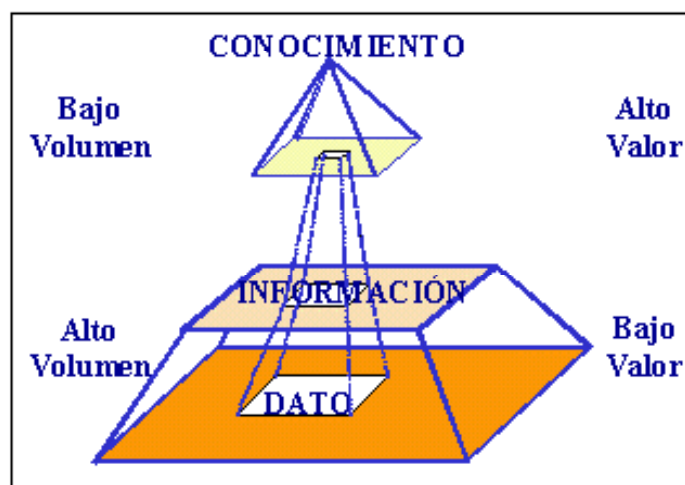


Universidad Nacional del Nordeste  
Facultad de Ciencias Exactas, Naturales y Agrimensura

Monografía de Adscripción  
Asignatura: Diseño y Administración de Datos

## MINERÍA DE DATOS



Mariana Inés Kubski - L.U.: 30.524  
Prof. Director: Mgter. David Luis la Red Martínez

Licenciatura en Sistemas de Información  
Corrientes - Argentina

2004



# Índice General

<b>1</b>	<b>Introducción</b>	<b>1</b>
1.1	Bases de Datos . . . . .	4
1.2	Data Warehouse . . . . .	5
1.2.1	Objetivos y Componentes del Data Warehouse . . . . .	8
1.2.2	Construcción del Data Warehouse . . . . .	9
1.2.3	Data Warehouse como Soporte de Decisión para los Ne- gocios . . . . .	10
1.2.4	Inteligencia de Negocios . . . . .	11
1.2.5	Principales Problemas del Data Warehouse . . . . .	12
1.3	Data Marts . . . . .	13
1.4	Sistemas OLAP . . . . .	13
1.5	Sistemas OLTP . . . . .	15
1.5.1	Diferencias entre Data Warehouse y OLTP . . . . .	15
<b>2</b>	<b>Descubriendo Conocimiento</b>	<b>19</b>
2.1	KDD: Knowledge Discovery Database . . . . .	21
2.2	Relación de KDD con otras áreas . . . . .	23
2.3	Componentes . . . . .	23
2.4	Etapas del Proceso . . . . .	25
2.5	Características de las Técnicas . . . . .	31
2.6	Técnicas de KDD . . . . .	32
2.6.1	Método Probabilístico . . . . .	33
2.6.2	Método Estadístico . . . . .	33
2.6.3	Método de Clasificación . . . . .	34
2.6.4	Desviación y Tendencia del Análisis . . . . .	34
2.6.5	Método Híbrido . . . . .	35
<b>3</b>	<b>Minería de Datos</b>	<b>37</b>
3.1	Antecedentes . . . . .	39

3.2	Por que “Minería de Datos” . . . . .	40
3.3	Características y Objetivos . . . . .	41
3.4	Arquitectura . . . . .	41
3.5	Aplicaciones Actuales . . . . .	42
3.6	Etapas Principales del Proceso de Minería . . . . .	43
3.7	Herramientas . . . . .	44
3.8	Modelos y Técnicas . . . . .	45
	3.8.1 Modelos Predictivos o Supervisados . . . . .	46
	3.8.2 Modelos de Descubrimiento o No Supervisados . . . . .	47
3.9	Extensiones . . . . .	48
	3.9.1 Web Mining . . . . .	48
	3.9.2 Text Mining . . . . .	50
3.10	Data Mining y Estadística . . . . .	50
3.11	Ventajas . . . . .	53
3.12	Desventajas . . . . .	54
	<b>Bibliografía</b>	<b>55</b>
	<b>Índice de Materias</b>	<b>57</b>

# Índice de Figuras

1.1	Proceso Clave en la Gestión del Conocimiento. . . . .	4
1.2	Arquitectura Data Warehouse. . . . .	7
1.3	Fases de Desarrollo de un Data Warehouse. . . . .	10
2.1	Evolución hacia la Minería de Datos. . . . .	20
2.2	Relación entre Dato, Información y Conocimiento. . . . .	22
2.3	Áreas Relacionadas con el KDD. . . . .	24
2.4	Componentes del KDD. . . . .	25
2.5	Proceso de Descubrimiento de Conocimiento. . . . .	31
3.1	Aproximación Estándar y de la Minería de Datos a la Detección de Información. . . . .	39
3.2	Evolución Histórica de la Minería de Datos. . . . .	40
3.3	Aplicaciones, Técnicas, y Algoritmos de la Minería de Datos. . . . .	49



# Capítulo 1

## Introducción

En estos últimos años ha sido posible optimizar en gran manera la gestión del almacenamiento de la información.

Como Gestión del Conocimiento se entiende al intento que se realiza por organizar los conocimientos en las bases de datos. Es la combinación de ciertas actividades como ser reunión, organización, división, análisis, y difusión del conocimiento con el fin de perfeccionar el desempeño de una organización.

De todos los tipos de capital intelectual, el del conocimiento es el más complejo y el más difícil de gestionar. Esta disciplina no es nueva, sino que sus raíces se remontan a la inteligencia artificial, que tiene como objetivo final la sintetización del comportamiento humano mediante ordenadores.

Desde hace algunos años se viene hablando de la gestión del conocimiento de forma creciente en medios de comunicación, mesas redondas, y conferencias. La atención que se le está prestando está creciendo a una velocidad impresionante.

¿ Porqué se habla tanto de gestión del conocimiento ?

Desde comienzos del 2000 se viene contemplando la emergente y paulatina importancia de la gestión del conocimiento. Surge la necesidad de un nuevo enfoque que integre modelos de gestión de negocio anteriores y que se centre en las personas como instrumentos capaces de incrementar los conocimientos y crear valores para la empresa.

Los gestores habituados a las reducciones de costes y al incremento con-

tinuo de la calidad, como instrumentos de gestión, vuelcan su atención al crecimiento a través del conocimiento y la innovación.

Las encuestas indican que la mayoría de las empresas consideran la gestión del conocimiento como un elemento decisivo de su estrategia. Pero el concepto mismo de gestión del conocimiento es problemático. La abundancia de términos que lo rodean crea confusión, y ponerlo en práctica exige la existencia previa de una sólida cultura del aprendizaje, y puede ser costosa.

La creciente aceptación de Internet permite a los trabajadores explorar más y más oportunidades laborales. Los trabajadores son más demandantes y tiene mayor movilidad que nunca. Las organizaciones reconocen que sus activos claves son los activos intelectuales y que sus prácticas de contratación y de retención deben reflejar esta situación. Este reconocimiento está ocasionando en las organizaciones alrededor de todo el mundo que desarrollen nuevos métodos de medida y gestión de estos activos.

Por lo tanto, se puede afirmar que la gestión del conocimiento ha despertado un gran interés entre los distintos agentes del mundo empresarial y económico.

Se la considera una herramienta de colaboración empresarial que los paquetes groupware utilizan para organizar, manejar y compartir las diversas formas de información empresarial generada por individuos y equipos en una organización.

Esta información se almacena en bibliotecas de documentos, bases de datos, depósito de documentos. Las bases de documentos forman parte de los sistemas de administración del conocimiento que día a día son desarrollados y utilizados por muchas empresas.

El groupware es un software de colaboración que ayuda a los equipos y grupos de trabajo a trabajar juntos en un variedad de formas. Proporciona muchas herramientas de software para comunicaciones electrónicas, conferencias y administración de trabajo en colaboración.

Las tecnologías de Gestión de Conocimiento deben permitir:

- Identificar los conocimientos necesarios.
- Identificar dónde y quién tiene el conocimiento o si necesita ser creado.
- Reunir y capturar el conocimiento identificado.

- Determinar su importancia.
- Resumir y sintetizar la información disponible.
- Distribuir la información a distintos niveles.
- Actualizar, eliminar y modificar el conocimiento obsoleto.
- Guardar y organizar el conocimiento obsoleto para futuras consultas.

La Gestión del Conocimiento es la nueva filosofía empresarial, está siendo aceptada por universidades, organizaciones e instituciones de todo tipo. El nuevo activo de las empresas, el capital intelectual, se basa en el conocimiento y la experiencia que toda organización tiene dentro de sí. Sin embargo, la estrategia de convertir datos en información, y esta a su vez en conocimiento para una correcta toma de decisiones, requiere el uso de una interfaz con el usuario. Esta interfaz se está configurando como un portal corporativo: el portal de conocimiento.

La planificación, diseño, construcción y mantenimiento de un portal de conocimiento requiere de tecnologías de la información y las comunicaciones que se convierten en la espina dorsal de los programas de gestión de conocimiento.

El proceso clave en la Gestión del Conocimiento se puede observar en la Fig. 1.1 de la pág. 4.

Existen ciertos factores que han despertado el interés por la gestión del conocimiento. Entre ellos, se pueden citar:

- Las reestructuraciones y reajustes que han disuelto redes personales y han ocasionado la pérdida de expertos que poseían conocimientos y experiencias para investigar, analizar problemas y encontrar soluciones.
- Las inversiones masivas en tecnologías de información y comunicaciones que permiten un mejor acceso a la información interna y externa a las corporaciones. Tecnologías que facilitan los medios para crear información y capturar el conocimiento.
- La necesidad de un contacto más estrecho con los clientes.
- La necesidad de difundir y compartir las experiencias con un mayor número de empleados.

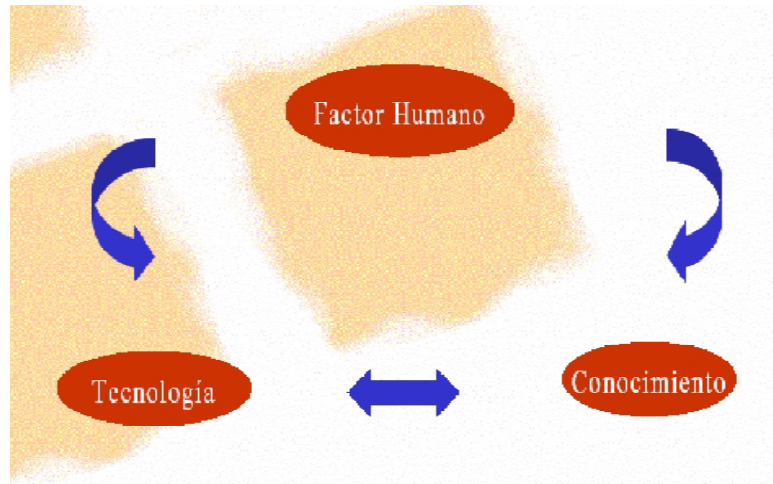


Figura 1.1: Proceso Clave en la Gestión del Conocimiento.

- La necesidad de unificar y adaptar respuestas ante la globalización del mercado.
- Una mayor capacidad de los empleados para aportar soluciones.
- Un mayor interés por las técnicas para incorporar las mejores prácticas existentes en otras corporaciones (benchmarking).
- La demanda de los clientes de aprovecharse de las experiencias desarrolladas en cualquier punto de las corporaciones proveedoras.
- La necesidad de colaboración de los trabajadores del conocimiento.
- La necesidad de reducir los tiempos de respuesta en desarrollo de productos y atención a clientes.

## 1.1 Bases de Datos

Una Base de Datos es un conjunto de información relacionada que se encuentra agrupada o estructurada. Proporciona la infraestructura necesaria para almacenar, recuperar, y manipular datos.

El continuo avance de la tecnología ha producido un gran crecimiento en lo que se refiere a capacidad de generación y almacenamiento de datos.

El abaratamiento de los sistemas de almacenamiento, la automatización de procesos, las mejoras en la confiabilidad y en la velocidad de transmisión, y las mejoras en la velocidad de cómputo de los procesadores, son algunas de las razones que han hecho que las bases de datos crezcan a una proporción fenomenal, excediendo nuestra habilidad para interpretar y comprender tanta información.

Por otra parte, generalmente los datos almacenados no siempre cuentan con una estructuración y coherencia específica; sobre todo si existen varias personas que son las responsables del almacenamiento de la información. Los problemas que se pueden presentar son:

- Que diferentes tipos de datos representen el mismo concepto, por ejemplo la fecha que puede guardarse con dos o cuatro dígitos.
- Que existan diferentes niveles de precisión al representar un dato, como números reales que no se almacenen siempre de igual manera.

Esta situación se agrava cuando se utilizan sistemas informáticos y soportes diferentes.

Surge entonces la necesidad de unificar los distintos ficheros y bases de datos para poder comprenderlos. Por ello, se necesita de tecnologías que sirvan de guía para comprender el contenido de las Bases de Datos.

## 1.2 Data Warehouse

Existen varias definiciones respecto a este término, pero básicamente se puede decir que *es la combinación de hardware, software especializado y datos provenientes de distintas fuentes, que sirve de apoyo a la administración para la toma de decisiones.*

Es una técnica para consolidar y administrar datos de diversas fuentes con el propósito de responder preguntas de negocios y tomar decisiones, de una forma que no era posible hasta ahora.

Es un almacén destinado específicamente para mantener datos organizados. También se lo denomina Bodegón de Datos, o simplemente Almacén de Datos.

Un Data Warehouse es una colección de datos orientados a temas integrados, no volátiles y variantes en el tiempo, organizados para soportar necesidades empresariales. De ello, se establece que un Data Warehouse se caracteriza por ser integrado, temático, histórico, y no volátil.

Integrado, es decir que al fluir del entorno operacional al entorno de almacén de datos, los datos asumen una codificación consistente.

Temático, debido a que almacena información resumida que se estructura en función de temas empresariales u organizacionales.

Histórico, dado que contiene suficiente espacio para almacenar datos que posean una antigüedad de cinco a diez años, o aun mayor.

No volátil, es decir, los datos no se modifican o cambian bajo ningún concepto una vez introducidos en el almacén de datos; únicamente pueden ser cargados, leídos y / o accedidos.

Dos beneficios claves que provee el Data Warehouse son, por un lado, la creación de una arquitectura de datos única para todas las aplicaciones, visualizada en la Fig. 1.2 de la pág. 7; y por el otro, la resolución de problemas de integridad y calidad de datos.

Su propósito es permitir que los Administradores de Bases de Datos redacten informes o analicen esas grandes cantidades de información, para así poder tomar decisiones según los resultados del análisis.

Un Administrador es aquel que se encarga del funcionamiento general del Sistema de las Bases de Datos dentro de una organización; debe contar con aptitudes técnicas para manejar el Sistema y con nociones de administración, pero, sobre todo, con un conocimiento profundo de las normas y políticas de la organización, así como con el criterio para saber cuando aplicarlas.

Básicamente, sus funciones son:

- Alimentar directamente la Base de Datos con aquella información que escape del dominio del usuario para asegurar su representatividad y utilidad para fines de análisis.
- Coordinar el diseño de aplicaciones con el área de informática para pre-

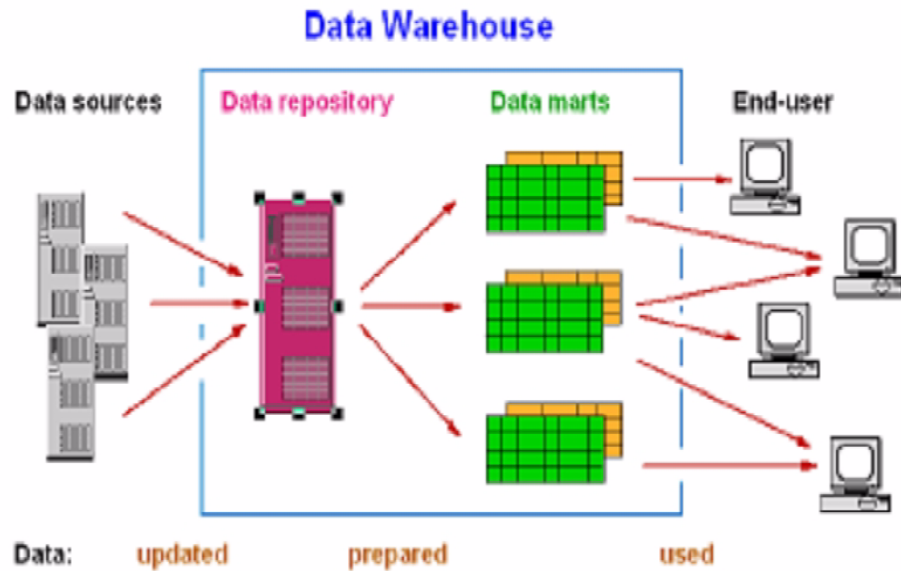


Figura 1.2: Arquitectura Data Warehouse.

servar la compatibilidad de los sistemas y facilitar el uso de la Base de Datos.

- Depurar continuamente la Base de Datos para garantizar su confiabilidad.
- Concientizar al Usuario sobre los usos y la utilidad de la Base de Datos para propiciar su máximo aprovechamiento.
- Brindar apoyo técnico al Usuario, Operador e Informático respecto al manejo y mantenimiento de la Base de Datos para evitar inconsistencias y contaminación de los datos.
- Analizar la información que emana periódicamente de la Base de Datos, cruzándola con aquella que generan los estudios de Mercados, para conformar alertas e informes oportunos.
- Elaborar los Informes o Reportes que sean acordados o aquellos que le sean solicitados con el propósito de informar a las Gerencias oportunamente.

Si bien un Data Warehouse es una Base de Datos, su modo de operar es muy diferente al de una Base de Datos en cuanto al soporte de transacciones y la actividad del negocio en línea.

Las Bases de Datos se diseñaron para las transacciones diarias. Almacenan la información de un sector de la organización, se actualizan a medida que llegan datos que deben ser almacenados y se operan mediante los cuatro mecanismos ya conocidos de “añadir, eliminar, modificar, imprimir”. Además manejan pequeños volúmenes de datos. Por ende, no son aptas como apoyo para la toma de decisiones.

El Data Warehouse almacena y resume información sobre transacciones cotidianas a lo largo del tiempo. Puede que contenga información que ya no es posible reproducir del sistema para la operación cotidiana porque es información primitiva, pero es útil porque indica el funcionar histórico de la organización.

Presenta una estructura multidimensional, con diferentes puntos de vista que reflejan los distintos aspectos de la organización.

Las consultas al almacén no son tan sistemáticas como las transacciones y usualmente demandan más recursos de cómputo. Inclusive resulta conveniente separar, por una parte, los equipos y sistemas de la operación cotidiana de transacciones, y por la otra, el Data Warehouse.

### 1.2.1 Objetivos y Componentes del Data Warehouse

- Colocar la mayor cantidad de información comercial posible en manos de tantos usuarios diferentes como sea posible.
- Mejorar el tiempo de espera que insumen los informes habituales.
- Monitorear el comportamiento de los clientes, competidores, y procesos internos.
- Mejorar la capacidad de respuesta a problemas comerciales.
- Incrementar la precisión de las mediciones.
- Aumentar la productividad.
- Incrementar y distribuir las responsabilidades.

Los principales componentes de un Data Warehouse son:

- Depósito para almacenar los datos.
- Herramientas para extraer, transformar y cargar fuentes de datos externos y opcionales.
- Herramientas para hacer referencia y analizar los datos en el depósito.

### 1.2.2 Construcción del Data Warehouse

Un Data Warehouse se genera a partir de otras bases de datos, su construcción y desarrollo requiere integrar varios componentes de tecnología y la habilidad para hacerlos funcionar todos juntos.

El objetivo fundamental es transformar datos en conocimiento. Para ello es necesario ensamblar datos existentes siguiendo instrucciones precisas para obtener un óptimo resultado.

Para su construcción se debe considerar en primer lugar el hardware necesario, dado que a mayor tamaño del almacén, mayor deberá ser la capacidad de almacenamiento y el poder de procesamiento. Luego el software y los datos ha utilizar.

Básicamente, las fases para la construcción del Data Warehouse son:

- Extracción: se crean los archivos de la base de datos para transacciones y se guardan en el servidor que mantendrá el almacén de datos. (Se extrae la información operacional).
- Depuración: se unifica la información de los datos de manera que se pueda insertar en el almacén de datos. (Se transforma la información a formatos consistentes).
- Carga: se transfiere los archivos depurados a la base de datos que servirá como almacén de datos. Se comparan los datos del almacén con los originales y se comprueba que estén completos.

En síntesis: ¿Cómo trabaja el Data Warehouse?:

- Extrae la información operacional.
- Transforma la operación a formatos consistentes.

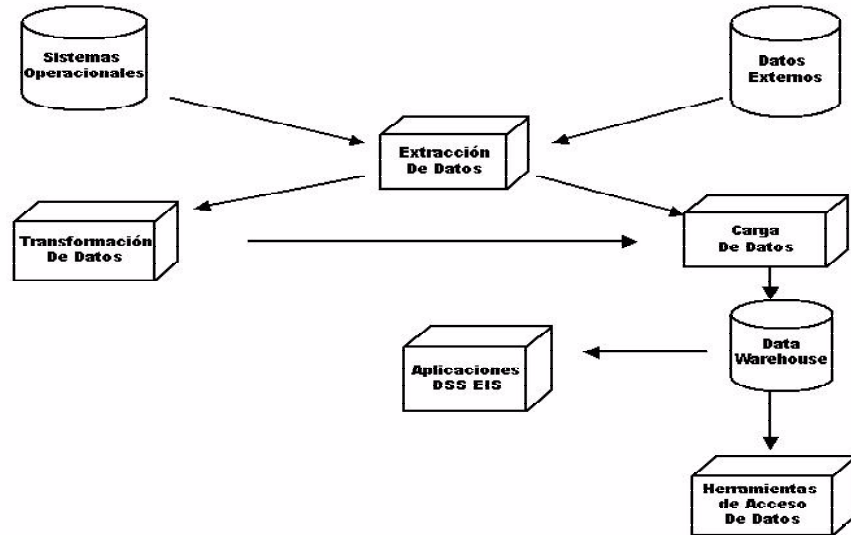


Figura 1.3: Fases de Desarrollo de un Data Warehouse.

- Automatiza las tareas de la información para prepararla a un análisis eficiente.

En la Fig. 1.3 de la pág. 10 se visualizan las fases que se deben llevar a cabo para la construcción de un Data Warehouse.

De todas maneras, el éxito de Data Warehouse no está en su construcción, sino en saber utilizarlo para mejorar procesos empresariales, operaciones y decisiones.

### 1.2.3 Data Warehouse como Soporte de Decisión para los Negocios

Los negocios necesitan aprovechar las posibilidades que les ofrece la actual tecnología para permanecer competitivos y rentables.

El conocimiento del mercado y de los clientes se ha convertido en un factor de supervivencia para las empresas. Y el Data Warehouse se perfila como la tecnología para lograr manejarlo.

Las organizaciones necesitan información renovada acerca de las tendencias presentes para mantener su competitividad. Precisan saber qué es lo que está pasando por las mentes de sus clientes.

Asimismo, necesitan determinar los requerimientos corporativos y traducirlos en consultas que puedan ser respondidas a través del Data Warehouse. Para ello, el Data Warehouse conserva información histórica y actual sobre un negocio, y permite recuperar datos que, bajo la forma de informes, facilitan el descubrimiento y la comprensión de patrones de comportamiento y tendencias de las cuales resultan conclusiones o recomendaciones para los futuros cursos de acción. Sintetiza algunos datos muy importantes, otorgando al usuario nuevo conocimiento comercial.

#### 1.2.4 Inteligencia de Negocios

Se refiere a un conjunto de productos y servicios para acceder a los datos, analizarlos y convertirlos en información. Se lo suele utilizar como sinónimo de soporte de decisiones.

La Inteligencia de Negocios es una manera de manejar la información histórica de una empresa a través de la construcción de un Data Warehouse y explotarla con fines de análisis para la mejor toma de decisiones. A través de la creación de modelos de información multidimensionales una organización puede beneficiarse al conocer de manera óptima cómo su negocio se ha comportado a lo largo del tiempo, cómo se comporta en el presente y cómo se estima se comportará en el futuro.

Algunos de los beneficios que obtienen las organizaciones al implementar este sistema son: capacidad de análisis, reducción de costos, reducción de tiempos de proceso, búsqueda de patrones desconocidos que sólo aparecen al momento de que los datos son analizados, generación de pronósticos, presupuestación y planeación.

La inteligencia en el negocio electrónico, incluye actividades como el procesamiento analítico en línea (OLAP) y aprovechamiento de datos, también llamada extracción de datos o Minería de Datos.

En Organizaciones que se ha utilizado el Data Warehouse como soporte para las decisiones se han notado varias diferencias con respecto a aquellas que no lo han hecho:

- El poder evaluar rápidamente las tendencias complejas, patrones y relaciones se convierte en una ventaja competitiva.
- Los usuarios se hacen cargo de tomar decisiones basadas en la información que obtienen. Además de que la propia organización los convierte en los responsables directos de la obtención de sus propios datos.
- Se admiten las debilidades internas y se corrigen los procesos que no funcionan. Existe una mayor disposición a modificar los procesos comerciales.
- Se realiza un procesamiento en ciclo cerrado. Es decir, se aprovecha la inteligencia de negocios no solamente para realizar el análisis necesario para emprender estas acciones, sino también para monitorear las acciones una vez realizadas.

### 1.2.5 Principales Problemas del Data Warehouse

Como en toda ciencia, uno de los principales problemas es creer que una nueva tecnología es tan buena o mejor que la tecnología que presume reemplazar, y a veces quizás el Data Warehouse no es la mejor solución a un determinado problema.

Invertir mucho tiempo en analizar las funciones y características de una tecnología dada no servirá de mucho si antes no se percibe si dicha tecnología realmente podrá satisfacer los requerimientos. Y adaptar el soporte de decisión si no se pueden soportar las decisiones posteriormente tampoco tiene sentido.

El modelado de datos recibe una cuota de atención desproporcionada, esto ha de resultar en tiempos de respuesta más lentos, generación de SQL más complejo, mayor dependencia de los metadatos referidos al software, y por ende, la reelaboración del modelo de datos.

El tamaño del equipo de desarrollo del Data Warehouse suele ser inversamente proporcional en su capacidad de producción. Muchos grupos de desarrollo de Data Warehouse tienen demasiado personal y poca dirección.

Los programadores son responsables del procesamiento que se lleva a cabo en el Data Warehouse, lo usan para cargar datos de prueba o probar prototipos de nuevas aplicaciones, y esto puede hacer mucho más daño. Una buena arquitectura, incluye un entorno de prueba separado, de modo tal de prote-

ger a los usuarios finales de las complejidades y los experimentos típicos de actividades de desarrollo.

### 1.3 Data Marts

Es un pequeño Data Warehouse, para un determinado número de usuarios, para un área funcional, específica de la organización. Es un subconjunto de una bodega de datos para un propósito específico.

El Data Marts puede extraerse del Data Warehouse de la organización aunque también es posible que el Data Warehouse se construya a partir de los Data Marts que se hayan diseñado.

Su funciones:

- Apoyar a otros sistemas para la toma de decisiones.
- Proporcionar dimensionalidad, permitiendo combinar múltiples entradas con una jerarquía reutilizable y extensible.
- Escalar a través de la empresa, desde soluciones para departamentos hasta plataformas empresariales, soportando todas las mayores bases de datos.
- Proporcionar un movimiento de datos consistente y la transformación para crear, cargar y desplegar modelos e informes de negocio.

### 1.4 Sistemas OLAP

On- Line Analytical Processing.

El procesamiento analítico en línea se define como el análisis rápido de información compartida. Aparece en contraposición al concepto tradicional OLTP ( On-Line Transactional Processing), que designa el procesamiento operacional de los datos, orientado a conseguir la máxima eficacia y rapidez en las transacciones individuales de los datos.

Es una aplicación de bases de datos orientada a array que permite visualizar, manipular y analizar bases de datos multidimensionales.

Permite a los usuarios analizar datos corporativos críticos para descubrir los factores decisivos que influyen en el negocio. Realiza todas las tareas analíticas y de reporte incluyendo informes de medidas de rendimiento del negocio que resaltan indicadores de rendimiento clave.

Beneficios claves:

- Permite a los usuarios de entender no solo lo que está pasando, sino cuando, donde, por qué y cómo.
- Resuelve todas las necesidades de análisis con una herramienta a velocidad electrónica.
- Proporciona capacidades de análisis para todos los tipos de usuario así como para clientes y proveedores.

Un server multidimensional OLAP permite que un modelo de negocios más sofisticado pueda ser aplicado cuando se navega por el Data Warehouse. Las estructuras multidimensionales permiten que el usuario analice los datos de acuerdo a como quiera mirar el negocio, por línea de producto, u otras perspectivas claves para su negocio. Considera unas variables en relación con otras y no de forma independiente entre sí. De esta manera, se pueden verificar hipótesis y resolver complejas consultas.

El usuario busca entonces nueva información, nuevos patrones que le sugieran relaciones entre diferentes aspectos apreciables de su actividad cotidiana. Por lo tanto, estas herramientas aun requieren de una alta participación del mismo, es él quien plantea las consultas, las dirige, y el análisis queda limitado por las ideas preconcebidas que éste pueda tener.

Estas herramientas ofrecen un mayor poder para revisar, graficar y visualizar información multidimensional. Los lenguajes restringidos y estructurados como SQL no son suficientes para el carácter explorador del OLAP.

El server de Data Mining debe estar integrado con el Data Warehouse y el server OLAP para insertar el análisis de negocios directamente en esta infraestructura. La integración con el Data Warehouse permite que decisiones operacionales sean implementadas directamente y monitoreadas.

A diferencia de OLAP, la Minería de Datos permite razonar de forma inductiva partiendo de los datos para llegar a una hipótesis general que modele

el problema. Es el propio sistema el que descubre nuevas hipótesis y relaciones. Además, la Minería de Datos trabaja con datos concretos, individuales, descubriendo patrones y generalizando a partir de allí. Las aplicaciones OLAP trabajan con datos agregados.

## 1.5 Sistemas OLTP

On-Line Transactional Processing.

Los sistemas de Procesamiento de Transacción En Línea (OLTP) están diseñados para mejorar la eficiencia operacional de una empresa. Se encargan del registro de las transacciones que reflejan el estado actual de los negocios. Los datos de transacción constituyen normalmente la mayor parte de la información en un Data Warehouse.

Una base de datos que soporta procesos transaccionales en línea (OLTP) se diseña para maximizar la capacidad transaccional de sus datos y típicamente tiene cientos de tablas.

Su diseño también se halla condicionado por los procesos operacionales que deberá soportar para la óptima actualización de sus datos, normalmente muchas de sus tablas sufren constantes y continuos cambios. Por ende, puede no ser adecuada para un sistema Data Warehouse. Estos últimos están orientados a procesos de consultas en contraposición con los procesos transaccionales.

### 1.5.1 Diferencias entre Data Warehouse y OLTP

Los sistemas tradicionales de transacciones y las aplicaciones de Data Warehouse son polos opuestos en cuanto a sus requerimientos de diseño y sus características de operación. Es muy importante comprender perfectamente estas diferencias para evitar caer en el diseño de un Data Warehouse como si fuera una aplicación de transacciones en línea (OLTP).

Las aplicaciones de OLTP están organizadas para ejecutar las transacciones para las cuales fueron creadas, como por ejemplo: mover dinero entre cuentas, un cargo o abono, una devolución de inventario, entre otras. Por otro lado, un Data Warehouse está organizado en base a conceptos, como por ejemplo: clientes, facturas, productos, etc.

Otra diferencia radica en el número de usuarios. Normalmente, el número de usuarios de un Data Warehouse es menor al de un OLTP. Es común encontrar que los sistemas transaccionales son accedidos por cientos de usuarios simultáneamente, mientras que los Data Warehouse sólo por decenas.

Los sistemas de OLTP realizan cientos de transacciones por segundo mientras que una sola consulta de un Data Warehouse puede tomar varios minutos.

Frecuentemente los sistemas transaccionales son menores en tamaño a los Data Warehouses, debido a que un Data Warehouse puede estar formado por información de varios OLTP.

Existen también diferencias en el diseño, mientras que el de un OLTP es extremadamente normalizado, el de un Data Warehouse tiende a ser desnormalizado. El OLTP normalmente está formado por un número mayor de tablas, cada una con pocas columnas, mientras que en un Data Warehouse el número de tablas es menor, pero cada una de éstas tiende a ser mayor en número de columnas.

Los OLTP son continuamente actualizados por los sistemas operacionales día a día, mientras que los Data Warehouse son actualizados en batch de manera periódica.

Las estructuras de los OLTP son muy estables, rara vez cambian, mientras las de los Data Warehouses sufren cambios constantes derivados de su evolución. Esto se debe a que los tipos de consultas a los cuales están sujetos son muy variados y es imposible preverlos todos de antemano.

En cuanto a los recursos humanos, las personas de negocios necesitan disponer de un enfoque fuerte sobre el conocimiento del área de la empresa y de los procesos empresariales. Es muy importante considerar las cualidades de las personas, ya que el desarrollo del Data Warehouse requiere participación tanto de estas como de los especialistas tecnológicos; estos dos grupos de personas deben trabajar juntos, compartiendo su conocimiento y destrezas en un espíritu de equipo de trabajo, para enfrentar los desafíos de desarrollo del Data Warehouse.

No sólo son los gerentes de Sistemas deben participar en la decisión de adoptar un Data Warehouse; hacerlo así implica que el proyecto no avance más allá del plano tecnológico. También deben involucrarse los ejecutivos que saben que es necesario contar con información confiable para conocer su mercado.

Se debe establecer el tiempo no solamente para la construcción y entrega de resultados del Data Warehouse, sino también para la planeación del proyecto y la definición de la arquitectura. Estos últimos establecen un marco de referencia y un conjunto de estándares que son críticos para la eficacia del Data Warehouse.

Respecto a las tecnologías hay que considerar que muchas tecnologías nuevas son introducidas por el Data Warehouse, y el costo de ésta puede ser tan sólo la inversión inicial del proyecto.

Existen costos evolutivos, ajustes continuos del Data Warehouse a través del tiempo, que generalmente se deben a cambios de expectativas.

Los costos de crecimiento que implican incrementos en el tiempo en volúmenes de datos, del número de usuarios del Data Warehouse, conlleva a un incremento de los recursos necesarios.

El Data Warehouse requiere soportar cambios que ocurren tanto en el origen de datos que éste usa, como en las necesidades de la información que éste soporta.

Se producen cambios en la tecnología que pueden afectar la manera que los datos operacionales son almacenados. Un cambio en el ambiente operacional puede cambiar el formato, estructura o significado de los datos usados como origen para el Data Warehouse. De esta manera serían impactados los procesos de Extracción, Transformación y Carga de datos.

Aun así , se debe tratar de:

- Mejorar la entrega de Información: información completa, correcta, consistente, oportuna y accesible. Información que las personas necesitan, en el tiempo que la necesitan y en el formato que la necesitan.
- Mejorar el Proceso de toma de decisiones: con un mayor soporte de información se obtienen decisiones más rápidas; así también, la gente de negocios adquiere mayor confianza en sus propias decisiones y las del resto, y logra un mayor entendimiento de los impactos de sus decisiones.
- Lograr un impacto positivo sobre los Procesos Empresariales: cuando a la gente se le da acceso a una mejor calidad de información, la organización puede lograr por sí sola:

- Eliminar los retardos de los procesos empresariales que resultan de información incorrecta, inconsistente y /o no existente.
- Integrar y optimizar procesos empresariales a través del uso compartido e integrado de las fuentes de información.
- Eliminar la producción y el procesamiento de datos que no son usados ni necesarios, producto de aplicaciones mal diseñadas o ya no utilizadas.

## Capítulo 2

# Descubrimiento del Conocimiento en Bases de Datos

Día tras día e inagotablemente en todas partes del mundo se generan cantidades inconcebibles de información. Se estima que la cantidad de datos en el mundo almacenados en bases de datos se duplica cada 20 meses. Es así que hoy día, las organizaciones tienen gran cantidad de datos almacenados y organizados, pero no pueden sacarles provecho, si no disponen de herramientas para ello.

La capacidad de generar y almacenar información ha crecido considerablemente en los últimos tiempos.

Básicamente, lo que ha permitido poder generar tanta información es el desarrollo tecnológico a niveles exponenciales tanto en el área de cómputo como en la de transmisión de datos. 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.

Lamentablemente, las técnicas tradicionales de análisis de información no han tenido un desarrollo equivalente, por lo tanto, la velocidad en que se almacenan los datos es muy superior a la velocidad en que se analizan.

Aunque los datos en sí mismos no constituyen información hasta que algún

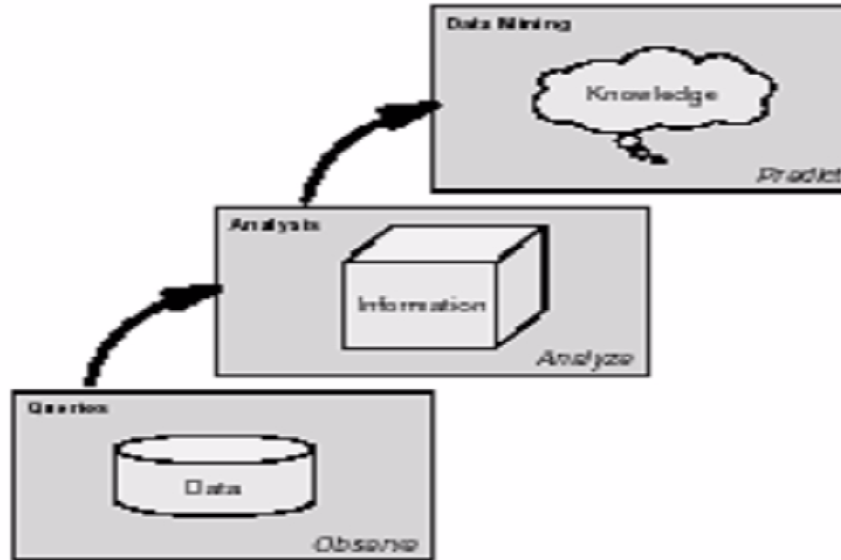


Figura 2.1: Evolución hacia la Minería de Datos.

usuario les atribuya algún significado especial. Pero cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación de la información y ese modelo representen un valor agregado, entonces nos referimos al conocimiento.

En la Fig. 2.1 de la pág. 20 se puede visualizar la evolución de los datos.

Existen datos que se almacenan en las empresas denominados dato-escritura, ya que sólo se graban o escriben en el disco duro, pero nunca se hace uso de ellos. Todas las empresas usan un dato llamado dato-escritura-lectura, que utilizan para hacer consultas dirigidas. Pero ha surgido un nuevo tipo de dato al cual se ha denominado dato-escritura-lectura-análisis. Éste proporciona en conjunto un verdadero conocimiento y es útil para el apoyo en la toma de decisiones. Pero esto no ha de resultar si no se cuenta con ciertas tecnologías que ayuden a explotar el potencial de este tipo de datos.

Desde hace ya un buen tiempo, existe un gran interés comercial por explotar los grandes volúmenes de información almacenada. Se considera que se está perdiendo una gran cantidad de información y conocimiento valioso que

se podrían extraer de los datos.

Existe la necesidad de generar nuevas 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.**

En la Fig. 2.2 de la pág. 22 se ilustra la jerarquía que existe en una base de datos entre datos, información y conocimiento. Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento.

El Data Mining trabaja en el nivel superior buscando patrones, comportamientos, agrupaciones, secuencias, tendencias o asociaciones que puedan generar algún modelo que nos permita comprender mejor el dominio para ayudar en una posible toma de decisión.

## 2.1 KDD: Knowledge Discovery Database

El descubrimiento del conocimiento se define como el *“Proceso de extracción no trivial para identificar patrones que sean válidos, novedosos, potencialmente útiles y entendibles, a partir de datos”*.

Es un proceso, ya que involucra varios pasos y es interactivo, al encontrar información útil en los datos. Los patrones deben ser:

- Válidos: deben brindar la posibilidad de poder ser aplicados en el futuro.
- Novedosos: descubrir información valiosa, para poder explorar así nuevos caminos.
- Útiles: para brindar conocimientos valiosos que asistan al usuario a la hora de tomar decisiones.
- Entendibles: que lleven a la comprensión del contenido de las bases de datos.

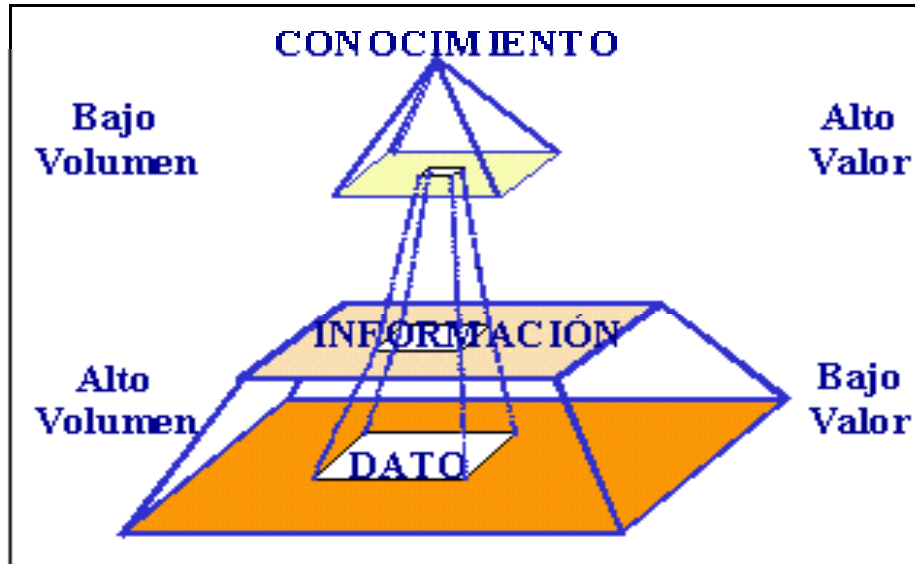


Figura 2.2: Relación entre Dato, Información y Conocimiento.

Al Descubrimiento de Conocimiento de Bases de Datos (KDD) también se lo conoce como Minería de Datos (Data Mining). Pero 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, es decir, el pre-procesamiento, minería de datos propiamente, y post-procesamiento.

Su objetivo fundamental es encontrar conocimiento útil, válido, relevante y nuevo sobre un fenómeno mediante algoritmos eficientes. También existe un profundo interés por presentar los resultados de manera visual o al menos de manera que se puedan interpretar claramente.

La interacción humano-máquina debe ser flexible, dinámica y colaboradora.

El resultado de la exploración debe ser interesante y la calidad no debe ser afectada por mayores volúmenes de datos o por ruido en los datos. En este sentido, los algoritmos de descubrimiento de información deben ser altamente robustos.

En síntesis, KDD pretende procesar automáticamente grandes cantidades

de datos, identificar los patrones mas relevantes y significativos, y finalmente presentarlos como conocimiento apropiado para así satisfacer las metas del usuario.

## 2.2 Relación de KDD con otras áreas

Debido a la aplicación potencial del descubrimiento del conocimiento, en diversas áreas hay un crecimiento de oportunidades en la investigación sobre este campo.

En el área de Tecnologías de bases de datos y bodegas de datos, por las formas eficientes de almacenar, acceder y manipular los datos.

En el de Aprendizaje computacional, estadística, computación suave (redes neuronales, lógica difusa, algoritmos genéticos, razonamiento probabilístico, entre otros), mediante el desarrollo de técnicas para extraer conocimiento a partir de los datos.

También en el área de Reconocimiento de Patrones mediante el desarrollo de herramientas de clasificación.

En el de Visualización de Datos, que permite por un extremo, el uso de una interfaz entre los humanos y los datos, y por el otro, entre los humanos y los patrones.

Y en el de Cómputo de Alto Desempeño, para mejorar el desempeño de los algoritmos debido a la complejidad y a la cantidad de datos que manejan los mismos.

En la Fig. 2.3 de la pág. 24 se indican las áreas relacionadas con el KDD.

## 2.3 Componentes

Dentro de los componentes que integran el Descubrimiento de Conocimiento en Bases de Datos podemos rescatar:

- Conocimiento del dominio y preferencias del usuario: incluye el diccionario de datos, información adicional de las estructuras de los datos,

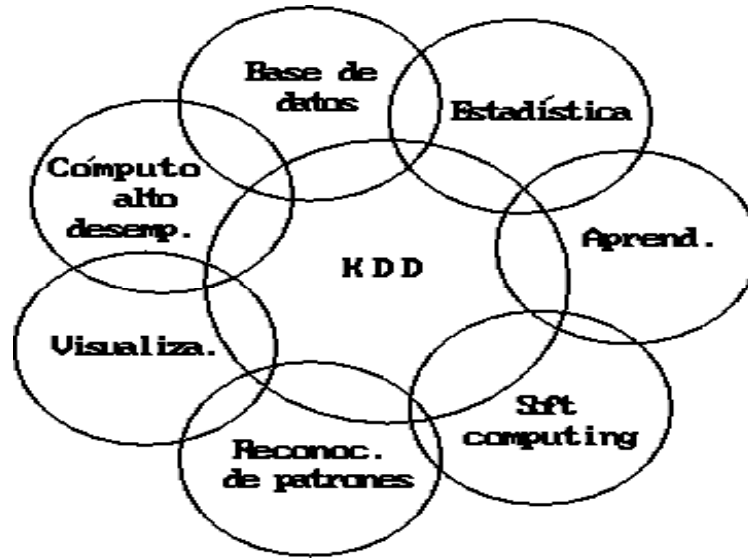


Figura 2.3: Áreas Relacionadas con el KDD.

restricciones entre campos, metas o preferencias del usuario, campos relevantes, listas de clases, jerarquías de generalización, modelos causales o funcionales, entre otros.

El objetivo del conocimiento del dominio es orientar y ayudar en la búsqueda de patrones interesantes, aunque a veces esto pueda causar resultados contraproducentes.

- Control del descubrimiento: toma el conocimiento del dominio, lo interpreta y decide qué hacer. En la mayoría de los sistemas el control lo realiza 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: consiste en juntar valores.

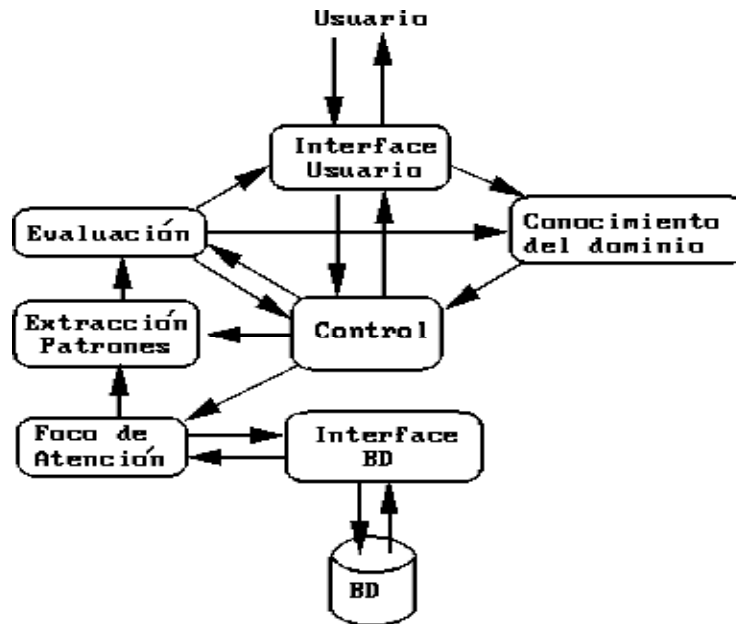


Figura 2.4: Componentes del KDD.

- Partición de datos: se particionan los datos en base a los valores que asuman los atributos.
- Proyección: consiste en ignorar algún o algunos atributos.
- Extracción de patrones.
- Evaluación.

La partición y proyección implican menos dimensiones. Y agregación y proyección implican menos dispersión. Los patrones son cualquier relación entre los elementos de la base de datos. Pueden incluir medidas de incertidumbre.

## 2.4 Etapas del Proceso

El proceso de descubrimiento del conocimiento toma los resultados tal como vienen de los datos, en forma cuidadosa y con precisión, y los transforma en información útil y entendible. No es información que habitualmente se pueda

recuperar mediante técnicas normales, pero se descubre mediante el uso de técnicas de Inteligencia Artificial.

Involucra varias fases:

- **Comprender el Dominio de la Aplicación:** se debe identificar cual es el problema que se desea resolver, aun con la suposición, o bien conociendo, que la respuesta se halla sepultada en alguna parte de los datos, aunque no se sabe exactamente donde está.

Es indispensable entonces contar con una clara descripción del problema a resolver; comprender sobre el total de los datos, cuales pueden ser los más importantes, y tener una visión lo más acabada posible de que tan lejos se está dispuesto a llegar, una vez que se conozcan los resultados. Es más, saber que tan lejos se está dispuesto a llegar permite muchas veces clarificar cual es el problema, e inclusive determinar que datos se podrán utilizar.

En el momento de identificar el problema, se debe pensar en términos de patrones y relaciones, que es con lo que opera el Data Mining.

Es muy importante comprender los datos pues en esta fase se determina que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original. Se debe determinar si los datos con los que se cuenta son suficientes para hallar la solución, si son realmente útiles. Algunas veces no resulta obvio que esos datos no pueden proveer la respuesta que se está buscando, por ello la importancia de prestar total atención a este punto.

- **Diseñar el Almacén de Datos:** se debe diseñar un esquema del almacén de datos que consiga unificar de manera operativa toda la información recogida.

La información que se quiere investigar generalmente se encuentra en bases de datos muy diversas, tanto internas como externas, que suelen utilizarse para el trabajo transaccional.

Si bien no todas las arquitecturas Data Warehouse son las mismas, una manera en que pueden ser usadas eficientemente para soportar aplicaciones es como se visualiza en la Fig.1.2 de la pág.7.

En este caso, cada aplicación de usuario final es soportada por su propio Data Marts, el cual se actualiza a intervalos regulares o cuando determinados datos cambian.

En esta estructura, cada Data Marts contiene datos concretos, y retiene conocimientos acerca de cuantos datos fueron derivados, el formato usado, que agregados fueron realizados, entre otros.

Se debe seleccionar un conjunto o subconjunto de bases de datos, enfocar la búsqueda en subconjuntos de variables, y seleccionar muestras de datos en donde realizar el proceso de descubrimiento.

Los datos y metadatos conforman lo que se denomina el Modelo de Datos.

Típicamente, un modelo de datos define la fuentes de datos utilizadas, los tipos de datos, su contenido, descripción, y el uso de los mismos.

La fuente de datos indica la ubicación física donde éstos fueron almacenados.

El tipo define como están estructurados los mismo, por ejemplo, el formato de tiempo que utilizan.

El contenido de datos lista las tablas o archivos de datos, y los campos que éstos contienen.

El uso de datos considera la pertenencia de las tablas y campos, que usuarios comprenden su contenido, y cuales lo explotan.

El modelo de datos también contiene información acerca de cuando se considera al dato como válido.

Cada registro puede contener una o más variables, donde cada variable puede derivar de un cierto número de fuentes diferente. En la mayoría de las aplicaciones, los tipos de datos más comunes son:

- Datos transaccionales: son datos propios de las operaciones diarias realizadas. Algunas de estas transacciones incluyen cierto grado de detalle respecto a su contenido.
- Datos relacionales: son datos de contenido relativamente no volátil. Por lo general se trata de información estable sobre clientes, equipos, productos, y procesos de trabajo.
- Datos demográficos, provenientes de fuentes externas.

Cuando el modelo de datos es requerido para soportar una aplicación que contiene requerimientos específicos, entonces los datos a utilizar pueden ser definidos por los usuarios finales. Se los debe interrogar acerca de cual es la información que están necesitando, y en base a ello, realizar las agregaciones necesarias para soportar estos requerimientos.

Generar un modelo de datos para aplicaciones de Minería de Datos puede consumir mucho más tiempo del que se cree.

La tendencia es utilizar manejadores de bases de datos y almacenes de datos que están optimizados para realizar un proceso analítico.

- **Seleccionar y Preprocesar los Datos:** una vez definido el modelo de datos que proveerá la estructura necesaria, en términos de las variables que se van a minar, es necesario suministrar los datos a utilizar.

Se debe identificar, coleccionar, filtrar y agregar los datos en el formato requerido por el modelo de datos.

El proceso que abarca desde la identificación de los datos hasta su preparado, es el que mayor tiempo consume en cualquier proceso de minería.

Al identificar las fuentes se debe considerar su origen y su contenido.

La Minería de Datos, al igual que otras herramientas de análisis, requiere que los datos estén consolidados en una única tabla o archivo. Si las variables requeridas están distribuidas alrededor de un número variado de fuentes, entonces se debe realizar la consolidación de las mismas, de tal manera que se obtenga un conjunto de registro de datos consistentes.

Si los datos a utilizar no provienen de un Data Warehouse, entonces se deben aplicar funciones de preproceso para la limpieza, agregación, transformación y filtrado.

Las herramientas de Minería de Datos usualmente proveen capacidades limitadas para limpiar los datos debido a que este es un proceso especial pero existe una gran variedad de productos que pueden cumplir con este propósito eficientemente.

Agregación y filtrado se pueden realizar por diferentes vías, según la estructura precisa de las fuentes de datos.

Se debe diseñar una estrategia adecuada para manejar ruido, valores faltan-

tes, valores incompletos, valores no existentes, secuencias de tiempo, y casos extremos de ser necesario.

- **Evaluar el Modelo de Datos:** se realiza una evaluación inicial del modelo de datos. Para ello se lleva a cabo una serie de pasos:

El primer paso radica en la inspección visual, consiste en “curiosear” los datos de entrada con herramientas de visualización. Esto permite principalmente, detectar distribuciones inverosímiles.

El segundo paso tiene que ver con la identificación de inconsistencias y la resolución de errores. Excepcionalmente, las distribuciones halladas en el primer paso se originan por la mala recolección de los datos. Los valores distantes o ausentes producen resultados parciales.

El último paso consiste en la selección final de las variables para comenzar a correr el proceso de minería.

Las variables dependientes o altamente correlacionadas pueden ser halladas mediante tests estadísticos, o regresión lineal o polinómica. Luego del chequeo estadístico, no todas las variables son nominadas a permanecer, sino únicamente aquellas que posean una clara interpretación, y aquellas que tengan sentido para el usuario final.

Este paso se puede simplificar si se utiliza un modelo de datos previamente examinado.

- **Seleccionar la Técnica de Minería de Datos:** se debe seleccionar la técnica de Minería de Datos más satisfactoria .

Este paso no implica únicamente definir la técnica apropiada o la combinación de técnicas a utilizar, sino también la vía a través de la cual se aplicará la misma.

Existen varios tipos de técnicas o algoritmos disponibles: clasificación, asociación, clustering, predicción de valores, patrones similares, y secuencias similares.

La selección del método a utilizar a menudo resulta obvio. Dependerá del tipo de conocimiento que se desea extraer. Usualmente, el desafío no está en que técnica utilizar, sino en el camino a seguir para aplicarla.

Las técnicas de Minería de Datos requieren la selección de algunos parámetros, pero para ello hay que saber como funcionan estas técnicas, y que realiza cada uno de los parámetros. Esto se analiza detalladamente en el Capítulo Funciones de Minería.

Se tiene que especificar un criterio de preferencia para seleccionar un modelo dentro de un conjunto posible de modelos, y se debe especificar la estrategia de búsqueda a utilizar, aunque generalmente se halla predeterminada en el algoritmo de minería.

- Interpretar los Resultados: los resultados que se obtienen al aplicar alguna técnica de Minería de Datos pueden proporcionar una gran riqueza de información. Pero esta información muchas veces resulta difícil de interpretar. Por ello, es muy importante presentar estos resultados mediante formas fáciles de comprender.

Para esta etapa es necesario disponer de un conjunto de herramientas que ayuden a visualizar los resultados y así proveer la información que se desee.

De ser necesario, es posible regresar a pasos anteriores. Esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias.

La interpretación puede servir también para eliminar patrones redundantes o irrelevantes.

- Desplegar los Resultados: Quizás esta sea la etapa más importante de todas.

Si a la Minería de Datos se la considera únicamente como una herramienta de análisis, desde ya se está fracasando al realizar el potencial análisis que ésta tiene para ofrecer.

La Minería de Datos crea representaciones de los datos que son llamados modelos. Estos modelos son muy importantes, porque no solamente proveen un profundo conocimiento de la organización en sí, sino que también se puede desplegar para utilizarse en otros procesos de negocios.

El conocimiento se obtiene para realizar acciones, la idea es incorporarlo dentro de un sistema de desempeño o simplemente almacenarlo y reportarlo a las personas interesadas.

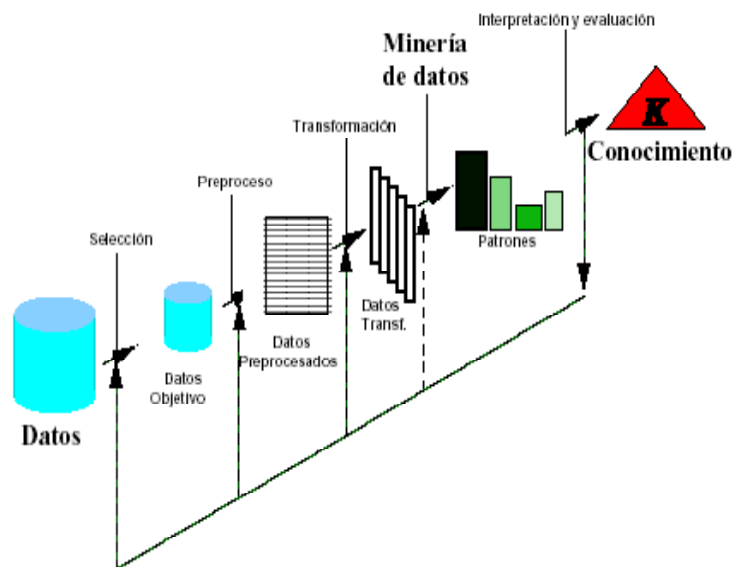


Figura 2.5: Proceso de Descubrimiento de Conocimiento.

La Fig. 2.5 de la pág. 31 muestra las etapas del proceso para el Descubrimiento de Conocimiento.

En este sentido, KDD implica un proceso interactivo e iterativo involucrando la aplicación de varios algoritmos de Minería de Datos.

## 2.5 Características de las Técnicas

Existen seis elementos esenciales que califican al Descubrimiento del Conocimiento como una técnica.

1. Todos los enfoques tratan con grandes cantidades de datos: se necesitan grandes cantidades de datos que proporcionen información suficiente para derivar un conocimiento adicional.
2. Se requiere de eficiencia debido al volumen de datos: es esencial que el proceso sea eficiente debido a la cantidad de datos con la que se trabaja.

3. Se considera a la Exactitud como un elemento esencial: para asegurar que el descubrimiento del conocimiento sea válido, es indispensable que exista exactitud.
4. Se requiere del uso de un lenguaje de alto nivel: los resultados deben ser presentados de una manera entendible para el usuario.
5. Todos los enfoques deben utilizar alguna forma de aprendizaje automatizado: Una de las mayores premisas de KDD es que el conocimiento se descubre usando técnicas de aprendizaje inteligente que van examinando los datos a través de procesos automatizados. KDD proporciona la capacidad para descubrir información nueva y significativa usando los datos existentes.
6. Se producen algunos resultados interesantes: si realmente se la considera una técnica útil para el Descubrimiento del Conocimiento, entonces el descubrimiento debe ser interesante; es decir, debe tener un valor potencial para el usuario.

La cantidad de datos que requieren procesamiento y análisis en grandes bases de datos exceden las capacidades humanas. La dificultad de transformar los datos con precisión es un conocimiento que va más allá de los límites de las bases de datos tradicionales. Por consiguiente, la utilización plena de los datos almacenados depende del uso que se otorgue a las técnicas del Descubrimiento del Conocimiento.

La utilidad de aplicaciones futuras en KDD es de largo alcance. KDD se puede usar como un medio de recuperación de información. Nuevos modelos o tendencias en los datos se pueden descubrir usando estas técnicas.

KDD también se puede usar como una base para las interfaces inteligentes del mañana, agregando un componente del Descubrimiento del Conocimiento a una máquina de bases de datos.

## **2.6 Técnicas de KDD**

Las técnicas pueden ser supervisadas o no supervisadas. En general, las técnicas de aprendizaje supervisadas disfrutan de la importancia del éxito definido

por la utilidad del Descubrimiento del Conocimiento. Los algoritmos de aprendizaje son complejos y generalmente considerados como la parte más difícil de cualquier técnica KDD.

Mientras el descubrimiento de una máquina confía solamente en métodos autónomos para el descubrimiento de la información, KDD combina métodos automatizados con la interacción humana para asegurar resultados exactos, útiles, y entendibles.

Existen muchos métodos diferentes que se clasifican como las técnicas de KDD. Hay métodos cuantitativos, como los probabilísticos y los estadísticos. Hay métodos que utilizan técnicas de visualización. Hay métodos de clasificación como la clasificación Bayesiana, lógica inductiva, descubrimiento de modelado de datos y análisis de decisión. Otros métodos incluyen la desviación y tendencia al análisis, algoritmos genéticos, redes neuronales y métodos híbridos que combinan dos o más técnicas.

Debido a las maneras en que estas técnicas pueden usarse y combinarse, existe una falta de acuerdos de cómo deben categorizarse.

### 2.6.1 Método Probabilístico

Esta familia de técnicas KDD utiliza modelos de representación gráfica para comparar las diferentes representaciones del conocimiento.

Estos modelos se basan en las probabilidades e independencia de los datos. Son útiles para aplicaciones que involucran incertidumbre y aplicaciones estructuradas tal que una probabilidad se puede asignar a cada uno de los resultados. Las técnicas probabilísticas pueden usarse en los sistemas de diagnóstico, planeación y sistemas de control.

### 2.6.2 Método Estadístico

El método estadístico utiliza la regla del descubrimiento y se basa en las relaciones de los datos. El algoritmo de aprendizaje inductivo puede seleccionar automáticamente trayectorias útiles y atributos para construir las reglas de una base de datos con muchas relaciones. Este tipo de inducción se utiliza para generalizar los modelos en los datos y construir las reglas de los modelos nombrados. El proceso analítico en línea (OLAP) es un ejemplo de un método

orientado a la estadística.

### 2.6.3 Método de Clasificación

La clasificación es probablemente el método más antiguo y mayormente usado de todos los métodos de KDD. Este método agrupa los datos de acuerdo a similitudes o clases. Existen muchos tipos de clasificación de técnicas y numerosas herramientas disponibles que son automatizadas.

El método Bayesian es un modelo gráfico que usa directamente arcos, exclusivamente para formar una gráfica acíclica. Aunque el método Bayesian usa los medios probabilísticos y gráficos de representación, también es considerado un tipo de clasificación.

Se usan muy frecuentemente las redes de Bayesian cuando la incertidumbre que se asocia con un resultado puede expresarse en términos de una probabilidad. Este método cuenta con un dominio del conocimiento codificado y ha sido usado para los sistemas de diagnóstico.

El descubrimiento de patrones y de datos es otro tipo de clasificación que sistemáticamente reduce una base de datos grande a unos cuantos archivos informativos. Si el dato es redundante y poco interesante se elimina, y la tarea de descubrir los patrones en los datos se simplifica. Este método trabaja en la premisa de un viejo dicho, “menos es más”. El descubrimiento de patrones y las técnicas de limpieza de datos son útiles para reducir enormes volúmenes de datos en las aplicaciones.

### 2.6.4 Desviación y Tendencia del Análisis

El método de detección por filtrado tiende a ser importante como base para este método de KDD. Normalmente las técnicas de análisis y desviación son aplicadas temporalmente en las bases de datos. Una buena aplicación para este tipo de KDD es el análisis de tráfico en las grandes redes de telecomunicaciones.

El volumen total de datos que requieren análisis generan una técnica imperativa automatizada. Este tipo de tendencia de análisis también podría demostrar utilidad en los datos astronómicos y oceanográficos, ya que sus datos se basan en el tiempo y son voluminosos.

### 2.6.5 Método Híbrido

También se lo conoce como método multi-paradigmático. Aunque la implementación puede ser más difícil, las herramientas híbridas son capaces de combinar la potencia de varios métodos. Algunos de los métodos comúnmente usados combinan técnicas de visualización, inducción, redes neuronales y sistemas basados en reglas para llevar a cabo el descubrimiento de conocimiento deseado. También se han usado bases de datos deductivas y algoritmos genéticos en los métodos híbridos.



## Capítulo 3

# Minería de Datos

Hace tan solo unos años, los datos de las empresas estaban orientados principalmente a alimentar sus sistemas contables, financieros, de inventarios, de producción, de recursos humanos y de ventas. En la medida que los negocios mundiales se hicieron más competitivos y complejos, los datos cada vez cobraron más vida y se convirtieron en información vital para la toma de decisiones de los gerentes.

El consumidor empieza a tener rostro y la diversidad prevaleciente en el mercado le ha cambiado el rostro al Mercadeo.

Entender al nuevo consumidor es una tarea cada vez más compleja, pues la antigua noción de desarrollar un producto e inducir su compra a un cliente potencial desprevenido mediante el uso de la publicidad masiva ya no existe. Para cada producto o servicio hay numerosas opciones de mercados posibles. Seleccionar el mercado y luego segmentarlo es una actividad colosal.

Es hora de comenzar a cavar y a construir un túnel en un escenario de mercado. Este sistema de excavación se denomina Data Mining y es la aplicación de las técnicas de la inteligencia artificial a grandes cantidades de datos para descubrir relaciones, tendencias y trayectorias ocultas con el propósito de convertir estos resultados en planes de negocios ejecutables, como redireccionar los esfuerzos de mercadeo o evaluar los centros de utilidades.

Al igual que para Data Warehouse, para el término Data Mining también existen varias definiciones. Algunas de ellas:

“Paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados” [2].

“Es 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” [3].

“La minería de datos es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos” [4].

Es un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos. Está muy ligada a las bodegas de datos ya que las mismas proporcionan la información histórica con la cual los algoritmos de minería de datos tienen la información necesaria para la toma de decisiones.

Es el análisis de archivos de transacciones, trabaja a nivel del conocimiento con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones. Ayuda a descubrir información rápidamente. Es una herramienta relacionada directamente al negocio.

Es un proceso que invierte la dinámica del método científico ya que en él, primero se formula la hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten la hipótesis.

En la minería de datos, en cambio, se coleccionan los datos y se espera que de ellos emerjan hipótesis. Se busca que los datos describan o indiquen por qué son como son. Luego se valida esa hipótesis inspirada por los datos. Por ello es que la minería de datos debe presentar un enfoque exploratorio, y no confirmador. Usar la minería de datos para confirmar las hipótesis formuladas puede ser peligroso, pues se está haciendo una inferencia poco válida.

Es incorrecto aceptar dichas hipótesis como explicaciones o relaciones causa-efecto. Es necesario coleccionar nuevos datos y validar las hipótesis generadas ante los nuevos datos, y después descartar aquellas que no son confirmadas por los mismos.

La Fig. 3.1 de la pág. 39 refleja la diferencia entre la Minería de Datos y el Método Científico.

No se la debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que

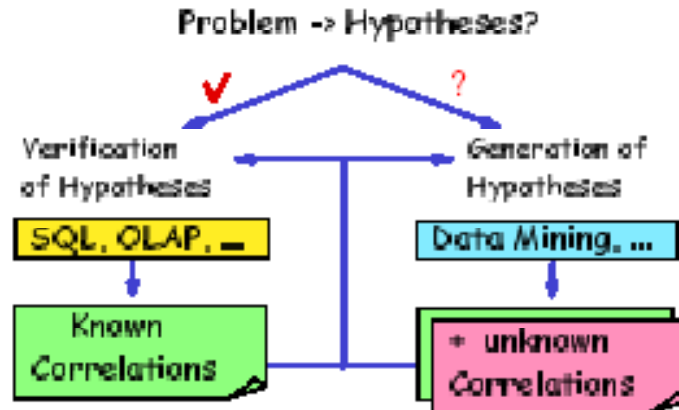


Figura 3.1: Aproximación Estándar y de la Minería de Datos a la Detección de Información.

pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente.

Se comienza a hablar de minería de datos cuando en el mercado se pone atención en el producto y en el cliente.

### 3.1 Antecedentes

La idea de minería de datos 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 minería de datos y KDD. A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología.

Actualmente son muchísimas las empresas en el mundo que ofrecen soluciones. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

La evolución de sus herramientas en el transcurso del tiempo, reflejada en la Fig. 3.2 de la pág. 40, puede dividirse en cuatro etapas principales:

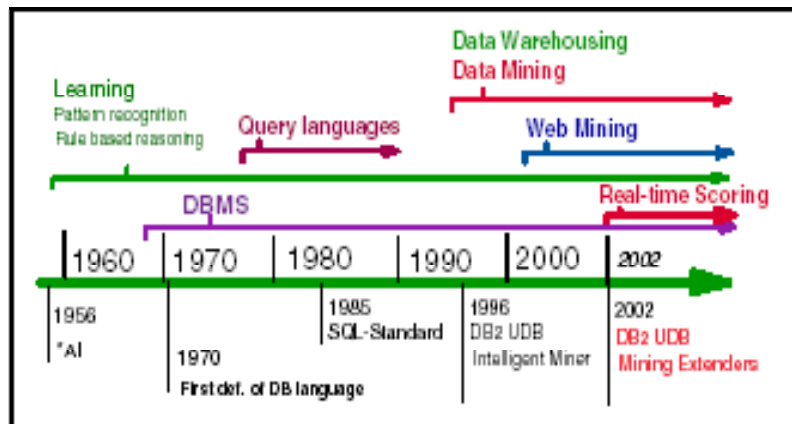


Figura 3.2: Evolución Histórica de la Minería de Datos.

- Colección de datos (1960).
- Acceso de datos (1980).
- Almacén de datos y apoyo a las decisiones (principios de la década de 1990).
- Minería de datos inteligente (finales de la década de 1990).

### 3.2 Por que “Minería de Datos”

Mediante la Excavación de Datos, se convierte una plataforma tecnológica en un sistema de información sobre el que se construyen soluciones de negocios.

Para que se entienda: demás está decir que el punto de partida es que las montañas de datos deben ser “de oro” y no “de chatarra”. De lo contrario no vale la pena excavar. Se necesitan equipos de cómputo, sistemas operativos y la infraestructura necesaria para apoyar ese proyecto minero. Luego vendrán técnicas de inteligencia artificial y de análisis estadístico que permitirán extraer el oro de la mina.

### 3.3 Características y Objetivos

- Los datos que realmente le interesan se encuentran en las profundidades de las bases de datos.
- Es más efectiva cuando los datos tienen elementos que pueden permitir una interpretación y explicación en concordancia con la experiencia humana. Estos elementos son el espacio y el tiempo.
- 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.
- El entorno de la minería de datos suele trabajar en una arquitectura cliente-servidor.
- El “minero” es muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por poderosas herramientas indagatorias para efectuar preguntas y obtener rápidamente respuestas.
- Al “hurgar y sacudir” los datos muchas veces se produce el descubrimiento de resultados valiosos e inesperados.
- Las herramientas de la minería de datos se combinan fácilmente y se pueden analizar y procesar rápidamente. Ayudan a extraer el “mineral” de la información que se halla enterrado en archivos corporativos o en registros archivados.
- Debido a la gran cantidad de datos con la cual se trabaja, algunas veces resulta necesario usar procesamiento en paralelo.
- La minería de datos produce diferentes tipos de información: asociaciones, secuencias, clasificaciones, agrupamientos, y pronósticos.
- La combinación de técnicas de almacenamiento de datos y el software de minería de datos facilita la predicción del futuro con base en patrones descubiertos en datos históricos.

### 3.4 Arquitectura

Para aplicar mejor las técnicas de Minería de Datos, éstas deben estar totalmente integradas con el Data Warehouse así como con herramientas flexibles

e interactivas para el análisis de negocios. Varias herramientas de Data Mining actualmente operan fuera del Warehouse, requiriendo pasos extra para extraer, importar y analizar los datos. Además, cuando nuevos conceptos requieren implementación operacional, la integración con el Warehouse simplifica la aplicación de los resultados desde Data Mining. El Data Warehouse analítico resultante puede ser aplicado para mejorar procesos de negocios en toda la organización, cualquiera sea el área.

El punto de inicio ideal es un Data Warehouse que contenga una combinación de datos de seguimiento interno junto con datos externos de mercado. Este Warehouse puede ser implementado en una variedad de sistemas de bases relacionales y debe ser optimizado para un acceso a los datos flexible y rápido.

La integración con el Data Warehouse permite que decisiones operacionales sean implementadas directamente y monitoreadas. A medida que el Data Warehouse crece con nuevas decisiones y resultados, la organización puede minar las mejores prácticas y aplicarlas en futuras decisiones.

### 3.5 Aplicaciones Actuales

En el mercado actual, Data Mining se encuentra disponible para ser aplicado en la comunidad de negocios porque está soportado por tres tecnologías, ya suficientemente desarrolladas, como ser la Recolección masiva de datos, las potentes computadoras con multiprocesadores, y los Algoritmos de Data Mining.

La madurez de estas técnicas, junto con los motores de bases de datos relacionales de alta performance, han hecho que estas tecnologías sean prácticas para los entornos de Data Warehouse actuales. En la actualidad, existe una gran variedad de situaciones en las cuales se puede aplicar la Minería de Datos:

- Análisis de canastas de mercado para mejorar la organización de tiendas, segmentación de mercado.
- Modelos para análisis de riesgos.
- Evaluación de campañas publicitarias.
- Análisis de la fidelidad de clientes: identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, selección de

sitios de tiendas, afinidad de productos, etc.

- Análisis de valores de bolsa y banca: análisis de clientes, aprobación de préstamos, determinación de montos de crédito, etc.
- Modelos de tráfico a partir de datos GPS.
- Perfiles de usuarios de redes.
- Detección de intrusos en redes.
- Astronomía: clasificación de cuerpos celestes.
- Aspectos climatológicos: predicción de tormentas, etc.
- Medicina: caracterización y predicción de enfermedades, probabilidad de respuesta satisfactoria a tratamiento médico.
- Industria y manufactura: diagnóstico de fallas.
- Detección de fraudes y comportamientos inusuales: telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, electricidad, etc.
- Determinación de niveles de audiencia de programas televisivos.
- Normalización automática de bases de datos.

### 3.6 Etapas Principales del Proceso de Minería

1. Determinación de los objetivos: delimitar los objetivos que el cliente desea bajo la orientación del especialista en Data Mining.
2. Preprocesamiento de los datos: especificar los datos de entrada que se desean explorar y analizar. Puede que una fuente no contenga todos los datos que se quieran utilizar, o bien puede que contenga datos irrelevantes. Por ello, se deben seleccionar, limpiar, enriquecer, reducir y transformar las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de Data Mining.
3. Determinación del modelo: se comienza realizando un análisis estadístico para analizar y explorar los datos, y luego 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: se verifica si los resultados obtenidos son coherentes y se los coteja con los obtenidos por el análisis estadístico y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

### 3.7 Herramientas

Las herramientas con las que opera el Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas y a las cuales los usuarios de esta información casi no están dispuestos a aceptar.

Muchas compañías ya colectan y refinan cantidades masivas de datos. Las técnicas de Data Mining pueden ser implementadas rápidamente en plataformas ya existentes de software y hardware para acrecentar el valor de las fuentes de información existentes. A su vez, puede ser implementadas en sistemas nuevos a medida que las plataformas existentes se actualizan y nuevos productos son desarrollados.

Cuando las herramientas de Data Mining son implementadas en sistemas de procesamiento paralelo de alta performance, pueden analizar bases de datos masivas en minutos. Procesamiento más rápido significa que los usuarios pueden automáticamente experimentar con más modelos para entender datos complejos. Alta velocidad permite que sea práctico para los usuarios analizar inmensas cantidades de datos. Grandes bases de datos, a su vez, producen mejores predicciones.

Una vez que las herramientas de Data Mining se implementan en computadoras cliente servidor de alta performance o de procesamiento paralelo, pueden analizar bases de datos masivas para brindar respuesta a cualquier tipo de preguntas.

Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer capacidades de:

- Predicción automatizada de tendencias y comportamientos. Data Mining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos.
- Descubrimiento automatizado de modelos previamente desconocidos. Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso.

### 3.8 Modelos y Técnicas

Modelado es el acto de construir un modelo en una situación donde se conoce la respuesta y luego se la aplica en otra situación de la cual se desconoce la respuesta. Esto se ha estado haciendo desde hace mucho tiempo, aun antes del auge de la tecnología de Data Mining.

El proceso de minería involucra ajustar modelos o determinar patrones a partir de datos. Este ajuste normalmente es de tipo estadístico, en el sentido que se permite un cierto ruido o error dentro del modelo.

Las computadoras se cargan con mucha información acerca de una variedad de situaciones y luego el software de Data Mining debe correr a través de los datos y, mediante alguna técnica, distinguir las características de los datos que llevarán al modelo. Una vez que el modelo se construye, puede ser usado en situaciones similares donde no se conoce la respuesta.

Con Data Mining, la mejor manera para realizar esto es dejando de lado ciertos datos para aislarlos del proceso. Una vez que el proceso está completo, los resultados pueden ser comparados contra los datos excluidos para confirmar la validez del modelo. Si el modelo funciona, las observaciones deben mantenerse para los datos excluidos.

La aplicación de los algoritmos de minería de datos requiere realizar una serie de actividades previas encaminadas a preparar los datos de entrada debido a que, en muchas ocasiones dichos datos proceden de fuentes heterogéneas, no tienen el formato adecuado o contienen ruido.

Luego se debe realizar el Análisis Preliminar de datos usando Herramientas de Consulta.

Se debe aplicar una consulta SQL a un conjunto de datos, para rescatar algunos aspectos visibles antes de aplicar las técnicas. La gran mayoría de la información, aproximadamente un 80 %, puede obtenerse con SQL. El 20 % restante conforma la información oculta, para la misma es donde se requiere utilizar de técnicas avanzadas.

Este primer análisis en SQL es para conocer cual es la distribución de los valores posibles de los atributos. Recién después se puede visualizar la performance del algoritmo correspondiente.

Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento [1].

### 3.8.1 Modelos Predictivos o Supervisados

Predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. A partir de datos cuyo atributo se conoce se induce una relación entre dicho atributo y otra serie de atributos. Básicamente entonces, hallan relaciones entre atributos.

Cuando una aplicación no es lo suficientemente madura y no tiene el potencial necesario para una solución predictiva, entonces hay que recurrir a los métodos no supervisados o de descubrimiento.

Las siguientes son algunas técnicas que se utilizan para crear modelos predictivos:

- **Clasificación:** se asignan los registros de datos en categorías predefinidas. Los datos son objetos caracterizados por atributos que pertenecen a diferentes clases. La meta es inducir un modelo para poder predecir una clase dados los valores de los atributos. Se utilizan:

Algoritmos genéticos: técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y selección natural en un diseño basado en los conceptos de evolución. Inspirados en el principio de la supervivencia de los más aptos. La recombinación de soluciones buenas en promedio produce mejores soluciones. Es una analogía con la evolución natural.

Redes neuronales artificiales: modelos predecible no-lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.

Árboles de decisión: estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Los métodos específicos de árboles de decisión incluyen:

CART (Árboles de clasificación y regresión): técnica utilizada para la clasificación de un conjunto de datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo conjunto de datos, sin clasificar, para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos creando dos divisiones. Requiere menos preparación de datos que CHAID.

CHAID Detección de interacción automática de Chi cuadrado: técnica similar a la anterior, pero segmenta un conjunto de datos utilizando tests de chi cuadrado para crear múltiples divisiones.

- Regresión o Estimación: se predice el valor de atributos continuos partiendo de otros atributos conocidos. La meta es inducir un modelo para poder predecir el valor de la clase dados los valores de los atributos. También utiliza árboles de regresión, regresión lineal, redes neuronales.

### 3.8.2 Modelos de Descubrimiento o No Supervisados

Descubren patrones y tendencias en los datos sin tener ningún tipo de conocimiento previo acerca de cuales son esos patrones buscados. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener beneficios científicos o de negocios.

El rango de éxito de estos modelos depende de la utilidad del descubrimiento de conocimiento.

Las siguientes son algunas de las técnicas que se utilizan para crear modelos descriptivos:

- Clustering: agrupa los datos en determinadas categorías llamados clusters, basándose en las similitudes de los mismos. Existen diferentes téc-

nicas de clustering y cada una de las mismas tiene sus propias aproximaciones para descubrir las aproximaciones que existen entre sus datos. Las aproximaciones las deben determinar los clusters. Las técnicas incluyen árboles de decisión, redes neuronales.

- **Análisis de Enlace (Link analysis):** describe una familia de técnicas que determinan asociaciones entre los registros de datos. El tipo de análisis de enlace más conocido es el Análisis de la canasta de mercado. El análisis de la Canasta de Mercados descubre la combinación de artículos que fueron comprados por diferentes consumidores, y por asociación o enlace se puede determinar que tipos de productos son comprados juntos. La canasta es un grupo de registros de datos, por lo tanto la técnica puede ser usada en cualquier situación donde haya un número grande de grupos de registros de datos. Incluye reglas de asociación, patrones secuenciales, secuencias similares.
- **Análisis de Frecuencia (Frequency analysis):** comprende aquellas técnicas de minería de datos que son aplicadas al análisis de registros ordenados en el tiempo o cualquier conjunto de datos que puedan ser ordenado en el tiempo. Estas técnicas de minería de datos intenta detectar secuencias o subsecuencias similares en los datos ordenados. Incluye patrones secuenciales, secuencias similares.

En la Fig. 3.3 de la pág. 49 se presentan las diferentes técnicas y algoritmos que se pueden utilizar para diferentes aplicaciones de la Minería de Datos.

## 3.9 Extensiones

### 3.9.1 Web Mining

Consiste en aplicar las técnicas de Minería de Datos a documentos y servicios del Web, con el fin de descubrir información o conocimiento potencialmente útil y previamente desconocido.

Todos los que visitan un sitio en Internet dejan huellas digitales como ser direcciones de IP, navegador, entre otras, que los servidores automáticamente almacenan en una bitácora de accesos (Log).

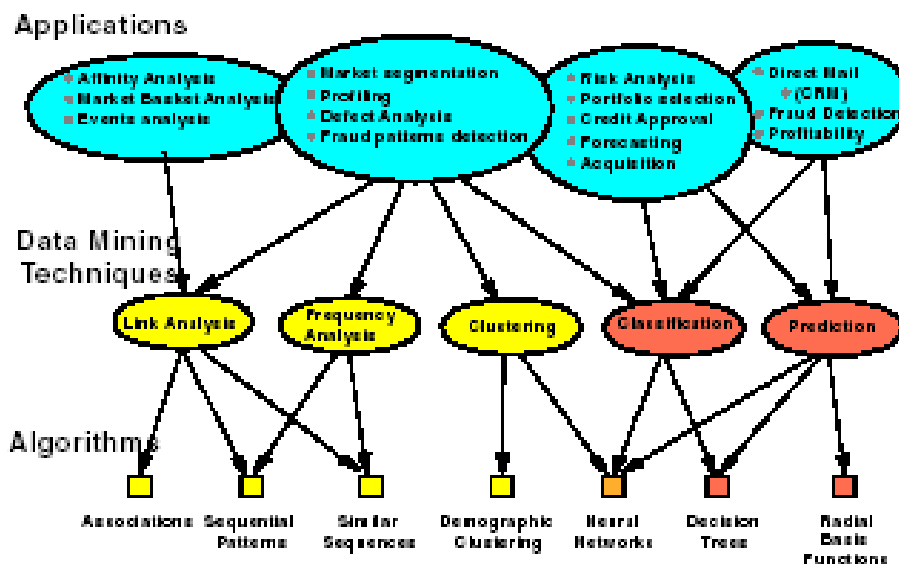


Figura 3.3: Aplicaciones, Técnicas, y Algoritmos de la Minería de Datos.

Las herramientas de Web Mining analizan y procesan estos logs para producir información significativa. Debido a que los contenidos de Internet consisten en varios tipos de datos, como texto, imagen, vídeo, o metadatos, se utiliza el término “minería de datos multimedia” como una instancia del Web Mining para tratar ese tipo de datos. Los accesos totales por dominio, horarios de accesos más frecuentes y visitas por día, entre otros datos, se registran mediante herramientas estadísticas que complementan todo el proceso de análisis del Web Mining.

El Web Mining se puede estructurar en fases. Las fases son:

- Descubrimiento de recursos: consiste en localizar los documentos relevantes o no usuales en la red. Esta función la realizan índices buscadores o índices temáticos.
- Extracción de Información: consiste en extraer determinada información a partir de un documento, sea HTML, XML, texto, PDF, u otro.
- Generalización: consiste en descubrir patrones generales a partir de sitios web individuales, mediante clustering, y asociaciones entre documentos.

- Análisis, validación e interpretación de los patrones.

Normalmente, el Web Mining puede clasificarse en tres dominios de extracción de conocimiento de acuerdo con la naturaleza de los datos:

1. Web Content Mining (minería de contenido web). Es el proceso que consiste en la extraer conocimiento del contenido de documentos o sus descripciones. La localización de patrones en el texto de los documentos, el descubrimiento del recurso basado en conceptos de indexación, o la tecnología basada en agentes, también pueden formar parte de esta categoría.
2. Web Structure Mining (minería de estructura web). Es el proceso de inferir conocimiento de la organización del WWW y la estructura de sus ligas.
3. Web Usage Mining (minería de uso web). Es el proceso de extracción de modelos interesantes usando los logs de los accesos al web.

### 3.9.2 Text Mining

El Text Mining se refiere a examinar una colección de documentos y descubrir información no contenida en ningún documento individual de la colección; en otras palabras, trata de obtener información sin haber partido de algo.

Dado que el ochenta por ciento de la información de una compañía está almacenada en forma de documentos, las técnicas como la categorización de texto, el procesamiento de lenguaje natural, la extracción y recuperación de la información o el aprendizaje automático, apoyan a la minería de texto.

En ocasiones se confunde el Text Mining con la recuperación de la información. Esta última consiste en la recuperación automática de documentos relevantes mediante indexaciones de textos, clasificación, y categorización.

## 3.10 Data Mining y Estadística

La estadística es una herramienta poderosa, y es un elemento crucial en el análisis de datos. Sin embargo, a veces se enfrentan problemas muy serios

en la interpretación de sus resultados, dado que no se recuerda que estos resultados se aplican a grupos y no a individuos.

Existe la posibilidad de ser engañados por la estadística. No todos tienen un sólido entendimiento de la matemática, los supuestos y el modelado para entender a la perfección el riesgo o margen de error en un ejercicio de inferencia estadística. Cuando se dice que una gran decisión se basó en la información disponible, típicamente es una serie de promedios y estimadores estadísticos que presentan una generalización de un gran volumen de datos, donde se realiza una inferencia.

Estos peligros se ven amplificados en el uso de software de Minería de Datos. Dichas herramientas informáticas ponen a disposición de un analista o minero de datos, la posibilidad de crear fácilmente indicadores, resúmenes, gráficas, y aparentes tendencias, sin un verdadero entendimiento de lo que se está reflejando.

¿ Es más fácil equivocarse con la Minería de Datos? La respuesta es sí.

Primero: porque aun con la estadística, el hallar una correlación no significa haber encontrado una relación causa-efecto. El software de Minería de Datos esta diseñado para hallar correlaciones, para olfatearlas. Su tarea consiste en encontrar aquella proyección de los datos, aquella perspectiva donde aparece una correlación y, lamentablemente, en muchos casos, presentarla como una relación causa-efecto.

Aun así, ambas ciencias tienen el mismo objetivo: mejorar la toma de decisiones mediante un conocimiento del entorno. Este entorno lo facilitan los datos almacenados en la organización, cuantitativos o cualitativos, y mediante información de terceras empresas.

Las técnicas estadísticas se centran generalmente en técnicas confirmatorias. Las técnicas de Data Mining son generalmente exploratorias. Así, cuando el problema al que se pretende dar respuesta es refutar o confirmar una hipótesis, se pueden utilizar ambas ciencias. Pero cuando el objetivo es exploratorio, es decir, cuando se desea concretar un problema o definir cuales son las variables más interesantes en un sistema de información, surge la necesidad de delegar parte del conocimiento analítico de la empresa en técnicas de aprendizaje, utilizando para ello el Data Mining.

Hasta aquí, ya se detecta una primera diferencia de aplicación de ambas herramientas: Data Mining se utiliza siempre y cuando no se parta de su-

puestos de partida, únicamente cuando lo que se pretenda es buscar algún conocimiento nuevo y susceptible de proporcionar información novedosa en la toma de decisiones.

Las técnicas de Data Mining son menos restrictivas que las técnicas Estadísticas. Una vez que se encuentre un punto de partida interesante y se esté dispuesto a utilizar algún análisis estadístico en particular, puede suceder que los datos no satisfagan los requerimientos del análisis estadístico. Entonces, las variables se deben examinar para determinar que tratamiento permite adecuarlas al análisis, no siendo posible o conveniente en todos los casos. Data Mining en cambio, permite ser utilizado con los mínimos supuestos posibles.

Otra. Cuando los datos de la empresa son muy dinámicos, las técnicas de Data Mining inciden sobre la inversión y la actualización del conocimiento. Un almacén de datos poco dinámico permite que una inversión en un análisis estadístico se justifique, dado que las conclusiones van a tener un ciclo de vida largo. Sin embargo, en un almacén muy dinámico, se pueden explorar cambios y determinar cuando una regla de negocio ha cambiado. Esto permite abordar diferentes cuestiones a corto y medio plazo.

Cuanto mayores sean las dimensiones del problema, Data Mining ofrece mejores soluciones. Cuantas más variables entran en el problema, más difícil resulta encontrar hipótesis de partida interesantes. O bien, aunque se pueda, el tiempo necesario nunca justificará la inversión. En ese caso, utilizar técnicas de Data Mining como árboles de decisión permiten encontrar relaciones inéditas para luego concretar la investigación sobre las variables más interesantes.

### **Contextos en los cuales es más adecuado el Análisis Estadístico**

El objetivo de la investigación es encontrar causalidad. Si se pretende determinar cuales son las causas de ciertos efectos, se deben utilizar técnicas de estadística. Las relaciones complejas que subyacen a técnicas de Data Mining impiden una interpretación certera de diagramas causa-efecto.

Si las conclusiones se han de extender a otros elementos de poblaciones similares se deben utilizar técnicas de inferencia estadística. Esto viene relacionado con situaciones en las que se dispone exclusivamente de muestras. En Data Mining, se generan modelos y luego se validan con otros casos conocidos de la población, utilizando como significación el ajuste de la predicción sobre una población conocida.

Un proyecto de Data Mining contiene referencias a la estadística en dos partes destacables del proceso: en la preparación de los datos, y en la aproximación a las variables de estudio.

Ambas perspectivas constituyen una sinergia y no son excluyentes una de la otra.

Por lo tanto, Data Mining y estadística son técnicas complementarias que permiten obtener conocimiento inédito de los almacenes de datos o dar respuestas a cuestiones concretas de negocio.

### 3.11 Ventajas

Si bien la Minería de Datos se presenta como una tecnología emergente, posee ciertas ventajas:

- resulta un buen punto de encuentro entre los investigadores y las personas de negocios.
- ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios.
- contribuye a la toma de decisiones tácticas y estratégicas proporcionando un sentido automatizado para identificar información clave desde volúmenes de datos generados por procesos tradicionales.
- permite a los usuarios dar prioridad a decisiones y acciones mostrando factores que tienen un mayor en un objetivo, qué segmentos de clientes son desechables y qué unidades de negocio son sobrepasados y por qué.
- genera Modelos descriptivos: en un contexto de objetivos definidos en los negocios permite a empresas, sin tener en cuenta la industria o el tamaño, explorar automáticamente, visualizar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales de la cuenta de resultados.
- genera Modelos predictivos: permite que relaciones no descubiertas e identificadas a través del proceso del Data Mining se expresen como reglas de negocio o modelos predictivos.

### 3.12 Desventajas

El desarrollo de la tecnología de Minería de Datos está en un momento crítico. Existen elementos que la hacen operable, sin embargo, existen algunos factores que pueden crear un deslustre a la Minería de Datos, como ser:

- que los productos a comercializar son, en la actualidad, significativamente costosos, y los consumidores pueden hallar una relación costo/beneficio improductiva,
- que se requiera de mucha experiencia para utilizar herramientas de la tecnología, o que sea muy fácil hallar patrones equívocos, triviales o no interesantes, y
- que no sea posible resolver los aspectos técnicos de hallar patrones en tiempo o en espacio.

Otro factor que se debe considerar es la privacidad. Antes se pensaba que la Minería de Datos no presentaba ningún peligro o riesgo para la privacidad de los clientes. Hoy en día, se piensa todo lo contrario, sin embargo, no existe un marco jurídico que haya mantenido el paso con el avance tecnológico.

# Bibliografía

- [1] S. M. Weiss;Ñ. Indurkha. *Predictive Data Mining*. M. Kaufmann, Harcourt Intl., USA, 1998.
- [2] U. M. Fayyad; G. Piatetsky-Shapiro; P. Smith; U. Ramasasmy. *Advances in Knowledge Discovery and Data Mining*. The MIT Press, USA, 1996.
- [3] D.Hand; H. Mannila; P. Smyth. *Principles of Data Mining*. The MIT Press, USA, 2000.
- [4] U. M. Fayyad; G. Grinstein; A. Wierse. *Data Mining and Knowledge Discovery*. M. Kaufmann, Harcourt Intl., USA, 2001.



# Índice de Materias

- Bases de datos
  - administrador de, 6
  - concepto, 4
  - funcionamiento, 8
- Conocimiento
  - descubrimiento del, 19
  - gestión del, 1
  - portal de, 3
- Data Marts
  - concepto, 13
  - funciones, 13
- Data Warehouse
  - beneficios, 6
  - características, 6
  - componentes, 8
  - concepto, 5
  - construcción, 9
  - costos, 17
  - diferencias con OLTP, 15
  - diseño, 26
  - funcionamiento, 8
  - objetivos, 8
  - problemas, 12
  - soporte de decisión, 10
- Datos
  - demográficos, 27
  - preproceso, 28, 43
  - relacionales, 27
  - selección, 28
  - transaccionales, 27
- Gestión del Conocimiento
  - antecedentes, 1
  - concepto, 1
  - proceso, 3
  - propósitos, 2
- Groupware
  - concepto, 2
- Inteligencia de Negocios, 11
- KDD
  - áreas relacionadas, 23
  - componentes, 23
  - concepto, 21
  - etapas, 25
  - objetivo, 22
  - técnicas, 32
    - características, 31
    - desviación y tendencia del análisis, 34
    - método de clasificación, 34
    - método estadístico, 33
    - método híbrido, 35
    - método probabilístico, 33
- Minería de datos
  - antecedentes, 39
  - aplicaciones, 42
  - arquitectura, 41
  - capacidades, 44
  - características, 41
  - concepto, 37

- desventajas, 54
  - etapas, 43
  - evolución, 39
  - extensiones, 48
  - herramientas, 44
  - modelos, 45
  - objetivos, 41
  - ventajas, 53
  - y el método científico, 38
  - y estadística, 50
- Modelo de datos
- concepto, 27
  - evaluación, 29
- Modelos
- de descubrimiento, 47
  - predictivos, 46
- OLAP
- beneficios, 14
  - características, 14
  - concepto, 13
- OLTP
- características, 15
  - concepto, 15
  - diferencias con DW, 15
- Resultados
- análisis, 44
  - desplegar, 30
  - interpretar, 30
- Sistemas
- OLAP, 13
  - OLTP, 15
- Técnicas
- análisis de enlace, 48
  - análisis de frecuencia, 48
  - clasificación, 46
  - regresión, 47
- Text Mining
- concepto, 50
  - técnicas, 50
- Web Mining
- concepto, 48
  - dominios, 50
  - fases, 49