

Universidade da Coruña

Máster en Técnicas Estadísticas

Asignatura de Análisis Exploratorio de Datos

Primer Cuatrimestre 2009/10

Trabajo:

Data Warehouse

Alumnos:

- **Ana Karina López Pallas**
- **Mónica Rodríguez Teijeiro**
- **José Benito Pérez López**

Contenido

1.	Introducción	3
1.1.	Objetivo	3
1.2.	Organización del documento	3
2.	Contexto del Data Warehousing	4
2.1.	Data Warehouse en el contexto del Business Intelligence.....	4
2.2.	Data Warehouse en el contexto de la Minería de Datos y KDD	6
3.	Data Warehousing	11
1.1	Objetivos del Data Warehouse.....	11
1.2	Infraestructura de un Data Warehouse	12
1.3	Elementos del Data Warehouse	13
1.4	Arquitectura del Data Warehouse	15
1.5	Diseño de los datos de un Data Warehouse.....	17
1.6	Proyecto de creación de un Data Warehouse	21
4.	Referencias.....	22

1. Introducción

1.1. Objetivo

El objetivo del trabajo es entender el significado de Data Warehouse y demás conceptos asociados, como parte de un proyecto de minería de datos.

Para ello el informe presentará los conceptos básicos asociados a un Data Warehouse: objetivos, estructura, arquitectura, elementos, enfoques de diseño, problemática de un proyecto destinado a su establecimiento. Se incluyen conceptos relacionados como Data Mart, DSS, EIS, OLTP, OLAP.

1.2. Organización del documento

En el siguiente apartado describe el contexto del Data Warehouse, desde un enfoque de Business Intelligence y como elemento de un proyecto de Minería de Datos, y describiendo los principales conceptos relacionados con el Data Warehouse.

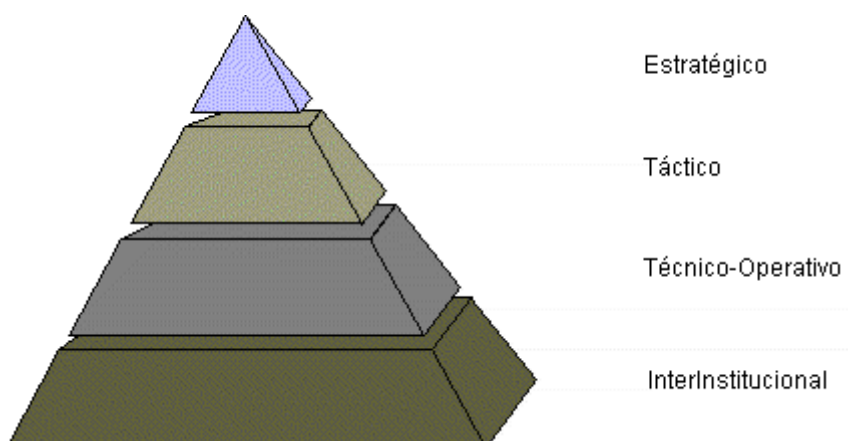
El capítulo tercero detallará los principales aspectos del Data Warehouse y el cuarto incluye las principales referencias utilizadas en el trabajo.

2. Contexto del Data Warehousing

2.1. Data Warehouse en el contexto del Business Intelligence

Uno de los principales activos de las organizaciones es su información, y a medida que cobra importancia su uso, aumentan las necesidades de los sistemas de información corporativos.

Los sistemas de información de una organización se pueden dividir en cuanto a su función y perfil de uso en los siguientes niveles:



- Sistemas Estratégicos, orientados a soportar la toma de decisiones, facilitan la labor de la dirección, proporcionándole un soporte básico, en forma de mejor información, para la toma de decisiones. Se caracterizan porque son sistemas sin carga periódica de trabajo, es decir, su utilización no es predecible, al contrario de los casos anteriores, cuya utilización es periódica.
 - Destacan entre estos sistemas: los Sistemas de Información Gerencial (MIS), Sistemas de Información Ejecutivos (EIS), Sistemas de Información Georeferencial (GIS), Sistemas de Simulación de Negocios (BIS y que en la práctica son sistemas expertos o de Inteligencia Artificial - AI).
- Sistemas Tácticos, diseñados para soportar las actividades de coordinación de actividades y manejo de documentación, definidos para facilitar consultas sobre información almacenada en el sistema, proporcionar informes y, en resumen, facilitar la gestión independiente de la información por parte de los niveles intermedios de la organización.
 - Destacan entre ellos: los Sistemas Ofimáticos (OA), Sistemas de Transmisión de Mensajería (Correo electrónico y Servidor de fax), coordinación y control de tareas (Work Flow) y tratamiento de documentos (Imagen, Trámite y Bases de Datos Documentales).

- Sistemas Técnico - Operativos, que cubren el núcleo de operaciones tradicionales de captura masiva de datos (Data Entry) y servicios básicos de tratamiento de datos, con tareas predefinidas (contabilidad, facturación, almacén, presupuesto, personal y otros sistemas administrativos). Estos sistemas están evolucionando con la irrupción de sensores, autómatas, sistemas multimedia, bases de datos relacionales más avanzadas y data warehousing.
- Sistemas Interinstitucionales, este último nivel de sistemas de información recién está surgiendo, es consecuencia del desarrollo organizacional orientado a un mercado de carácter global, el cual obliga a pensar e implementar estructuras de comunicación más estrechas entre la organización y el mercado (Empresa Extendida, Organización Inteligente e Integración Organizacional), todo esto a partir de la generalización de las redes informáticas de alcance nacional y global (INTERNET), que se convierten en vehículo de comunicación entre la organización y el mercado, no importa dónde esté la organización (INTRANET), el mercado de la institución (EXTRANET) y el mercado (Red Global).

Todos estos tipos de sistemas de información requieren el almacenamiento y uso de datos, con lo que uno de los elementos principales de su arquitectura son las bases de datos.

Los requerimientos de esas bases de datos varían grandemente por el tipo de procesamiento de la información almacena, que puede ser de dos tipos principalmente:

- OLTP (*On-Line Transactional Processing*): procesamiento transaccional en tiempo real, como el que se produce en los sistemas Interinstitucionales, Operativos y Tácticos; generalmente con poca profundidad histórica y cobertura vertical (departamental). El tipo de bases de datos que requieren suele denominarse Bases de Datos Operacionales
- OLAP (*On-Line Analytical Processing*): procesamiento analítico en tiempo real, como el que se produce en los sistemas estratégicos y técnicos, que requiere alta profundidad histórica y cobertura horizontal (diferentes departamentos o áreas). El tipo de base de datos que requieren son conocidas como Data Warehouse.

De esta formas las principales características de un Data Warehouse frente a una Base de Datos Operacional son:

Base de Datos Operacional	Data Warehouse
Datos Operacionales	Datos del negocio para Información
Orientado a la aplicación	Orientado al sujeto
Actual	Actual + histórico
Detallada	Detallada + más resumida
Cambia continuamente	Estable

De esta forma el concepto de **Data Warehouse** o Almacén de Datos surge en la década de los noventa [Inmon 1992] con la definición “es una colección de datos orientado a temas, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones de dirección”. Por tanto surge como una herramienta de apoyo al proceso de toma de decisiones, DSS (*Decision Support System*), y ligado principalmente a los Sistemas de Información para Ejecutivos de una organización o EIS (*Executive Information System*).

Este enfoque del Data Warehouse corporativo ha ido calando en la arquitectura de los sistemas de información corporativos, con lo que se ha ido enriqueciendo con tecnologías (software y hardware) y metodologías (diseño de bases de datos y de software, planificación, ...) específicas, con lo que ha dado en llamar Business Intelligence (BI).

2.2. Data Warehouse en el contexto de la Minería de Datos y KDD

Pero la definición anterior deja fuera otros entornos (científico, ingenieril, ...) y otros perfiles de uso (Minería de Datos) en los que el Data Warehouse tiene hoy en día una importancia capital.

Tal y como detallaremos más adelante, en este trabajo vamos a definir Data Warehouse como un elemento de la Minería de Datos y del proceso de descubrimiento de conocimiento a partir de bases de datos (KDD).

La Minería de Datos surge como un proceso o metodología que busca el descubrimiento de conocimiento procesable en los datos, utilizando las técnicas de modelización estadística. Su enfoque principal es el de aprovechar el conocimiento implícito en las bases de datos, y está influido por los avances en otras disciplinas, como los sistemas de apoyo a las decisiones, el aprendizaje automático, las bases de datos o la recuperación de información.

Se denomina Minería de Datos¹ al conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos; y describir de forma automatizada modelos previamente desconocidos².

La Minería de Datos en [Morales, 2003] se presenta como un proceso completo de descubrimiento de conocimiento que involucra varios pasos [Morales, 2003]:

1. Entendimiento del dominio de aplicación, el conocimiento relevante a utilizar y las metas del usuario.
2. Seleccionar un conjunto de datos en donde realizar el proceso de descubrimiento.
3. Limpieza y pre-procesamiento de los datos, diseñando una estrategia adecuada para manejar ruido, valores incompletos, valores fuera de rango, valores inconsistentes, etc.
4. Selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, reglas de asociación, etc.
5. Selección de los algoritmos a utilizar.
6. Transformación de los datos al formato requerido por el algoritmo específico de explotación de datos, hallando los atributos útiles, reduciendo las dimensiones de los datos, etc.
7. Llevar a cabo el proceso de minería de datos para encontrar patrones interesantes.
8. Evaluación de los patrones descubiertos y presentación de los mismos mediante técnicas de visualización. Quizás sea necesario eliminar patrones redundantes o no interesantes, o se necesite repetir algún paso anterior con otros datos, con otros algoritmos, con otras metas o con otras estrategias.
9. Utilización del conocimiento descubierto, ya sea incorporándolo dentro de un sistema o simplemente para almacenarlo y reportarlo a las personas interesadas.

Es este proceso es muy importante la etapa del pre-procesamiento de los datos y su transformación al formato requerido por el algoritmo, ya que dependiendo de cómo se realicen estas tareas, va a depender la calidad final de los patrones descubiertos. Un patrón es interesante si es fácilmente entendible por las personas, potencialmente útil, novedoso o valida alguna hipótesis que el usuario busca confirmar. Un patrón interesante representa conocimiento [Ale, 2005].

¹ Servente y García Martínez, 2002; Perichinsky y García-Martínez, 2000; Perichinsky

et al., 2000; Perichinsky et al., 2001; Perichinsky et al., 2003

² Piatetski-Shapiro et al., 1991; Chen et al., 1996; Mannila, 1997

En [Witten & Frank 2000] se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos, y el término Minería de Datos Inteligente³ refiere específicamente a la aplicación de métodos de aprendizaje automático⁴, para descubrir y enumerar patrones presentes en los datos, para estos, se desarrollaron un gran número de métodos de análisis de datos basados en la estadística⁵. En la medida en que se incrementaba la cantidad de información almacenada en las bases de datos, estos métodos empezaron a enfrentar problemas de eficiencia y escalabilidad y es aquí donde aparece el concepto de minería de datos. Una de las diferencias entre al análisis de datos tradicional y la minería de datos es que el primero supone que las hipótesis ya están construidas y validadas contra los datos, mientras que el segundo supone que los patrones e hipótesis son automáticamente extraídos de los datos [Hernández Orallo, 2000].

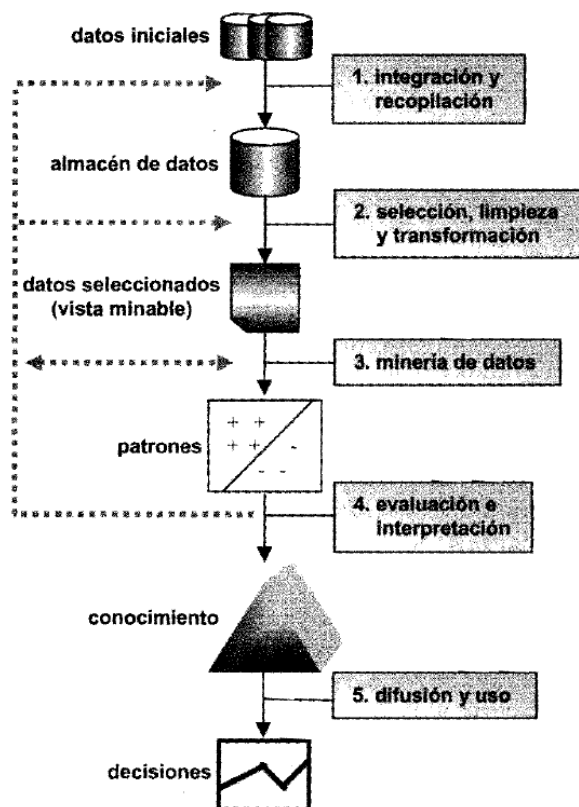
En [Fayyad et al. 1996a] se define el descubrimiento de conocimiento a partir de datos (KDD sus siglas en inglés) como “el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos” que consta de diferentes fases, en la que la Minería de Datos propiamente dicha sería una fase que aglutinaría las técnicas de modelización estadística y aprendizaje automático.

El proceso KDD tendría entonces según [Hernández, J., Ramírez, M. J., Ferri, C., 2005] como entrada datos y como salida conocimiento para la toma de decisiones, y constaría de las siguientes 5 fases principales:

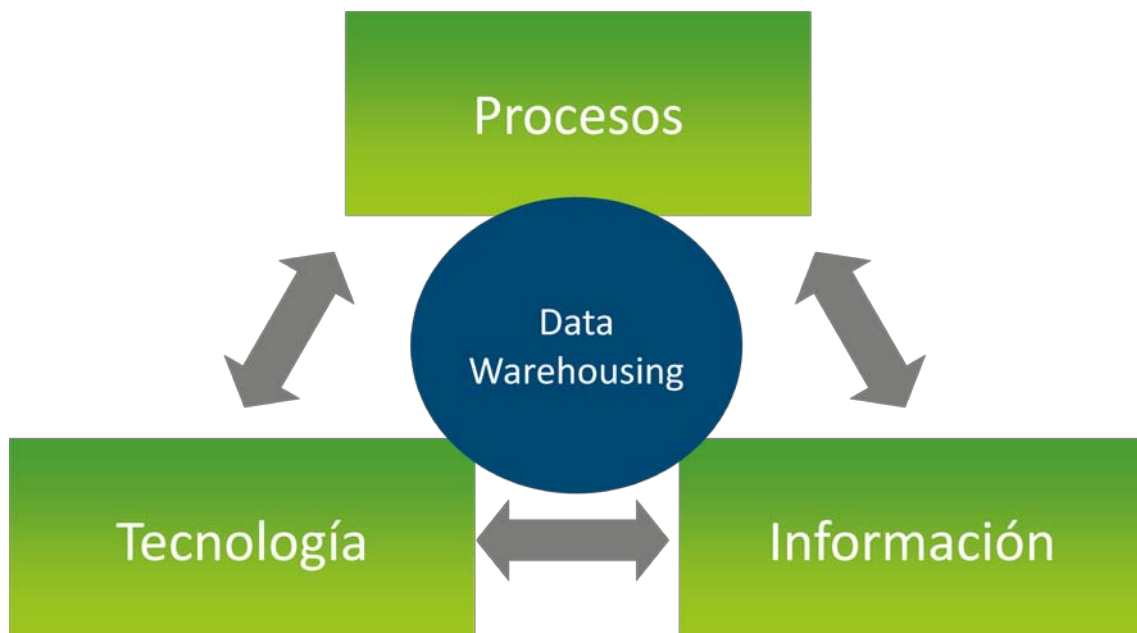
³ Evangelos y Han, 1996; Michalski et al., 1998

⁴ Michalski et al., 1983; Holsheimer y Siebes, 1991

⁵ Michalski et al, 1982



De esta forma el Data Warehousing surge como un elemento del proceso de KDD, formado por los siguientes elementos básicos (que serán analizados con mayor detalle más adelante):



Los procesos principales serían los de ETL (Extracción, Transformación y Carga de Datos), los de uso y los de mantenimiento y operación.

La tecnología se refiere a la infraestructura hardware y software necesario para el almacenamiento de los datos y la ejecución de los procesos.

Por último la información se refiere a los datos almacenados y su calidad.

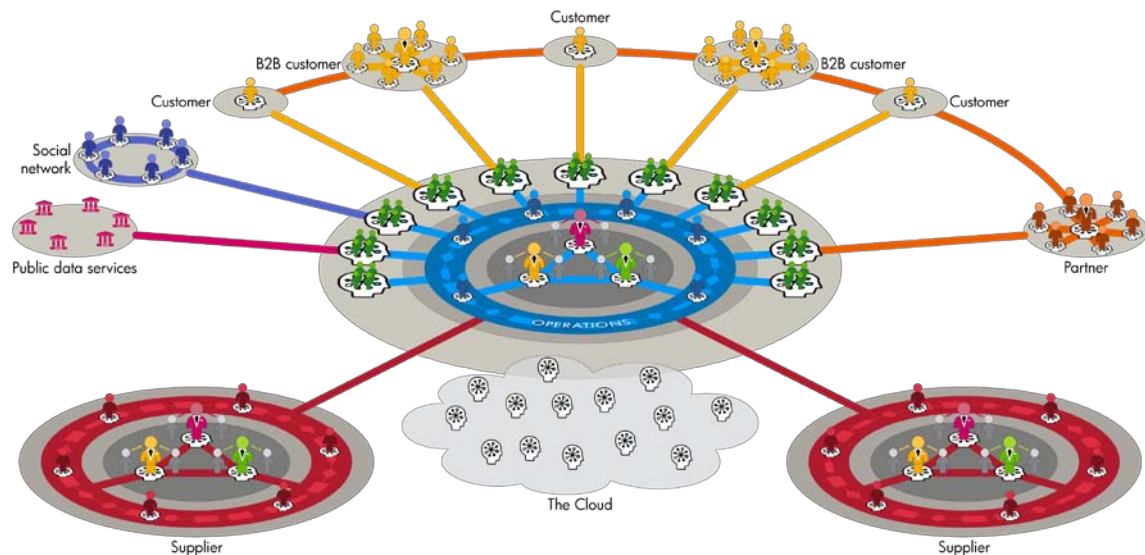
3. Data Warehousing

1.1 Objetivos del Data Warehouse

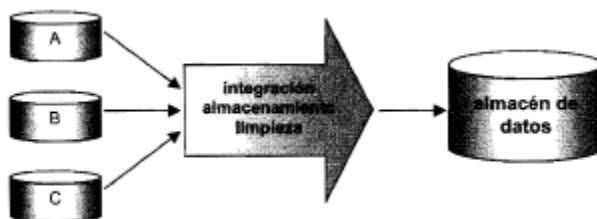
El primer paso para el proceso de extracción de conocimiento a partir de bases de datos es reconocer y reunir los datos con los que se va a trabajar.

Si esta recopilación es puntual o no involucra muchas cantidades y variedades de datos simples, seguramente sin más que una serie de buenas prácticas y sentido común se pueda obtener un conjunto de datos con la calidad suficiente.

Pero en caso contrario será necesario hacer un almacén de datos adaptado a las necesidades.



Dado que el uso que le vamos a dar es OLAP, el almacén de datos debería disponer de un diseño, tecnología y procesos específicos para el uso que se le va a dar, ya que los requerimientos son muy diferentes al de los sistemas operacionales, y con capacidad para cumplir los objetivos de rendimiento, calidad y seguridad que se marquen.



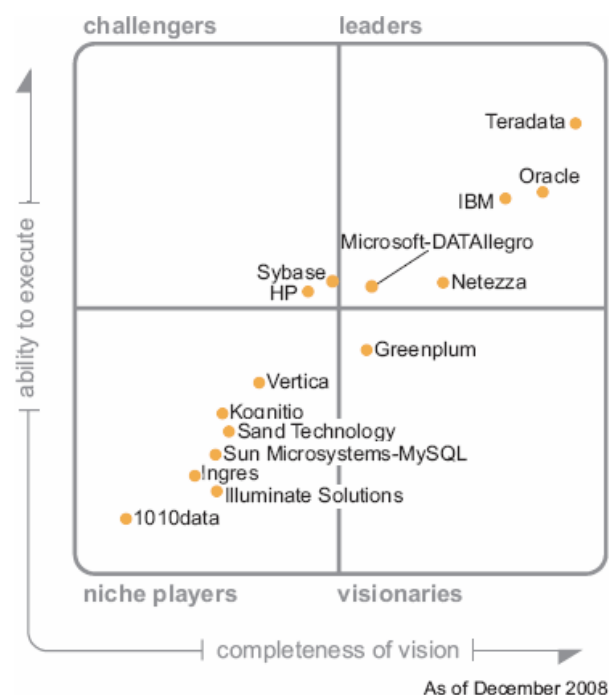
En un almacén de datos lo que se quiere es contener datos que son necesarios o útiles para una organización, es decir, que se utiliza como un repositorio de datos para posteriormente transformarlos en información útil

para el usuario. Un almacén de datos debe entregar la información correcta a la gente indicada en el momento óptimo y en el formato adecuado. El almacén de datos da respuesta a las necesidades de usuarios expertos, utilizando Sistemas de Soporte a Decisiones (DSS), Sistemas de información ejecutiva (EIS) o herramientas para hacer consultas o informes. Los usuarios finales pueden hacer fácilmente consultas sobre sus almacenes de datos sin tocar o afectar la operación del sistema.

1.2 Infraestructura de un Data Warehouse

Para cumplir los objetivos del Data Warehouse primeramente se deberá definir la infraestructura hardware y software necesaria y específica, en especial teniendo en cuenta la naturaleza histórica del Data Warehouse y el granulado o nivel de detalle de la información almacenada, será habitual que el volumen de información sea muy superior al de otros sistemas.

La enorme difusión de esta arquitectura ha hecho que el Data Warehousing y el Business Intelligence sea un sub-sector propio dentro del sector de las Tecnologías de la Información y la Comunicación (TIC), en el que hay un gran número de proveedores, tal y como muestra por ejemplo el siguiente Gartner Magic Quadrant for Data Warehouse Database Management Systems:



Un elemento importante a tener en cuenta es la incompatibilidad de los sistemas OLTP y OLAP para compartir infraestructura, ya que el tipo de consultas del OLTP son sobre un único dato pero muy a menudo, y el tipo de respuesta ha de ser lo más inmediato posible. Por el contrario los OLAP hacen un tipo de consultas masivas de un gran número de datos con

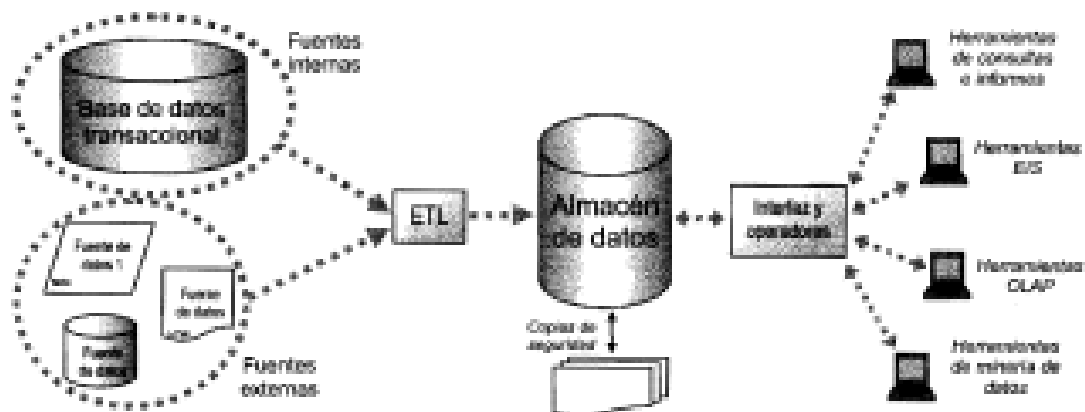
grandes requerimientos de cálculo (en memoria) pero de los que nos se espera una respuesta inmediata.

Además en función de los requerimientos y la infraestructura del Data Warehouse se pueden diseñar diferentes aspectos, como los esquemas físicos utilizados:

- ROLAP (Relational OLAP): físicamente el Data Warehouse se construye sobre una base de datos relacional
- MOLAP (Multidimensional OLAP): físicamente el Data Warehouse se construye sobre estructuras basadas en matrices multidimensionales.

1.3 Elementos del Data Warehouse

Los elementos principales de un Data Warehouse son:

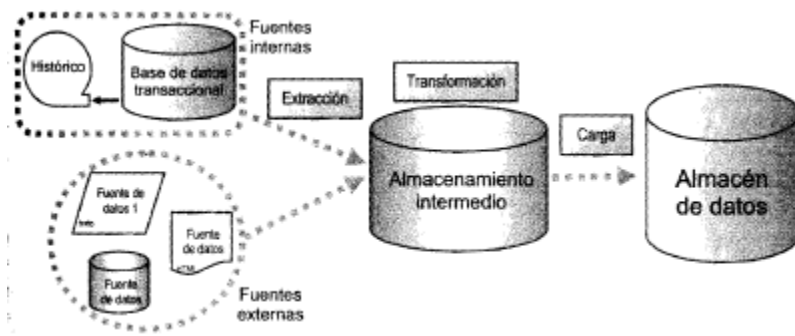


Sistemas fuente, internos o externos a la organización, de donde se recopilan los datos que poblarán el Data Warehouse.

Procesos ETL (Extracción, Transformación y Carga de Datos) que realizan la extracción de los datos de los sistemas fuente, las transformaciones previstas por los analistas para adaptar los datos a las necesidades de la organización y la carga de los mismos en el Data Warehouse.

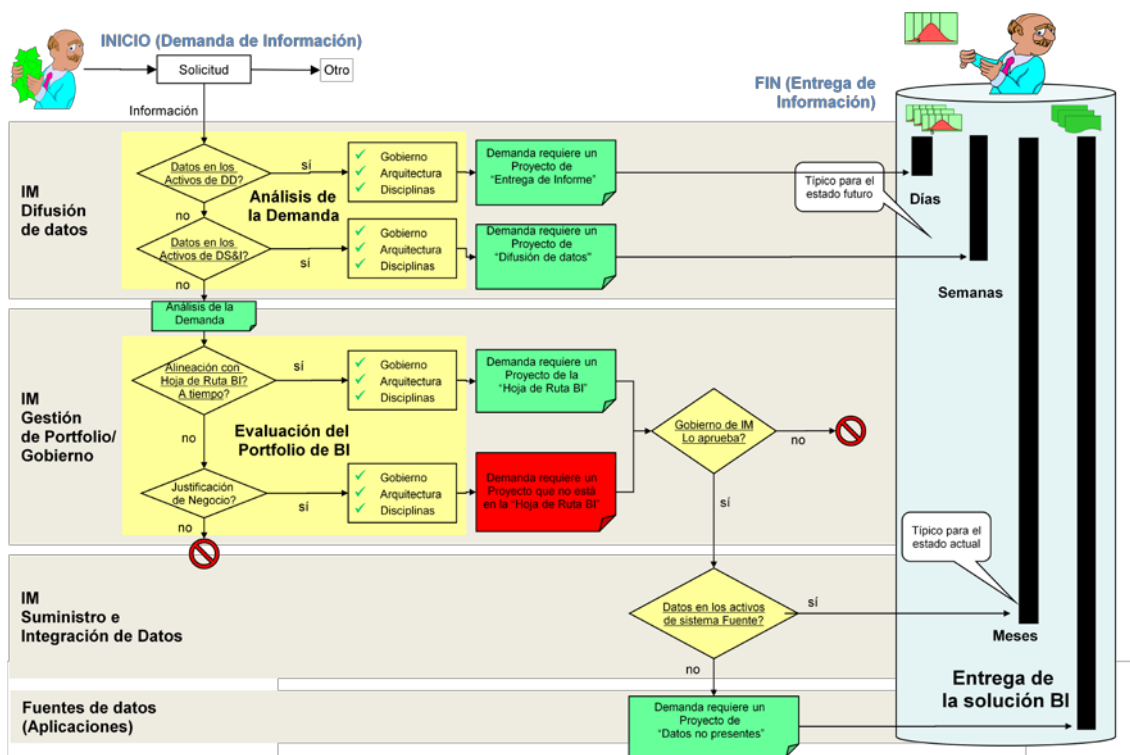
Estos procesos pueden ser periódicos (para incorporar información que se recopila periódicamente) o puntuales, y pueden ser lanzados manualmente o planificados automáticamente (hoy día hay sistemas específicos que aglutinan estas planificaciones), y son gestionados por el entorno de operación del Data Warehouse.

Los procesos ETL son el elemento más costoso en tiempo en la creación de un Data Warehouse, y en función de las necesidades propias de cada uno puede contener elementos adicionales, como un área de almacenamiento intermedio (especialmente indicado cuando existen fuentes externas de datos).



Los Data Warehouse incorporan interfaces para la exportación o consulta de la información contenida en los diferentes entornos y perfiles de uso.

Por último es importante destacar la función de mantenimiento del Data Warehouse. En este ámbito incluimos tanto las habituales de infraestructura de sistemas (copias de seguridad, gestión del rendimiento, gestión de usuarios y permisos, ...) como otras más específicas como la gestión de la demanda de información:



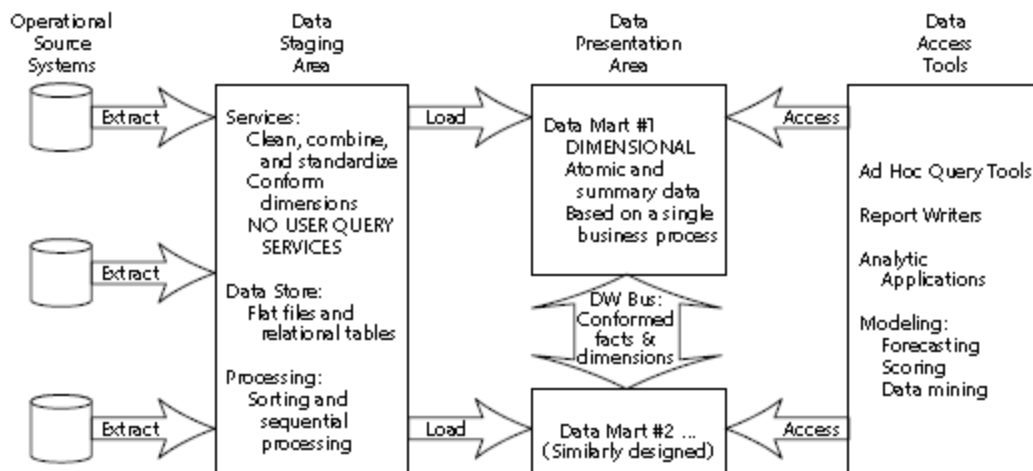
Además de los datos es importante la gestión de los metadatos, que se define comúnmente como "datos acerca de los datos", en el sentido de que se trata de datos que describen cuál es la estructura de los datos que se van a almacenar y cómo se relacionan.

El metadato documenta, entre otras cosas, qué tablas existen en una base de datos, qué columnas posee cada una de las tablas y qué tipo de datos se pueden almacenar. Los datos son de interés para el usuario final, el metadato es de interés para los programas que tienen que manejar estos

datos. Sin embargo, el rol que cumple el metadato en un entorno de almacén de datos es muy diferente al rol que cumple en los ambientes operacionales. En el ámbito de los data warehouse el metadato juega un papel fundamental, su función consiste en recoger todas las definiciones de la organización y el concepto de los datos en el almacén de datos, debe contener toda la información concerniente a:

- Tablas
- Columnas de tablas
- Relaciones entre tablas
- Jerarquías y Dimensiones de datos
- Entidades y Relaciones

Se suele presentar el Data Warehouse como una serie de áreas compuestas por la tecnología, procesos y datos que componen el Data Warehouse:

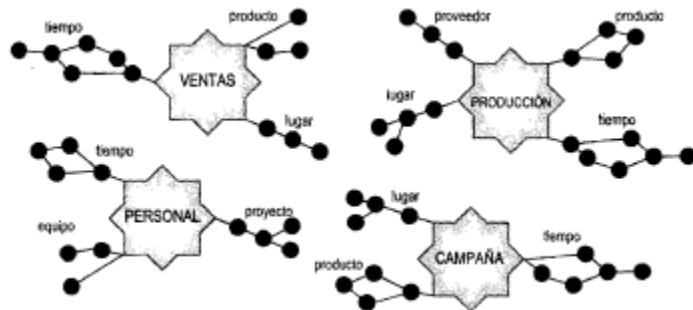


1.4 Arquitectura del Data Warehouse

El Data Warehouse pretende integrar toda la información útil para el análisis y la toma de decisiones, y por ello los primeros proyectos de construcción de Data Warehouse se plantearon como un diseño monolítico. El problema principal con el que se encontraron es que tenían que construirlo en un solo paso, con lo que pasaba mucho tiempo hasta que se ponían en uso, y en algunos casos incluso el proyecto fracasaba; no siempre debido a problemas técnicos sino más bien por problemas de coordinación o por la dificultad para adaptarse a los cambios que se producían a lo largo del proyecto.

Por ellos se empezó a diseñar los Data Warehouse como una agregación de almacenes de datos OLAP especializados o departamentales, llamados Data Marts (esta nomenclatura es utilizada en [Hernández Orallo, 2000] para las estrellas del diseño dimensional que veremos más adelante, pero a nuestro

parecer está más justificado su uso aquí), que requerían mucho menos esfuerzo de coordinación y podían ser puestos en marcha en mucho menos tiempo.



Esta metodología de diseño de los Data Warehouse se ha ido perfeccionando con un enfoque incremental planificado y estandarizado, lo que ha permitido la utilización de arquitecturas BUS de dimensiones que facilitan la integración de los diferentes Data Marts.

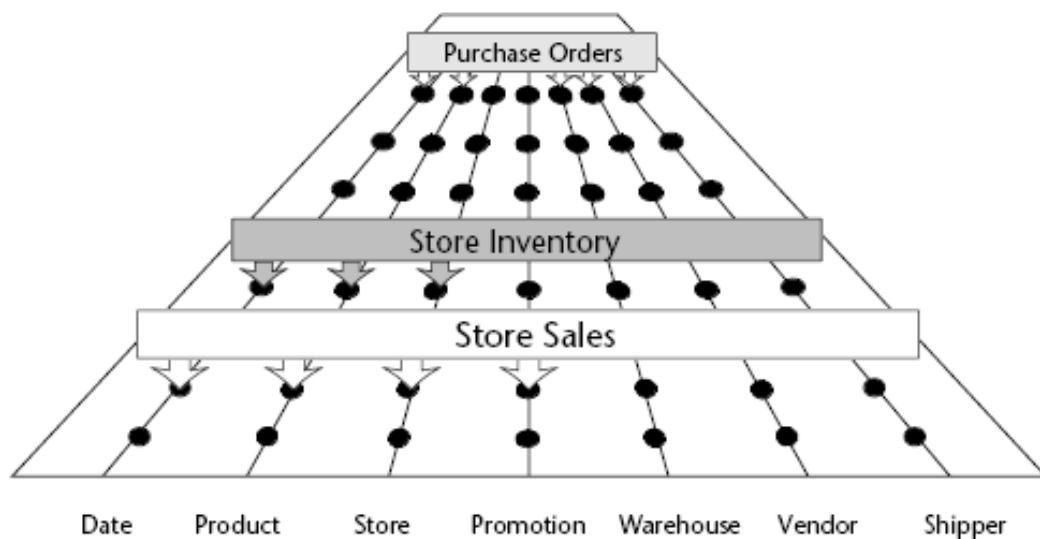
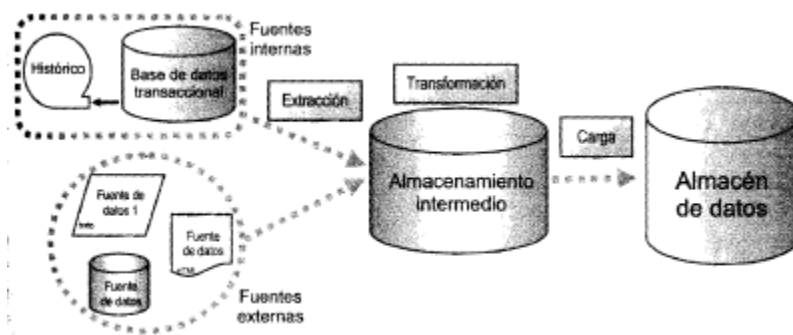


Figure 3.7 Sharing dimensions across the value chain.



1.5 Diseño de los datos de un Data Warehouse

El diseño del modelo de datos almacenado en el Data Warehouse tiene dos enfoques principales:

- Normalizado (Bill Inmon)
- Dimensional (Ralph Kimball)

En este trabajo vamos a seguir el enfoque dimensional de Ralph Kimball, reflejando en el libro "The data warehouse toolkit: the complete guide to dimensional modeling" que se ha convertido en un estándar de facto en el entorno del Business Intelligence.

El diseño dimensional o multidimensional se basa en dos tipos de tablas, tablas de hechos y tablas de dimensiones.

La tabla de hechos es la tabla primaria en un modelo dimensional y almacena los indicadores numéricos. Generalmente se trata de almacenar un dato (resultante de un proceso) en un solo Data Mart, aunque en muchas ocasiones se puede duplicar.

Cada fila de una tabla de hechos corresponde a una medida, y viceversa. Normalmente estas medidas son numéricas y aditivas (para permitir la agregación dinámica del procesamiento OLAP).

Las tablas de hechos expresan relaciones de muchos a muchos entre dimensiones, y pueden llegar a tener un enorme número de filas. Por el contrario suelen tener un pequeño número de columnas.

Daily Sales Fact Table
Date Key (FK)
Product Key (FK)
Store Key (FK)
Quantity Sold
Dollar Sales Amount

Figure 1.2 Sample fact table.

Por contra las tablas de dimensión contienen descriptores textuales de la actividad. En los modelos dimensionales bien diseñados las tablas dimensionales tienen muchas columnas o atributos, que describen cada registro de la tabla (se deben incluir tantos descriptores textuales como sea posible). En ocasiones las tablas dimensionales incluyen 50 o 100 atributos

descriptores, y en cambio suelen tener un número pequeño de registros en comparación con las tablas de hechos.

Cada registro de una tabla dimensional está unívocamente determinado por una clave primaria, que le sirve como integridad referencial con los registros de las tablas de hechos. Estas claves, llamadas surrogadas, son clave primaria de las dimensiones y clave foránea en las tablas de hechos, conformando las relaciones del modelo entidad-relación de los modelos dimensionales.

Las claves surrogadas suelen tener las siguientes características:

- Número entero secuencial
- No tiene significado
- No depende de criterios del operacional (cambiantes)
- Mejora rendimiento

Los atributos de las tablas de dimensiones sirven de fuente primaria de las restricciones de las consultas, agregaciones y etiquetas de informes.

Las tablas de dimensiones son el punto de entrada a las tablas de hechos, por lo que atributos de dimensiones robustos permiten análisis robustos.

Product Dimension Table
Product Key (PK)
Product Description
SKU Number (Natural Key)
Brand Description
Category Description
Department Description
Package Type Description
Package Size
Fat Content Description
Diet Type Description
Weight
Weight Units of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
... and many more

Figure 1.3 Sample dimension table.

Una dimensión básica y presente en cualquier Data Warehouse es la dimensión fecha. Suele contener un gran número de atributos de manera explícita, para así independizarse de funciones nativas SQL no estándar:

Date Dimension
Date Key (PK)
Date
Full Date Description
Day of Week
Day Number In Epoch
Week Number In Epoch
Month Number In Epoch
Day Number In Calendar Month
Day Number In Calendar Year
Day Number In Fiscal Month
Day Number In Fiscal Year
Last Day In Week Indicator
Last Day In Month Indicator
Calendar Week Ending Date
Calendar Week Number In Year
Calendar Month Name
Calendar Month Number In Year
Calendar Year-Month (YYYY-MM)
Calendar Quarter
Calendar Year-Quarter
Calendar Half Year
Calendar Year
Fiscal Week
Fiscal Week Number In Year
Fiscal Month
Fiscal Month Number In Year
Fiscal Year-Month
Fiscal Quarter
Fiscal Year-Quarter
Fiscal Half Year
Fiscal Year
Holiday Indicator
Weekday Indicator
Selling Season
Major Event
SQL Date Stamp
... and more

Las tablas de hechos y dimensiones junto con sus relaciones se unen formando un modelo en estrella, que son fácilmente extensibles para adaptarse a los cambios.



Figure 1.4 Fact and dimension tables in a dimensional model.

A medida que las dimensiones de un modelo en estrella se van desnormalizando se va convirtiendo en un modelo en copo de nieve.

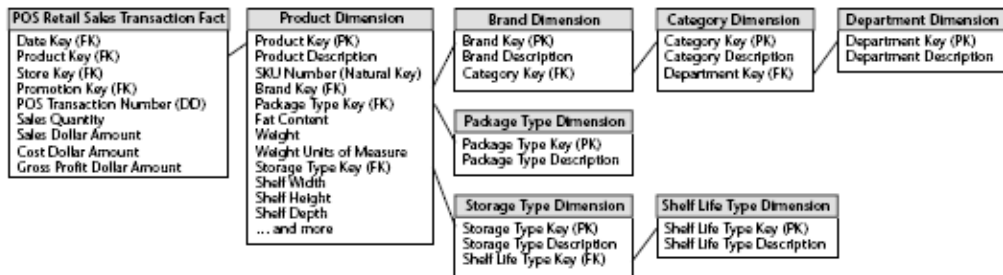


Figure 2.12 Partially snowflaked product dimension.

Las herramientas OLAP sobre de un modelo de datos dimensional pueden realizar fácilmente consultas e informes que de otra forma sería de gran complejidad.

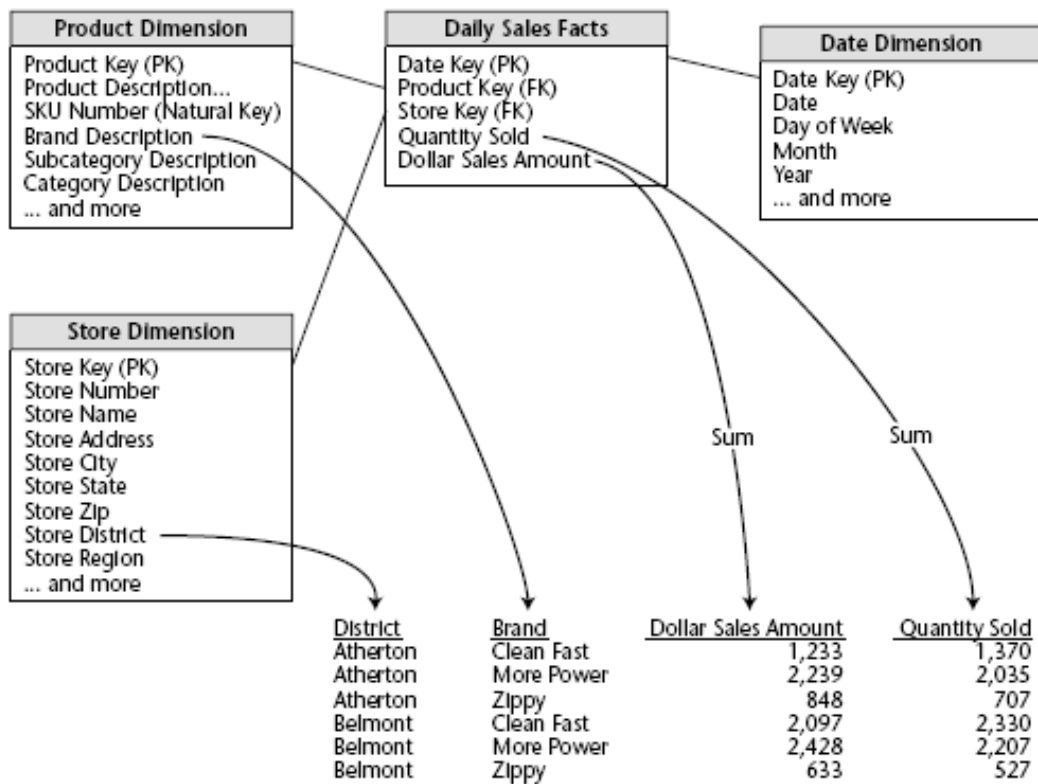


Figure 1.5 Dragging and dropping dimensional attributes and facts into a simple report.

La metodología de diseño dimensional propuesta por Ralph Kimball se basa en cuatro etapas de diseño:

1. Seleccionar el proceso de negocio o actividad
2. Definir el nivel de detalle (granulado)
3. Elegir las dimensiones
4. Identificar los hechos

1.6 Proyecto de creación de un Data Warehouse

Un Data Warehouse no se compra ni se instala, sino que se construye mediante un proyecto, que ha de tener en cuenta los diferentes aspectos del entorno en el que va a funcionar y de los usos que se le van a dar:

	Definición	Requerimientos	Arquitectura	Análisis	Diseño	Construcción	Pruebas	Despliegue	Uso
Usuarios									
Datos									
Tecnología									
Metadatos									

4. Referencias

- Tutorial <http://www.programacion.com/bbdd/tutorial/warehouse/> o cualquier otro que pueda encontrarse en Internet.
- Hernández Orallo et al. Introducción a la minería de datos. Pearson. Capítulo 3.
- Wikipedia (inglés) y conceptos incluidos en la sección "See also".
- The data warehouse toolkit : the complete guide to dimensional modeling / Ralph Kimball, Margy Ross. — 2nd ed, 2002.
- www.businessintelligence.info