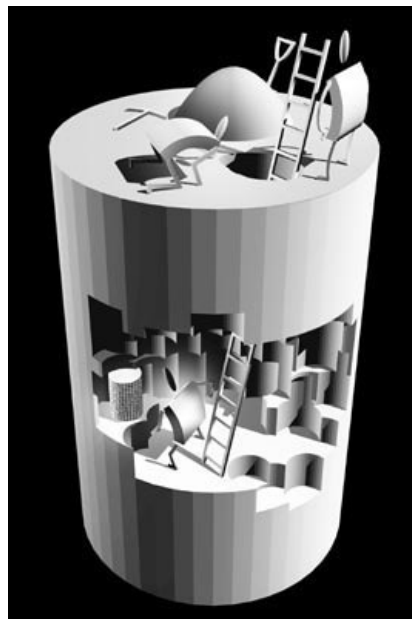




Universidad Nacional del Nordeste  
Facultad de Ciencias Exactas, Naturales y Agrimensura

Monografía de Adscripción:

# **“Minería de Datos”**



Alumna: Maneiro, Mariela Yanina – L.U.: 36.382

Director: Mgter. La Red Martínez, David Luis

Licenciatura en Sistemas de Información

-2008-

# INDICE

| Temas  | Páginas |
|--|---------|
| 1. La Sociedad de la Información y del Conocimiento.....                 | 1       |
| 1.1. Diferencia entre Conocimiento e Información.....                    | 2       |
| 1.2. Gestión del Conocimiento.....                                       | 2       |
| 2. Conceptos y Herramientas de Inteligencia de Negocios.....             | 3       |
| 2.1. Definición de Inteligencia de Negocios.....                         | 3       |
| 2.2. Características.....  | 4       |
| 2.3. Componentes.....  | 5       |
| 2.4. Aplicando el conocimiento.....                                      | 6       |
| 3. Sistemas de Información.....  | 6       |
| 3.1. Finalidad de los Sistemas de Información.....                       | 6       |
| 3.2. Sistemas de Soporte de Decisiones.....                              | 7       |
| 3.2.1. Definición.....   | 7       |
| 3.3. Data Warehouse.....   | 7       |
| 3.3.1. Definición.....   | 8       |
| 3.3.2. Implementación.....   | 8       |
| 3.4. DataMart.....   | 9       |
| 3.4.1. Definición.....   | 9       |
| 3.4.2. Tipos de DataMart.....  | 9       |
| 3.5. Sistemas OLAP.....  | 10      |
| 3.5.1. Definición.....   | 11      |
| 3.5.2. Beneficios de OLAP.....   | 11      |
| 3.5.3. Medidas, Dimensiones, Hechos.....                                 | 11      |
| 3.6. Herramientas para la Toma de Decisiones.....                        | 12      |
| 3.6.1. Diferencias entre las distintas Herramientas.....                 | 12      |
| 4. Descubrimiento del Conocimiento.....                                  | 13      |
| 4.1. Definición.....   | 14      |
| 4.2. Metas.....  | 15      |
| 4.3. Áreas relacionadas.....   | 15      |
| 4.4. Componentes.....  | 15      |
| 4.5. Proceso de Descubrimiento.....                                      | 16      |
| 4.6. Fases de KDD.....   | 17      |
| 5. Minería de Datos.....   | 23      |
| 5.1. Historia de la Minería de Datos.....                                | 23      |
| 5.2. Concepto de minería de datos.....                                   | 24      |
| 5.3. Los Fundamentos de Minería de Datos.....                            | 24      |
| 5.4. Principales características y objetivos de la minería de datos..... | 25      |
| 5.5. Etapas principales del proceso de Minería.....                      | 26      |
| 5.6. Tipología de Patrones de Minería de Datos.....                      | 26      |
| 5.7. Herramientas de Minería de Datos.....                               | 27      |
| 5.8. Glosario de Términos de Minería de Datos.....                       | 28      |
| 6. Bibliografía.....   | 31      |

## **1. La Sociedad de la Información y del Conocimiento**

Los increíbles avances tecnológicos han hecho posible lo que hace apenas algunos años era considerado como ciencia-ficción. El mundo se ha empequeñecido virtualmente: ahora es posible compartir ideas, proyectos y resultados, sin importar las distancias o los horarios. En consecuencia, las organizaciones se ven inmersas en un proceso continuo intentando anticipar, reaccionar y responder a un medio ambiente de cambio, duda y complejidad. Considerando el grado de interdependencia y diversidad a que se ha llegado en los tiempos actuales, tal mecánica es indispensable para garantizar un mínimo de condiciones de supervivencia.

Nada parece detener el cambio ni se anticipa una posible disminución en su velocidad, por el contrario, da la impresión de incrementarse cada vez más, añadiendo severas presiones a las estructuras internas de cualquier organización. Se requiere saber más en menos tiempo con el fin de tomar las decisiones correctas.

Para mantener la adaptabilidad, es necesario crear y preservar un estado de cambio permanente en estructuras, procesos, objetivos y metas; como ejemplo, la capacitación tiene que considerarse de acuerdo con los parámetros dinámicos del medio: es preciso un aprendizaje organizacional donde el personal mejore continuamente sus capacidades, porque podría ser la única fuente para lograr ventajas competitivas y de eficiencia en los productos y servicios.

Este entorno global no es la única fuerza que impulsa a las organizaciones hacia el cambio, las presiones internas también son casi tan poderosas. Por tales razones, cada vez es más requerido un nuevo estilo de trabajar, individuos capaces de superar las limitantes de espacio, tiempo o ubicación geográfica: los trabajadores de la información y el conocimiento, así como personas en las que cualidades como la creatividad y la innovación siempre estén presentes.

La aceleración de las innovaciones tecnológicas han transformado la sociedad industrial, alteraron las clásicas relaciones socio-políticas y reorientaron la lógica de la dinámica social, debido entre otros rasgos, al modo diferente de producción y tratamiento de la información, resultado de las TICs.

El siglo XXI será por excelencia el siglo de la sociedad de la información y del conocimiento, de hecho, ya nos encontramos inmersos en este tipo de sociedad que desempeñará un papel decisivo en el desarrollo económico de los Estados, y a su vez, en la construcción y afirmación de la personalidad individual.

El concepto de la sociedad de la información surge en realidad en los años noventa del siglo XX, coincidiendo con la implantación en los países desarrollados de las TIC (Tecnologías de la Información y Comunicación). Y alcanza su apogeo en el momento en el que las distintas Administraciones públicas se hacen eco de la importancia que tendrán en un futuro inmediato las industrias de la tecnología informática y el universo de las telecomunicaciones.

Nos encontramos en una sociedad basada en una economía fundada en el conocimiento. El conocimiento, por lo tanto, se encuentra ahora ocupando el lugar central del crecimiento económico y de la elevación progresiva del bienestar social.

### **1.1. Diferencia entre Conocimiento e Información**

En esta economía del conocimiento, debemos distinguir el concepto de conocimiento del de información.

*“Poseer conocimiento, sea en la esfera que sea, es ser capaz de realizar actividades intelectuales o manuales. El conocimiento es por tanto fundamentalmente una capacidad cognoscitiva. La información, en cambio, es un conjunto de datos, estructurados y formateados pero inertes e inactivos hasta que no sean utilizados por los que tienen el conocimiento suficiente para interpretarlos y manipularlos”.*

A pesar de que el conocimiento se basa en la información, ésta por sí sola no genera conocimiento.

Conocer y pensar no es simplemente almacenar, tratar y comunicar datos. Serán procesos de generalización de distinto tipo y sus resultados, los que determinarán el saber cómo actuar sobre algo en una situación dada. El desarrollar procesos de pensamiento alternativos, creativos e idiosincrásicos. La información no es en sí conocimiento. El acceso a ella no garantiza en absoluto desarrollar procesos originales de pensamiento.

### **1.2. Gestión del Conocimiento**

Así surge el concepto de “Gestión del Conocimiento”, que es la obtención del conocimiento necesario por las personas adecuadas, en el tiempo, forma y lugar adecuados (Ackermans, Speel & Ratcliffe).

Es un proceso sistemático e intencionado de creación, compartición y aplicación de conocimiento crítico para el desarrollo de la estrategia de negocio, las decisiones u operaciones que conlleva.

Son procesos preacordados que permiten mejorar la utilización del conocimiento y de la información que manejan las personas y los grupos.

No es un proceso aleatorio, sino intencionado, que permite que las organizaciones que desean alcanzar mayores niveles de logro en sus resultados, lo hagan mediante una inversión consciente en la gestión del conocimiento que involucra a personas, nuevas instancias de trabajo colaborativo, recursos materiales y técnicos, etc.

El objetivo que persigue es lograr primeramente mentalizar a la organización del valor que efectivamente tiene para la empresa el desarrollo del conocimiento, transformándolo a sí en un nuevo y óptimo ACTIVO, un patrimonio, un capital efectivo de la organización.

En la medida que las personas viven procesos de formación permanente, ligados a sus tareas organizacionales y actualizan sus conocimientos y sus prácticas laborales, la empresa podrá obtener mejores resultados, sean estos productivos, afectivos, de inserción social, de bien común, etc.

Así y no perdiendo esta perspectiva se puede afrontar y enfrentar a la evolución y el progreso de las nuevas tecnologías de tal forma que en un futuro se cree una sociedad más humana y justa donde lo tecnológico y lo humano se integren.

## 2. Conceptos y Herramientas de Inteligencia de Negocios

En la actualidad los negocios en todo el mundo generan enormes cantidades de datos, como resultado de las transacciones que se dan para soportar las operaciones que se realizan.

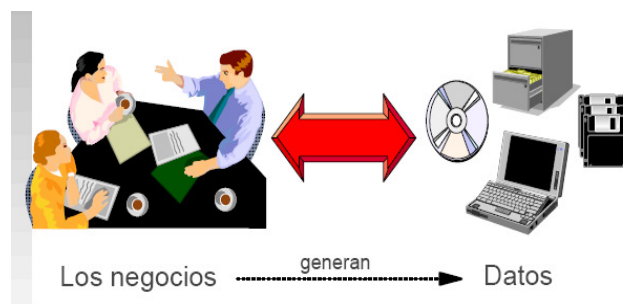
Entonces surgen nuevos problemas, como por ejemplo:

- Sobrecarga de información.
- Mucha información genérica.
- Ausencia de información personalizada y/o relevante para los distintos perfiles que existen en un negocio.
- Falta de retroalimentación oportuna para la mejora de los negocios.

Las empresas necesitan ser competitivas hoy en día, y para ello han comprendido que uno de sus principales activos es la *información* con la que cuentan y aún más, la administración de esta información, es por ello que requieren tener información personalizada de acuerdo a las necesidades de los perfiles que trabajan en ella, que sea fácil de entender y que se pueda obtener oportunamente.

Debido a esto, muchas empresas demandan y tienen la necesidad de sistemas que administren de manera inteligente la información.

La información que se encuentra informatizada en bases de datos ha crecido espectacularmente en volumen y en variedad en la última década. Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido. Aparte de su función de “memoria de la organización”, la información histórica es útil para predecir la información futura.



### 2.1. Definición de Inteligencia de Negocios

Por definición, la inteligencia de negocios (BI) busca explorar información y analizarla para obtener nuevos conocimientos que permitan mejorar la gestión de las empresas y organizaciones. Para lograrlo, se requiere de la implementación de software, traducido en diversas herramientas y técnicas de extracción y estructuración de los datos.

Sin embargo, la implementación de la BI no es una tarea sencilla, puesto que supone desafíos que involucran no sólo aspectos técnicos sino también de negocios. Esto significa que las herramientas de software no sirven de nada si no existe una misión y objetivos claros y si no se cuenta con datos confiables.

Por ello, un desafío clave asociado al éxito de una implementación de BI será superar los errores que afectan la calidad de la información, ya que de ello dependerá, en definitiva, el que brinde frutos positivos. En esta tarea, que puede ser muy ardua, habrá que identificar los datos inapropiados, duplicados, inconsistentes, perdidos y erróneos a través de un diagnóstico riguroso que permita posteriormente efectuar su validación y depuración.

También, se denomina inteligencia de negocios (business intelligence, BI) al conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa. Este conjunto de herramientas y metodologías tienen en común ciertas características.

## 2.2. Características

Las siguientes son las características que tienen en común las herramientas y metodologías de la Inteligencia de Negocios:

- Accesibilidad a la información. Los datos son la fuente principal de este concepto. Lo primero que deben garantizar este tipo de herramientas y técnicas será el acceso de los usuarios a los datos con independencia de la procedencia de estos.
- Apoyo en la toma de decisiones. Se busca ir más allá en la presentación de la información, de manera que los usuarios tengan acceso a herramientas de análisis que les permitan seleccionar y manipular sólo aquellos datos que les interesen.
- Orientación al usuario final. Se busca independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

De acuerdo a su nivel de complejidad se pueden clasificar las soluciones de inteligencia de negocios en:

- Consultas e informes simples (Queries y reports)
- Cubos OLAP (On Line Analytic Processing)
- Data Mining o minería de datos
- Sistemas de previsión empresarial [Estimación de series temporales, ...]

Para que una empresa sea competitiva, las personas que toman las decisiones necesitan acceder rápida y fácilmente a la información de la empresa y esto se realiza por medio del Business Intelligence. La inteligencia de negocios es el proceso de análisis de datos de la empresa para poder extraer conocimiento de ellos. Con BI se puede, por ejemplo: crear una base de datos de clientes, prever ventas y devoluciones, compartir información entre diferentes departamentos, mejorar el servicio al cliente.

### 2.3. Componentes

La inteligencia organizacional (BI) se compone al menos de:

- **Multidimensionalidad:** esta información se encuentra en hojas de cálculo, bases de datos, etc. Reúne información dispersa en toda la empresa y en diferentes fuentes para proveer a los departamentos de la accesibilidad, poder y flexibilidad que necesiten para analizar información.
- **Data Mining:** las empresas suelen recabar información sobre producción, mercados y clientes, pero en realidad el éxito del negocio depende de la visión para intuir cambios o nuevas tendencias. Las aplicaciones de data mining identifican tendencias y comportamientos para extraer información y descubrir las relaciones en bases de datos que revelen comportamientos poco evidentes.
- **Agentes:** son programas que piensan y que pueden realizar tareas muy básicas sin que intervenga el ser humano.

En BI se manejan conceptos como: Datos, Información y Conocimiento.

**Datos:** son hechos objetivos fáciles de capturar, estructurar y transferir.

**Información:** son datos que tienen relevancia y un propósito, donde la intervención humana es necesaria. Derivada de un conjunto de datos mediante agrupación de temas, resumen, comparaciones y otras actividades.

**Conocimiento:** es información que ha sido conectada lógicamente con usos aplicados a ella. Es el entendimiento que se da en la mente y requiere reflexión y síntesis. Es difícil de estructurar, transferir y capturar en las máquinas y es frecuentemente tácita.

Luego hay que aplicar el conocimiento respecto a:

#### **Datos**

Un dato puro, es decir un valor discreto o un hecho aislado, como el valor de una nota tienen poco uso. Cuando este dato es puesto con otros datos puros se le añade algún significado.

#### **Información**

Cuando un dato puro se pone en un contexto específico y es relacionado con otros datos en el mismo contexto se le añade relevancia lo que resulta en lo que hemos definido como información. El conjunto de datos ahora tiene un propósito y puede ser utilizada para cambiar algún procedimiento, tomar una decisión, entre otras cosas.

#### **Conocimiento**

La información es el origen del conocimiento. La información es revisada por una persona que la puede interpretar, aplicar juicios de valor, extrapolar otros significados, aplicarla a una situación específica dentro de un negocio o derivar nueva información. Este proceso de transformar información en conocimientos involucra cualidades, actitudes, habilidades cognitivas y experiencias previas de las personas.

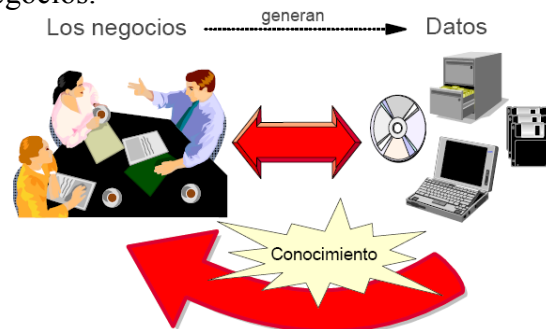
Pero el conocimiento no es suficiente. El capturar datos, hacerlos disponibles, transformarlos en información y el derivar conocimiento de la información, no mejora el negocio por sí solo. El tener varios empleados con conocimiento es bueno, pero si ellos no toman ningún tipo de acción basados en el conocimiento adquirido, éste se desperdicia.

El conocimiento adquirido debe ser aplicado a los problemas del negocio, de modo que se puedan tener más clientes satisfechos, un mejor servicio, más ganancia, empleados más satisfechos, entre otras cosas más.

## 2.4. Aplicando el conocimiento

El objetivo de toda empresa es utilizar el conocimiento que se tiene no solo tal cual, sino para derivar cosas nuevas en base a ellos. No sólo debemos usarlo nosotros, sino capturarlo de modo que otros puedan usarlo sin necesidad de derivarlo nuevamente. Este proceso es el que corresponde al área de la administración del conocimiento.

Como resultado al implementar el proceso de captura, almacenar, organizar y transformar datos en información, manteniéndola y permitiendo acceso flexible a ella, guardando el conocimiento obtenido y dando las facilidades para ubicar y obtener el conocimiento almacenado, puede incrementar el valor de los datos antiguos de los sistemas operacionales y/o transaccionales. Esto debe mejorar la habilidad de competir en el mundo de los negocios.



## 3. Sistemas de Información

### 3.1. Finalidad de los Sistemas de Información

La información reduce nuestra incertidumbre (sobre algún aspecto de la realidad) y, por tanto, nos permite tomar mejores decisiones.

Actualmente, con la informatización de las organizaciones y la aparición de aplicaciones de software operacionales sobre el sistema de información (SI), la finalidad principal de los sistemas de información es dar soporte a los procesos básicos de la organización: ventas, producción, personal, etc.

Primero se busca satisfacer la necesidad de tener un soporte informático para los procesos básicos de la organización: **sistemas de información para la gestión.**

Luego las organizaciones exigen nuevas prestaciones de los SI: **sistemas de información para la toma de decisiones.**

### **3.2. Sistemas de Soporte de Decisiones**

En el proceso de tomar decisiones hay dos partes en general: tener los datos y obtener respuestas de los datos.

- Tener los datos: Ambiente Data Warehouse, es la recolección, transformación, organización, almacenamiento y mantenimiento de información, de los servicios de administración de data warehouse y el mantenimiento de la metadato.
- Obtener respuestas de los datos: Sistema de Soporte de Decisiones, contienen todos los servicios o procesos para elegir, manipular y analizar información y presentar los resultados. Debería brindar acceso transparente a los datos en varias partes del data warehouse y proveer de interfaces comunes a un grupo de usuarios asociados.

#### **3.2.1. Definición de Sistemas de Soporte de Decisiones**

Un *sistema de soporte de decisiones* puede ser definido de una manera general como un sistema de computación diseñado para el proceso de soporte de decisiones (planeamiento, administración y operacional) en un negocio. Esto incluye muchos tipos de sistemas, que van de queries sencillos a funciones complejas de data mining. Este tipo de sistemas es la ventana del usuario a la información almacenada en el ambiente Data Warehouse.

### **3.3. Data Warehouse**

La necesidad de poder disponer de una forma rápida y sencilla de toda la información histórica presente en los sistemas operacionales y su uso para la toma de decisiones ha empujado a las empresas y a la comunidad científica a buscar nuevas formas de estructuración y acceso a estos datos de forma eficiente para, de esta forma, conseguir una ventaja con sus competidores.

Existe un acuerdo en que los sistemas tradicionales de bases de datos no resultan adecuados para realizar consultas analíticas sobre ellos desde una perspectiva multidimensional, que es la forma en la que los analistas de negocio ven los datos de la organización.

Una posible solución consiste en la implantación de un sistema de almacén de datos, que soporta un repositorio de información procedente fundamentalmente de sistemas operacionales que proporciona los datos para el procesamiento analítico y la toma de decisiones.

Un sistema de Data Warehouse contiene datos refinados, históricos, resumidos y no volátiles, y ofrece a los analistas un entorno integrado de información organizada de acuerdo a sus requisitos. El proceso de desarrollar un almacén de datos es, como cualquier tarea que implique algún tipo de integración de recursos pre-existentes,

sumamente complejo, sujeto a errores, generalmente frustrante, y que lleva a que muchos proyectos se abandonen antes de su terminación.

Para mantenerse competitiva una organización necesita una buena gestión de datos, que minimice las duplicidades en su tratamiento y que asegure la calidad de los mismos, de manera que puedan servir como fuente para la toma de decisiones estratégicas y tácticas. Este es precisamente el enfoque del almacén de datos (datawarehouse), que pretende servir como un área de almacenamiento de datos integrados para la toma de decisiones.

El entorno del almacén de datos está formado por datos, sistemas aplicativos, tecnología, facilidades y personas, por lo que puede ser auditado con la ayuda de los COBIT ("Objetivos de control para la información y tecnología relacionada") publicados por la ISACF (Information System Audit and Control Foundation) (ISACF, 1996), en los que se analizan sistemáticamente los aspectos relativos a la eficacia, eficiencia, confidencialidad, integridad, disponibilidad, cumplimiento y fiabilidad de la información. Se entiende por objetivo de control: "una declaración del resultado deseado o del propósito a alcanzar implementando procedimientos de control en una actividad particular de las tecnologías de la información".

### 3.3.1. Definición

Una definición dice que un data warehouse es: "Un almacenamiento no volátil de datos, transacciones y eventos". Que incluye datos operacionales y externos. También contiene información acerca de las transacciones del negocio y los eventos que hicieron que se produzca dicha transacción. En lugar de cambiar la información en el almacén, se le añade la nueva información o se actualizan las correcciones, manteniendo un histórico. Esto provee relevancia y contexto a lo largo del tiempo, lo que es necesario para el análisis que los usuarios necesitan realizar.

La información y los datos en el data warehouse deben estar integrados, consolidados, asegurados y depurados, de modo que sean el soporte de decisiones corporativas.

### 3.3.2. Implementación

Un data warehouse puede ser implementado en una o dos capas:

- **Dos capas:** Es para ambientes de implementación medianos o grandes, en la mayoría de los caso distribuidos y que tengan una gran variedad de tipos de usuarios, de requerimientos de análisis. Involucra tener todo el data warehouse en un conjunto de bases de datos y luego mover subconjuntos de ellas en un segundo almacenamiento de datos al que finalmente ingresan los usuarios.

Frecuentemente los sistemas de una capa son migrados a sistemas de dos capas.

Otra posibilidad es establecer un Warehouse virtual dando a los usuarios acceso directo al origen de datos en lugar de transformarlo en un warehouse. Esto se conoce como Sistemas de Administración de información (Management Information System, MIS por sus siglas en inglés). El cuál es un sistema de reporte para realizar agregaciones y resúmenes directamente de la data operacional, en este caso los usuarios sólo estarían interesados en el estado actual de la data. Este tipo de data

warehouse requiere el desarrollo de queries complicados y también tienen el riesgo de que los usuarios probablemente estén analizando data que no está depurada y que no sea del todo útil en la toma de decisiones.

### Capas lógicas en un data warehouse de dos capas

En un data warehouse de dos capas, el origen de datos es procesado y colocado en bases de datos que se llaman Data Warehouse Central (DWC).

Un DWC está diseñado para soportar grandes cantidades de información detallada, incluyendo la información histórica que sea necesaria. Usualmente es una base de datos relacional con poca redundancia cuyo propósito es soportar las necesidades de todos los usuarios del negocio. Muy pocos usuarios, o ninguno, tienen acceso a esta capa del data warehouse.

El DataMart es específico para un grupo de usuarios y por lo general tiene agregaciones adicionales e información derivada.

- **Una capa:** Es recomendado para pequeñas y medianas empresas, inicio de proyectos, prototipos y proyectos de pruebas de conceptos. El origen de datos relevantes es almacenado y transformado dentro del warehouse el cual es utilizado por todos los usuarios.

## **3.4. DataMart**

### **3.4.1. Definición**

El DataMart es un subconjunto de información corporativa con formato adicional a la medida de un usuario específico del negocio, como resultado el DataMart por sí solo no es un data warehouse, no se le debe confundir como un pequeño data warehouse o una implementación de una capa del mismo. Del mismo modo, varios DataMarts en un negocio no se pueden confundir con un data warehouse. Un DataMart puede tener más cantidad de información que un data warehouse, pero siempre será menor en complejidad y alcance de la data. Un data warehouse tiene más usuarios y más temas que un DataMart y provee de una vista más comprensiva entre múltiples áreas.

Los DataMarts pueden existir sin estar conectados a un data warehouse, esto se puede dar cuando un grupo o departamento necesita información para análisis e implementa su propio ambiente, éstos se construyen antes de un data warehouse ya que toma menor tiempo.

### **3.4.2. Tipos de DataMarts**

Hay dos tipos de DataMarts:

- **Independiente:** Un DataMart que funciona sólo y tiene su propio sistema único para extraer y transformar el origen de datos. Si uno de estos datos o valores es el mismo en cualquier otro DataMart es por pura casualidad.
- **Dependiente o Federado:** Uno o un conjunto de DataMarts que usan los mismos procesos de extracción y transformación de datos y tiene el mismo contenido

para la data compartida. Este permite que los usuarios de un DataMart para realizar de su data y la de otros DataMarts juntos.

Un DataMart independiente puede funcionar bien en una situación especial (por ejemplo que exista un área altamente descentralizado en el negocio). Sin embargo, en la mayoría de los casos, éstos deberían ser parte de una arquitectura mayor, la que debería soportar que el data warehouse sea implementado por fases: Primero estableciendo los DataMarts y luego construyendo todo el ambiente de dos capas. Este ambiente con DataMarts conectados minimizará situaciones en las que personas de diferentes partes del negocio tengan resultados diferentes de su análisis de datos.

Lo que se recomienda es una implementación de una arquitectura de dos capas con un data warehouse central construido directamente desde el origen de datos y uno o más DataMarts federados que son construidos desde la data existente en el data warehouse central. Algunos de los DataMarts pueden tener su propia base de datos o pueden acceder al data warehouse central a través de una vista u otros mecanismos de filtro.

### **3.5. Sistemas OLAP**

Actualmente los data warehouse y las técnicas olap son las maneras más efectivas y tecnológicamente más avanzadas para integrar, transformar y combinar los datos para facilitar al usuario o a otros sistemas el análisis de la información.

La tecnología OLAP generalmente se asocia a los almacenes de datos, aunque se puede tener almacenes de datos sin OLAP y viceversa.

Habitualmente se utilizan herramientas OLAP (On-line Analytical Processing) como herramientas frontales para el acceso a los datos. Las herramientas OLAP, como los almacenes de datos y bases de datos multidimensionales, están basadas en el modelo multidimensional. Las técnicas de modelado conceptual y los modelos conceptuales utilizados para las aplicaciones OLTP (On-line Transaction Processing) no son adecuados para las aplicaciones OLAP ya que no son capaces de representar los requisitos básicos de este tipo de aplicaciones.

En los últimos años, se ha pasado de tener todos los ordenadores de la empresa manejados por un grupo reducido de expertos en informática, a que prácticamente cada empleado disponga como mínimo de su propio terminal. Concretamente, en lo que se refiere a las bases de datos, se ha pasado de la necesidad de depender de los informáticos para la obtención de la información requerida por los usuarios a la posibilidad de que sea el propio usuario final de la información el que acceda a ella utilizando las herramientas adecuadas.

Claramente, el reto es conseguir que esto les resulte tan fácil como sea posible. El modelado multidimensional trata precisamente de conseguirlo, acercándose a la concepción que ya tienen los propios analistas de la empresa, según la cual dividen los elementos de análisis en "hechos" a analizar y "dimensiones" utilizadas para hacerlo.

### 3.5.1. Definición de Sistemas OLAP

Es un método para buscar en los datos de diferentes maneras. Con OLAP los datos son clasificados en diferentes dimensiones las que pueden ser vistas unas con otras en cualquier combinación para obtener diferentes análisis de los datos que contienen.

### 3.5.2. Beneficios de OLAP

- Es de fácil uso y acceso flexible para el usuario.
- Los datos están organizados en varias dimensiones lo que permite que los usuarios hagan un mejor análisis.
- Ahorro generado por la productividad de personal altamente profesional y caro que usa permanentemente software y sistemas de información.
- Permite encontrar la historia en los datos

*¿Qué es multidimensionalidad?*

Multidimensionalidad es convertir los datos de varias fuentes, tablas relacionales o archivos planos en un estructura donde la data es agrupada en dimensiones separadas y heterogéneas (a esto generalmente se le llama 'cubo').

Las dimensiones son perspectivas de alto nivel de los datos que representan la información más importante de un negocio. En un banco se tendrán Cuentas, Clientes, Tiempo, Productos, Agencias, Regiones, etc.

En una aplicación OLAP estas dimensiones tienden a no cambiar durante el tiempo. Cada dimensión tiene componentes que son llamados 'miembros'. Por ejemplo el primer trimestre del año es un miembro de la dimensión Tiempo. Cada dimensión puede tener jerarquías entre sus miembros, por ejemplo un mes se puede considerar dentro de un trimestre.

*¿Qué es un modelo estrella?*

Un modelo estrella es un conjunto de tablas en una base de datos relacional diseñada para representar datos de manera multidimensional.

Una tabla de hechos almacena los datos numéricos y está unida a otras tablas dimensionales que almacenan la información descriptiva acerca de los nombres de la dimensión y sus miembros.

Los sistemas OLAP que utilizan este tipo de modelo de base de datos son llamados sistemas ROLAP (Relational OLAP por sus siglas en inglés).

### 3.5.3. Medidas, Dimensiones, Hechos

Para construir un modelo multidimensional se deben identificar las **medidas** candidatas. Estas corresponden a ítems de datos que los usuarios utilizan en sus queries para medir la performance de un elemento dentro del negocio. Por ejemplo, los usuarios finales de un negocio de bebidas toman decisiones importantes basándose en la cantidad de bebidas vendidas.

Las medidas candidatas son los datos numéricos, pero no cada atributo numérico es una medida candidata.

Son parte de dominio de valor continuo, se deben distinguir las medidas de atributos discretos que son parte de las dimensiones.

Son las que están involucradas en cálculos de resúmenes.

El número y tipo de **dimensiones** para cada medida del modelo debe ser determinada cuidadosamente. El significado de una medida está influenciado por las definiciones de los tipos de medidas que tiene. Cuando se trata de definir la dimensiones, el añadir, eliminar o cambiar propiedades particulares de las dimensiones candidatas cambia el contexto y en consecuencia el significado de la medida candidata.

Los **hechos** contienen:

- Un identificador de hechos
- Llaves de dimensión, que lo enlaza con las dimensiones.
- Medidas
- Varios tipos de atributos, los que usualmente se derivan de otros datos en el modelo.

Cada hecho debería tener un equivalente en el mundo real de los negocios. Los hechos relacionados al negocio representan una de las siguientes cosas:

- Objetos del negocio cuyo estado es de interés del analista de información.
- Objetos del negocio cuyos cambios de estado son de interés del analista de información.
- Transacciones o eventos del negocio.

### 3.6. Herramientas para la Toma de Decisiones

Han aparecido diferentes herramientas de negocio o DSS que coexisten: EIS, OLAP, consultas & informes, minería de datos, etc.

#### 3.6.1. Diferencias entre las distintas Herramientas

*¿Cuál es la diferencia entre EIS y OLAP?*

Un EIS (*Executive Information System*) es un sistema de información y un conjunto de herramientas asociadas que proporciona a los directivos acceso a la información de estado y sus actividades de gestión. Está especializado en analizar el estado diario de la organización (mediante indicadores clave) para informar rápidamente sobre *cambios* a los directivos.

La información solicitada suele ser, en gran medida, numérica (*ventas semanales, nivel de stocks, balances parciales, etc.*) y representada de forma gráfica al estilo de las hojas de cálculo.

Las herramientas OLAP (*On-Line Analytical Processing*) son más genéricas funcionan sobre un sistema de información (transaccional o almacén de datos) y permiten realizar

agregaciones y combinaciones de los datos de maneras mucho más complejas y ambiciosas, con objetivos de análisis más estratégicos.

*¿Cuál es la diferencia entre “informes avanzados” y OLAP?*

Los sistemas de informes o consultas avanzadas están basados, generalmente, en sistemas *relacionales u objeto-relacionales*, utilizan los operadores clásicos: concatenación, proyección, selección, agrupamiento (en SQL y extensiones) y el resultado se presenta de una manera tabular.

Las herramientas OLAP están basadas, generalmente, en sistemas o *interfaces multidimensionales*, utilizando operadores específicos (además de los clásicos): *drill, roll, pivot, slice & dice*. Y el resultado se presenta de una manera matricial o híbrida.

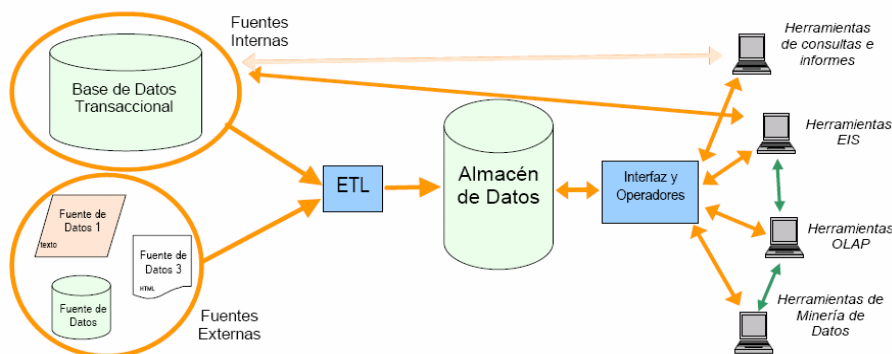
*¿Cuál es la diferencia entre OLAP y minería de datos?*

Las herramientas OLAP proporcionan facilidades para “manejar” y “transformar” los datos. Producen otros “datos” (más agregados, combinados) y ayudan a analizar los datos porque producen *diferentes vistas* de los mismos.

Las herramientas de Minería de Datos son muy variadas: permiten “extraer” patrones, modelos, descubrir relaciones, regularidades, tendencias, etc. Y producen “reglas” o “patrones” (“conocimiento”).

*¿Qué interrelaciones existen entre todas estas herramientas?*

Se indican en la gráfica la interrelación existente. Algunas herramientas han hecho cambiar la manera de trabajar de otras.



## 4. Descubrimiento del Conocimiento

En los últimos años se ha visto un gran crecimiento en la capacidad de generación y almacenamiento de información, debido a la creciente automatización de procesos y los avances en las capacidades de almacenamiento de información. Desafortunadamente, no se ha visto un desarrollo equivalente en las técnicas de análisis de información, por lo que existe la necesidad de una nueva generación de técnicas y herramientas

computacionales con la capacidad de asistir a usuarios en el análisis automático e inteligentes de datos. El procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil para un usuario y satisfacerle sus metas, es el objetivo principal del área de Descubrimiento de Conocimiento en Bases de Datos.

La tecnología actual nos permite capturar y almacenar una gran cantidad de datos. Tratar de encontrar patrones, tendencias y anomalías es uno de los grandes retos de vida moderna.

Código de barras, automatización de procesos en general, avances tecnológicos en almacenamiento de información y abaratamiento de precios en memoria, son algunos de los factores que han contribuido a la generación masiva de datos.

Se ha estimado que la cantidad de datos almacenados en el mundo en bases de datos se duplica cada 20 meses.

Las técnicas tradicionales de análisis de información no han tenido un desarrollo equivalente. La velocidad en que se almacenan datos es muy superior a la velocidad en que se analizan. Existe un gran interés comercial por explotar los grandes volúmenes de información almacenada.

Se cree que se está perdiendo una gran cantidad de información y conocimiento valioso que se podría extraer de los datos.

Al Descubrimiento de Conocimiento de Bases de Datos (KDD) a veces también se le conoce como minería de datos (*Data Mining*).

Sin embargo, muchos autores se refieren al proceso de minería de datos como el de la aplicación de un algoritmo para extraer patrones de datos y a KDD al proceso completo (pre-procesamiento, minería, post-procesamiento).

#### **4.1. Definición**

*Descubrimiento de Conocimiento en Bases de Datos:*

*“Proceso de extracción no trivial para identificar patrones que sean válidos, novedosos, potencialmente útiles y entendibles, a partir de datos”.*

- Proceso: KDD involucra varios pasos y es interactivo, al encontrar información útil en los datos, se realizan mejores preguntas.
- Válido: se utilizan principalmente los datos y se espera que los patrones puedan aplicarse en el futuro.
- Novedoso: desconocido con anterioridad.
- Útil: aplicable y cumpliendo las metas del usuario.
- Entendible: que nos lleve a la comprensión, muchas veces medido por el tamaño.

A veces se hace una mezcla de medidas de utilidad, novedad, simplicidad y validez para establecer que tan *interesantes* pueden ser los patrones. Esta medida, generalmente está definida por el usuario, y es parte de los parámetros de operación de los algoritmos de minería.

En este sentido, podemos decir que un patrón representa conocimiento si su medida de interés rebasa un cierto umbral.

El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos.

Se estima que la extracción de patrones (minería) de los datos ocupa solo el 15% - 20% del esfuerzo total del proceso de KDD.

#### 4.2. Metas

- Procesar automáticamente grandes cantidades de datos *crudos*,
- identificar los patrones más significativos y relevantes, y
- presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

#### 4.3. Áreas relacionadas

KDD es un nuevo campo interdisciplinario que involucra investigación de áreas tales como:

- Tecnologías de bases de datos y bodegas de datos: maneras eficientes de almacenar, acceder y manipular datos
- Aprendizaje computacional, estadística, computación *suave* (redes neuronales, lógica difusa, algoritmos genéticos, razonamiento probabilístico): desarrollo de técnicas para extraer conocimiento a partir de datos
- Reconocimiento de patrones: desarrollo de herramientas de clasificación
- Visualización: interfaz entre humanos y datos, y entre humanos y patrones
- Cómputo de alto desempeño: mejora de desempeño de algoritmos debido a su complejidad y a la cantidad de datos

#### 4.4. Componentes

- *Conocimiento del dominio y preferencias del usuario:*

Incluye el diccionario de datos, información adicional de las estructuras de los datos, restricciones entre campos, metas o preferencias del usuario, campos relevantes, listas de clases, jerarquías de generalización, modelos causales o funcionales, etc.

El objetivo del conocimiento del dominio es orientar y ayudar en la búsqueda de patrones interesantes (aunque a veces puede causar resultados contraproducentes). Se tiene que hacer un balance entre eficiencia y completos del conocimiento.

- *Control del descubrimiento:*

Toma el conocimiento del dominio, lo interpreta y decide qué hacer (en la mayoría de los sistemas el control lo hace el usuario).

- *Interfaces:*

Con la base de datos y con el usuario.

- *Foco de atención:*

Especifica qué tablas, campos y registros acceder. Tiene que tener mecanismos de selección aleatoria de registros tomando muestras estadísticamente significativas, puede usar predicados para seleccionar un subconjunto de los registros que comparten cierta característica, etc.

Algunas técnicas para enfocar la atención incluyen:

- *Agregación:* junta valores (por ejemplo, los más bajos y los más altos)
- *Partición de datos:* en base a valores de atributos (por ejemplo, sólo aquellos datos que tengan ciertos valores)
- *Proyección:* ignorar algún(os) atributo(s)
- *Partición y proyección implican menos dimensiones.* Agregación y proyección implican menos dispersión.

- *Extracción de patrones:*

Donde patrón se refiere a cualquier relación entre los elementos de la base de datos. Pueden incluir medidas de incertidumbre. Aquí se aplican una gran cantidad de algoritmos de aprendizaje y estadísticos.

- *Evaluación:*

Un patrón es interesante en la medida que sea confiable, novedoso y útil respecto al conocimiento y los objetivos del usuario. La evaluación normalmente se le deja a los algoritmos de extracción de patrones que generalmente están basados en significancia estadística (sin embargo, no es ni debe ser el único criterio).

#### **4.5. Proceso de Descubrimiento**

El proceso de descubrimiento de conocimiento en bases de datos involucra varios pasos:

1. Entendimiento del dominio de aplicación, el conocimiento relevante a usar y las metas del usuario. Esta es la tarea que puede llegar a consumir el mayor tiempo.
2. Seleccionar un conjunto o subconjunto de bases de datos, seleccionar y enfocar la búsqueda en subconjuntos de variables, y seleccionar muestras de datos (instancias) en donde realizar el proceso de descubrimiento.

Los datos tradicionalmente han sido tablas ASCII y la tendencia es utilizar manejadores de bases de datos y almacenes de datos que están optimizados para realizar un proceso analítico.

3. Limpieza y preprocesamiento de datos, diseñando una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, casos extremos (si es necesario), etc.
4. Selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc.
5. Selección de él o de los algoritmos a utilizar.
6. Transformación de datos al formato requerido por el algoritmo específico de minería de datos.
7. Llevar a cabo el proceso de minería de datos. Se buscan patrones que pueden expresarse como un modelo o simplemente que expresen dependencias de los datos.

El modelo encontrado depende de su función (e.g, clasificación) y de su forma de representarlo (e.g., árboles de decisión, reglas, etc.).

Se tiene que especificar un criterio de preferencia para seleccionar un modelo dentro de un conjunto posible de modelos.

Se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está predeterminada en el algoritmo de minería).

8. Interpretar los resultados y posiblemente regresar a los pasos anteriores.

Esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias.

Este es un paso crucial en donde se requiere tener conocimiento del dominio.

La interpretación puede beneficiarse de procesos de visualización, y sirve también para borrar patrones redundantes o irrelevantes.

9. Incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) lo cual puede incluir resolver conflictos potenciales con el conocimiento existente.

10. El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas. En este sentido, KDD implica un proceso interactivo e iterativo involucrando la aplicación de varios algoritmos de minería de datos.

#### **4.6. Fases de KDD**

##### **1) Recolección de datos:**

Las primeras fases del KDD determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original.

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra en bases de datos y otras fuentes muy diversas, tanto *internas* como *externas*; muchas de estas fuentes son las que se utilizan para el trabajo *transaccional*.

El análisis posterior será mucho más sencillo si la fuente es *unificada*, *accesible* (interna) y desconectada del trabajo *transaccional*.

El proceso subsiguiente de minería de datos depende mucho de la *fente* (**Olap** u **oltp**, **Datawarehouse** o copia con el esquema original, **Rolap** o **molap**). Depende también del tipo de *usuario*:

- “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.

Aparte de información interna de la organización, los almacenes de datos pueden recoger información externa:

- Demografías (censo), páginas amarillas, psicografías (perfiles por zonas), uso de internet, 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, etc.
- Bases de datos externas compradas a otras compañías.

## 2) Selección, limpieza y transformación de datos:

Se realiza la *Limpieza* (data cleansing) y *criba* (selección) de datos. Se deben eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba). Para ello se utilizan métodos estadísticos casi exclusivamente:

- Histogramas (detección de datos anómalos).
- Selección de datos (muestreo, ya sea verticalmente, eliminando atributos, u horizontalmente, eliminando tuplas).
- Redefinición de atributos (agrupación o separación).

Ante la presencia de *datos anómalos* (outliers) las acciones a seguir podrían ser las siguientes:

- *Ignorar*: Algunos algoritmos son robustos a datos anómalos (p.ej. Árboles).
- *Filtrar (eliminar o reemplazar) la columna*: Solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna es reemplazarla por una columna

discreta diciendo si el valor era normal u outlier (por encima o por debajo).

- *Filtrar la fila*: Puede sesgar los datos, porque muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.
- *Reemplazar el valor*: Por el valor “nulo” si el algoritmo lo trata bien o por máximos o mínimos, dependiendo por donde es el outlier, o por medias. A veces se puede *predecir* a partir de otros datos, utilizando cualquier técnica de ml.
- *Discretizar*: Transformar un valor continuo en uno discreto (p.ej. Muy alto, alto, medio, bajo, muy bajo) hace que los outliers caigan en “muy alto” o “muy bajo” sin mayores problemas.

Ante la presencia de *datos faltantes* (missing values) las acciones a seguir podrían ser las siguientes:

- *Ignorar*: Algunos algoritmos son robustos a datos faltantes (p.ej. Árboles).
- *Filtrar (eliminar o reemplazar) la columna*: Solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna es reemplazarla por una columna booleana diciendo si el valor existía o no.
- *Filtrar la fila*: Claramente sesga los datos, porque muchas veces las causas de un dato faltante están relacionadas con casos o tipos especiales.
- *Reemplazar el valor*: Por medias. A veces se puede *predecir* a partir de otros datos, utilizando cualquier técnica de ml.
- *Segmentar*: Se segmentan las tuplas por los valores que tienen disponibles. Se obtienen modelos diferentes para cada segmento y luego se combinan.
- *Modificar la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles*.

Respecto a las razones sobre *datos faltantes* (missing values). A veces es importante examinar las razones tras datos faltantes y actuar en consecuencia:

- *Algunos valores faltantes expresan características relevantes*:
  - P.ej. La falta de teléfono puede representar en muchos casos un deseo de que no se moleste a la persona en cuestión, o un cambio de domicilio reciente.
- *Valores no existentes*: Muchos valores faltantes existen en la realidad, pero otros no. P.ej. El cliente que se acaba de dar de alta no tiene consumo medio de los últimos 12 meses.

- *Datos incompletos*: Si los datos vienen de fuentes diferentes, al combinarlos se suele hacer la unión y no la intersección de campos, con lo que muchos datos faltantes representan que esas tuplas vienen de una/s fuente/s diferente/s al resto.

Respecto a la Transformación del *esquema*, se tiene:

- *Esquema original*:
  - Ventajas: Las r.i. se mantienen (no hay que reaprenderlas, no despistan).
  - Desventajas: Muchas técnicas no se pueden utilizar.
- *Tabla universal*: cualquier esquema relacional se puede convertir (en una correspondencia 1 a 1) a una tabla universal:
  - Ventajas: Modelos de aprendizaje más simples (proposicionales).
  - Desventajas: Muchísima redundancia (tamaños ingentes). La información del esquema se pierde. Muchas dependencias funcionales se vuelven a re-descubrir. Se debe añadir metainformación.
- *Desnormalizado tipo estrella o copo de nieve (datamarts)*:
  - Ventajas: Se pueden buscar reglas sobre información sumariada y si resultan factibles se pueden comprobar con la información detallada. Se utilizan operadores propios: *roll-up*, *drill-down*, *slicing and dicing*.
  - Desventajas: Orientadas a extraer un tipo de información (granjeros).
- *Intercambio de dimensiones*: (filas por columnas):
  - Ejemplo: Una tabla de cestas de la compra, donde cada atributo indica si el producto se ha comprado o no. El *objetivo es* ver si dos productos se compran conjuntamente (*regla de asociación*). Es muy costoso: hay que mirar al menos la raíz cuadrada de todas las relaciones (cestas):
    - Puede haber millones en una semana...
    - Sin embargo... Productos sólo hay unos 10.000.
  - Ejemplo: Sólo es necesario hace *xor* entre dos filas para saber si hay asociación.
- *Transformación de los campos*:
  - *Numerización / etiquetado*:
    - Ventajas:
      - Se reduce espacio. Ej: *apellido*  $\Rightarrow$  *entero*.
      - Se pueden utilizar técnicas más simples.

- Desventajas:
  - Se necesita meta-información para distinguir los datos inicialmente no numéricos (la cantidad no es relevante) de los inicialmente numéricos (la cantidad es relevante: precios, unidades, etc.).
  - A veces se puede “sesgar” el modelo (*biasing*).
- *Discretización:*
  - Ventajas: Se reduce espacio. Ej.  $0..10 \Rightarrow$  (*pequeño, mediano, grande*). Se pueden utilizar árboles de decisión y construir reglas discretas.
  - Desventajas: Una mala discretización puede invalidar los resultados.

### 3) Selección de la Técnica de minería de datos a utilizar:

Esto incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc. La selección de él o de los algoritmos a utilizar. La transformación de los datos al formato requerido por el algoritmo específico de minería de datos. Y llevar a cabo el proceso de minería de datos, se buscan patrones que puedan expresarse como un modelo o simplemente que expresen dependencias de los datos, el modelo encontrado depende de su función (clasificación) y de su forma de representarlo (árboles de decisión, reglas, etc.).

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

Algunas *Características especiales* de los datos.

- Aparte del gran volumen, ¿por qué las técnicas de aprendizaje automático y estadística no son *directamente* aplicables?:
  - Los datos residen en el disco; no se pueden escanear múltiples veces.
  - Algunas técnicas de muestreo no son compatibles con algoritmos no incrementales.
  - Muy alta dimensionalidad (muchos campos).
  - Evidencia positiva.
  - Datos imperfectos.
- Aunque algunos se aplican casi directamente, el interés en la investigación en minería de datos está en su adaptación.

*Patrones* a descubrir:

Una vez recogidos los datos de interés, un explorador puede decidir qué tipo de patrón quiere descubrir. El tipo de conocimiento que se desea extraer va a marcar

claramente la *técnica* de minería de datos a utilizar. Según *como* sea la búsqueda del conocimiento se puede distinguir entre:

- *Directed data mining*: se sabe claramente lo que se busca, generalmente predecir unos ciertos datos o clases.
- *Undirected data mining*: no se sabe lo que se busca, se trabaja con los datos (*¡hasta que confiesen!*).

En el primer caso, algunos sistemas de minería de datos se encargan generalmente de elegir el *algoritmo* más idóneo entre los disponibles para un determinado tipo de patrón a buscar.

#### 4) Evaluación y validación:

La fase anterior produce una o más hipótesis de modelos. Para seleccionar y validar estos modelos es necesario el uso de **criterios de evaluación de hipótesis**. Por ejemplo:

- 1ª fase: Comprobación de la precisión del modelo en un **banco de ejemplos independiente** del que se ha utilizado para aprender el modelo. Se puede elegir el mejor modelo.
- 2ª fase: Se puede realizar una **experiencia piloto** con ese modelo. Por ejemplo, si el modelo encontrado se quería utilizar para predecir la respuesta de los clientes a un nuevo producto, se puede enviar un mailing a un subconjunto de clientes y evaluar la *fiabilidad del modelo*.

#### 5) Interpretación y difusión:

El despliegue del modelo a veces es trivial pero otras veces requiere un proceso de *implementación* o *interpretación*. El modelo puede requerir **implementación** (P.ej. Tiempo real de detección de tarjetas fraudulentas).

El modelo es descriptivo y requiere **interpretación** (P.ej. Una caracterización de zonas geográficas según la distribución de los productos vendidos). El modelo puede tener muchos usuarios y necesita **difusión**, puede requerir ser expresado de una manera comprensible para ser distribuido en la organización (P.ej. Las cervezas y los productos congelados se compran frecuentemente en conjunto ⇒ ponerlos en estantes distantes).

#### 6) Actualización y monitorización:

Los procesos derivan en un *mantenimiento* y producen realimentaciones en el proceso KDD.

- *Actualización*: Un modelo válido puede dejar de serlo por un cambio de contexto: Cambios económicos, en la competencia, en las fuentes de datos, etc.

- *Monitorización*: Consiste en ir revalidando el modelo con cierta frecuencia sobre nuevos datos. El objetivo es detectar si el modelo requiere una actualización.

## 5. Minería de Datos

### 5.1. Historia de la Minería de Datos

La minería de datos, entendida como la búsqueda de patrones dentro de grandes bases de datos utilizando para ello métodos estadísticos y de aprendizaje basado en computadora, está empezando a extenderse en nuestro país. Empresas en el sector de telecomunicaciones, financiero y de autoservicio están en el proceso de adquirir alguna solución tecnológica en este campo, por lo que surge una demanda por recursos humanos con conocimientos en minería de datos.

Además, al enfrentar un ambiente más competitivo las empresas requieren de tecnologías que les permitan pronosticar, dentro de un marco probabilística, el comportamiento de sus clientes y prospectos a fin de desarrollar estrategias de atracción o retención.

Aunque desde un punto de vista académico el término data mining es una etapa dentro de un proceso mayor llamado extracción de conocimiento en bases de datos, (mencionado en el capítulo anterior) en el entorno comercial, así como en este trabajo, ambos términos se usan de manera indistinta.

Lo que en verdad hace el data mining es reunir las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo, principalmente usando como materia prima las bases de datos.

Una definición tradicional es la siguiente: Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos (Fayyad y otros, 1996). Desde el punto de vista empresarial, se define como: La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión (Molina y otros, 2001).

La idea de data mining no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como data fishing, data mining o data archaeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de data mining y KDD. A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología; en 2002 existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones.

Las listas de discusión sobre este tema las forman investigadores de más de ochenta países. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

El data mining es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de data mining muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta.

La data mining es la etapa de descubrimiento en el proceso de KDD: Paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados (Fayyad et al., 1996). Aunque se suelen usar indistintamente los términos KDD y Minería de Datos.

## **5.2. Concepto de minería de datos**

Una definición de Minería de Datos es “el descubrimiento eficiente de información valiosa, no-obvia de una gran colección de datos”, cuyo objetivo “es ayudar a buscar situaciones interesantes con los criterios correctos, complementar una labor que hasta ahora se ha considerado “intelectual” y de alto nivel, privativa de los gerentes, planificadores y administradores. Además, de realizar la búsqueda fuera de horas pico, usando tiempos de máquina excedentes”.

La utilidad de la Minería de Datos ya no se pone a discusión, por lo cual está tecnología esta siendo aplicada por muchas herramientas de software. Las técnicas de aplicación varían de acuerdo a la herramienta, algunas la instrumentan haciendo uso de redes neuronales (SPSS Neural Connection), otras con generación de reglas (Data Logic) o Arboles de Decisión (XpertRule Profiler).

La minería de datos es definida por Kopanmakis & Theodoluidis (2003), como el proceso de descubrimiento de conocimiento sobre almacenes de datos complejos mediante la extracción oculta y potencialmente útil en forma de patrones globales y relaciones estructurales implícitas entre datos. Otros como Written & Frank (2003) apuntan que la minería de datos como aquel proceso en que se extrae conocimiento útil y comprensible, previamente desconocido, y a partir de grandes conjuntos de datos almacenados en distintos formatos.

Hernández (2005) aclara que existen muchos términos que se relacionan o utilizan como sinónimos de la minería de datos, una de ellas es el KDD el cual Fayyad (2002) define como el proceso no trivial de identificar patrones validos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos. A diferencia de la minería de datos es un proceso mas complejo que lleva no solo a obtención de modelos o patrones, que es el objetivo de la minería de datos, sino que incluye además una evaluación y una posible interpretación de los mismos.

## **5.3. Los Fundamentos de la Minería de Datos**

Las técnicas de Minería de Datos son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron

almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real.

Minería de Datos toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. La Minería de Datos está lista para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

- Recolección masiva de datos.
- Potentes computadoras con multiprocesadores.
- Algoritmos de Data Mining.

#### **5.4. Principales características y objetivos de la minería de datos**

- Explorar los datos se encuentran en las profundidades de las bases de datos, como los almacenes de datos, que algunas veces contienen información almacenada durante varios años.
- En algunos casos, los datos se consolidan en un almacén de datos y en mercados de datos; en otros, se mantienen en servidores de Internet e Intranet.
- El entorno de la minería de datos suele tener una arquitectura cliente servidor.
- Las herramientas de la minería de datos ayudan a extraer el mineral de la información enterrado en archivos corporativos o en registros públicos, archivados.
- El minero es, muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por barrenadoras de datos y otras poderosas herramientas indagatorias para efectuar preguntas adhoc y obtener rápidamente respuestas.
- Hurgar y sacudir a menudo implica el descubrimiento de resultados valiosos e inesperados.
- Las herramientas de la minería de datos se combinan fácilmente y pueden analizarse y procesarse rápidamente.
- Debido a la gran cantidad de datos, algunas veces resulta necesario usar procesamiento en paralelo para la minería de datos.
- La minería de datos produce cinco tipos de información:
  - Asociaciones.
  - Secuencias.
  - Clasificaciones.
  - Agrupamientos.
  - Pronósticos.

- Los mineros de datos usan varias herramientas y técnicas.

La minería de datos es un proceso que invierte la dinámica del método científico en el siguiente sentido:

En el método científico, primero se formula la hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten la hipótesis.

Si esto se hace con la formalidad adecuada (cuidando cuáles son las variables controladas y cuáles experimentales), se obtiene un nuevo conocimiento.

### **5.5. Etapas principales del proceso de Minería**

Podemos decir que "en Minería de Datos cada caso es un caso". Sin embargo, en términos generales, el proceso se compone de cuatro etapas principales:

1. Determinación de los objetivos. Trata de la delimitación de los objetivos que el cliente desea.
2. Preprocesamiento de los datos. Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Es la etapa que consume más de la mitad del tiempo del proyecto.
3. Determinación del modelo. Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
4. Análisis de los resultados: Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

### **5.6. Tipología de Patrones de Minería de Datos**

Existen diferentes Tipos de Conocimiento los cuales son:

- Asociaciones: Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta.  
Ejemplo: en un supermercado se analiza si los pañales y los potitos de bebé se compran conjuntamente.
- Dependencias: Una dependencia funcional (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Existen muchas dependencias nada interesantes (causalidades inversas).  
Ejemplo: que un paciente haya sido ingresado en maternidad determina su sexo.

- **Clasificación:** Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas.  
Ejemplo: se sabe (por un estudio de dependencias) que los atributos edad, grado de miopías y astigmatismo han determinado los pacientes para los que su operación de cirugía ocular ha sido satisfactoria. Podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.
- **Agrupamiento / Segmentación:** El agrupamiento (o clustering) es la detección de grupos de individuos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto.
- **Tendencias/Regresión:** El objetivo es predecir los valores de una variable continua a partir de la evolución sobre otra variable continua, generalmente el tiempo, o sobre un conjunto de variables.  
Ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.
- **Información del Esquema:** (descubrir claves primarias alternativas, R.I.).
- **Reglas Generales:** patrones no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

### 5.7. Herramientas de Minería de Datos

- **Redes Neuronales (Neural Networks):** Grupo de unidades no-lineales interconectadas y organizadas por capas. Estas pueden ser funciones matemáticas y números almacenados en computadoras digitales, pero pueden ser elaboradas también mediante dispositivos analógicos como los transistores a efecto de campo (FET). A pesar del incremento en velocidad y de la escala de integración en los semiconductores, la mejor contribución de las redes neuronales tendrá que esperar por computadoras más rápidas, masivas y paralelas.
- **Mapas característicos de Kohonen (Self-organizing Maps):** Es una red neuronal del tipo de entrenamiento no-supervisado. Los datos son mostrados a la estructura y esta se sensibiliza a los patrones presentes. Una vez entrenada es capaz de identificar tales patrones en nuevos datos.
- **Reconocimiento de patrones (Pattern Recognition):** Se trata de un grupo de técnicas orientadas a evaluar la similitud y las diferencias entre señales. Se involucran en esto a varios tipos de pre-procesamiento tales como la transformada de Fourier.

- k-nearest neighbor: Un procedimiento para clasificar a los "records" de un archivo mediante la identificación de grupos (clusters) y decidiendo a cual grupo pertenece cada uno de los "records".
- Algoritmo Genético (Genetic Algorithm): Imitando la evolución de las especies mediante la mutación, reproducción y selección, estos algoritmos proporcionan programas y optimizaciones que pueden ser utilizados en la construcción y entrenamiento de otras estructuras como las redes neuronales.

### 5.8. Glosario de Términos de Minería de Datos

- Algoritmos genéticos: Técnicas de optimización que usan procesos tales como combinación genética, mutación y selección natural en un diseño basado en los conceptos de evolución natural.
- Análisis de series de tiempo (time-series): Análisis de una secuencia de medidas hechas a intervalos específicos. El tiempo es usualmente la dimensión dominante de los datos.
- Análisis prospectivo de datos: Análisis de datos que predice futuras tendencias, comportamientos o eventos basado en datos históricos.
- Análisis exploratorio de datos: Uso de técnicas estadísticas tanto gráficas como descriptivas para aprender acerca de la estructura de un conjunto de datos.
- Análisis retrospectivo de datos: Análisis de datos que provee una visión de las tendencias, comportamientos o eventos basado en datos históricos.
- Árbol de decisión: Estructura en forma de árbol que representa un conjunto de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos.
- Base de datos multidimensional: Base de datos diseñada para procesamiento analítico on-line (OLAP). Estructurada como un hipercubo con un eje por dimensión.
- CART: (Árboles de clasificación y regresión) Una técnica de árbol de decisión usada para la clasificación de un conjunto de datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos creando 2 divisiones. Requiere menos preparación de datos que CHAID.
- CHAID Detección de interacción automática de Chi cuadrado: Una técnica de árbol de decisión usada para la clasificación de un conjunto de datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos utilizando tests de chi cuadrado para crear múltiples divisiones. Antecede, y requiere más preparación de datos, que CART.

- **Clasificación:** Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a variable(s) específica(s) las cuales se están tratando de predecir. Por ejemplo, un problema típico de clasificación es el de dividir una base de datos de compañías en grupos que son lo más homogéneos posibles con respecto a variables como "posibilidades de crédito" con valores tales como "Bueno" y "Malo".
- **Clustering (agrupamiento):** Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a todas las variables disponibles.
- **Computadoras con multiprocesadores:** Una computadora que incluye múltiples procesadores conectados por una red. Ver procesamiento paralelo.
- **Data cleansing:** Proceso de asegurar que todos los valores en un conjunto de datos sean consistentes y correctamente registrados.
- **Data Mining:** La extracción de información predecible escondida en grandes bases de datos.
- **Data Warehouse:** Sistema para el almacenamiento y distribución de cantidades masivas de datos.
- **Datos anormales:** Datos que resultan de errores (por ej.: errores en el tipeado durante la carga) o que representan eventos inusuales.
- **Dimensión:** En una base de datos relacional o plana, cada campo en un registro representa una dimensión. En una base de datos multidimensional, una dimensión es un conjunto de entidades similares; por ej.: una base de datos multidimensional de ventas podría incluir las dimensiones Producto, Tiempo y Ciudad.
- **Modelo analítico:** Una estructura y proceso para analizar un conjunto de datos. Por ejemplo, un árbol de decisión es un modelo para la clasificación de un conjunto de datos.
- **Modelo lineal:** Un modelo analítico que asume relaciones lineales entre una variable seleccionada (dependiente) y sus predictores (variables independientes).
- **Modelo no lineal:** Un modelo analítico que no asume una relación lineal en los coeficientes de las variables que son estudiadas.
- **Modelo predictivo:** Estructura y proceso para predecir valores de variables especificadas en un conjunto de datos.

- Navegación de datos: Proceso de visualizar diferentes dimensiones, "fetas" y niveles de una base de datos multidimensional.
- OLAP Procesamiento analítico on-line (On Line Analytic processing): Se refiere a aplicaciones de bases de datos orientadas a array que permite a los usuarios ver, navegar, manipular y analizar bases de datos multidimensionales.
- Outlier: Un ítem de datos cuyo valor cae fuera de los límites que encierran a la mayoría del resto de los valores correspondientes de la muestra. Puede indicar datos anormales. Deberían ser examinados detenidamente; pueden dar importante información.
- Procesamiento paralelo: Uso coordinado de múltiples procesadores para realizar tareas computacionales. El procesamiento paralelo puede ocurrir en una computadora con múltiples procesadores o en una red de estaciones de trabajo o PCs.
- Regresión lineal: Técnica estadística utilizada para encontrar la mejor relación lineal que encaja entre una variable seleccionada (dependiente) y sus predicados (variables independientes).
- Regresión logística: Una regresión lineal que predice las proporciones de una variable seleccionada categórica, tal como Tipo de Consumidor, en una población.
- SMP Multiprocesador simétrico (Symmetric multiprocessor): Tipo de computadora con multiprocesadores en la cual la memoria es compartida entre los procesadores.

## 6. Bibliografía

### Básica:

- Alur, N.; Haas, P.; Momirovska, D.; Read, P.; Summers, N.; Totanes, V.; Zuzarte, C. DB2 UDB's High Function Business Intelligence in e-business - 1/E. IBM Corp, USA, 2002. SG24-6546-00.
- Adams, J.; Koushik, S.; Vasudeva, G.; Galambos, G. Patterns for e-business. A Strategy for Reuse. IBM Press, USA, 2001.
- Berry, M. J.; Linoff, G. Data Mining Techniques: for Marketing, Sales, and Customer Support. John Wiley & Sons, USA, 1997.
- Kiselev, M. V.; Ananyan, S. M.; Arseniev, S. B. Regression-Based Classification Methods and Their Comparison with Decision Tree Algorithms, In: Proceedings of 1st European Symposium on Principles of Data Mining and Knowledge Discovery. Springer, pp 134-144, Trondheim, Norway, 1997.
- Kimball, R.; Merz, R. The Data Warehouse Toolkit: Building The Web-Enabled Data Warehouse. John Wiley & Sons, USA, 2000.
- Kohonen, T. Self-Organizing Maps. Springer-Verlag, Berlín, Alemania, 1995.
- Rud, O. The Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management. John Wiley & Sons, USA, 2000.
- Witten, I. H.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (Morgan Kaufmann Series in Database Management Systems). Morgan Kaufmann Pub, USA, 1999.

### Complementaria:

- International Business Machines Corporation. WebSphere Studio Application Developer Version 4 for use with the Windows 2000 and Windows NT Operating System. First Edition. IBM Press, USA, 2002.
- Coulouris, G.; Dollimore, J.; Kindberg, V. Sistemas Distribuidos – Conceptos y Diseño – 3/E. Addison Wesley, España, 2001. ISBN 84-7829-049-4.

### Páginas Web:

1. IBM Scholars Programs Education Materials: <http://www.ibm.com/university/> S/ consulta del 27/05/07.
2. IBM Corporation: <http://www.ibm.com/software/soul/wsinfo> S/ consulta del 27/05/07.
3. The Queen's University of Belfast: [http://bbdd.escet.urjc.es/documentos/data\\_mining/dm-OHP-final\\_1.html](http://bbdd.escet.urjc.es/documentos/data_mining/dm-OHP-final_1.html) S/ consulta del 27/05/07.
4. Instituto Tecnológico de Monterrey, México: <http://www-cia.mty.itesm.mx/~fcantu> S/ consulta del 27/05/07.