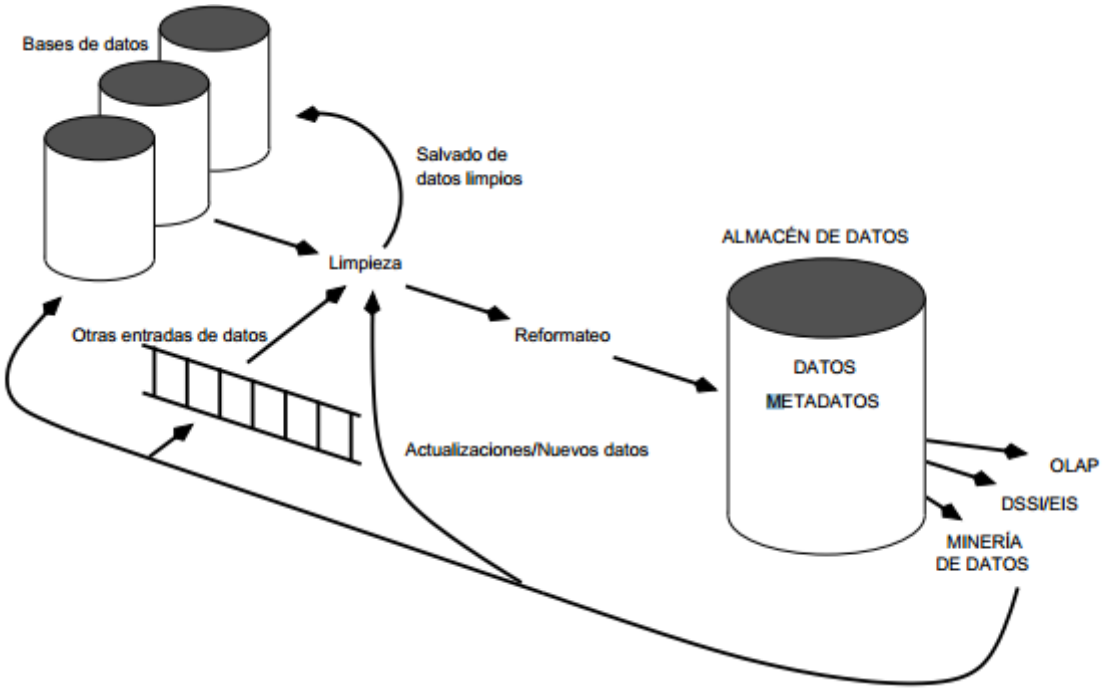
OLAP (on-line analitical processing): análisis de datos complejos del almacén de datos.

Los DSS (decission support systems) proporcionan a las personas que han de tomar decisiones importantes dentro de una organización, datos de nivel superior para la toma de decisiones complejas.

La minería de datos se emplea para el descubrimiento de conocimiento: es un proceso de búsqueda, a partir de los datos, de conocimientos nuevos y no anticipados

Características del Almacén de Datos

Perspectiva general de la estructura conceptual de un almacén de datos:



Los almacenes de datos tienen un orden de magnitud (a veces dos) superior al de las bases de datos fuente.

Este inmenso volumen de datos (probablemente de terabytes) ha sido tratado mediante:

Los almacenes de datos en grandes empresas son proyectos de gran tamaño que requieren una enorme inversión de tiempo y recursos.

Los almacenes de datos virtuales proporcionan vistas de bases de datos operacionales que se materializan para un acceso eficiente.

Los data marts tienen generalmente como objetivo un subconjunto de la organización

Diseño de un almacén de datos

Para la construcción de un **Data Warehouse** se necesitan herramientas para ayudar a la migración y a la transformación de los datos hacia un Almacén.

Data Mart son subconjuntos de datos de un **Data warehouse** para áreas especificas.

Entre las características de un **data mart** destacan:

Usuarios limitados.
Área especifica.
Tiene un propósito especifico.
Tiene una función de apoyo.

Diseño de un Almacén de datos

- ✓Situación actual de partida
- ✓Tipo y característica del negocio
- ✓Entorno técnico
- ✓Expectativas de los usuarios
- ✓Etapa de desarrollo
- ✓Prototipo
- ✓Piloto
- ✓Prueba del concepto tecnológico

Ventajas e inconvenientes de los almacenes de datos

Ventajas

Los almacenes de datos hacen más fácil el acceso a una gran variedad de datos a los usuarios finales.

Facilitan el funcionamiento de las aplicaciones de los sistemas de apoyo a la decisión tales como Informes de tendencia.

Trabajan en conjunto y por lo tanto aumentan el valor operacional de las aplicaciones empresariales, en especial la gestión de relaciones con clientes.

Inconvenientes

Pueden suponer altos costos

Pueden quedar obsoletos relativamente pronto

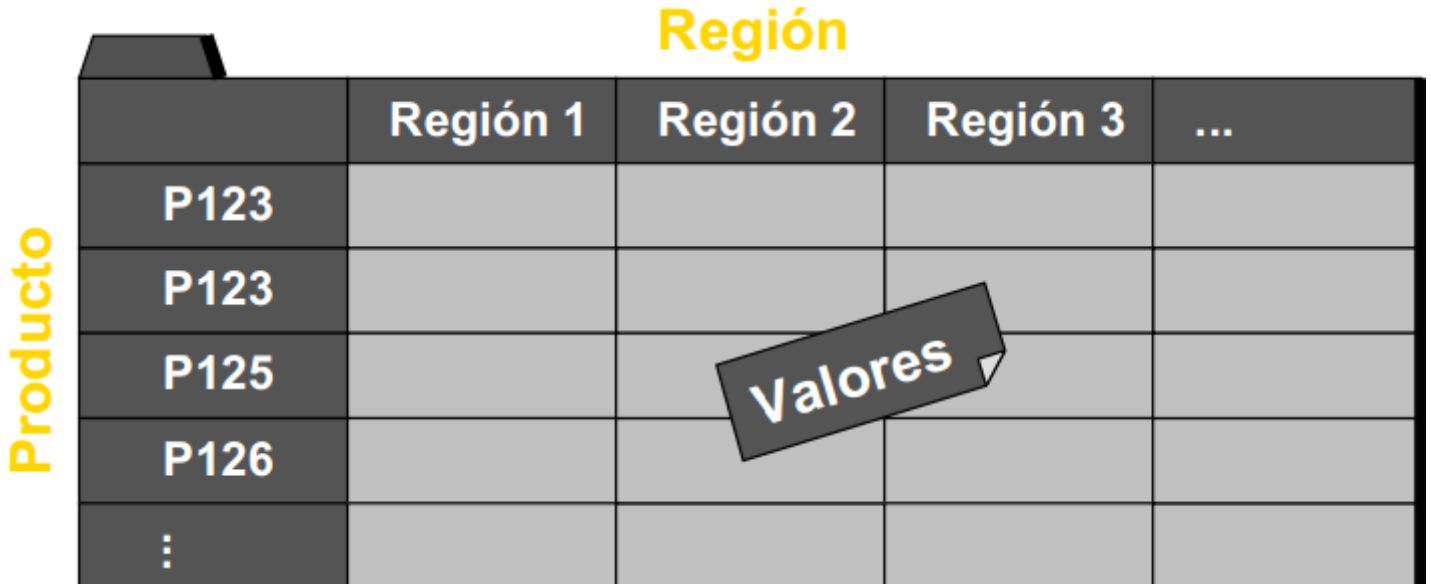
Data Warehousing

Data Warehousing es el proceso de extraer y filtrar los datos de las operaciones comunes a la organización, procedentes de los distintos sistemas de información y/o sistemas externos, para transformarlos, integrarlos y almacenarlos en un depósito o almacén de datos (**Data Warehouse**) con el fin de acceder a ellos para dar soporte en el proceso de toma de decisiones de una organización.

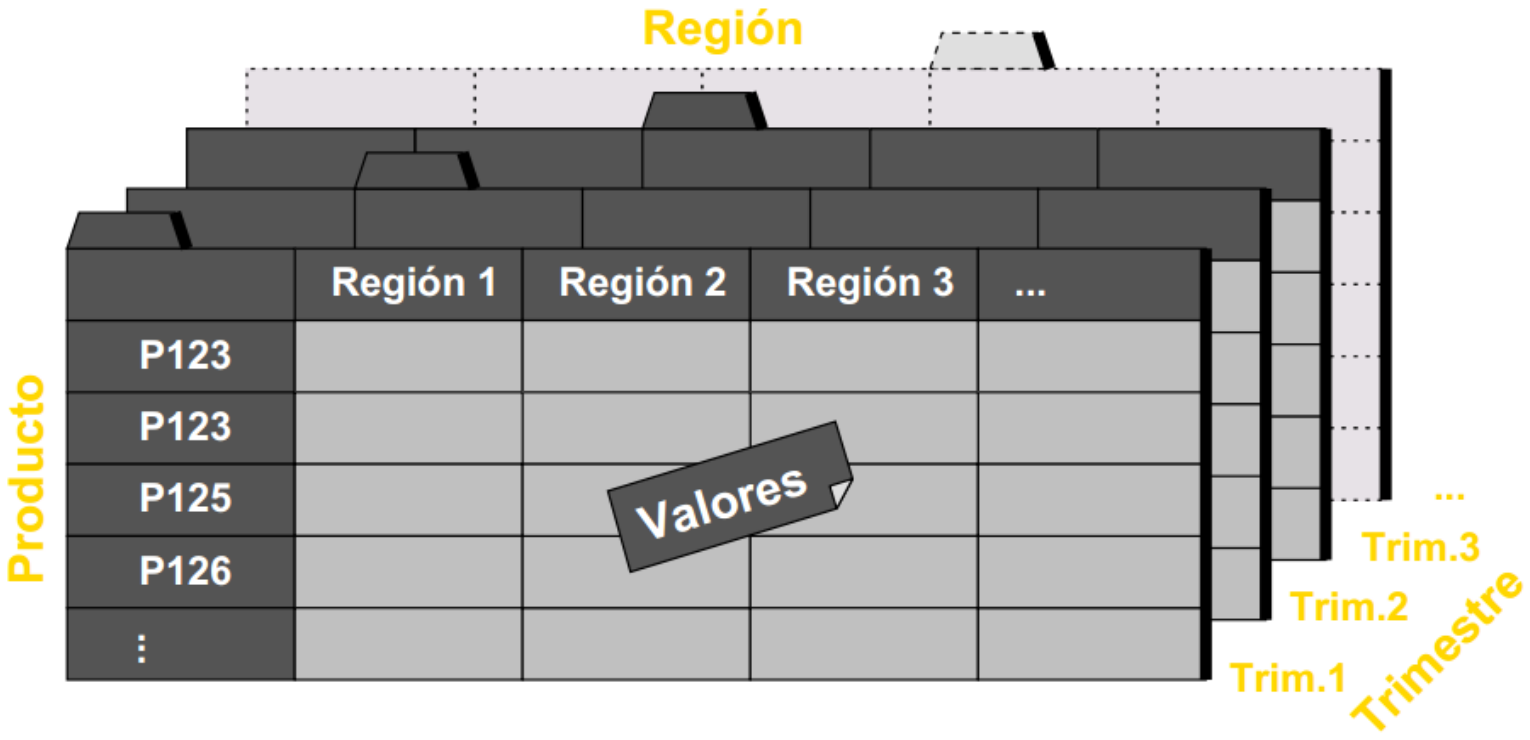
¿Qué diferencia hay entre Data Warehousing y Data Warehouse?

Modelado de datos en almacenes de datos

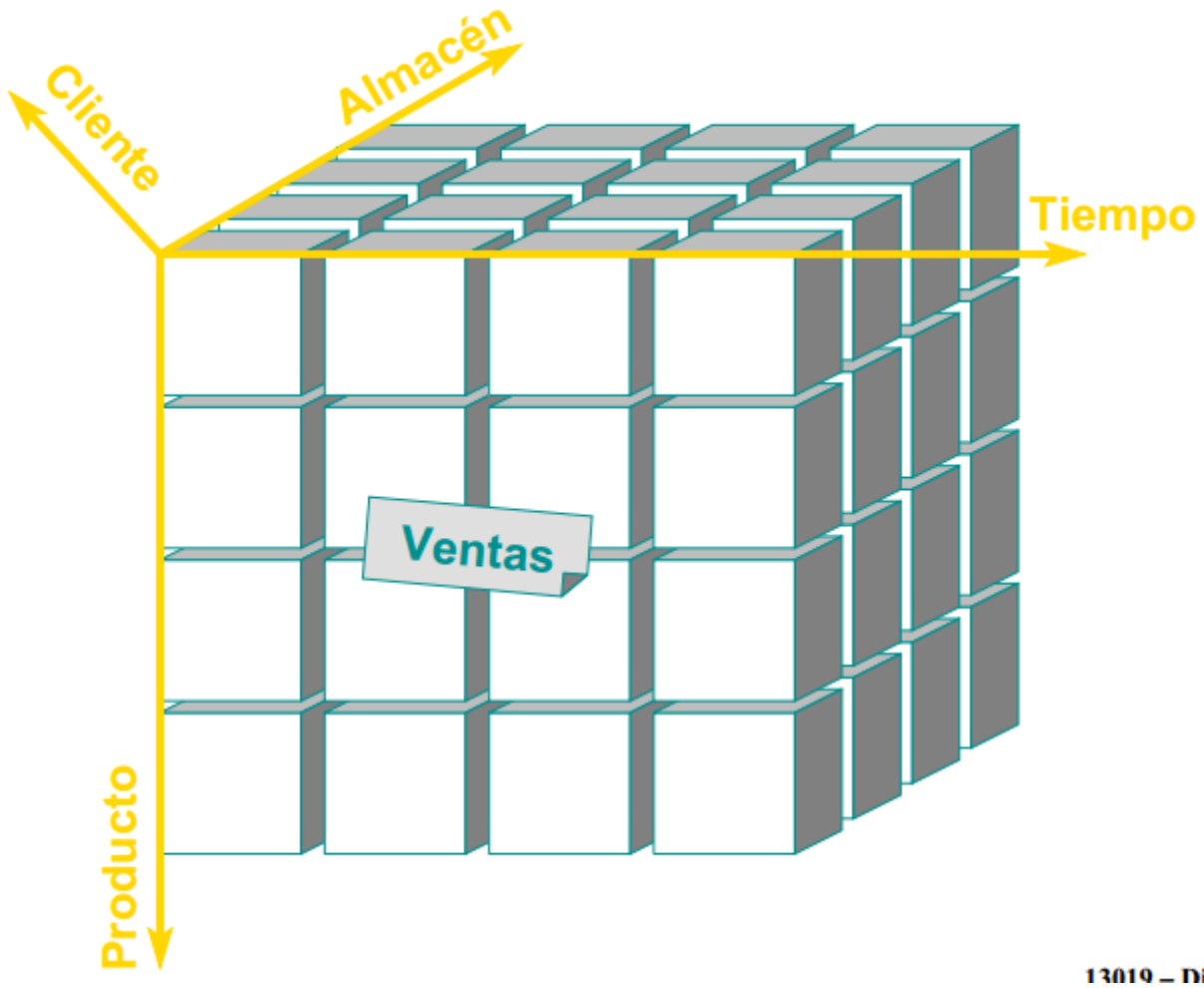
Una hoja de cálculo estándar constituye una matriz bidimensional.



Si añadimos una dimensión temporal tendríamos una matriz tridimensional.



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



13019 – Diseño

Los modelos multidimensionales se prestan fácilmente a representaciones jerárquicas en lo que se conoce como exploración ascendente (roll-up) y exploración descendente (drill- down)

El diseño multidimensional es un método de diseño de bases de datos basado en el modelo relacional.

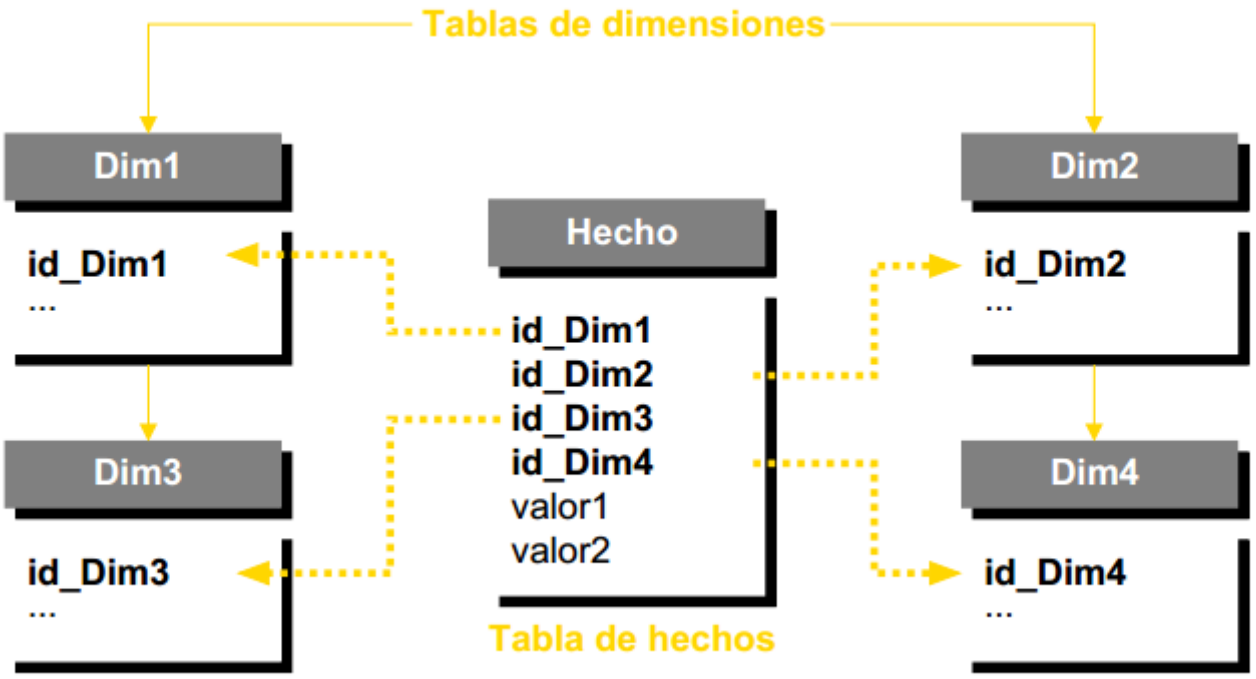
Está compuesto por dos tipos de tablas:

Varias tablas de dimensiones, cada una formada por tuplas de atributos de la dimensión.

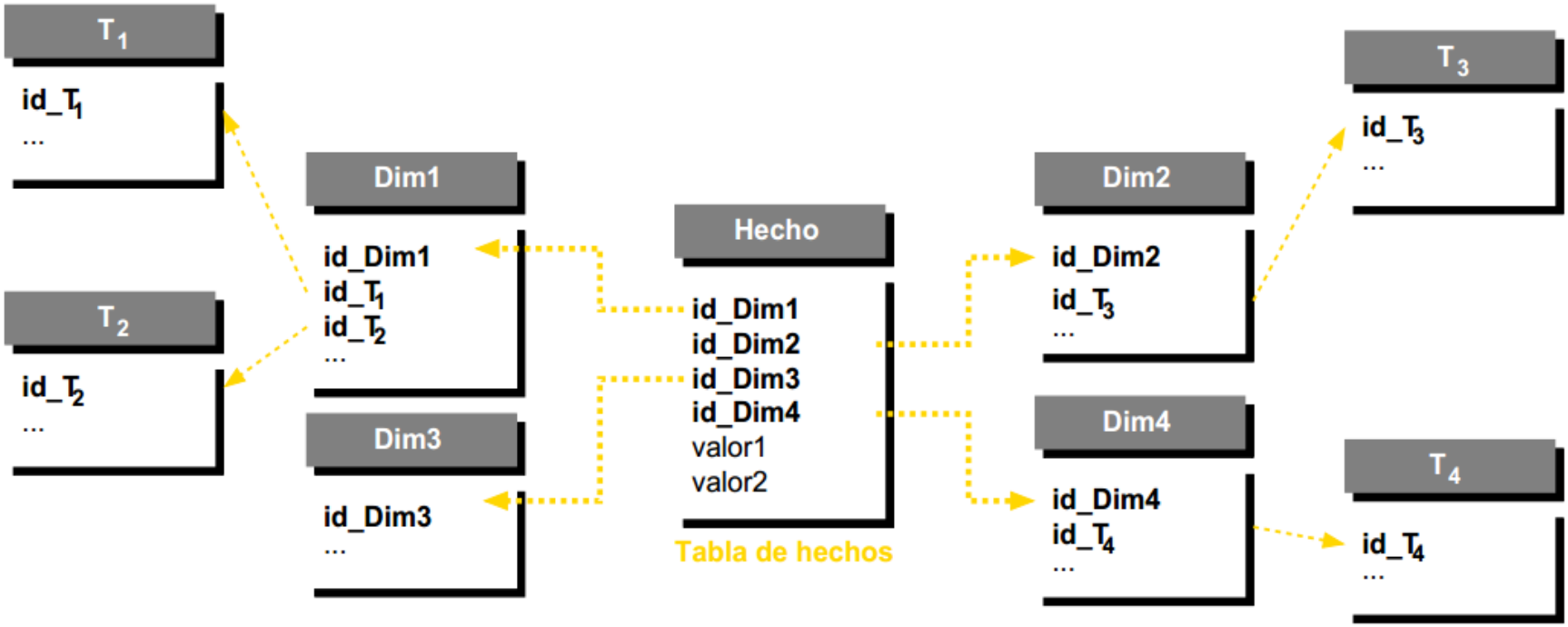
Una tabla de hechos, compuesta por tuplas, una por cada hecho registrado. este hecho contiene alguna variable o variables medidas u observadas y las idéntifica con punteros a las tablas de dimensiones.

Tres son los esquemas multidimensionales comunes:

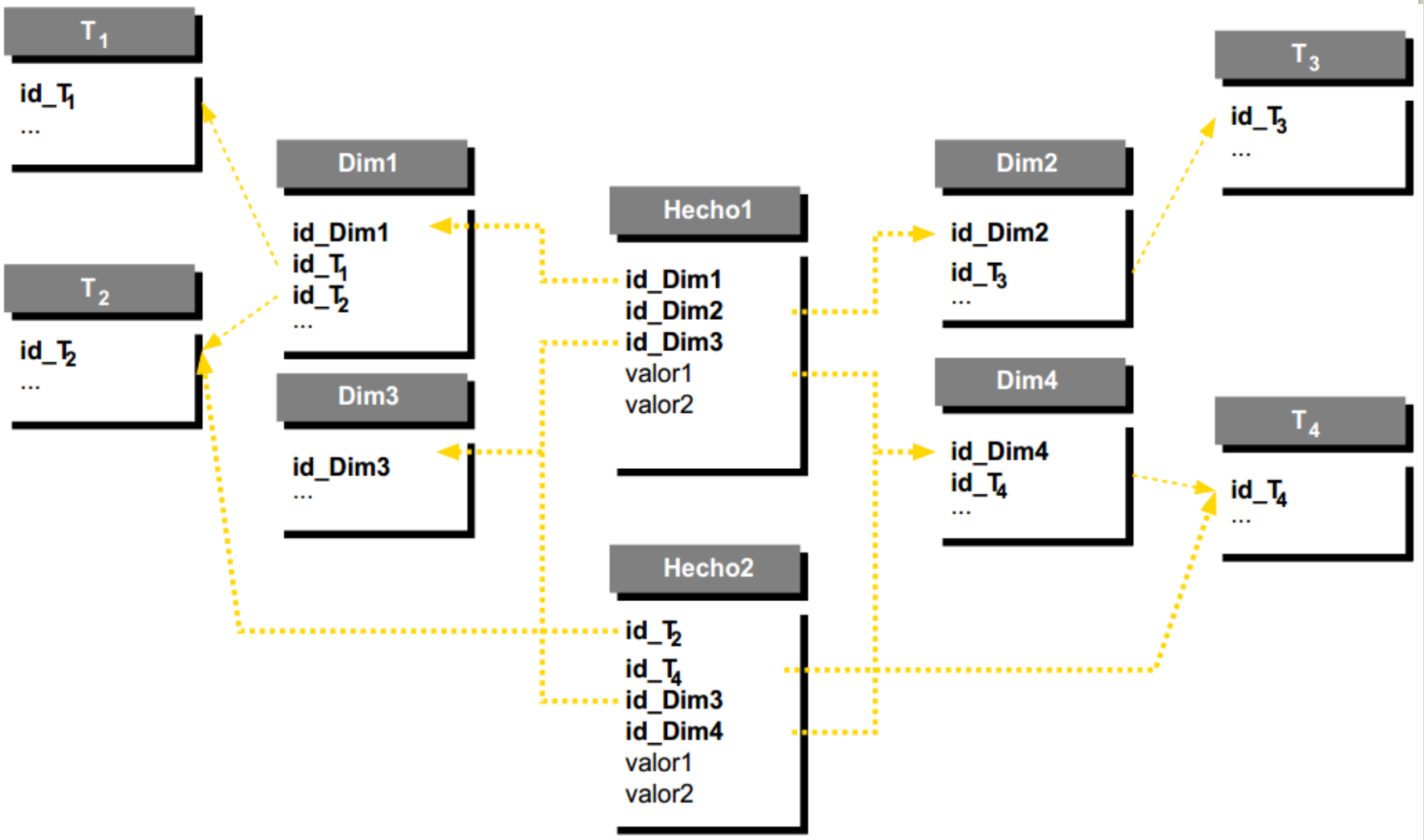
Esquema en estrella: formado por una tabla de hechos con una única tabla para cada dimensión.



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.



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



Construcción de un almacén de datos

Los diseñadores deben tener una amplia perspectiva del uso que se espera del almacén.

No existe un modo de anticipar todas las consultas o análisis posibles durante la fase de diseño.

Sin embargo, el diseño debería soportar específicamente las consultas ad hoc.

Ejemplo: una empresa de productos de consumo con un gran soporte demarketing necesita organizar el almacén de datos de forma diferente a como lo hace otra basada en la recaudación de fondos con fines no lucrativos.

Es necesario seleccionar un esquema adecuado que refleje el uso previsto.

Preparación de los datos

Muchas de las cuestiones que rodean a los sistemas de apoyo para la toma de decisiones, se refieren en primer lugar a las tareas de obtener y preparar los datos.

Los datos deben ser extraídos de diversas fuentes, limpiados, transformados y consolidados en la base de datos de apoyo para la toma de decisiones. Posteriormente, debe ser actualizados periódicamente.

Cada una de estas operaciones involucra sus propias consideraciones especiales.

Extracción

La extracción es el proceso de capturar datos de las bases de datos operacionales y otras fuentes.

Hay muchas herramientas disponibles para ayudar en esta tarea, incluyendo herramientas proporcionadas por el sistema, programas de extracción personalizados y productos de extracción comerciales (de propósito general).

El proceso de extracción tiende a ser intensivo en E/S y por lo tanto, puede interferir con las operaciones críticas.

Limpieza

Pocas fuentes de datos controlan adecuadamente la calidad de los datos.

Los datos requieren frecuentemente de una limpieza antes de que puedan ser introducidos.

Las operaciones de limpieza típicas incluyen:

✓El llenado de valores ausentes, la corrección de errores tipográficos y otros de captura de datos.

✓El establecimiento de abreviaturas y formatos estándares.

✓El reemplazo de sinónimos por identificadores estándares, etcétera.

Los datos que son erróneos y que no pueden ser limpiados, serán reemplazados.

La información obtenida durante el proceso de limpieza puede ser usada para identificar la causa de los errores en el origen y por tanto, mejorar la calidad de datos.

Funcionalidad de los almacenes de datos.

Los almacenes de datos existen para facilitar las consultas complejas, que involucran a gran cantidad de datos y que son con frecuencia ad hoc.

Por lo tanto, deben proporcionar un soporte de consulta mucho mayor y más eficaz que el exigido por las bases de datos transaccionales.

El componente de acceso de los almacenes de datos soporta una funcionalidad de hoja de cálculo extendida, un procesamiento de consultas eficiente, consultas estructuradas, consultas ad hoc y minería de datos.

La funcionalidad de hoja de cálculo extendida incluye un soporte para lo más novedoso en aplicaciones de hojas de cálculo.

Funcionalidad…

También proporciona soporte para programas de aplicaciones OLAP:

Exploración ascendente (roll up): los datos se resumen con una generalización en aumento.

Exploración descendente (drill down): se muestran niveles de detalle cada vez mayores.

Pivotación (rotación): se realiza una tabulación cruzada.

Rodaja y cubo: ejecución de operaciones de proyección en las dimensiones.

Clasificación: los datos se ordenan por valor ordinal.

Atributos derivados (calculados): los atributos se calculan mediante operaciones con valores almacenados y derivados

OLAP (Procesamiento analítico en línea)

La tecnología OLAP facilita el análisis de datos en línea en un DW, proporcionando respuestas rápidas a consultas analíticas complejas.

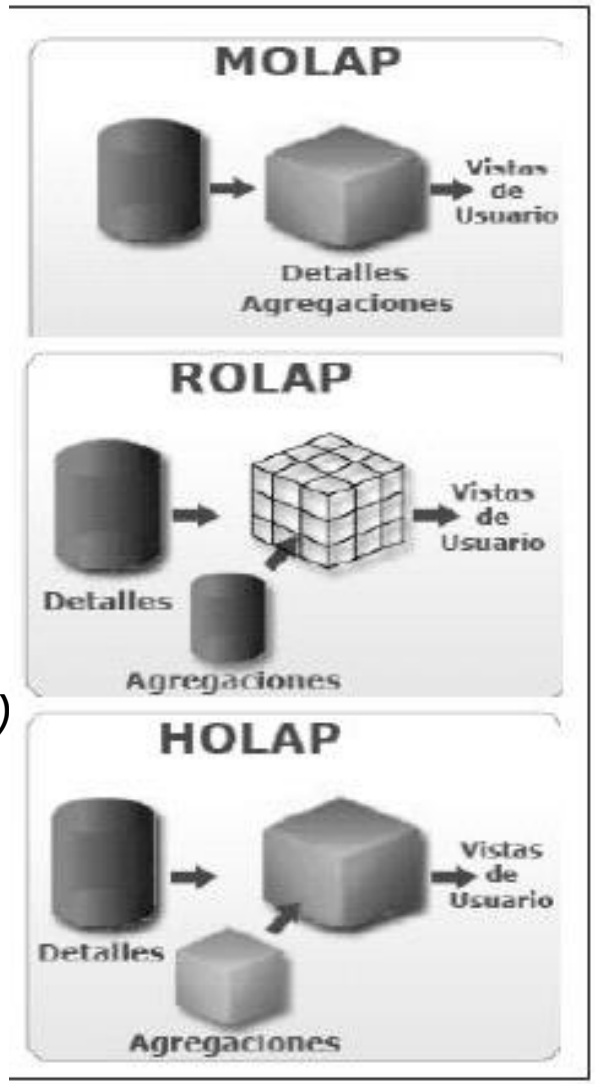
Modos de almacenamiento de OLAP

OLAP puede trabajar con tres tipos de almacenamiento**:**

Almacenamiento MOLAP (multidimensional OLAP)

Almacenamiento ROLAP (Relational OLAP)

Almacenamiento HOLAP (Hybrid OLAP)



pos de almacenamiento

Integración de datos (ETL)

Extract, Transform and Load (*Extraer, transformar y cargar*, frecuentemente abreviado a **ETL**)

Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data marts o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

Extraer , Transformar y cargar.

Procesamiento paralelo

Un desarrollo reciente en el software ETL es la aplicación de procesamiento paralelo. Esto ha permitido desarrollar una serie de métodos para mejorar el rendimiento general de los procesos ETL cuando se trata de grandes volúmenes de datos.

Hay 3 tipos principales que se pueden implementar en las aplicaciones de ETL

- **De datos**
- **De segmentación (pipeline)**
- **De componente**

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. Es recomendable realizar un examen completo de la validez de los datos (**Data profiling**) del sistema de origen durante el análisis para identificar las condiciones necesarias para que los datos puedan ser tratados adecuadamente por las reglas de transformación especificadas. Esto conducirá a una modificación de las reglas de validación implementadas en el proceso ETL.

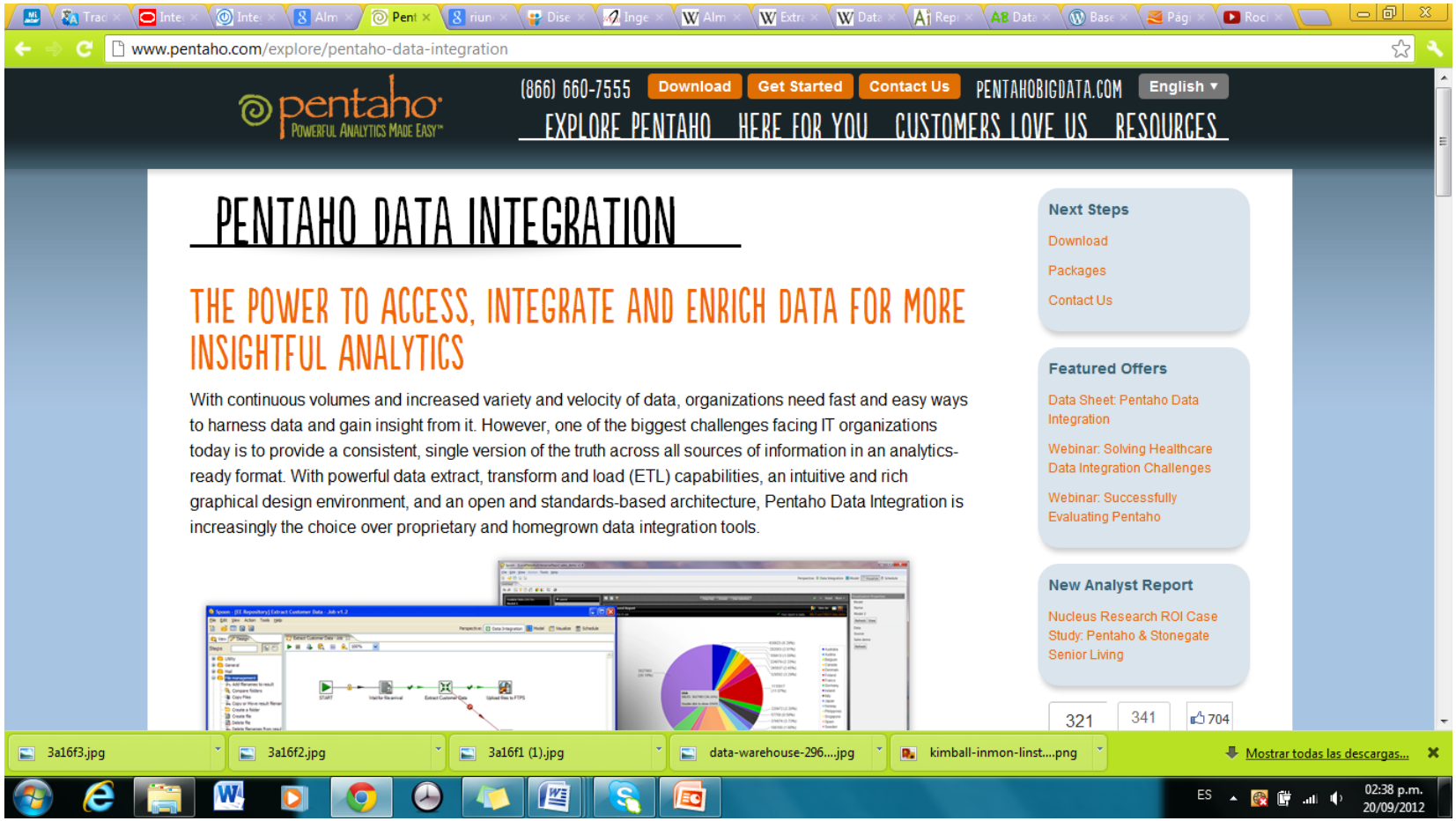
Ejemplo de ETL

XMLoader

Es una completa herramienta diseñada para la Extracción, Transformación y Carga de información entre sistemas informáticos (ETL por su sigla en inglés). Está orientada a facilitar la interacción de MS-Excel y/o archivos planos con bases de datos y sistemas corporativos de mensajería XML

http://www.analitica.com.co/website/es/productos-analitica/representaciones

Pentaho Data Integration es cada vez más la elección sobre las herramientas de datos de propiedad y de cosecha propia integración



Bibliografía.

Audifilm Grupo Brime., Oracle Express technology., Reporte técnico, 2003

Date, C. J., Introducción a los sistemas de bases de datos., Prentice Hall, 2001.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals., Data Mining and Knowledge Discovery, 1997

Anca Vaduva, Klaus R. Dittrich, "Metadata Management for Data Warehousing: Between Vision and Reality", 2001 International Database Engineering & amp; Aplications Symposium (IDEAS´01), Gronoble France.

Effy Oz , Administración de los sistemas de información 5ª. Edición Cengage Learning

James A. Senn Analisis y diseño de sistemas de información 2da edición MC Graw Hill

http://riunet.upv.es/bitstream/handle/10251/2505/tesisUPV2842.pdf

http://es.wikipedia.org/wiki/Almac%C3%A9n_de_datos

Gracias!