CAPITULO IV

DATA WAREHOUSE Y DATAMINING

4.1 Data Warehouse

4.2 DataMining

Un concepto importante dentro de la integración de sistemas y, además, relacionado con los sistemas gerenciales es DataWarehouse.

En la integración de sistemas DataWarehouse funciona como un **depósito inteligente de datos**, reestructurados a partir de bases de datos operacionales, que pueden ser utilizados por diferentes sistemas dentro de la empresa, atendiendo a una planificación bien realizada dentro de la tecnología de la información.

Además, como una **aplicación** que forma parte de la automatización de cualquier empresa, puede encontrarse un sistema de información gerencial, que funciona como una herramienta para accesar o explorar un DataWarehouse y transformar los datos en información.

También se puede decir que un sistema de información gerencial a gran escala debe poseer las denominadas técnicas Datamining que ofrecen una alternativa poderosa como parte del desarrollo de las funciones de los sistemas gerenciales.

Para dar mayor información sobre este aspecto, se pone a consideración este capitulo:

4.1 DATA WAREHOUSE

El Data Warehouse, es el centro de atención de las grandes instituciones, porque provee un ambiente para que las organizaciones hagan un mejor uso de la información que está siendo administrada por diversas aplicaciones operacionales.

Un Data Warehouse es una "colección" de datos en la cual se encuentra integrada la información de la empresa y que se usa como soporte para el proceso de toma de decisiones gerenciales.

Las aplicaciones para soporte de decisiones basadas en un data warehouse, pueden hacer más práctica y fácil la explotación de datos para una mayor eficacia de la empresa, que no se logra cuando se usan sólo los datos que provienen de las aplicaciones operacionales.

Un Data Warehouse se crea al extraer datos desde una o más bases de datos de aplicaciones operacionales. La data extraída es transformada para eliminar inconsistencias y resumida si es necesario y luego, cargada en el Data Warehouse. El proceso de transformar, resumir y combinar los extractos de datos, ayudan a crear el ambiente para el acceso a la información empresarial ó institucional.

Las organizaciones tienen que aprovechar sus recursos de información para crear la información de la operación del negocio, pero deben considerarse las estrategias tecnológicas necesarias para la implementación de una arquitectura completa de Data Warehouse.

Las aplicaciones mas importantes o software que se usa sobre un data warehouse se encuentran en los sistemas de información gerencial.

4.1.1 Concepto Data Warehouse y base de datos operacionales

Data Warehouse soporta el procesamiento informático al proveer una plataforma sólida, a partir de los datos históricos para hacer el análisis. Facilita la integración de sistemas de aplicación no integrados. Organiza y almacena los datos que se necesitan para el procesamiento analítico, informático sobre una amplia perspectiva de tiempo. **[WWW004]**

Un Data Warehouse o Depósito de Datos es una colección de datos orientado a temas, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones gerenciales.

Se puede caracterizar un Data Warehouse haciendo un contraste de cómo los datos de un negocio almacenados en un Data Warehouse, difieren de los datos operacionales usados por las aplicaciones de producción.

BASE DE DATOS OPERACIONAL	DATA WAREHOUSE
Datos Operacionales	Datos del negocio para Información
Orientado a la aplicación	Orientado al sujeto
Actual	Actual + histórico
Detallada	Detallada + más resumida
Cambia continuamente	Estable

Tabla 4.1 Comparación Entre Base De Datos Operacional Y Data Warehouse.

El ingreso de datos en el Data Warehouse viene desde el ambiente operacional en casi todos los casos. El Data Warehouse es siempre un almacén de datos transformados y separados físicamente de la aplicación donde se encontraron los datos en el ambiente operacional.

4.1.2 Características De Un Data Warehouse

Entre las principales características de Data Warehouse se tiene las siguientes:

- Orientado al tema.
- Integrado.
- De tiempo variante.
- No volátil.

4.1.2.1 Orientado a Temas

Una primera característica del Data Warehouse es que la información se clasifica basándose en los aspectos que son de interés para la empresa. Siendo así, los datos tomados están en contraste con los clásicos procesos orientados a las aplicaciones.

El ambiente operacional se diseña alrededor de las aplicaciones y funciones tales como préstamos, ahorros, tarjeta bancaria y depósitos para una institución financiera.

En el ambiente Data Warehouse se organiza alrededor de sujetos tales como cliente, vendedor, producto y actividad. La alineación alrededor de las áreas de los temas afecta el diseño y la implementación de los datos encontrados en el Data Warehouse.

En el Data Warehouse se excluye la información que no será usada por el proceso de sistemas de soporte de decisiones, mientras que la información de las orientadas a las aplicaciones, contiene datos para satisfacer de inmediato los requerimientos funcionales y de proceso, que pueden ser usados o no por el analista de soporte de decisiones.

Otra diferencia importante está en la interrelación de la información. Los datos operacionales mantienen una relación continua entre dos o más tablas basadas en una regla comercial que está vigente.

4.1.2.2 Integración

El aspecto más importante del ambiente Data Warehouse es que la información encontrada al interior está siempre integrada.

La integración de datos se muestra de muchas maneras: en convenciones de nombres consistentes, en la medida uniforme de variables, en la codificación de estructuras consistentes, en atributos físicos de los datos consistentes, fuentes múltiples y otros.

El contraste de la integración encontrada en el Data Warehouse con la carencia de integración del ambiente de aplicaciones, se muestra en la Figura 4.1, con diferencias bien marcadas.

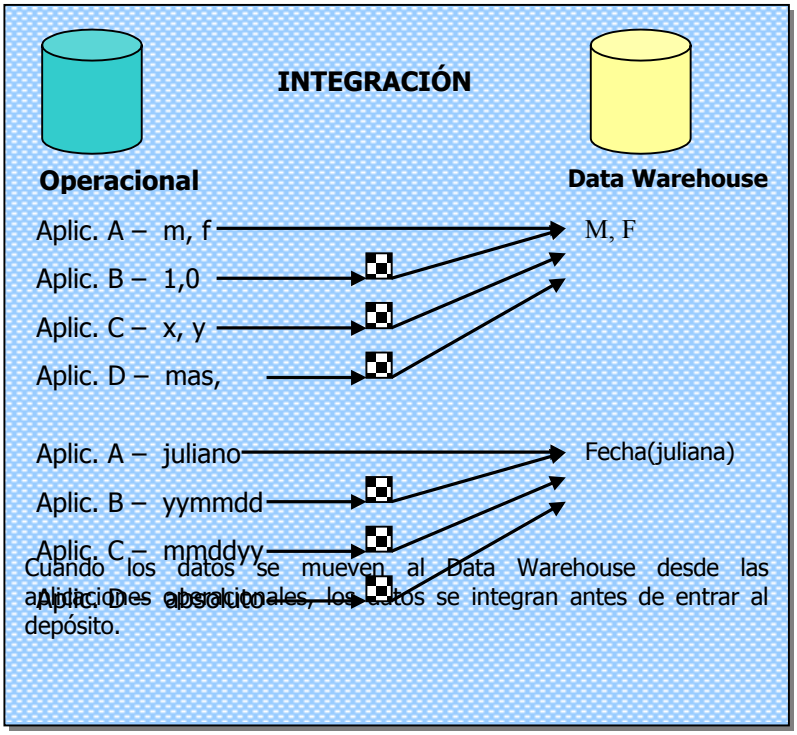

Figura 4.1 Integración Data Warehouse.

Codificación: Los diseñadores de aplicaciones codifican el campo GENERO en varias formas. Un diseñador representa GENERO como una "M" y una "F", otros como un "1" y un "0", otros como una "X" y una "Y" e inclusive, como "masculino" y "femenino".

Por lo tanto, cuando él GENERO se carga en el Data Warehouse desde una aplicación, donde ha sido representado en formato "M" y "F", los datos deben convertirse al formato del Data Warehouse.

Medida de atributos: Al dar medidas a los atributos, la transformación traduce las diversas unidades de medida usadas en las diferentes bases de datos para transformarlas en una medida estándar común.

Convenciones de nombramiento: El mismo elemento es frecuentemente referido por nombres diferentes en las diversas aplicaciones. En el proceso de transformación se debe usar un mismo nombre.

Fuentes Múltiples: El mismo elemento puede derivarse de diferentes fuentes. En este caso, el proceso de transformación debe asegurar que la fuente apropiada sea usada.

Cualquiera que sea la forma del diseño, el resultado es el mismo - la información necesita ser almacenada en el Data Warehouse en un modelo globalmente aceptable y singular, aun cuando los sistemas operacionales subyacentes almacenen los datos de manera diferente.

4.1.2.3 De tiempo variante

Toda la información del Data Warehouse es requerida en algún momento. Esta característica básica de los datos en un depósito, es muy diferente de la información encontrada en el ambiente operacional. En éstos, la información se requiere al momento de acceder. Como la información en el Data Warehouse es solicitada en cualquier momento (es decir, no "ahora mismo"), los datos encontrados en el depósito se llaman de "tiempo variante".

El tiempo variante se muestra de varias maneras:

- La más simple es que la información representa los datos sobre un horizonte largo de tiempo - desde cinco a diez años.
- La segunda manera en la que se muestra el tiempo variante en el Data Warehouse está en la estructura clave. Cada estructura clave en el Data Warehouse contiene, implícita o explícitamente, un elemento de tiempo como día, semana, mes, etc.
- La tercera manera en que aparece el tiempo variante es cuando la información del Data Warehouse, una vez registrada correctamente, no puede ser actualizada. La información del Data Warehouse es, para todos los propósitos prácticos, una serie larga de "snapshots" (vistas instantáneas).

4.1.2.4 No volátil

La información es útil sólo cuando es estable. Los datos operacionales cambian sobre una base momento a momento. La perspectiva mayor, esencial para el análisis y la toma de decisiones, requiere una base de datos estable.

La actualización (insertar, borrar y modificar), se hace regularmente en el ambiente operacional sobre una base de registro por registro. Pero la manipulación básica de los datos que ocurre en el Data Warehouse es mucho más simple. Hay dos únicos tipos de

operaciones: la carga inicial de datos y el acceso a los mismos. No hay actualización de datos (en el sentido general de actualización) en el depósito, como una parte normal de procesamiento.

La fuente de casi toda la información del Data Warehouse es el ambiente operacional. La primera impresión de muchas personas se centra en la gran redundancia de datos, entre el ambiente operacional y el ambiente de Data Warehouse. Dicho razonamiento es superficial y demuestra una carencia de entendimiento con respecto a qué ocurre en el Data Warehouse. De hecho, hay una mínima redundancia de datos entre ambos ambientes.

Los datos se filtran cuando pasan desde el ambiente operacional al de depósito. Existe mucha data que nunca sale del ambiente operacional. Sólo los datos que realmente se necesitan ingresarán al ambiente de Data Warehouse.

El horizonte de tiempo de los datos es muy diferente de un ambiente al otro. La información en el ambiente operacional es más reciente con respecto a la del Data Warehouse. Desde la perspectiva de los horizontes de tiempo únicos, hay poca superposición entre los ambientes operacional y de Data Warehouse.

El Data Warehouse contiene un resumen de la información que no se encuentra en el ambiente operacional.

Los datos experimentan una transformación fundamental cuando pasa al Data Warehouse. Dicho de otra manera, la mayoría de los datos se alteran física y radicalmente cuando se mueven al depósito. No es la misma data que reside en el ambiente operacional desde el punto de vista de integración.

4.1.3 Arquitectura de un Data Warehouse

A fin de comprender como se relacionan todos los componentes involucrados en una estrategia data warehouse es esencial tener una Arquitectura Data Warehouse.

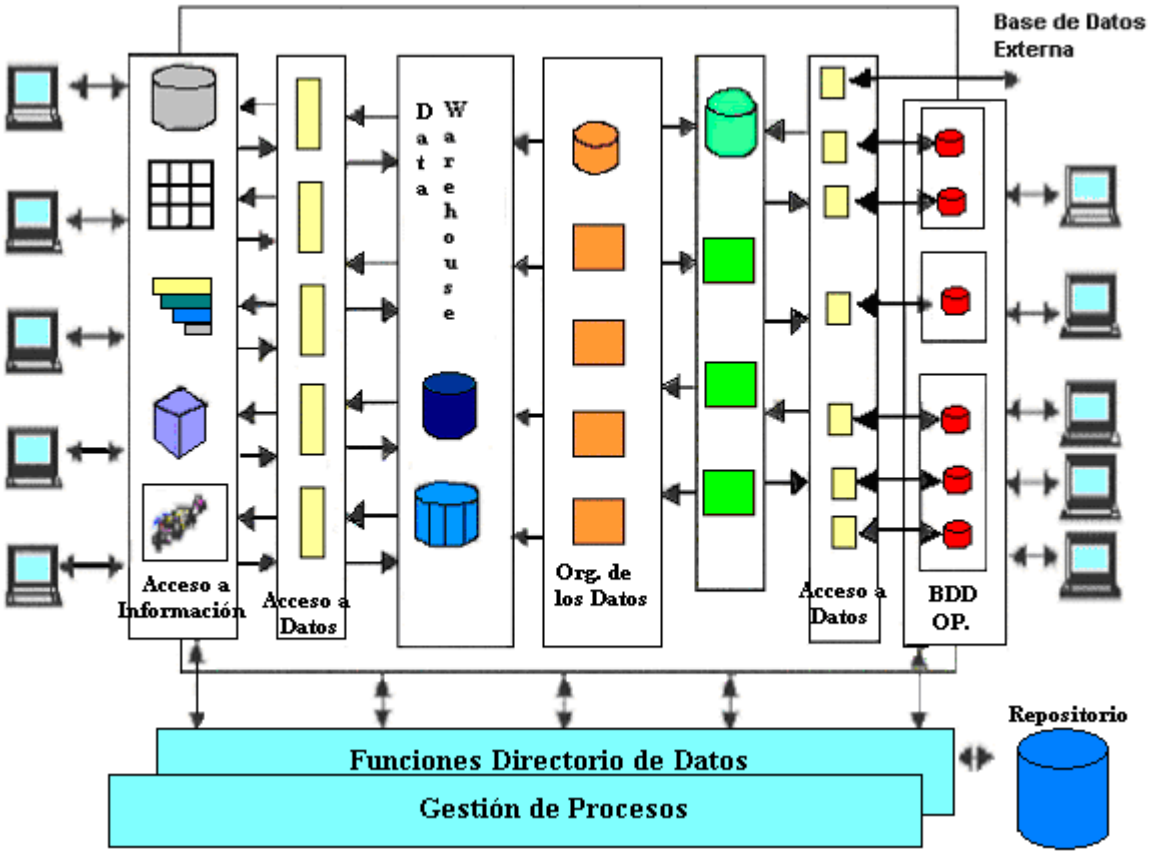

Figura 4.2 Elementos Constituyentes De Una Arquitectura Data Warehouse.

Una Arquitectura Data Warehouse (Data Warehouse Architecture - DWA) es una forma de representar la estructura total de datos, comunicación, procesamiento y presentación, que existe para los usuarios finales que disponen de una computadora dentro de la empresa.

La arquitectura se constituye de un número de partes interconectadas:

- Base de datos operacional / Nivel de base de datos externo.
- Nivel de acceso a la información.
- Nivel de acceso a los datos.
- Nivel de directorio de datos (Metadata).
- Nivel de gestión de proceso.
- Nivel de mensaje de la aplicación.
- Nivel de Data Warehouse.
- Nivel de organización de datos.

4.1.3.1 Base de datos operacional / nivel de base de datos externo

Los sistemas operacionales procesan datos para apoyar las necesidades operacionales críticas. Sin embargo, a causa del enfoque limitado de los sistemas operacionales, las bases de datos diseñadas para soportar estos sistemas, tienen dificultad al acceder a los datos para otra gestión o propósitos informáticos.

Esta dificultad en acceder a los datos operacionales es amplificada ya que muchos de estos sistemas tienen de 10 a 15 años de antigüedad. El tiempo de algunos de estos sistemas significa que la tecnología de acceso a los datos disponible para obtener los datos operacionales, es así mismo antigua.

Ciertamente, la finalidad del data Warehouse es liberar la información que es almacenada en bases de datos operacionales y combinarla con la información desde otra fuente de datos, generalmente externa.

Cada vez más, las organizaciones grandes adquieren datos adicionales desde bases de datos externas. Esta información incluye tendencias demográficas, econométricas, adquisitivas y competitivas (que pueden ser proporcionadas por Instituciones Oficiales - INEI). Internet provee el acceso a más recursos de datos todos los días.

4.1.3.2 Nivel de acceso a la información

El nivel de acceso a la información de la arquitectura Data Warehouse, es el nivel del que el usuario final se encarga directamente. En particular, representa las herramientas que el usuario final normalmente usa día a día. Por ejemplo: Excel, Lotus 1-2-3, Access, etc.

Este nivel también incluye el hardware y software involucrados en mostrar información en pantalla y emitir reportes de impresión, hojas de cálculo, gráficos y diagramas para el análisis y presentación. Hace dos décadas que el nivel de acceso a la información se ha expandido enormemente, especialmente a los usuarios finales quienes se han volcado a los PCs monousuarios y los PCs en redes.

4.1.3.3 Nivel de acceso a los datos

El nivel de acceso a los datos de la arquitectura Data Warehouse está involucrado con el nivel de acceso a la información para conversar en el nivel operacional. En la red mundial de hoy, el lenguaje de datos común que ha surgido es SQL. Originalmente, SQL fue

desarrollado por IBM como un lenguaje de consulta, pero en los últimos veinte años ha llegado a ser el estándar para el intercambio de datos.

El nivel de acceso a los datos no solamente conecta DBMSs diferentes y sistemas de archivos sobre el mismo hardware, sino también a los fabricantes y protocolos de red. Una de las claves de una estrategia Data Warehouse es proveer a los usuarios finales con "acceso a datos universales".

4.1.3.4 Nivel de directorio de datos (Metadata)

A fin de proveer el acceso a los datos universales, es absolutamente necesario mantener alguna forma de directorio de datos o repositorio de la información metadata. La metadata es la información alrededor de los datos dentro de la empresa.

Las descripciones de registro en un programa COBOL son metadata. También lo son las sentencias DIMENSION en un programa FORTRAN o las sentencias a crear en SQL.

A fin de tener un depósito totalmente funcional, es necesario tener una variedad de metadata disponibles, información sobre las vistas de datos de los usuarios finales e información sobre las bases de datos operacionales. Idealmente, los usuarios finales deberían de acceder a los datos desde el Data Warehouse (o desde las bases de datos operacionales), sin tener que conocer dónde residen los datos o la forma en que se han almacenados.

4.1.3.5 Nivel de gestión de procesos

El nivel de gestión de procesos tiene que ver con la programación de diversas tareas que deben realizarse para construir y mantener el Data Warehouse y la información del directorio de datos. Este nivel puede depender del alto nivel de control de trabajo para muchos procesos (procedimientos) que deben ocurrir para mantener el Data Warehouse actualizado.

4.1.3.6 Nivel de mensaje de la aplicación

El nivel de mensaje de la aplicación tiene que ver con el transporte de información alrededor de la red de la empresa. El mensaje de aplicación se refiere también como "subproducto", pero puede involucrar sólo protocolos de red. Puede usarse por ejemplo, para aislar aplicaciones operacionales o estratégicas a partir del formato de datos exacto,

recolectar transacciones o los mensajes y entregarlos a una ubicación segura en un tiempo seguro.

4.1.3.7 Nivel Data Warehouse (Físico)

En el Data Warehouse (núcleo) es donde ocurre la data actual, usada principalmente para usos estratégicos. En algunos casos, uno puede pensar del Data Warehouse simplemente como una vista lógica o virtual de datos. En muchos ejemplos, el Data Warehouse puede no involucrar almacenamiento de datos.

En un Data Warehouse físico, copias, en algunos casos, muchas copias de datos operacionales y/o externos, son almacenadas realmente en una forma que es fácil de acceder y es altamente flexible. Cada vez más, los Data Warehouse son almacenados sobre plataformas cliente/servidor, pero por lo general se almacenan sobre mainframes.

4.1.3.8 Nivel de organización de datos

El componente final de la arquitectura Data Warehouse es la organización de los datos. Se llama también gestión de copia o réplica, pero de hecho, incluye todos los procesos necesarios como seleccionar, editar, resumir, combinar y cargar datos en el depósito y acceder a la información desde bases de datos operacionales y/o externas.

La organización de datos involucra con frecuencia una programación compleja, pero cada vez más, están creándose las herramientas Data Warehouse para ayudar en este proceso. Involucra también programas de análisis de calidad de datos y filtros que identifican modelos y estructura de datos dentro de la data operacional existente.

4.1.4 Operaciones en un Data Warehouse

En la Figura 4.3 se muestra algunos de los tipos de operaciones que se efectúan dentro de un ambiente Data Warehouse.

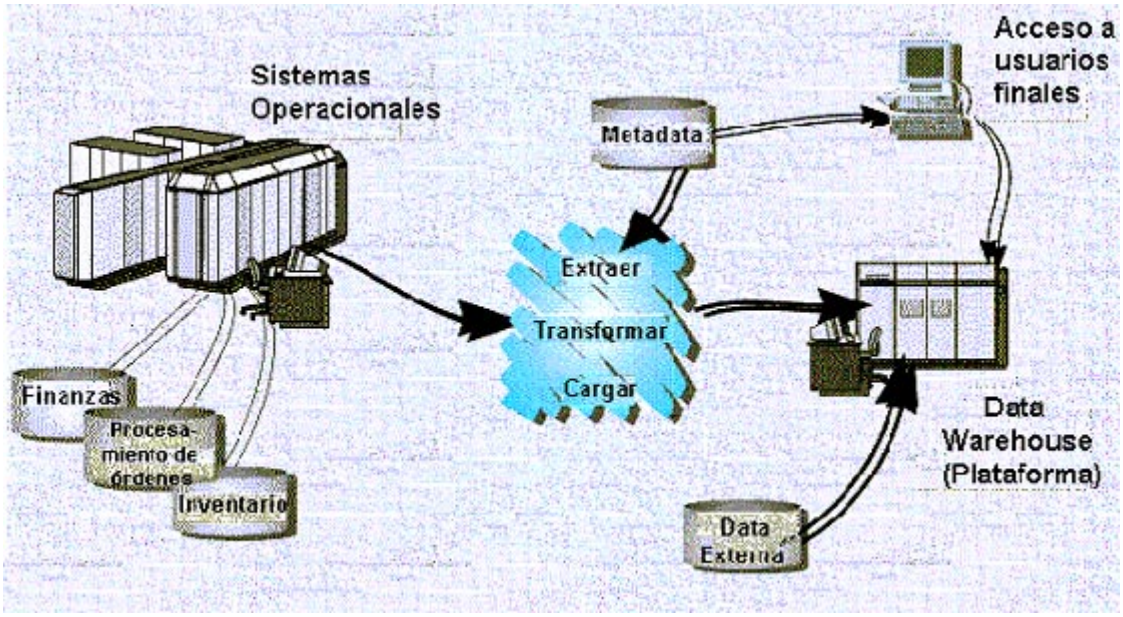

Figura 4.3 Operaciones En Data Warehouse.

4.1.4.1 Sistemas Operacionales

Los datos administrados por los sistemas de aplicación operacionales son la fuente principal de datos para el Data Warehouse.

Las bases de datos operacionales se organizan como archivos indexados (UFAS, VSAM), bases de datos de redes/jerárquicas (I-D-S/II, IMS, IDMS) o sistemas de base de datos relacionales (DB2, Oracle, Informix, etc.).

4.1.4.2 Extracción, transformación y carga de los datos

Se requieren herramientas de gestión de datos para extraer datos desde bases de datos y/o archivos operacionales, luego es necesario manipular o transformar los datos antes de cargar los resultados en el Data Warehouse.

Tomar los datos desde varias bases de datos operacionales y transformarlos en datos requeridos para el depósito, se refiere a la transformación o a la integración de datos. Todas las inconsistencias deben resolverse antes que los elementos de datos sean almacenados en el Data Warehouse.

4.1.4.3 Metadata

Otro paso necesario es crear la metadata. La metadata (es decir, datos acerca de datos) describe los contenidos del Data Warehouse.

4.1.4.4 Acceso de usuario final

Los usuarios acceden al Data Warehouse por medio de herramientas de productividad basadas en GUI (Graphical User Interface - Interfase gráfica de usuario). Puede proveerse a los usuarios del Data Warehouse muchos de estos tipos de herramientas: software de consultas, generadores de reportes, procesamiento analítico en línea, herramientas Data/Visual Mining, etc., dependiendo de los tipos de usuarios y sus requerimientos particulares.

Sin embargo, una sola herramienta no satisface todos los requerimientos, por lo que es necesaria la integración de una serie de herramientas.

4.1.4.5 Plataforma del Data Warehouse

La plataforma para el Data Warehouse es casi siempre un servidor de base de datos relacional. Cuando se manipulan volúmenes muy grandes de datos puede requerirse una configuración en bloque de servidores UNIX con multiprocesador simétrico (SMP) o un servidor con procesador paralelo masivo (MPP) especializado.

Los extractos de la data integrada/transformada se cargan en el Data Warehouse. Uno de los más populares RDBMSs disponibles para Data Warehouse sobre la plataforma UNIX (SMP y MPP) generalmente es Teradata. La elección de la plataforma es crítica. El depósito crecerá y hay que comprender los requerimientos después de 3 o 5 años.

El sistema de depósito ejecuta las consultas que se pasa a los datos por el software de acceso a los datos del usuario. Aunque un usuario visualiza las consultas desde el punto de vista de un GUI, las consultas típicamente se formulan como pedidos SQL, porque SQL es un lenguaje universal y el estándar de hecho para el acceso a datos.

4.1.4.6 Datos externos

Dependiendo de la aplicación, el alcance del Data Warehouse puede extenderse por la capacidad de acceder a la data externa. Por ejemplo, los datos accesibles por medio de servicios de computadora en línea (tales como CompuServe y America On Line) y/o vía Internet, pueden estar disponibles a los usuarios del Data Warehouse.

4.1.4.7 Evolución del depósito

La construcción de un Data Warehouse es una tarea grande. No es recomendable emprender el desarrollo del Data Warehouse de la empresa como un proyecto cualquiera. Más bien, se recomienda que los requerimientos de una serie de fases se desarrollen e implementen en modelos consecutivos que permitan un proceso de implementación más gradual e iterativo.

No existe ninguna organización que haya triunfado en el desarrollo del Data Warehouse de la empresa, en un sólo paso. Muchas, sin embargo, lo han logrado luego de un desarrollo paso a paso. Los pasos previos evolucionan juntamente con la materia que está siendo agregada.

Los datos en el Data Warehouse no son volátiles y es un repositorio de datos de sólo lectura (en general). Sin embargo, pueden añadirse nuevos elementos sobre una base regular para que el contenido siga la evolución de los datos en la base de datos fuente, tanto en los contenidos como en el tiempo.

Uno de los desafíos de mantener un Data Warehouse, es idear métodos para identificar datos nuevos o modificados en las bases de datos operacionales. Algunas maneras para identificar estos datos incluyen insertar fecha/tiempo en los registros de base de datos y entonces crear copias de registros actualizados y copiar información de los registros de transacción y/o base de datos diarios.

Estos elementos de datos nuevos y/o modificados son extraídos, integrados, transformados y agregados al Data Warehouse en pasos periódicos programados. Como se añaden las nuevas ocurrencias de datos, los datos antiguos son eliminados. Por ejemplo, si los detalles de un sujeto particular se mantienen por 5 años, como se agregó la última semana, la semana anterior es eliminada.

4.1.5 Software en un Data Warehouse

Se necesita software especializado que permita capturar los datos relevantes en forma rápida y pueda verse a través de diferentes dimensiones de los datos. El software no debería limitarse únicamente al acceso a los datos, sino también, al análisis significativo de los datos. En efecto, transforma los datos de la información cruda o no procesada, en información útil para la empresa.

El software o las herramientas de negocios inteligentes se colocan sobre la plataforma data warehouse y proveen este servicio. Debido a que son el punto principal de contacto entre la aplicación del depósito y la gente que lo usa, estas herramientas pueden constituir la diferencia entre el éxito o fracaso de un depósito.

Las herramientas de negocio inteligentes se han convertido en los sucesores de los sistemas de soporte de decisión, pero tienen un alcance más amplio. No solamente ayudan en las decisiones de soporte sino, en muchos casos, estas herramientas soportan muchas funciones operacionales y de misión-crítica de la compañía. Sin embargo, estos productos no son infalibles ya que sólo se consigue el máximo provecho del Data Warehouse, si se elige las herramientas adecuadas a las necesidades de cada usuario final.

El software usado en un Data Warehouse se clasifican en Herramientas de Consulta y Reporte, Herramientas de Base de Datos Multidimensionales/ OLAP (On Line Analytical Processing), Sistemas de Información Ejecutivos, Herramientas Data Mining y los Sistemas de Gestión de Bases de Datos propiamente.

4.1.5.1 Herramientas de consulta y reporte

Existe una gran cantidad de poderosas herramientas de consulta y reporte en el mercado. Las más simples de estas herramientas son productos de reporte y consultas básicas. Ellos proporcionan desde pantallas gráficas a generadores SQL. Más que aprender SQL o escribir un programa para acceder a la información de una base de datos, las herramientas de consulta al igual que la mayoría de herramientas visuales, le permiten apuntar y dar un click a los menús y botones para especificar los elementos de datos, condiciones, criterios de agrupación y otros atributos de una solicitud de información.

La herramienta de consulta genera entonces un llamado a una base de datos, extrae los datos pertinentes, efectúa cálculos adicionales, manipula los datos si es necesario y presenta los resultados en un formato claro.

El procesamiento estadístico se limita comúnmente a promedios, sumas, desviaciones estándar y otras funciones de análisis básicas.

Para hacer consultas más accesibles a usuarios no-técnicos, los productos tales como Crystal Reports de Seagate, Impromptu de Cognos, Reportsmith de Borland, Intelligent

Query de IQ Software, Esperant de Software AG y GQL de Andyne, ofrecen interfases gráficas para seleccionar, arrastrar y pegar.

4.1.5.2 Herramientas de base de datos multidimensionales/OLAP

Los generadores de reporte tienen sus limitaciones cuando los usuarios finales necesitan más que una sola vista estática de los datos, que no sean sujeto de otras manipulaciones. Para estos usuarios, las herramientas del procesamiento analítico en línea (OLAP - On Line Analytical Processing), proveen capacidades "Slide y Dice" que contestaría "¿qué sucedió?". Al analizar por qué los resultados están como están.

Las primeras soluciones OLAP estuvieron basadas en bases de datos multidimensionales (MDDBS). Un cubo estructural (dos veces un hipercubo o un arreglo multidimensional) almacenaba los datos para que se puedan manipular intuitivamente y claramente ver las asociaciones a través de dimensiones múltiples. Los productos pioneros tal como Essbase de Arbor Software soportan directamente las diferentes vistas y las manipulaciones dimensionales requeridas por OLAP.

4.1.5.3 Sistemas de información ejecutivos o gerenciales

Las herramientas de sistemas de información ejecutivos (Executive Information Systems - EIS), proporcionan medios sumamente fáciles de usar para consulta y análisis de la información confiable. Generalmente se diseñan para el usuario que necesita conseguir los datos rápidamente, pero quiere utilizar el menor tiempo posible para comprender el uso de la herramienta.

También, permiten a los desarrolladores de sistemas colocar el contexto del negocio alrededor de información diversa. Un uso típico de un EIS es facilitar al usuario la recuperación y análisis de la métricas, de rendimiento de la organización.

4.1.5.4 Herramientas Data Mining

El proceso de data Mining extrae los conocimientos guardados o información predictiva desde el data warehouse sin requerir pedidos o preguntas específicas.

Las herramientas Mining usan algunas de las técnicas de computación más avanzadas para generar modelos y asociaciones. Mining es un dato-conducido, no una aplicación-conducida.

El Intelligent Miner de IBM para AIX soporta sofisticadas técnicas Mining, así como las funciones de preparación de los datos para extraer información desde bases de datos Oracle o Sybase y cargarlos en DB2 para Mining. Otros ejemplos de herramientas Data Mining comerciales incluyen Darwin de Thinking Machines, herramientas de visualización de datos en MDDB de SAS Institute, SGI MineSet y Focus 6 Serie de Visualización y Análisis de Information Builders.

4.1.5.5 Sistemas de gestión de bases de datos

Este software proporciona procesamiento en paralelo y/o algo fuera de los aspectos ordinarios, que puedan ser especialmente interesantes para la gente de desarrollo de data warehouse y de sistemas de soporte de decisiones.

4.2 DATAMINING

Con el crecimiento dramático de los almacenes de datos(Data Warehouse) y la necesidad de obtener beneficios de estos recursos disponibles, las empresas tienen la necesidad de información que va más allá de la que los sistemas tradicionales de base de datos pueden entregar. Así el Datamining o KDD ofrece grandes beneficios a las empresas.

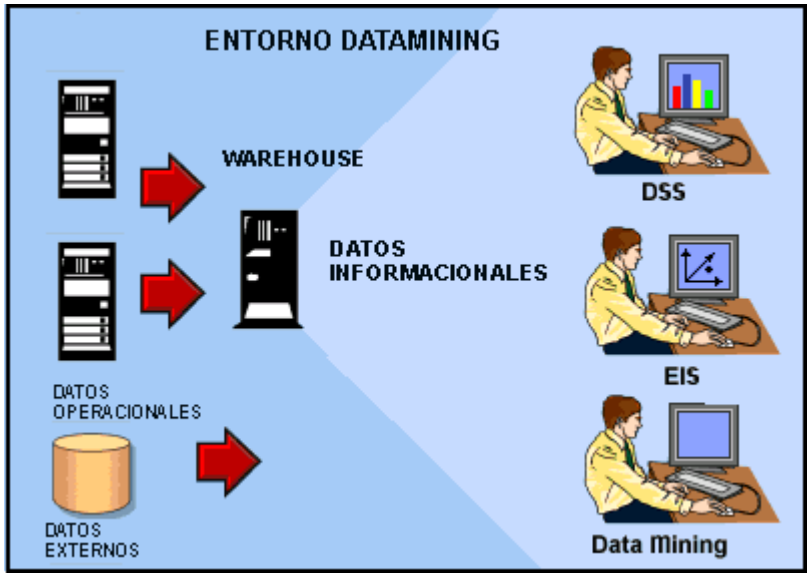

Figura 4.4 Entorno De Datamining.

El **Data Mining** es un **proceso** que, a través del descubrimiento y cuantificación de relaciones predictivas en los datos, permite transformar la información disponible en conocimiento útil de negocio. Esto es debido a que no es suficiente "navegar" por los datos para resolver los problemas de negocio, sino que se hace necesario seguir una metodología ordenada que permita obtener rendimientos tangibles de este conjunto de herramientas y técnicas de las que dispone el usuario. Constituye, por tanto, una de las vías clave de explotación del Data Warehouse, dado que es este su entorno natural de trabajo. **[WWW005]**

Se trata de un concepto de explotación de naturaleza radicalmente distinta a la de los sistemas de información de gestión, dado que se basa en la información de detalle contenida en el almacén. Adicionalmente, el usuario no se conforma con la mera visualización de datos, sino que trata de obtener una relación entre los mismos que tenga repercusiones en su negocio.

En el Data Mining, **la extracción de información oculta y predecible de grandes bases de datos**, es una tecnología para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse).

La llegada del Data Mining se considera como una etapa de la introducción de métodos cuantitativos, científicos en el mundo del comercio, industria y negocios.

Muchas compañías ya recaudan 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 y pueden ser integradas con nuevos productos y sistemas pues son traídas en línea (on-line). Una vez que las herramientas de Data Mining implementadas en computadoras cliente-servidor de alto rendimiento o de procesamiento paralelo, pueden analizar bases de datos masivas para brindar respuesta a preguntas tales como, "¿Cuáles clientes tienen más probabilidad de responder al próximo mailing promocional, y por qué? y presentar los resultados en forma de tablas, con gráficos, reportes, texto, hipertexto, etc.

4.2.1 Definiciones

Entre algunas definiciones acerca del Datamining podemos citar las siguientes:

- La extracción no superficial de información implícita, previamente desconocida y potencialmente útil acerca de los datos.
- La búsqueda de relaciones y patrones globales que existen en grandes bases de datos, pero que están ocultos.
- Uso de una variedad de técnicas para identificar trozos de información o conocimiento relativo a la toma de decisiones en los datos, y la extracción de ellos de tal manera que puedan ser utilizados en áreas que soporten la toma de decisiones y predicción.
- Es un proceso que, a través del descubrimiento y cuantificación de relaciones predictivas en los datos, permite transformar la información disponible en conocimiento útil de negocio.
- El Data Mining es un proceso continuo e iterativo que implica el uso de un software específico, una metodología propia y la creatividad humana para

conseguir información muy valiosa, patrones, relaciones, anomalías y dependencias a través de la exploración de datos.

4.2.2 Modelos de Data Mining

Entre los diferentes modelos de Data Mining podemos mencionar los siguientes:

Modelo De Verificación: que toma una hipótesis del usuario y verifica su validez en los datos. El énfasis de este modelo está en que el usuario es el responsable por la formulación de la hipótesis y efectuar la consulta sobre los datos para afirmar o negar su hipótesis.

Modelo De Descubrimiento: este modelo se caracteriza por que es el sistema el que automáticamente descubre información importante oculta en los datos. El proceso se centra, entonces, en la búsqueda de patrones, tendencias y generalizaciones acerca de los datos, sin la intervención del usuario.

4.2.3 Ventajas de Data Mining

La tecnología del Data Mining aporta dos beneficios clave en los negocios:

- **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 (tales como el aumento de los ingresos, incremento de los beneficios, contención de costes y gestión de riesgos).
- **Modelos Predictivos:** Permite que relaciones no descubiertas e identificadas a través del proceso del Data Mining sean expresadas como reglas de negocio o modelos predictivos. Estos resultados pueden comunicarse en formatos tradicionales (presentaciones, informes, información electrónica compartida, contenidos en aplicaciones, etc.) para guiar la estrategia y planificación de la empresa.

4.2.4 Fundamentos de Data Mining

Las técnicas de Data Mining son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. Data Mining toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y pro-activa. Data Mining está lista para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

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

Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que consistentemente tienen más rendimiento que métodos estadísticos clásicos.

En la evolución desde los datos de negocios a información de negocios, cada nuevo paso se basa en el previo. Por ejemplo, el acceso a datos dinámicos es crítico para las aplicaciones de navegación de datos, y la habilidad para almacenar grandes bases de datos es crítica para Data Mining.

Los componentes esenciales de la tecnología de Data Mining han estado bajo desarrollo por décadas, en áreas de investigación como estadísticas, inteligencia artificial y aprendizaje de máquinas. Hoy, la madurez de estas técnicas, junto con los motores de bases de datos relacionales de alto rendimiento, hicieron que estas tecnologías fueran prácticas para los entornos de Data Warehouse actuales.

4.2.5 Técnicas de Data Mining

Las técnicas de Data Mining pueden producir los beneficios de automatización en las plataformas de hardware y software existentes y puede ser implementadas en sistemas nuevos a medida que las plataformas existentes se actualicen y nuevos productos sean

desarrollados. Cuando las herramientas de Data Mining son implementadas en sistemas de procesamiento paralelo de alto rendimiento, 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 hace que sea práctico para los usuarios analizar inmensas cantidades de datos. Grandes bases de datos, a su vez, producen mejores predicciones.

Para soportar el proceso de Data Mining, el usuario dispone de una extensa gama de técnicas que le pueden ayudar en cada una de las fases de dicho proceso, las cuales se describen a continuación:

4.2.5.1 Análisis estadístico:

Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información. Los **análisis prospectivos** automatizados ofrecidos por un producto así van más allá de los eventos pasados provistos por herramientas retrospectivas típicas de sistemas de soporte de decisión. Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas y a los cuales los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

ANOVA: o Análisis de la Varianza, compara si existen diferencias significativas entre las medidas de una o más variables continuas en grupos de población distintos.

Regresión: define la relación entre una o más variables y un conjunto de variables predictoras de las primeras.

Ji cuadrado: contrasta la hipótesis de independencia entre variables.

Componentes Principales: permite reducir el número de variables observadas a un menor número de variables artificiales, conservando la mayor parte de la información sobre la Varianza de las variables.

Análisis Cluster: Permite clasificar una población en un número determinado de grupos, basándose en semejanzas y diferencias de perfiles existentes entre los diferentes componentes de dicha población.

Análisis Discriminante: método de clasificación de individuos en grupos que previamente se han establecido, y que permite encontrar la regla de clasificación de los elementos de estos grupos, y, por tanto, identificar cuáles son las variables que mejor definan la pertenencia al grupo.

4.2.5.2 Métodos basados en árboles de decisión:

El método Chaid (Chi Squared Automatic Interaction Detector): es un análisis que genera un árbol de decisión para predecir el comportamiento de una variable, a partir de una o más variables predictoras, de forma que los conjuntos de una misma rama y un mismo nivel son disjuntos. Es útil en aquellas situaciones en las que el objetivo es dividir una población en distintos segmentos basándose en algún criterio de decisión.

El árbol De Decisión: se construye partiendo el conjunto de datos en dos o más subconjuntos de observaciones a partir de los valores que toman las variables predictoras. Cada uno de estos subconjuntos vuelve después a ser particionado utilizando el mismo algoritmo. Este proceso continúa hasta que no se encuentran diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta.

La raíz del árbol es el conjunto de datos íntegro, los subconjuntos y los subsubconjuntos conforman las ramas del árbol. Un conjunto en el que se hace una partición se llama nodo.

El número de subconjuntos en una partición puede ir de dos hasta el número de valores distintos que puede tomar la variable usada para hacer la separación. La variable de predicción usada para crear una partición es aquella más significativamente relacionada con la variable de respuesta de acuerdo con el test de independencia de la Chi cuadrado sobre una tabla de contingencia.

Algoritmos Genéticos: Son métodos numéricos de optimización, en los que aquella variable o variables que se pretenden optimizar junto con las variables de estudio constituyen un segmento de información. Aquellas configuraciones de las variables de análisis que obtengan mejores valores para la variable de respuesta, corresponderán a segmentos con mayor capacidad reproductiva. A través de la reproducción, los mejores segmentos perduran y su proporción crece de generación en generación. Se puede además introducir elementos aleatorios para la modificación de las variables

(mutaciones). Al cabo de cierto número de iteraciones, la población estará constituida por buenas soluciones al problema de optimización.

Redes Neuronales: Genéricamente son métodos de proceso numérico en paralelo, en el que las variables interactúan mediante transformaciones lineales o no lineales, hasta obtener unas salidas. Estas salidas se contrastan con los que tenían que haber salido, basándose en unos datos de prueba, dando lugar a un proceso de retroalimentación mediante el cual la red sé reconfigura, hasta obtener un modelo adecuado.

Lógica Difusa: Es una generalización del concepto de estadística. La estadística clásica se basa en la teoría de probabilidades, a su vez ésta en la técnica conjuntista, en la que la relación de pertenencia a un conjunto es dicotómica (el 2 es par o no lo es). Si establecemos la noción de conjunto borroso como aquel en el que la pertenencia tiene una cierta graduación (¿un día a 20ºC es caluroso?), se dispondrá de una estadística más amplia y con resultados más cercanos al modo de razonamiento humano.

Series Temporales: Es el conocimiento de una variable a través del tiempo para, a partir de ese conocimiento, y bajo el supuesto de que no van a producirse cambios estructurales, poder realizar predicciones. Suelen basarse en un estudio de la serie en ciclos, tendencias y estacionalidades, que se diferencian por el ámbito de tiempo abarcado, para por composición obtener la serie original. Se pueden aplicar enfoques híbridos con los métodos anteriores, en los que la serie se puede explicar no sólo en función del tiempo sino como combinación de otras variables de entorno más estables y, por lo tanto, más fácilmente predecibles.

4.2.6 Metodología de Aplicación

Para utilizar estas técnicas de forma eficiente y ordenada es preciso aplicar una metodología estructurada, al proceso de Data Mining. A este respecto proponemos la siguiente metodología, siempre adaptable a la situación de negocio particular a la que se aplique:

Muestreo: Extracción de la población muestral sobre la que se va a aplicar el análisis. En ocasiones se trata de una muestra aleatoria, pero puede ser también un subconjunto de datos del Data Warehouse que cumplan unas condiciones determinadas. El objeto de trabajar con una muestra de la población en lugar de toda ella, es la simplificación del

estudio y la disminución de la carga de proceso. La muestra más óptima será aquella que teniendo un error asumible contenga el número mínimo de observaciones.

En el caso de que se recurra a un muestreo aleatorio, se debería tener la opción de elegir

- El nivel de confianza de la muestra (usualmente el 95% o el 99%).
- El tamaño máximo de la muestra (número máximo de registros), en cuyo caso el sistema deberá informar del error cometido y la representatividad de la muestra sobre la población original.
- El error muestral que está dispuesto a cometer, en cuyo caso el sistema informará del número de observaciones que debe contener la muestra y su representatividad sobre la población original.
- Para facilitar este paso se debe disponer de herramientas de extracción dinámica de información con o sin muestreo (simple o estratificado). En el caso del muestreo, dichas herramientas deben tener la opción de, dado un nivel de confianza, fijar el tamaño de la muestra y obtener el error o bien fijar el error y obtener el tamaño mínimo de la muestra que nos proporcione este grado de error.

Exploración: Una vez determinada la población que sirve para la obtención del modelo se deberá determinar cuales son las variables explicativas que van a servir como "inputs" al modelo. Para ello es importante hacer una exploración por la información disponible de la población que nos permita eliminar variables que no influyen y agrupar aquellas que repercuten en la misma dirección.

El objetivo es simplificar en lo posible el problema con el fin de optimizar la eficiencia del modelo. En este paso se pueden emplear herramientas que nos permitan visualizar de forma gráfica la información utilizando las variables explicativas como dimensiones.

También se pueden emplear técnicas estadísticas que nos ayuden a poner de manifiesto relaciones entre variables. A este respecto resultará ideal una herramienta que permita la visualización y el análisis estadístico integrados.

Manipulación: Tratamiento realizado sobre los datos de forma previa a la modelización, en base a la exploración realizada, de forma que se definan claramente los inputs del modelo a realizar (selección de variables explicativas, agrupación de variables similares, etc.).

Modelización: Permite establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibilitan inferir el valor de las mismas con un nivel de confianza determinado.

Valoración: Análisis de la bondad del modelo contrastando con otros métodos estadísticos o con nuevas poblaciones muestrales.

Método Del Vecino Más Cercano: una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases del/de los *k* registro (s) más similar/es a él en un conjunto de datos históricos (donde *k* ³ 1). Algunas veces se llama la técnica del vecino *k*-más cercano.

Regla De Inducción: la extracción de reglas if-then de datos basados en significado estadístico.

Muchas de estas tecnologías han estado en uso por más de una década en herramientas de análisis especializadas que trabajan con volúmenes de datos relativamente pequeños. Estas capacidades están ahora evolucionando para integrarse directamente con herramientas OLAP y de Data Warehouse.

Las soluciones que aporta el Data Mining están basadas en la implementación, a través de la programación, de interfaces de uso general y algoritmos propios y disponibles para todos que permiten una eficiente exploración y organización de los datos. Estos algoritmos apoyan la identificación de patrones, relaciones y anomalías de interés potencial para los que toman las decisiones en los negocios.

Además de implementar estos algoritmos en un método accesible para el usuario la tecnología del Data Mining requiere una comprensión de varias bases de datos e implementación de soluciones de Data Mining para aprovechar las características de dichas bases de datos (si hay alguna) y que hacen que las tareas del Data Mining sean más eficientes en grandes volúmenes de datos.

4.2.7 El Alcance de Data Mining

El nombre de Data Mining deriva de las similitudes entre buscar valiosa información de negocios en grandes bases de datos como por ejemplo encontrar información de la venta de un producto entre grandes montos de Gigabytes almacenados y minar una montaña para encontrar una veta de metales valiosos. Ambos procesos requieren examinar una inmensa cantidad de material, o investigar inteligentemente hasta encontrar exactamente donde residen los valores. Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer estas capacidades:

- **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. Un típico ejemplo de problema predecible es el marketing apuntado a objetivos (targeted marketing). Data Mining usa datos en mailing promocionales anteriores para identificar posibles objetivos para maximizar los resultados de la inversión en futuros mailing. Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.
- **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. Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar datos anormales que pueden representar errores de tipeado en la carga de datos.

4.2.8 Cómo trabaja Data Mining

¿Cuán exactamente es capaz Data Mining de decir cosas importantes que el usuario desconoce o que van a pasar? La técnica usada para realizar estas hazañas en Data Mining se llama **Modelado**. Modelado es simplemente el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta. Por ejemplo, si busca un galeón español hundido en los

mares lo primero que podría hacer es investigar otros tesoros españoles que ya fueron encontrados en el pasado. Este acto de construcción de un modelo es algo que la gente ha estado haciendo desde hace mucho tiempo, seguramente desde antes del auge de las computadoras y de la tecnología de Data Mining. Lo que ocurre en las computadoras, no es muy diferente de la manera en que la gente construye modelos. Las computadoras son cargadas con mucha información acerca de una variedad de situaciones donde una respuesta es conocida y luego el software de Data Mining en la computadora debe correr a través de los datos y distinguir las características de los datos que llevarán al modelo. Una vez que el modelo se construyó, puede ser usado en situaciones similares donde no se conoce la respuesta.

Si alguien le dice que tiene un modelo que puede predecir el uso de los clientes, ¿Cómo puede saber si es realmente un buen modelo? La primera cosa que puede probar es pedirle que aplique el modelo a su base de clientes - donde usted ya conoce la respuesta. Con Data Mining, la mejor manera para realizar esto es dejando de lado ciertos datos para aislarlos del proceso de Data Mining. Una vez que el proceso está completo, los resultados pueden ser testeados contra los datos excluidos para confirmar la validez del modelo. Si el modelo funciona, las observaciones deben mantenerse para los datos excluidos.

4.2.9 Data Mining y Estadística

El Data Mining es el descendiente y según algunos el sucesor de la estadística tal y como ésta se utiliza actualmente.

Estadística y Data Mining conducen al mismo objetivo, el de efectuar "modelos" compactos y comprensibles que rindan cuenta de las relaciones establecidas entre la descripción de una situación y un resultado (o un juicio) relacionado con dicha descripción. Fundamentalmente, la diferencia entre ambas reside en que las técnicas del Data Mining construyen el modelo de manera automática mientras que las técnicas estadísticas "clásicas" necesitan ser manejadas - y orientadas - por un estadístico profesional.

Las técnicas de Data Mining permiten ganar tanto en rendimiento como en manejabilidad e incluso en tiempo de trabajo. La posibilidad de realizar uno mismo sus propios modelos sin necesidad de subcontratar ni ponerse de acuerdo con un estadístico proporciona una gran libertad a los usuarios profesionales.

El Data Mining es una buena idea ya que al construir espontáneamente un modelo de dependencias en lugar de verificar las hipótesis de un estadístico, es posible, a veces, a través de las técnicas de Data Mining remontar tesoros a la superficie.

Las técnicas del Data Mining nos hacen prescindir de un estadístico, sin embargo, todavía es indispensable dominar el oficio. Las principales ventajas de Data Mining son su rapidez y su sencillez.

Además, dichas técnicas permiten trabajar con grandes cantidades de ejemplos sin ningún inconveniente. También permiten tratar una gran cantidad de variables predictivas.

Como en todo lo producido por la máquina, las predicciones estadísticas fabricadas por el Data Mining deben ser inspeccionadas por personas familiarizadas con el asunto, de manera a comprender y verificar lo que fue producido. Por ende, es importante que dichas predicciones dispongan de una forma ampliamente legible y, en la medida de lo posible, que ya sea conocida en otro campo.

Existe un término medio entre la claridad del modelo y su poder de predicción, mientras más sencilla sea la forma del modelo, más fácil será su comprensión, pero tendrá menor capacidad para tomar en cuenta dependencias sutiles o demasiado variadas (no lineales).

4.2.10 Una arquitectura para Data Mining

Para aplicar mejor estas técnicas avanzadas, 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, en áreas tales como manejo de campañas promocionales, detección de fraudes, lanzamiento de nuevos productos, etc.

El punto de inicio ideal es un Data Warehouse que contenga una combinación de datos de seguimiento interno de todos los clientes junto con datos externos de mercado acerca de la actividad de los competidores. Información histórica sobre potenciales clientes también provee una excelente base para explorar. 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.

Un servidor 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 - resumido por línea de producto, u otras perspectivas claves para su negocio. El Server de Data Mining debe estar integrado con el Data Warehouse y el servidor OLAP para insertar el análisis de negocios directamente en esta infraestructura. Un avanzado, metadata centrado en procesos define los objetivos del Data Mining para resultados específicos tales como manejos de campaña, exploración, y optimización de promociones. 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.

Este diseño representa una transferencia fundamental desde los sistemas de soporte de decisión convencionales. Más que simplemente proveer datos a los usuarios finales a través de software de consultas y reportes, el servidor de Análisis Avanzado aplica los modelos de negocios del usuario directamente al warehouse y devuelve un análisis proactivo de la información más relevante. Estos resultados mejoran los metadatos en el servidor OLAP proveyendo un estrato de metadatos que representa una vista fraccionada de los datos. Generadores de reportes, visualizadores y otras herramientas de análisis pueden ser aplicadas para planificar futuras acciones y confirmar el impacto de esos planes.