Tabla de Contenidos

1 Introducción

1.1 Llevando la Tecnología Informática a niveles estratégicos de negocios

En el presente informe se abordan temas de actualidad en el área de la Inteligencia de Negocios. Se plantea ésta como un movimiento o tendencia dentro de la Tecnología de Información (TI), que pretende cubrir las necesidades de información de nivel estratégico de las empresas y/o instituciones que aprovechan y dependen de sus sistemas informáticos.

Si se tiene una visión muy simplista de la TI, limitándola exclusivamente a su perspectiva transaccional, se puede deducir, erróneamente, que la TI no es generadora de valor agregado para las empresas y se la percibe como una necesidad operativa; si se quiere, como "un mal necesario". No obstante, con una visión más amplia y realista, se debe reconocer que la TI ayuda a hacer eficientes muchos de los procesos de negocios de las empresas, y que facilita sus actividades normales, como la comunicación interna, los cierres contables, el manejo de los inventarios, etc. Finalmente, si analizamos la evolución de la economía mundial en las últimas décadas, debemos reconocer que el papel de la TI en "la aldea global" es imprescindible, ya que muchos de los negocios y las relaciones humanas actuales que no se podrían llevar a cabo sin una infraestructura de TI eficiente y adecuada (nos referimos a infraestructura como una concepción amplia y que engloba a los sistemas informáticos, las telecomunicaciones, la investigación, etc.), tal es el caso de los negocios de transferencias de dinero, los negocios electrónicos (o su generalización dentro del e-business), los sistemas de cámaras de compensación bancarios, el correo electrónico, etc.[1]

Ahora bien, sea cual fuere la visión que se tenga de la TI, tendemos a ubicarla en ámbitos transaccionales; es decir, la TI permite, facilita y/o genera procesos transaccionales. Las aplicaciones, las bases de datos, las telecomunicaciones y en general, toda la tecnología informática se han diseñado y optimizado pensando en necesidades y objetivos transaccionales.

En el otro extremo, siempre han existido las necesidades de información estratégica, resumida e histórica, es decir: información para la alta gerencia. De hecho estas necesidades son más antiguas que la propia Informática. Los sistemas transaccionales generan y almacenan datos, pero no se diseñaron ni optimizaron para analizarlos. Podemos extraer alguna información del sistema, pero siempre con capacidades y rendimientos muy limitados. Dadas tales necesidades, muchas veces se intenta almacenar información gerencial, por medio de tablas históricas, vistas, mini-repositorios de datos, etc. Ahora bien, estos "sub-sistemas estratégicos" tienen grandes limitaciones de rendimiento y capacidad de almacenamiento, por lo que no pocas veces ponen en aprietos al sistema "base".

Cuando estos "sub-sistemas estratégicos" se formalizan como un "producto" o un "sujeto de consultoría", nos encontramos ante los tradicionales "Sistemas de Información Gerencial", éstos generan y presentan informes y gráficos, para los gerentes, con datos extraídos de los sistemas de

[1] Está fuera del alcance de este informe el valorar el papel de la TI en la economía actual. No obstante, debemos reconocer el hecho de que buena parte de la economía mundial globalizada se soporta sobre una gran infraestructura informática y de telecomunicaciones, la cual es dinámica y está en una constante evolución.

información transaccional. Estos sistemas constituyen una formalización de las "consultas gerenciales", por lo que sufren de todas las debilidades y limitaciones mencionadas, posiblemente en mayor medida.

Dada esta situación, ¿cómo responde la TI a esta necesidad de información estratégica gerencial? Por medio de la especialización, se ha creado dentro de la TI un área para la generación de información de alto nivel, que se conoce como Inteligencia de Negocios (IN). Los sistemas de IN tienen objetivos y necesidades distintos de los transaccionales, se han desarrollado nuevos paradigmas para el almacenamiento y consulta de la información, nuevas técnicas para el diseño y modelación de los sistemas, novedosos métodos de visualización de la información, y nuevas metodologías para la administración de los proyectos de IN. Es decir, se trata de un enfoque integral que vas más allá de la propia técnica, tratándose de una nueva área de especialización dentro de la TI.

En el presente informe, se pone a disposición del lector una visión sintética y actualizada de la IN, de manera que permita tener claros sus alcances y limitaciones, así como contar con una visión actualizada de la IN. El informe no pretende ser un manual técnico ni un estudio teórico de la materia, si bien hace referencias a la teoría y recomienda bibliografía para que el lector cuente con una guía para investigar en mayor profundidad los temas tratados.

Se propone una visión global e integral de la Inteligencia de Negocios, seguida por capítulos para las distintas tecnologías de IN. Se presenta al final un conjunto de recomendaciones para enfrentar este tipo de proyectos.

2 La tendencia de la Inteligencia de Negocios en la Tecnología de la Información actual

El concepto de **Inteligencia de Negocios (IN)** se utiliza para hacer referencia a un conjunto de tecnologías de Bases de Datos, de Aplicaciones, de plataformas tecnológicas, así como de modelos de razonamiento y de modelación de sistemas informáticos, y en general, a una tendencia que pretende integrar precisamente a tales elementos en un todo coherente. Lo que se pretende por medio de la IN es generar información estratégica y ponerla a disposición de los tomadores de decisiones de las empresas y/o instituciones, logrando así un mayor valor agregado de la TI a la organización, al superar los aspectos operacionales[2].

Para Elizabeth Vitt, dado que la IN es un concepto multifacético, "debemos examinarlo desde sus diferentes perspectivas u objetivos" [Vitt 2002]:

1. **Tomar mejores decisiones y más rápidamente:** la meta principal de los sistemas de IN es mejorar el rendimiento de los negocios de las organizaciones. Hacerlas más eficientes y coherentes y finalmente que éstas logren tomar las decisiones más acertadas.
2. **Convertir los datos en información:** existe un gran camino entre los datos que proveen los sistemas transaccionales y la información que requieren quienes toman decisiones para hacer su trabajo. En IN se convierten estos datos en la información requerida por estos usuarios de un alto nivel.
3. **Utilizar un enfoque racional en la administración:** Vitt considera que "la IN se puede describir como un enfoque de la Administración, un estado mental organizacional, una filosofía de la Administración, …, en resumen, se trata de la actitud de la IN" [Vitt 2002]

En lo últimos años la tecnología relacionada con la IN ha estado evolucionando y expandiéndose en forma continua. Se han ido generando nuevas áreas de especialización, muchas de las cuales no pasan de ser tecnologías emergentes, mientras que otras se van consolidando a la vez que se siguen expandiendo y multiplicando.

Este desarrollo va de la mano con la evolución de aspectos de investigación académica y/o científica así como de la evolución del hardware y del software que utilizan las empresas para sus desarrollos de sistemas de IN.

En términos generales podemos decir que la IN ayuda a las empresas a generar información a partir de sus datos, en una forma rápida y dinámica, con lo cual se pueden tomar mejores decisiones de negocios de una manera más rápida que en el pasado. Con esta información se establecen y miden los resultados de los negocios de las empresas. Precisamente lo que motiva el desarrollo de sistemas de IN es el objetivo de contar con información actualizada que apoye la toma de decisiones estratégicas de negocios, y que dicha información esté disponible cuando se la requiera.

[2] Por medio de un proyecto de IN exitoso, el área de TI podría elevarse a un nivel gerencial o afianzarse en él. De esta manera, un beneficio colateral es que la IN puede a mejorar el status de la TI en la empresa.

Podemos entender la IN como aquellos sistemas de información que responden a la visión de la "Tecnología que genera Información Estratégica para la toma de decisiones". Desde este punto de vista, un sistema de IN puede estar compuesto por sofisticados sistemas de bases de datos, herramientas de consultas, interfaces de usuario final, complejos algoritmos para la generación de información, programas de carga de datos, etc. En el otro extremo, un simple Balanced Scorecard[3] en una hoja electrónica que se actualiza "manualmente" puede ser considerado como un sistema de IN.

Esta concepción de la IN va más allá del alcance mismo de la Informática, pues se requiere de un componente interdisciplinario que es fundamental (es una condición necesaria para el éxito): el conocimiento específico de los negocios desde el punto de vista gerencial y/o estratégico. A partir de dicho conocimiento es que se deben desarrollar los modelos de datos de IN y será a partir de este conocimiento que se validará la funcionalidad y utilidad del sistema desarrollado. Por lo tanto, el conocimiento interdisciplinario de cada tipo de negocio forma parte intrínseca de los sistemas de IN; no puede concebirse la existencia de un sistema de IN sin este componente.

Ahora bien, ¿cuáles tecnologías computacionales se consideran dentro de la IN? Si partimos de una visión histórica reciente, observamos que los sistemas y tecnologías de los Depósitos de Datos (Data Warehouse) y/o Mercados de Datos (Data Marts) responden a estos requerimientos de información[4.] Dado que los DW son quizá la tecnología de IN de mayor difusión y consolidación, es muy común el que se equiparen los conceptos de IN con los de DW, no obstante en este informe se adopta una concepción más amplia, como se ha venido comentado.

Desde hace varios años los sistemas y tecnología de los DW han dejado de ser una novedad y/o "tecnología emergente" para convertirse en grandes sistemas de información complementarios a los sistemas transaccionales de las empresas[5,] consolidados y en un constante crecimiento (tanto en volumen de datos como en requerimientos). De hecho hoy en día casi todos los fabricantes de tecnologías de bases de datos cuentan con motores orientados a los Data Warehouse, que son complementarios e integrados a sus motores de bases de datos transaccionales. Todas estas empresas ya han generado al menos dos generaciones de este tipo de tecnología. Igualmente, existen fabricantes de software totalmente especializados en el área, e incluso en lo que se pueden considerar como sub-áreas o bien componentes del DW, como por ejemplo:

- Modelaje de datos
- Migración y transformación de datos
- Almacenamiento (bases de datos)
- Visualización de la información
- Consultoría en el desarrollo e implementación y Administración de Proyectos
- Capacitación en diversas etapas de la IN, incluyendo especializaciones de nivel de post-grados.

[3] Frecuentemente se traduce como 'Cuadro de mando integral'. Dejamos el término en Inglés, dado su amplia utilización por los lectores en Español.

[4] Para un estudio a profundidad de estos sistemas véase "Depósitos de Datos" de Beatriz Jiménez y Rafael Ávalos, informe No.25 del Club de Investigación Tecnológica [Jiménez 1998].

[5] Esta aseveración es correcta al menos para los países industrializados y para algunos países Latinoamericanos. Sin embargo en otras naciones de Latinoamérica los *Data Warehouse* siguen considerándose algo novedoso, emergente y muchas veces se les observa con mucha desconfianza (tal es el caso de Costa Rica).

Si partiéramos de una visión histórica más amplia, podemos afirmar que incluso los primeros "Sistemas de Información Gerencial" de hace más de 30 años buscaban ya los objetivos arriba planteados; de hecho, los requerimientos de información gerencial-estratégica siempre han existido. No obstante, lo que se incluye dentro de la IN es tecnología reciente, básicamente de los años 90 en adelante.

Los Data Warehouse y Data Marts nos brindan un enfoque histórico de los resultados generados en la empresa. Si bien el enfoque es (y debe ser) totalmente gerencial, cualquier gerente sabe que la información histórica es solo parte de los insumos que él requiere para tomar sus decisiones. Se requiere de más información en otras áreas, "etapas" y/o niveles de los procesos de toma de decisiones. Recientemente han florecido otras tecnologías que van complementando estos requerimientos de datos, para citar algunos casos específicos:

- **Balanced Scorecards:** se utilizan para llevar un control en forma continua de los resultados de metas específicas de negocios (por medio de medidas), se "vigilan" estadísticas, razones financieras, etc. Estos sistemas trabajan en forma "casi" en línea, y responden a modelos de análisis de negocios a partir del estudio constante de la evolución de metas específicas de negocios, y de la interrelación de éstas. La gestión gerencial se divide en cuatro grandes áreas: Visión y Estrategia, Comunicación, Planificación y Aprendizaje.

  La teoría del Balanced Scorecard (BS) fue propuesta por Robert Kaplan y Edward Norton [Kaplan 1996], para quienes BS es más que un sistema de medidas, para pasar a convertirse en un sistema de administración basado en el análisis de dichas las medidas.
- Para obtener un análisis estadístico e inferencial de la información se aplican, dentro de la **Minería de Datos** algoritmos especializados que pretenden adelantarse al comportamiento de los clientes, del mercado y, en general, de los agentes económicos y variables involucradas en los procesos de negocios.
- En Minería de Datos se incluyen y coexisten varios tipos de tecnologías y algoritmos de diversas áreas de especialización. Entre las aplicaciones comunes están los sistemas y modelos predictivos, sistemas de "alarmas", análisis de riesgo, etc.
- Para darle seguimiento a los clientes de la empresa en forma puntual, se han desarrollado los sistemas de **Customer Relationship Management (CRM)[6].** Estos sistemas pretenden dar avisos sobre cuáles clientes deben requerir especial atención, para aprovechar las oportunidades de negocios que ellos ofrecen.
- **Visualización de la información**: Se trata de las herramientas del usuario final, que permiten obtener representaciones gráficas y/o tabulares de la información almacenada en los sistemas de IN. Existen muchas técnicas para visualizar los datos[7]: mapas geográficos, pictogramas, gráficos estadísticos, diferentes tipos de tablas. También hay diversos elementos para la interacción del usuario con el sistema: barras de herramientas, botones, objetos para seleccionar filtros, etc.

[6] Administración de las relaciones con los clientes.

[7] A lo largo del informe se presentan técnicas de visualización y al final aparece un capítulo destinado a las herramientas del usuario final.

La visión actual de la Inteligencia de Negocios integra todas estas tendencias en un *todo coherente*, dentro de una estrategia global de IN. Esta visión es, entre otras cosas, muy eficiente, ya que permite generar una estrategia *macro* de la IN para las empresas, con tecnologías de desarrollo integradas y con herramientas de visualización de datos (para los usuarios) con un punto común de acceso a la información. No obstante, los sistemas se pueden desarrollar en forma aislada, lo cual puede llevar a las empresas ha incurrir en la duplicidad de tareas y esfuerzos.

3 Sistemas de Data Warehouse, Data Marts y Tecnología OLAP

3.1 Sistemas de Data Warehouse y Data Marts

Los Data Warehouse (DW) se han definido de muchas maneras, en las que a veces se incluyen o excluyen elementos, e incluso se encuentran algunos que son contradictorios entre sí, por lo que es difícil llegar a una definición exacta o única. En términos muy generales, podemos decir que un DW está compuesto por una o mas Bases de Datos Históricas, las cuales se alimentan de los OLTP8 de las empresas y de otras fuentes. Los DW integran información de distintos sistemas y áreas de la empresa, y se espera que integren la información de todos los sistemas, unidades de negocios, oficinas, sucursales, regiones, etc. que conforman la empresa[9].

Al hablar de Datawarehousing es inevitable el hacer referencia a los dos autores más conocidos en la materia: Ralph Kimball y William H. Inmon. Ambos han desarrollado sus propios enfoques, modelos y arquitecturas de los DW, y establecido empresas de consultoría que se fundamentan en estos enfoques.

El término "Data Warehouse" fue utilizado por primera vez por William H. Inmon en uno de sus libros: "Building the Data Warehouse" publicado en 1992. Mucha gente ha adoptado la definición de Inmon a tal punto que la utilizan para determinar si un sistema "se trata realmente de un DW" solo si éste se apega a su definición. Para Inmon "un Data Warehouse es una colección de Datos con una orientación a temas específicos, integrados, no volátiles e históricos, tal que apoya los procesos de toma de decisiones" [Inmon 96]. En esta definición se involucran los cuatro elementos fundamentales de un DW que merecen atención individual:

1. Son orientados a temas específicos. Los DW se enfocan en temas o categorías de los datos, no en aspectos puntuales-transaccionales. Se trata más bien de resumir los datos por categorías tales como: clientes, agencias, oficinas, colores (de productos), etc.
2. Integrados: la información del DW involucra a toda la institución. Para tomar una decisión de alto nivel se requiere contar con toda la información y no con datos aislados. El concepto de integración es amplio: se trata de toda la organización tanto desde el punto de vista de sus unidades de negocios (finanzas, contabilidad, ventas, producción, etc.) como desde el punto de vista geográfico (todas las oficinas, sucursales, etc.).
3. No volátiles: La información del DW es de consulta, se trata de un repositorio de datos. En otras palabras, el sistema no es transaccional, no recibe actualizaciones en sus datos, presenta la información a manera de fotografías en distintos momentos del tiempo. Normalmente el DW solo recibe dos operaciones: la carga inicial de sus datos y la lectura posterior de éstos.
4. Históricos: El DW es histórico, es necesario comparar los resultados de la empresa en distintos períodos de tiempo. La antigüedad de la historia que se almacena en los DW es variable, normalmente no es necesario almacenar más de 3 años de historia, pero esto no

[8] **OLTP = On Line Transactional Processing**. Se trata de los sistemas operacionales o transaccionales de las empresas.
[9] Para un análisis conceptual más profundo de los Datawarehouses se recomienda leer [Jiménez 1998].

es una regla. Todos los reportes del DW siempre van a involucrar la variable Tiempo (explícita o implícitamente).

La información de los DW tiene un enfoque gerencial-estratégico, es decir, los modelos de datos subyacentes a estos sistemas responden a la necesidad de información estratégica histórica para tomar decisiones de alto nivel en las empresas. Dichos modelos de datos responden a los "procesos naturales de razonamiento" que utilizan los gerentes y/o ejecutivos de las empresas para tomar sus decisiones[10]. Este enfoque de diseño de datos es radicalmente distinto a los de los OLTP, en donde se busca llevar el registro de las transacciones operativas del negocio o bien generarlas.

Una definición de Data Mart nos dice que éste representa un subconjunto del DW, por lo tanto heredará los cuatro conceptos básicos de la definición de Inmon y los objetivos de diseño y requerimientos de información de los DW. Según Ralph Kimball, "un Data Mart es un trozo completo del pastel tomado del pastel completo que es el Data Warehouse" [Kimball 1998]. Por lo tanto, un Data Mart es un sistema de IN para una unidad de negocios específica (por ejemplo el Data Mart de Ventas o el Data Mart del Área de Crédito) o bien del sistema para un conjunto administrativo o regional específico (por ejemplo el Data Mart de la subsidiaria de Costa Rica). Ya sea el enfoque geográfico o el de unidades de negocios, los Data Marts deben ser completos en sí mismos y deben respetar las características de los DW:

1. Deben ser históricos,
2. No volátiles,
3. Orientados a temas específicos,
4. Integrados (dentro de sus alcances particulares pero además integrados con los demás Data Marts).

A partir de esta concepción de los Data Marts, Ralph Kimball define el Data Warehouse como "… la unión de todos los Data Marts que lo componen" [Kimball 1998]. De aquí se deriva su arquitectura denominada "The Data Warehouse Bus Structure" en la cual cada Data Mart es una parte del DW y todos están interconectados utilizando estructuras comunes (es decir, están integrados)."

Es en este punto donde se encuentran las principales diferencias entre estos autores: para Inmon el DW es el lugar donde las empresas alcanzan la integración de sus sistemas. Kimball describe su "Data Warehouse Bus" como la arquitectura que integra la información con Data Marts que comparten e integran los distintos elementos del sistema[11]. Para Inmon los Data Marts se derivan del DW y para Kimball los Data Marts conforman el DW [Peterson 2000].

[10] Algunos advierten que tales procesos de razonamiento no son universales, por lo que posiblemente sea necesario ajustar los modelos de datos cuando cambia un gerente.

[11] Los DW incluyen cubos de datos multidimensionales y/o bases de datos relacionales, en cuyo caso los cubos pueden compartir dimensiones y medidas mientras que los esquemas de bases de datos relacionales pueden compartir tablas de catálogos y/o de parámetros. Esto se analiza en la sección "3.4 Tecnología OLAP y Pensamiento Multidimensional"

3.2 Depósitos de Datos Operativos

En el enfoque de Inmon se incluye lo que él ha denominado un Depósito de Datos Operativo (DDO). Éste es un repositorio de información transaccional que incluye como mínimo los datos que se requiere cargar históricamente al DW. Es decir, los datos se cargan primero al DDO y de este se van trasladando periódicamente al DW.

Una de las ventajas de contar con un DDO es que él puede servir a otros propósitos además de ser la fuente de información del DW: es un repositorio único e integrado para generar los reportes transaccionales de la empresa, liberando de esta manera a los sistemas transaccionales de las tareas de reportería. El DDO resulta muy útil cuando en las empresas conviven distintos sistemas informáticos con plataformas de tecnología distintas, ya que el DDO integra y/o resume los datos en un solo lugar (es decir, que el DDO nos lleva en forma natural a alcanzar el objetivo de la integración del DW).

Otra ventaja es que el DDO es que es un sistema confiable: los datos han sido validados, depurados, corregidos, y además debe ser diseñado en forma eficiente.

Sin embargo la desventaja de construir un DW a partir de un DDO radica en el hecho de que los resultados se van obtener tras un período de desarrollo que puede ser largo, pues el DDO es en sí mismo un proyecto complejo y delicado. Es decir no existe el DW sino hasta que se haya construido el DDO.

En el otro extremo, el enfoque de los Data Marts nos permite desarrollar pequeños módulos del DW en el corto plazo, los cuales se van liberando y poniendo en producción en períodos cortos de tiempo. Con esta estrategia se generan resultados en el corto plazo. Los peligros de seguir esta práctica son, entre otros:

- No lograr una correcta integración de los Data Marts, así, al final se construirá con un conjunto de "Sistemas de Información Gerencial" aislados y desintegrados, lo cual contradice la concepción misma de lo que es un DW.
- Incurrir en la duplicación de esfuerzos (lo que implica el desperdicio de recursos), al repetir tareas en diferentes etapas del proyecto.

Sea cual sea la estrategia que se siga, se debe contar con una metodología de trabajo adecuada y con experiencia para lograr alcanzar los objetivos finales.

En resumen, y sin entrar en contradicciones, se puede concluir que un DW es un Repositorio de Datos histórico, con información institucional integrada, diseñado y construido con una visión gerencial y estratégica, que ayuda a mejorar los procesos de toma de decisiones al brindar información histórica y resumida de las principales medidas de interés de la empresa. El DW no es volátil, es decir, no se modifica ni recibe operaciones transaccionales, solamente cargas de datos que son "fotografías" de éstos en momentos específicos del tiempo. El DW puede estar compuesto por un conjunto de Data Marts que comparten estructuras y elementos comunes, o bien, se puede construir a partir de un DDO único del cual se pueden derivar diversos Data Marts (perfectamente integrados).

3.3 Diferencias entre los sistemas transaccionales y los depósitos de datos (Data Warehouse)

Una manera de comprender la naturaleza de un DW es comparándolo con los sistemas informáticos transaccionales. Las personas con formación informática están familiarizadas con este tipo de tecnologías más no con las de los DW. En esencia esta comparación es válida no solo con respecto de los DW sino con respecto de globalidad de la tecnología de la IN.

Para Jiawaei Han las operaciones básicas de los sistemas informáticos transaccionales son las de ejecutar transaciones en-línea y llevar a cabo el procesamiento de consultas. La mayoría de operaciones de negocios del día a día se registran o ejecutan por medio de estos sistemas: ventas, asientos contables, actualizaciones de inventarios, transferencias de fondos entre cuentas, retiros y/o depósitos bancarios, pagos de sueldos, etc. [Han 2001]. Por otra parte los DW no reciben actualizaciones constantes de datos, la información que se almacenan proviene de "cortes" de la información en momentos específicos del tiempo. Además el DW no brinda detalle alguno sobre transacciones específicas, sino que proveen información resumida y relevante para el apoyo a la toma de decisiones. La tecnología que permite generar este tipo de consultas a los DW se conoce como OLAP (On Line Analytical Procesing). En el cuadro 4.2.1. se presenta una comparación de las características de los sistemas transaccionales y los Data Warehouse.

Cuadro 4.2.1. Comparación de características de los sistemas transaccionales y los Data Warehouse

Característica	Sistemas transaccionales	Data Warehouse
Orientación del sistema	Ejecución y procesamiento de transacciones del 'día a día'	Generación de información estratégica e histórica.
Usuarios	Oficinistas, contadores, personal informático, jefes de departamentos operativos, clientes, etc.	Gerentes, ejecutivos, Juntas Directivas, analistas de información.
Tipo de diseño de base de datos	Modelos Entidad Relación, y/o sistemas de bases de datos orientados a las aplicaciones OLTP	Bases de datos Multidimensionales, Esquemas Relacionales del tipo Estrellas, con objetivos estratégicos en la información.
Nivel de detalle de los datos	Se almacenen con el mayor detalle, se trata de las transacciones específicas	Datos agregados en distintos niveles, no interesa el detalle sino los resúmenes de los datos.
Características del hardware y su configuración	Servidores de "pequeños a medianos" con sistemas de alta redundancia, configurados para tener recuperaciones inmediatas ante fallas y optimizados para realizar transacciones puntuales en línea y con "muchos usuarios".	Servidores de "grandes a gigantes", optimizados para almacenar grandes volúmenes de datos y responder a consultas complejas que involucran mucha información, y con "pocos usuarios".
Operaciones normales	Mucha lectura y escritura: actualizaciones, inserciones, sistemas de seguridad con alta redundancia, consultas.	Básicamente lectura de los datos: consultas complejas de los usuarios.

Volúmenes de datos	Dado que la información es siempre la actual, el volumen de datos no responde a la cantidad de transacciones que se almacenen. De 100 MB a 1 o 2 GB.	Dado que se acumula información histórica, los DW crecen constantemente. Los volúmenes se miden en Gigabytes a TeraBytes

3.4 Tecnología OLAP y Pensamiento Multidimensional

3.4.1 Cubos Multidimensionales de Datos, Dimensiones y Medidas

La mayoría de los sistemas de DW y las herramientas de consultas OLAP se basan en las Bases de Datos Multidimensionales para almacenar la información y generar sus consultas en forma dinámica. Este tipo de bases de datos utilizan el concepto de Cubo de Datos Multidimensional en vez de tablas o archivos planos. Un cubo de datos se construye a partir de un esquema de base de datos de estrella.

En los cubos multidimensionales se almacena la información organizándola en Medidas y Dimensiones. Las medidas son los datos cuantitativos, es todo lo que se puede sumar, contar, etc. Las Dimensiones son categorías descriptivas, es decir son datos cualitativos. Finalmente en un cubo multidimensional de datos sobre un tema o aspecto de negocios, se generan todas las posibles combinaciones de medidas y dimensiones respecto del tema.

Para aclarar el concepto de los cubos multidimensionales, se presenta a continuación un ejemplo con un cubo de ventas para una empresa de automóviles (Ver cuadro 4.3.1.1). El cubo tiene las siguientes dimensiones y medidas:

Dimensiones

- Marca del vehículo
- Fecha de la venta
- Categoría del vehículo (Sedan, Station Wagon, etc.)
- Estado del vehículo (Nuevos, Usados)

Medidas

- Costo total del vehículo
- Venta Bruta
- Margen Neto (=Venta Bruta - Costo total del vehículo)
- Comisión de vendedor

A partir del cubo de ventas de vehículos es posible construir un sinnúmero de reportes multidimensionales.

La "naturaleza" de los cubos es la de agregar y desagregar las medidas a través de las dimensiones. En el siguiente cuadro se muestra la agregación y desagregación por medio de los distintos niveles de las dimensiones. En el tiempo se muestran el detalle en el nivel del mes, luego se agregan en el trimestre, y los trimestres se agregan en el año. Además se agregan las

medidas agregando las dimensiones.

Es importante resaltar el hecho de que la información que se extrae de los cubos, ya sea con herramientas gráficas o tabulares, es generada por los usuarios en forma muy sencilla. Los reportes se pueden construir en unos minutos, liberando de esta forma al área de TI de las labores de 'reportería'.

Cuadro 4.3.1.1 Reporte OLAP de Ventas de Vehículos. Se muestra la dimensión de tiempo agregada y desagregada, la dimensión de Categoría de Vehículo, la de Marca. La medida es de la "Venta Bruta Colones". En el ejemplo se muestra una desagregación mensual del segundo cuatrimestre del 2003.

Ventas de vehículos usados, según Fecha de Venta, Categoría de Vehículo y Marca

Estado del Vehículo Usado

Venta Bruta Colones				Marca Vehiculo		
Año	Trimestre	Mes	Categotria Vehiculo	Honda	Nissan	Gran Total
2002						
			4 x 4		87.816.374	87.816.374
			Sedan	610.311.012	696.698.509	1.307.009.520
2002 Total				610.311.012	784.514.883	1.394.825.895
2003						
	Trim 1-2003					
			4 x 4		22.300.000	22.300.000
			Sedan	103.100.000	169.900.000	273.000.000
	Trim 1-2003 Total			103.100.000	192.200.000	295.300.000
	Trim 2-2003					
		Abril-2003				
			4 x 4		10.500.000	10.500.000
			Sedan	60.100.000	62.900.000	123.000.000
		Abril-2003 Total		60.100.000	73.400.000	133.500.000
		Mayo-2003				
			4 x 4		15.500.000	15.500.000
			Sedan	43.500.000	65.000.000	108.500.000
		Mayo-2003 Total		43.500.000	80.500.000	124.000.000
		Junio-2003				
			4 x 4		12.000.000	12.000.000
			Sedan	53.000.000	62.400.000	115.400.000
		Junio-2003 Total		53.000.000	74.400.000	127.400.000
	Trim 2-2003 Total			156.600.000	228.300.000	384.900.000
	Trim 3-2003					
			4 x 4		49.500.000	49.500.000
			Sedan	188.900.000	201.700.000	390.600.000
	Trim 3-2003 Total			188.900.000	251.200.000	440.100.000
	Trim 4-2003					
			4 x 4		46.000.000	46.000.000
			Sedan	269.100.000	228.700.000	497.800.000
	Trim 4-2003 Total			269.100.000	274.700.000	543.800.000
2003 Total				717.700.000	946.400.000	1.664.100.000
Gran Total				1.328.011.012	1.730.914.883	3.058.925.895

3.4.2 Esquemas de Estrella

Los cubos se construyen a partir de datos almacenados en tablas, con un esquema de estrella, el cual está compuesto por una tabla de hechos y un conjunto de tablas de dimensiones. La tabla de hechos tiene llaves foráneas que hacen referencia a las dimensiones, normalmente el conjunto de llaves foráneas se pueden utilizar como llave primaria. La tabla de hechos además tiene campos numéricos (las medidas), se trata de los valores que se desea analizar, agregar o resumir.

Son ejemplos de Medidas:

- Saldos de Cuentas
- Montos de Ventas
- Cantidad de Clientes
- Montos de Intereses devengados
- Etc.

Las Tablas de dimensiones contienen información descriptiva y se trata básicamente de tablas de categorías. Las dimensiones se organizan en Jerarquías, las cuales contienen niveles, en cada nivel de una jerarquía se almacenan los Miembros de la dimensión.

Ejemplos de Dimensiones y Jerarquías.

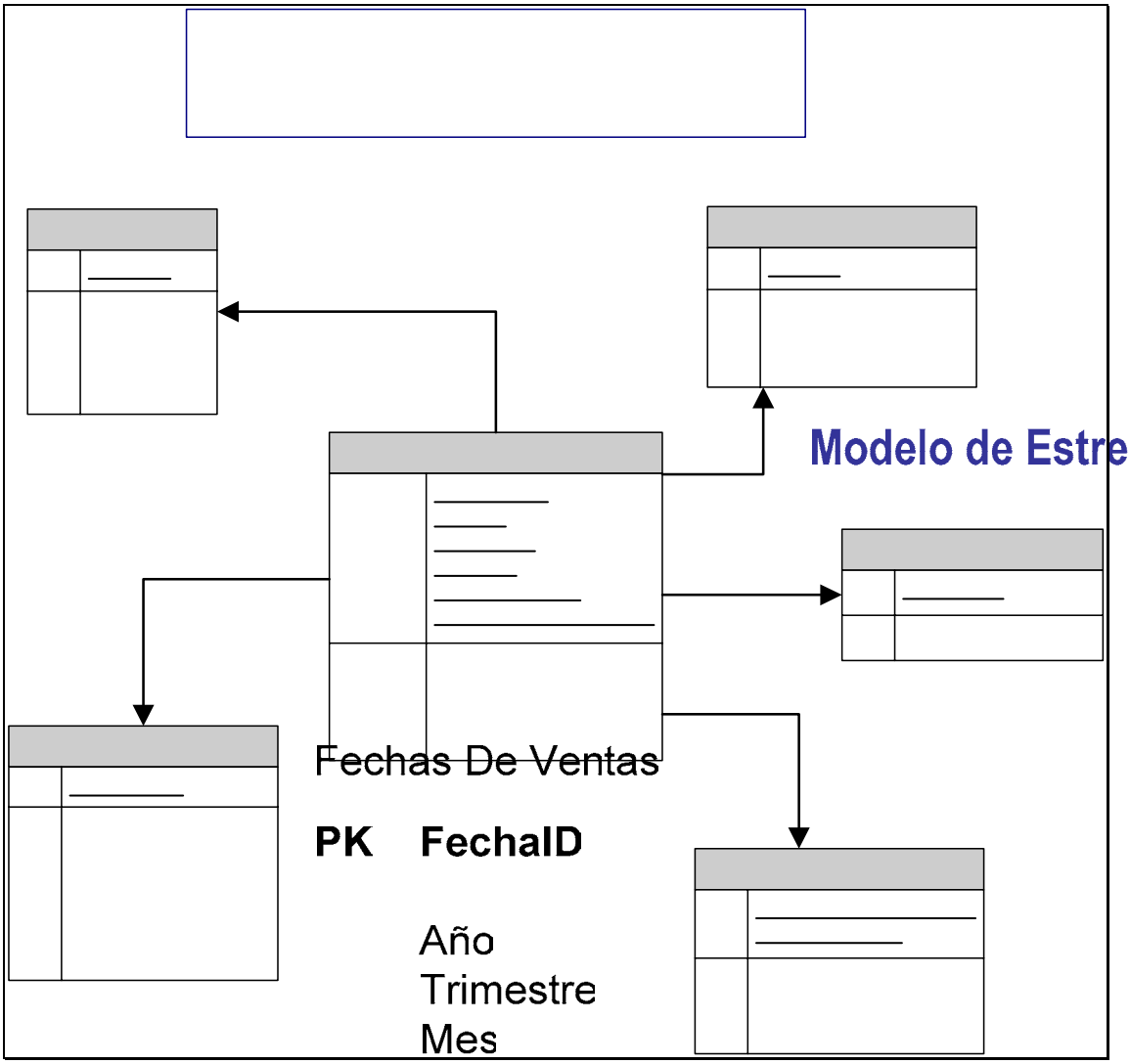
Cuadro 4.3.2.1. Dimensión de Tiempo

JERARQUÍA 1		**JERARQUÍA 2**	
Niveles	**Miembros**	**Niveles**	**Miembros**
Año	2000, 2001, 2002, …	Año	2000, 2001, 2002, …
Cuatrimestre	Cuat. 1-2000, Cuat. 2-2000, Cuat. 3-2000, …	Semestre	Sem. 1-2000, Sem. 2-2000, Sem. 1-2001, …
Mes	Enero-2001, Febrero-2001, Marzo-2001, …	Trimestre	Trim. 1-2000, Trim. 2-2000, Trim. 3-2000, …
		Mes	Enero-2001, Febrero -2001, Marzo-2001, …

Cuadro 4.3.2.2. Dimensión Geográfica

JERARQUÍA	
Niveles	**Miembros**
Provincia	San José, Alajuela, Cartago
Cantón	San José Central, Curridabat, Santa Ana
Distrito	Los Angeles, Pavas, Distrito Central
Barrio	Santa Catalina, Santa Fe, Lomas de Ayarco

Ilustración 4.3.2.1. A continuación se presenta un modelo de Estrella por medio de un diagrama Entidad Relación.



Kimball define el modelaje multidimensional de la siguiente manera: "Es una técnica de modelación lógica que busca presentar los datos en una estructura estándar, la cual es intuitiva y que facilita un alto rendimiento en el acceso a los datos. Es inherentemente dimensional y utiliza el Modelo Relacional con algunas importantes restricciones" [Kimball 98]. La estructura estándar es precisamente la de la estrella, nótese que la Tabla de Hechos representa un conjunto de relaciones de muchos a muchos. El modelo es intuitivo en el sentido de que se basa en la organización de las medidas según las categorías de los datos, de ahí el concepto de dimensionalidad ya que las categorías son precisamente las dimensiones. Obsérvese además que la estrella nos presenta los datos en una forma des-normalizada, si se quiere redundante, esto hace precisamente que la consulta de los datos sea muy eficiente ya que muchos de los joins se encuentran pre-almacenados de esta manera.

Un concepto fundamental es el de la Aditividad de las Medidas. Rara vez se analiza la información de un registro aislado, más bien lo que se consulta son agrupaciones de registros filtrados por valores específicos de las dimensiones. Por lo tanto, tal y como se ha comentado, las medidas se deben agregar y esto se realiza resumiendo los datos por medio de funciones básicas como sumas, máximos, mínimos, etc. Alternativamente, esto se puede hacer por medio de operaciones y/o algoritmos más complejos.

A partir del esquema estrella presentado se puede generar un cubo que genere reportes con información como:

El costo total de los productos que se han vendido en una sucursal específica, en un rango de días determinado, el nombre de estos productos y la moneda en que se han realizado estas ventas.

3.4.3 MOLAP, ROLAP y HOLAP

Los Motores de Bases de Datos Multidimensionales en realidad se alimentan o leen su información de las estrellas de datos, pero luego la almacenen en estructuras propias e independientes, específicamente diseñadas y optimizadas para almacenar y generar agregaciones, historia y reportes ad-hoc. Obsérvese que esto contrasta con los motores relacionales, los cuales se optimizan para registrar transacciones, para tener recuperaciones ante fallos, para trabajar en-línea, etc. Lo que normalmente se consulta son los cubos almacenados en estos motores multidimensionales, para esto se han construido lenguajes de consulta propios (como el MDX = Multidimensional Expressions), los cuales son complejos y requieren conocimientos técnicos. No obstante, las herramientas de usuario final facilitan la tareas de las consultas ya que son herramientas gráficas en las que el usuario no debe programar ni escribir código alguno; los datos se le presentan al usuario en forma de 'tablas pivote'[12] y con características gráficas.

Existen varios métodos de almacenamiento de los datos de los cubos:

- MOLAP: las dimensiones y medidas se almacenan en estructuras multidimensionales. Este método es el que da los mejores tiempos de respuesta en la consultas, pero es el que consume más espacio.
- Se utiliza MOLAP para almacenar los datos que se consultan más frecuentemente.
- ROLAP: se almacenan las medidas y dimensiones en tablas relacionales. Este método es el que da los tiempos de respuesta más lentos en la consultas, pero es el que consume menos espacio.
- Se utiliza ROLAP para almacenar los datos que se consultan con menos frecuencia, por lo general se trata de datos de años anteriores.
- HOLAP: se utilizan métodos híbridos, repartiendo el almacenamiento de los datos entre tablas y estructuras multidimensionales con tiempos de respuesta intermedios.

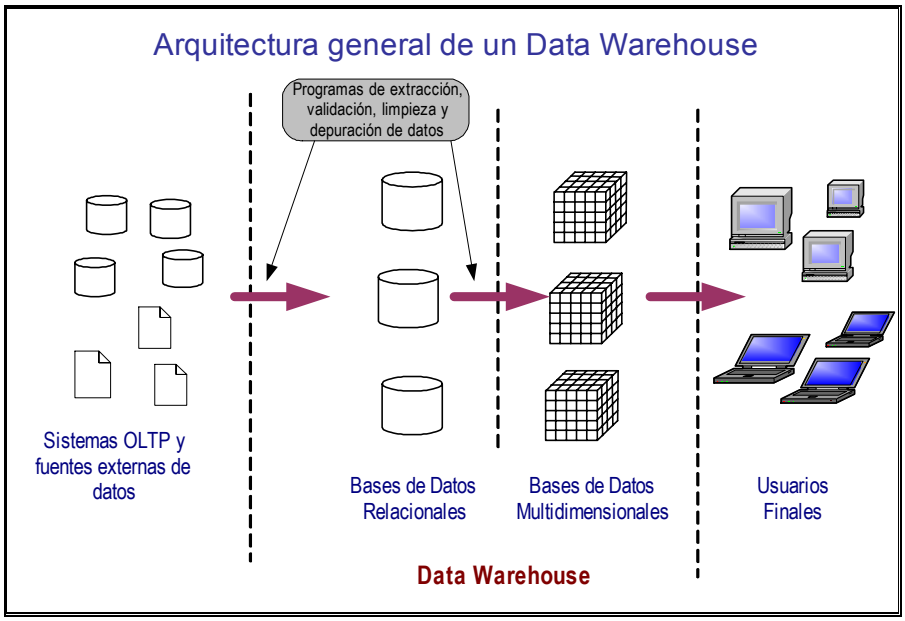
[12] Una tabla-pivote combina y compara grandes cantidades de datos, obteniendo totales y subtotales. Se pueden rotar sus renglones y columnas para generar diferentes subtotales y obtener así distintas perspectivas de los mismos datos. Son de uso común en las hojas electrónicas.

3.5 Arquitectura de los sistemas de Data Warehouse

Los componentes típicos de la arquitectura de un Data Warehouse conforman tres niveles básicos:

1. Bases de Datos Relacionales
2. Bases de Datos Multidimensionales
3. Herramientas dirigidas a los usuarios finales

Ilustración 4.4.1. Arquitectura general de un Data Warehouse



En el nivel de las Bases de Datos Relacionales se depositan los datos que se extraen desde las distintas fuentes de información: bases datos de los sistemas OLTP, archivos de texto, hojas electrónicas, etc. Durante esta extracción los datos sufren procesos de depuración que involucran validaciones, correcciones, limpieza y transformaciones para convertirlos en los datos que se requieren en el Data Warehouse.

Luego se almacenan los datos en forma histórica, en las estructuras de estrellas. Se van almacenando cortes de la información en momentos específicos del tiempo, es decir, en las estrellas se van insertando las fotografías de la información en los distintos cortes temporales (por día, semana, mes, etc.).

En el nivel de las bases de datos multidimensionales se almacenan los cubos de datos, se trata de las estructuras de datos que facilitan el análisis OLAP de la información. Los cubos de datos se alimentan directamente de las estrellas. Como ya se ha comentado, se utilizan motores de bases de datos específicos multidimensionales.

Finalmente, en el último nivel se encuentran los usuarios finales que exploran los datos de los cubos multidimensionales y de las estrellas. Las herramientas de usuario final son aplicaciones gráficas que permiten realizar consultas multidimensionales de manera ad-hoc; también se utilizan programas que construyen gráficos a partir de los datos extraídos de los cubos; finalmente, existen herramientas de visualización que presentan los datos en forma tabular y en forma de gráficos.

Recientemente se están construyendo portales para centralizar las consultas del Data Warehouse por medio de Internet y/o de las Intranets de las empresas. Esta modalidad agrega un nivel adicional a la estructura que se ha explicado. Se debe agregar precisamente una etapa o sección que administre las aplicaciones Web: un servidor Web en el que residen las aplicaciones que se utilizan en el Portal, los usuarios interactúan con este servidor Web, el cual asume el papel de intermediario (entre los usuarios y los datos) y el papel de Servidor de Aplicaciones, de esta manera los usuarios no acceden directamente las bases de datos del Data Warehouse sino que lo hacen a través del Portal Web.

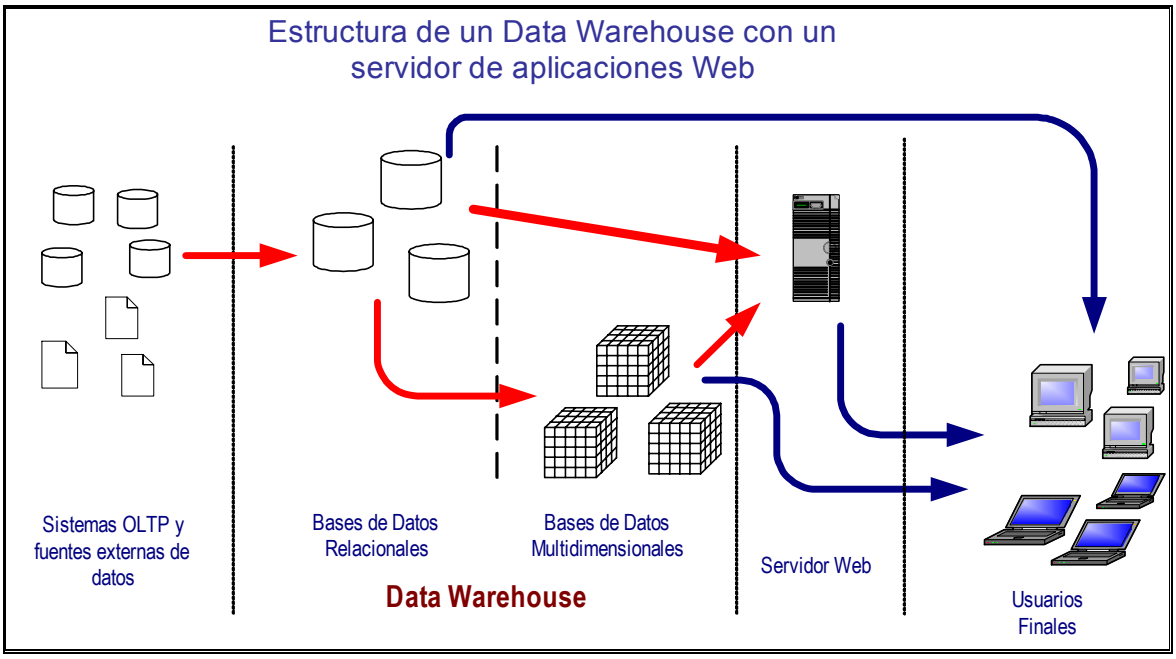
Las principales ventajas de contar con un ambiente que involucre un portal Web son las siguientes:

- Centralización de la información del Data Warehouse. Por medio del portal se genera una sensación de centralización de la información para el usuario. De otra manera se podrían estar utilizando distintas aplicaciones para explotar los datos, por lo que los reportes y gráficos podrían no ser compatibles entre sí, y/o encontrarse en formatos distintos.
- Control de versiones de los reportes. Al tener un portal, se presentarán en él un conjunto de reportes y gráficos pre-definidos, todos los usuarios que vean el portal verán precisamente lo mismo, lo que conduce a información estandarizada. Si se utilizan otras herramientas, cada usuario podría tener su propia versión de los reportes[13].

Finalmente, se puede contar con ambos mundos simultáneamente, pues es recomendable el contar con un Portal Web y con aplicaciones de escritorio. Se pueden utilizar aplicaciones poderosas que explotan la información tanto de los cubos de datos como de las estrellas, así como se puede contar también con un portal que unifique la información en el Web o la Intranet (ver ilustración 4.4.2).

[13] No necesariamente los usuarios verán lo mismo dado que se deben programar esquemas de seguridad y autenticación que provoquen una personalización de la información. Por ejemplo, puede establecerse que el gerente de ventas de autos usados solo vea sus ventas, viceversa con el gerente de autos nuevos. Sin embargo el gerente general debería tener acceso a toda la información.

Ilustración 4.4.2. Arquitectura de un Data Warehouse con un Portal WEB.



En este punto es muy importante considerar el hecho de que si bien las herramientas usuario final son fundamentales para el éxito de un proyecto de Data Warehouse, el esfuerzo mayor se da en las etapas de diseño y construcción de los procesos de extracción y depuración de datos, de hecho se puede afirmar que más de 75% del tiempo del proyecto se consume en esta etapa. Las aplicaciones del usuario final deben tener calidad, estabilidad y deben contar con las funcionalidades que permitan al usuario trabajar cómodamente respondiendo a sus necesidades y expectativas, además deber ser concisas y sencillas. No obstante, estas herramientas son inútiles si no se cuenta con modelos de datos diseñados en forma correcta, que respondan a las características específicas del negocio y que hayan sido implementados con una metodología adecuada.

Obsérvese que la mayor parte de los recursos de un proyecto de DW se consumen en la extracción/depuración de los datos. De una manera simplista y resumida, podemos asignar los recursos (incluyendo el tiempo) de un proyecto de DW de la siguiente manera:

- 75% Programación de la extracción, migración, pruebas con los datos
- 15% Diseño y construcción de modelos de datos: conceptuales, lógicos y físicos; tanto relacionales como multidimensionales.
- 10% Instalación de aplicaciones y capacitación de usuarios y técnicos

Si bien se puede contar con modelos de datos y cubos pre-diseñados (un "sistema enlatado"), viendo estos datos es claro que esto no será de gran ayuda para el proyecto, mucho menos si consideramos que estos modelos siempre se deben modificar para adaptarlos a la realidad particular de la empresa. Dada esta situación, con modelos conceptuales y estructuras de datos pre-construidos (por ejemplo bases de datos relacionales y/o cubos multidimensionales) no

contaríamos ni con el 15% del proyecto. Por lo tanto, lo más recomendable y natural es implementar un sistema de IN desarrollado "a la medida" de la empresa, no un "sistema enlatado"[14].

Más que contar con modelos conceptuales y estructuras de datos pre-construidos, desarrollar un sistema a la medida, o bien, hacer una combinación de ambos, lo que es fundamental en un proyecto de DW es el contar con conocimiento y experiencia de negocios, y contar además con una metodología de trabajo probada y que conduzca a la obtención de resultados (más adelante se hace una revisión de los Factores Críticos de Éxito para un proyecto de DW).

Otro punto fundamental, es la importancia del concepto de la integración de los datos en el DW, ya sea que se integre la información apegándose a la definición de Inmon con un DDO y Data Marts que se derivan de él, o bien apegándose a la de Kimball por medio de su arquitectura Data Warehouse de Bus. En ambos casos existe la integración de la información institucional, la cual se alcanza tras un esfuerzo por implementar una Arquitectura de Datos bien definida para el DW, la cual responde precisamente a las definiciones de DW que se presentaron antes.

Según Timothy Peterson, se puede seguir un desarrollo de sistemas de Data Marts sin una arquitectura definida, sin embargo esto lo que va a generar son islas de datos sin integración alguna. Con tal enfoque se pueden generar pequeños Data Marts en cuestión de unas pocas semanas, pero no se puede generar información con un enfoque empresarial amplio. La única ventaja de este enfoque es que efectivamente se pueden tener Data Marts en producción en períodos de desarrollo de unas semanas [Peterson 2000]. Al hacer esto se construirían lo que podríamos denominar "sistemitas de reportería" que no responden a los objetivos ni a las definiciones de los Data Warehouse, en otras palabras, no estamos ante una estrategia de Data Warehouse ni mucho menos de Inteligencia de Negocios.

El principal problema de este enfoque es la negación del objetivo fundamental de contar con datos empresariales unificados e históricos. Si se requiere contar con información de varios departamentos (por ejemplo), será necesario realizar complejas transformaciones y programación entre los distintos módulos (obsérvese que en este caso no les podemos llamar Data Marts); desafortunadamente mucho del trabajo realizado para los "sistemitas" originales no va a funcionar para estos propósitos [Peterson 2000]. Es importante aclarar que ni el enfoque de Kimball ni el de Inmon tienen relación alguna con una estrategia de trabajo o arquitectura de este tipo (si es que a eso se le puede considerar una arquitectura).

3.6 El problema de la obtención de los datos

Las fuentes básicas de datos para el DW son, como se ha mencionado, los sistemas OLTP. Uno de los objetivos de un DW es la integración de la información de diversos sistemas en un repositorio único y completo. Por lo tanto, se debe seguir una estrategia de extracción y traslado de datos desde los OLTP hacia los sistemas OLAP. Los datos que se requieren del OLTP son un

[14] Esta situación se confirma cuando la empresa cuenta con un sistema transaccional "enlatado", por ejemplo un ERP. En este caso se puede implementar un sistema de IN "enlatado". Esta aparente contradicción más bien concuerda con lo recomendado, ya que el sistema de IN estará construido a la medida de la empresa, o sea, a la medida del sistema transaccional.

subconjunto de su contenido. Además en el DW no se almacenan detalles sino resúmenes, por lo tanto la información que se extrae del OLTP es un subconjunto-agregado de sus datos.

Los datos por almacenar en el DW responden a requerimientos de información específicos, con una orientación estratégica e histórica, y tales que es posible que no se hayan considerado para los sistemas transaccionales. No obstante, muchos de estos datos "no interesantes" en el mundo transaccional serán requeridos en el DW. Por lo tanto, es muy normal que la información disponible en el sistema fuente esté incompleta, sea inconsistente o incluso inexistente, por lo que al tratar de generar datos para un DW nos encontramos con limitaciones y restricciones importantes en la disponibilidad de los datos fuente. Por ello la información del DW no solo proviene de los sistemas transaccionales, sino de la generación de datos a partir de otras fuentes tales como archivos de texto, hojas electrónicas, archivos XML, etc.

Finalmente, una vez que se han establecido las fuentes de datos, se procede a construir los procesos de extracción-traslado de datos al DW. Han y Kamber organizan estos procesos en un conjunto de técnicas de extracción de datos [Han 2001]:

1. **Limpieza de los datos**: En los sistemas actuales de producción, los datos tienden a estar incompletos, tener ruido e inconsistencias. Por medio de procesos de "limpieza de datos" se completan los datos faltantes, se suavizan los datos extremos (es decir, se reduce el ruido) y se pueden corregir inconsistencias de la información[15].

2. **Integración de datos**: Por medio de procesos de integración se va a unificar datos de distintas fuentes. Por ejemplo, es posible que existan distintos catálogos de clientes en distintos sistemas de la empresa, en los cuales hay clientes repetidos y clientes que residen en solo un sistema. En el DW se requiere unificar esta información en un catálogo único, que se construirá por medio de procesos de Integración. La integración se refiere no solo a la unificación de los datos de distintas fuentes, sino a su homogenización, dado que en distintas fuentes se van a obtener distintos tipos de datos, de llaves y en general, de características distintas.

3. **Transformación de datos**: La información extraída no siempre se encuentra en el estado requerido por el DW, por lo que se aplican técnicas de transformación para llevar los datos a las definiciones que se requieren. Las transformaciones típicas son:
    a. *Suavizado de datos extremos*: se utiliza para reducir ruido de los datos, al existir datos con valores exagerados que se saben erróneos, se aplica un valor 'default' (por omisión), un promedio, un indicador, etc., con el fin de corregir el error.
    b. *Agregación de información*: aplica básicamente a datos cuantitativos. Como se ha mencionado, el DW presenta la información en estado resumido; los datos fuente se encuentran en estado detallado por lo que se deben agrupar. Esto reduce la cantidad de registros de las tablas del DW.
    c. *Generalización*: se trata básicamente de modificaciones a los atributos. Por ejemplo, en el DW no interesa la hora y minuto de las transacciones de las tarjetas

[15] Si hay mucho problema de datos 'sucios' es probable que los sistemas transaccionales sean deficientes (poca previsión por revisar la integridad de los datos) o que una cultura de la 'calidad del dato' requiera ser instaurada en la organización. Para conocer más sobre calidad de los datos véanse [Muñoz 2000] y [Quirós 2000].

de crédito, sino su agrupación por día, semana y/o mes. Estos se lleva a cabo en los procesos de extracción aplicando rutinas de "generalización de los datos"

d. *Construcción de atributos (o datos derivados):* algunos datos no existen en las fuentes, pero se pueden obtener al aplicar algunos algoritmos y/o reglas de negocios. Por ejemplo, se puede determinar si una inversión es nueva si su fecha coincide con la fecha de hoy, con un algoritmo como este se construye una dimensión con los valores: "Inversión nueva / Inversión vieja".
e. *Normalización de información*: se pueden "organizar" los datos cuantitativos en rangos: de 1 a 10, de 11 a 20, etc. Con esto se pueden convertir datos cuantitativos en categorías descriptivas.

4. **Reducción de datos**: En la mayoría de las cargas de datos hacia el DW no se requiere toda la información de las fuentes, sino que se carga al DW un subconjunto de ella. Para mejorar el rendimiento y reducir la complejidad de las cargas de datos es buena idea el simplificarlos. Por ejemplo, de una tabla histórica con varios cientos de miles de registros se requieren tan solo algunas columnas y los registros del último mes, por lo que es más sencillo y eficiente cargar los datos de un query de dicha tabla filtrando por fechas y escogiendo las columnas requeridas (que realizar una carga de toda la tabla).

Los DW reciben cargas de datos periódicas (diarias, semanales y/o mensuales), por lo que es mejor considerar los procesos de reducir el volumen de datos que se van a cargar.

En algunos casos puede ser útil el pre-proceso de los datos llevándolos a tablas temporales y/o vistas. La idea es separar los procesos de cargas en varios procesos sencillos, eficientes y fáciles de administrar en vez de tener un proceso grande y complejo.

Existen herramientas y tecnologías para la construcción y calendarización de procesos de carga de datos. En general se utilizan combinaciones de herramientas y lenguajes de programación, herramientas gráficas y lenguajes propios de las bases de datos. Una vez que los procesos se encuentran estabilizados, éstos se pueden calendarizar para que se vayan ejecutando periódicamente y de esta manera se vaya poblando de información el DW.

La construcción de los procesos de carga es la labor más ardua y compleja del proyecto de DW. Es aquí donde se consumen la mayoría de los recursos. Es además, la etapa más delicada por lo que se merece precisamente este uso intensivo de recursos. Estamos hablando ni más ni menos que de la generación de los datos del DW, si ésta está mal al final todo estará mal.

Si la extracción de datos no está correctamente especificada ni bien definida, los procesos de carga alimentarán el DW con datos de dudosa calidad, por lo que una previa definición de variables de extracción es requerida.

3.7 Aspectos Metodológicos

El análisis o discusión de metodologías de desarrollo e implementación para sistemas de IN está fuera del alcance del presente informe, pero haremos referencia a algunos Factores Críticos de Éxito que se consideran fundamentales. También recomendaremos aspectos y/o elementos que se deberían incluir en cualquier metodología de trabajo para sistemas de Inteligencia de

Negocios.

3.7.1 Etapas de un proyecto de DW: El Ciclo de Vida del DW

Una de las metodologías más conocidas para el desarrollo de sistemas de DW es la denominada "El Ciclo de Vida Dimensional de los Negocios"[16] desarrollado por el Ralph Kimball y un conjunto de consultores (en adelante "ciclo de vida"). Este ciclo de vida es producto de la investigación e implementación a lo largo de varios años y a lo largo de muchos proyectos de DW; una de sus principales fortalezas es haberlo puesto en práctica muchas veces y recibir la retroalimentación que brinda la experiencia.

Muchas de las propuestas de metodología y/o estrategias para el desarrollo de sistemas de DW se basan precisamente en el Ciclo de Vida.

A continuación se hace un resumen de los elementos que conforman el Ciclo de Vida, los cuales pueden ejecutarse tanto en forma secuencial como paralela y frecuentemente en forma iterativa; estos componentes son complementos lógicos que consideran la mayoría de variables y situaciones importantes del Datawarehousing.

El Ciclo de Vida está compuesto por las siguientes etapas básicas [Kimball 1998]:

1. **Planeación del proyecto**: La primera etapa del DW es la planeación del proyecto. En esta etapa se define el DW, se establecen sus limitaciones y su alcance. Se establece un objetivo global por perseguir y constituye la columna vertebral del proyecto. La planeación es dependiente de los requerimientos de negocios para el DW.

2. **Administración del proyecto**: Por medio de la administración se le da seguimiento al proyecto, se vigila la aplicación de los elementos del ciclo de vida y se toman decisiones en forma apropiada. Como en cualquier otro tipo de proyecto es requerida su administración en forma centralizada. Sin este componente es imposible corregir errores, modificar tiempos metas, etc., y el proyecto estaría simplemente "a la deriva".

3. **Definición de requerimientos de negocios**: El DW responde a una necesidad de información particular por parte de los tomadores de decisiones de alto nivel. Por lo tanto, se debe tener un entendimiento muy claro y profundo de los aspectos de los negocios.
   Los desarrolladores de DW deben estar en capacidad de realizar una interpretación correcta de la información obtenida de sus usuarios, para poder trasladar estas ideas en términos de requerimientos para el DW. Muchísimos proyectos terminan sin brindar a sus usuarios la información que éstos realmente necesitan para tomar sus decisiones; en este tipo de proyectos pueden darse grandes desperdicios de tiempo y recursos, dado que el sistema resultante simplemente no es utilizado.
4. **Modelamiento Multidimensional**: Por medio de la definición de los requerimientos de negocio se determina qué datos e información se van a trasladar y transformar para el DW. En la etapa del Modelamiento Multidimensional se convierten los requerimientos

[16] Traducción libre para *"The Business Dimensional Lifecycle"*.

en medidas y dimensiones, se generan Modelos Conceptuales y Lógicos del sistema.

5. **Diseño físico**: Es la construcción física de las bases de datos a partir de los Modelos Lógicos desarrollados en la etapa anterior. A partir de las estrellas luego se van a construir los Cubos Multidimensionales.

6. **Preparación de datos, diseño, desarrollo e implementación**: En esta etapa se realiza la extracción y transformación de datos hacia el DW. Igualmente se llevan a cabo pruebas y ajustes con la información extraída. Esta es una etapa de fundamental importancia y es además iterativa, normalmente se van realizando cargas y ajustes hasta llegar a una versión final de los procesos de extracción.

7. **Diseño de la arquitectura técnica**: Se establece la arquitectura y estructura del sistema por desarrollar. Se determina qué sistemas operativos se van a utilizar, se toman decisiones en materia de Bases de Datos, aplicaciones de usuario final (aquí se decide si se utiliza o no un portal para el DW), etc. También se toman las primeras decisiones sobre la seguridad en cuanto al acceso a los datos de DW.

8. **Selección e instalación de productos**: Una vez que se ha determinado la Arquitectura tecnológica que va utilizar, se procede a realizar la instalación de la tecnología adquirida.

9. **Aplicaciones de usuario final**: Las aplicaciones de usuario final se deben especificar a partir de las necesidades y requerimientos de los usuarios. Se deben definir aplicaciones para al menos dos tipos de usuarios: los que realizan reportes ad-hoc (dinámicos y modificables) y los que utilizan reportes pre-diseñados (fijos, no cambian).
   Una vez que se han evaluado aplicaciones y se ha decidido cuáles se van a utilizar, a medida que se van liberando los distintos Data Marts, se van instalando las aplicaciones en los equipos de los usuarios. En una ambiente tipo Web se van publicando y habilitando nuevas opciones en el portal de manera que los resultados quedan disponibles para los usuarios.
   Las aplicaciones se pueden desarrollar para el proyecto, o bien se pueden adquirir aplicaciones ya construidas.
   En la sección 5.3. Visualización de la información, página 47, se trata el tema de las aplicaciones de usuario final con mayor profundidad.

10. **Mantenimiento y crecimiento**: Los DW tienen una naturaleza dinámica, pues evolucionan al ritmo de los negocios de la empresa. Si un DW cambia, esto es una buena señal.
    El mantenimiento se refiere a tareas post-implementación, es la administración del sistema. El crecimiento es la capacidad del sistema de recibir modificaciones, ajustes, nuevos requerimientos, etc.

11. **Capacitación y divulgación de resultados:** Finalmente debemos agregar una etapa de capacitación y divulgación de resultados. Se deben brindar capacitaciones formales, tanto a técnicos como a usuarios, en distintas etapas del proyecto. Además de esto, es imprescindible que a lo largo del proyecto se dé un proceso de "evangelización" en materia de Business Intelligence, Data Warehouse, pensamiento multidimensional, etc. Esto se debe dar en todos los niveles de usuarios: técnicos, usuarios expertos y usuarios pasivos. De hecho, como se verá a continuación, esta evangelización es un factor crítico de éxito.
    Finalmente, los buenos resultados se deben "mostrar" y promover en la empresa, por medio de una Intranet, publicaciones internas, correos electrónicos, etc. Es decir, el proyecto se debe vender internamente.

3.7.2 Factores Críticos de Éxito

Un Factor Crítico de Éxito (FCE) es una situación, característica y/o condición, que se considera necesaria mas no suficiente para el logro de las metas y objetivos del proyecto.

Los siguientes son FCE comunes para el desarrollo de Data Marts, se pueden aplicar a lo largo de todo el proyecto. Deben ser evaluados y atendidos constantemente durante el desarrollo del DW.

1. **No se puede trabajar sin una metodología**. Quienes van a desarrollar el sistema deben contar con una metodología de trabajo. Esto debería ser obvio pero desgraciadamente en nuestro medio existe mucha improvisación y deseos de lucro irracionales que llevan a consultores externos a prometer resultados imposibles. No pocas veces las empresas de desarrollo ni siquiera cuentan con una metodología de trabajo.
    Tanto si el desarrollo se hace a lo interno de la empresa o si se va realizar por medio de un proyecto de 'outsourcing', este punto es fundamental. Se debe evaluar con mucha seriedad la metodología del desarrollador, revisando sus documentos, presentaciones, etc. Si la metodología no existe, es inadecuada o no se ha probado en casos de éxito, nos encontramos ante un pronóstico de fracaso, por lo tanto el consultor debe ser automáticamente descalificado[17].

2. **Utilización de una metodología de trabajo consolidada**. La metodología de trabajo debe haber sido aplicada en otros proyectos y en forma exitosa. Es decir, debe ser una metodología probada. Dado que el Datawarehousing no es cosa nueva, existe mucha experiencia y teoría desarrollada, de manera que no tiene mucho sentido el improvisar o inventar lo ya inventado (sin dejar de lado la creatividad, por supuesto).

3. **Evaluación de los desarrolladores.** Si el desarrollo se realiza por medio de un outsourcing, la empresa adquiriente debe contar con un método de evaluación y control del consultor. En muchas ocasiones es tan responsable de la mala praxis el ejecutor como el cliente. En muchos proyectos fracasados, con tan solo un adecuado análisis previo del consultor, se pudieron evitar muchos fracasos y la pérdida de miles de dólares.
    Esta evaluación debería incluir la verificación profunda de las referencias del oferente del

[17] La metodología es importante, pero lo es más que el consultor demuestre contar con experiencia, comprensión de las necesidades del negocio y los gerentes involucrados, y mucho sentido común.

outsourcing, el análisis de la metodología, la evaluación de la tecnología propuesta, etc.

4. **Evaluación de opciones de desarrollo.** Existen muchos productos 'enlatados'. Teóricamente, estos productos tienen ventajas: se supone que los análisis de negocios ya se han llevado a cabo, las estructuras de bases de datos están listas, las aplicaciones de usuario final están ya programadas, ya se cuenta con un conjunto de reportes listos, etc. Todo esto suena muy tentador, pero estos sistemas tienen un alto costo y no siempre son lo que aparentan.
   El producto enlatado debe ser evaluado a la luz de los requerimientos específicos de la empresa. Es muy difícil que dos compañías de una misma industria tengan una estructura de análisis de negocios idéntica. No existe algo así como "Estándares de Data Warehouse" por industria (DW de Banca, DW Comercial, DW de producción, etc.). La cosa no es tan sencilla como suponer que el DW del Banco X va a funcionar en el Banco Y.
   Otro tanto es cierto con respecto de los reportes ya construidos: primero que nada para que sean de utilidad, los reportes deben de coincidir con las necesidades 'psicológicas' y técnicas de los gerentes de la empresa. Además es una característica de los DW el que las herramientas de usuario final permiten construir reportes ad-hoc en forma muy sencilla, dinámica y rápida. Además, esto es labor de los usuarios, por lo que proveer reportes preconstruidos es de un valor agregado mínimo y tiende a distraer la atención de los verdaderos valores agregados de la IN.
   Si se sigue una estrategia de desarrollo como la del Data Warehouse Bus de Kimball, en pocos meses se van a ir desarrollando y liberando Data Marts, los cuales se construyen a la medida de las necesidades y problemas específicos de la empresa. No obstante, si se deciden por un enlatado, este debe ser evaluado minuciosamente, lo mismo que la empresa que lo va a implementar.
   Otra ventaja de un enlatado es que se supone que "se puede instalar rápido". Pero la realidad es que no pocas veces la adaptación del producto a la realidad de la empresa es mucho más lenta que el desarrollo de un sistema nuevo y a la medida. En este caso es recomendable hacer un análisis y una valoración de los riesgos y costos.

5. **Confiabilidad de las especificaciones del sistema.** Deben existir reglas de negocio, fórmulas, mapeos de orígenes de datos, etc., que sean correctos, claros y concisos para que la extracción de datos sea confiable. Además, estos mapeos deben tener coherencia con la semántica de los datos y con el Modelo Conceptual del sistema.
   Es a partir de las especificaciones que se construirán los procesos de extracción de los datos, y se realizarán las pruebas técnicas sobre ellos. Por lo tanto, si las especificaciones son incorrectas o imprecisas y/o si el Modelo Conceptual es impreciso, el desarrollo de los programas de carga se volverá laborioso y muy ineficiente.
   Además, dado que las pruebas se realizan a la luz de las especificaciones técnicas y funcionales, las primeras serían inválidas si estas especificaciones son inadecuadas.

6. **Datos básicos confiables.** Los datos de las fuentes primarias de información deben ser correctos y consistentes. Un sistema de Inteligencia de Negocios refleja la situación de los datos de la empresa, el sistema presenta una imagen de los datos de los sistemas transaccionales de la compañía. Por lo tanto, aunque las especificaciones de extracción

de datos sean las adecuadas, si los datos que provienen de la fuente tienen errores y/o inconsistencias, el sistema final reflejará esta situación.

7. **Disponibilidad de personal técnico y de negocios para el proyecto.** El contar con estos recursos es fundamental para mantener los cronogramas al día y obtener resultados de calidad. El personal funcional (conocedor del negocio) debe formar parte del equipo de trabajo, aportando ideas, requerimientos, aclaraciones, etc. Este personal también debe aprobar los diseños finales de los módulos.

8. **Disponibilidad de infraestructura tecnológica.** Esto incluye hardware, software e infraestructura física. Sin esta infraestructura es imposible avanzar en el desarrollo del proyecto.

9. **La evangelización lleva a generar expectativas adecuadas de los usuarios**: Si los usuarios tienen expectativas distintas de lo que ofrece un sistema de IN, ellos simplemente no estarán satisfechos con los resultados del desarrollo. Por lo tanto, a lo largo del proyecto debe darse un proceso de 'evangelización' en materia de Inteligencia de Negocios: pensamiento multidimensional, conceptos de OLAP versus OLTP, etc. Esto se aplica tanto para el personal técnico como para el funcional.

10. **Pertinencia y relevancia de los datos.** Cuando hablamos de **pertinencia** nos referimos a que las variables cuantitativas y cualitativas que forman parte del modelo conceptual del sistema deben responder a las características de un sistema de IN, es decir, deben ser variables que permitan la generación de reportes históricos de carácter gerencial, en contraposición a la información transaccional.
    Al hacer referencia al concepto de **relevancia**, nos referimos a que las variables del sistema, además de ser pertinentes, deben ser intrínsecamente útiles y necesarias para los análisis que los usuarios deseen realizar. Existen muchas variables que si bien no son transaccionales, tampoco son relevantes.
    El peligro de incluir variables no pertinentes y/o irrelevantes es que se puede incurrir en problemas de bajo rendimiento del sistema, innecesarios y altos volúmenes de almacenamiento en discos duros, dificultades en el mantenimiento y administración del sistema, alargar mucho el tiempo de desarrollo y pruebas. Todo esto se traduce en altos costos que representan un verdadero peligro para el éxito del proyecto. Un adecuado estudio de requerimientos de información es imprescindible para asegurar que se incluyen todas las variables relevantes y pertinentes en el sistema.

11. **La experiencia en IN es fundamental:** La IN es un área de especialización en la TI, existe una gran acumulación de conocimientos, experiencias y técnicas específicos para desarrollar los sistemas de IN. Por lo tanto, no se debe improvisar – ni mucho menos inventar – lo ya existente; se debe más bien buscar y aprovechar recursos especializados, con conocimientos y experiencia en esta área (como se haría con cualquier otra área de especialización).

4 Sistemas de Minería de Datos

4.1 Justificación de la Minería de Datos

Los sistemas de DW y la tecnología OLAP nos muestran la información tal y como ésta se ha generado en los procesos de negocios de la empresa. Es decir, nos muestran las fotografías "crudas" de los resultados finales de los procesos de negocios: ventas, costos, producción, saldos de operaciones crediticias, captación, etc. Tales son los tipos de datos que se analizarán en los cubos de datos multidimensionales.

A partir de un sistema OLAP no podemos determinar las causas y/o efectos de acontecimientos específicos sobre los datos, por qué se comportan de una manera específica y/o cuál podría ser su resultado ante un escenario distinto. Con un DW, la experiencia y conocimiento de negocios de un usuario experto lo lleva a realizar análisis de sus medidas utilizando combinaciones de dimensiones tales que él sabe que tienen un alto grado de correlación entre sí. El conocimiento del usuario le indica cuáles dimensiones y en qué circunstancias tiene sentido mezclarlas entre sí y con cuáles medidas. Luego, el usuario puede experimentar con otras variables que quizá nunca hubiera utilizado pero, dado que el DW las pone a su disposición, tiene la oportunidad de hacerlo[18], incluso él puede "jugar" con algunos cálculos estadísticos básicos y con todo un instrumental gráfico para la presentación de los datos (todo esto dentro de la estructura de DW que se haya implementado). El tiempo de uso y la experimentación permiten sacar el mayor provecho y valor agregado al sistema de DW implementado.

No obstante, siempre van a existir combinaciones posibles de medidas y dimensiones a las que nunca se nos ocurriría consultar; se trata de aquellas combinaciones que no son obvias, incluso para el conocedor, pero que pueden ser muy importantes. ¿Podríamos utilizar técnicas de investigación de datos que vayan un poco mas allá, que nos presenten un panorama "mas allá de lo obvio" o no tan espontáneo? Si bien el conocimiento y experiencia en el negocio son fundamentales e imprescindibles, por qué no brindarle al conocedor herramientas de investigación que permitan inferir conocimientos más profundos y no tan obvios (¿por qué no llevar el conocimiento a niveles de mayor profundidad analítica?).

Se trata precisamente del estudio de la correlación entre variables, del análisis de la asociación de comportamientos, del impacto que tienen variables exógenas al negocio sobre sus resultados; en fin, de información que se encuentra en otro nivel de procesamiento fuera del mundo del OLAP. Precisamente para este tipo de situación surge la opción de la Minería de Datos, con todo un conjunto de tecnologías dirigidas al "Descubrimiento del Conocimiento"[19].

4.2 Qué es la Minería de Datos

En general, en un proceso de Minería de Datos (MD) se trata de llegar al descubrimiento de información valiosa a partir de grandes volúmenes de datos. Esto se realiza por medio de distintas técnicas en las que se aplican algoritmos de diversa naturaleza. El proceso de la minería

[18] Una de las ventajas usuales de un DW es la de contar con un conjunto de información amplia, que va mas allá de las expectativas comunes del usuario.
[19] Hay autores que utilizan el término "Knowledge Discovery" como sinónimo de "Minería de Datos".

es un paso o etapa hacia el descubrimiento del conocimiento, de aquí que muchos autores se refieren a este proceso completo como Knowledge Discovery. Sin embargo en la industria cada vez se utiliza más el término Data Mining para referenciar a todo el proceso. Para Jiawei Han [Han 2001], el proceso de descubrimiento del conocimiento involucra estas etapas (obsérvese que los puntos del 1 al 4 forman parte del desarrollo del DW):

1. **Depuración y/o limpieza de los datos.** Para eliminar ruido e información inconsistente.
2. **Integración de la información.** Cuando se combinen múltiples fuentes de datos.
3. **Selección de los datos por utilizar.** Se deben seleccionar datos que sean relevantes para los análisis que se van a ejecutar.
4. **Transformación de los datos.** Procesos de traslado, extracción y resumen de los datos.
5. **Minería sobre los datos.** Proceso en el que se aplican a los datos diversas técnicas estadísticas, matemáticas, etc.
6. **Evaluación de patrones.** Es el análisis de los resultados del proceso anterior.
7. **Presentación del conocimiento.** Se utilizan herramientas de visualización de datos para presentar los resultados alcanzados.

Siguiendo esta concepción, una arquitectura de un sistema de MD, como mínimo, estará compuesta de:

- Bases de datos, un Data Warehouse u otro repositorio de datos
- Un servidor de bases de datos
- Una base de conocimiento
- Un motor de minería
- Un módulo de evaluación de patrones y/o resultados
- Herramientas de visualización de resultados

Podemos ver la MD como una consecuencia 'natural' de los sistemas de DW y los motores de Bases de Datos OLAP. No obstante para desarrollar un sistema de MD no es estrictamente necesario el contar con un DW. Es posible incluso que un DW no ayude en un proceso de MD dado que la información de carácter resumido y/o agregado almacenada en las estructuras del DW puede ser distinta a la requerida para un análisis de MD. Para que el DW sea la base de datos del sistema de MD es requerido que en el diseño y desarrollo del primero se hayan considerado requerimientos de datos para el segundo.

Los límites y la definición misma de la MD son difusos, sobre todo por el hecho de se que trata de un área de investigación en constante evolución, en donde aparecen nuevos avances (algunos autores los catalogan como MD y otros cómo "otras cosas"). También confunde el hecho de que nos encontramos con tecnologías y algoritmos de muchas áreas de investigación distintas. Muchas veces algún fabricante de software desarrolla un producto que bien se podría clasificar dentro de la MD, no obstante éste decide "bautizar" su tecnología con otro nombre para efectos de mercadeo. Otro problema es el nombre mismo de ésta área, ya que hay autores que términos como: Descubrimiento del conocimiento (*Knowlegde Discovery*), Minería del Conocimiento (*Knowlegde Mining*), entre otros. Incluso a veces con diferentes acepciones.

4.3 Tipos de análisis

Existen varios tipos de análisis o estudios que se pueden llevar a cabo en MD, para problemas de diversa naturaleza con distintos requerimientos de información. Para realizar dichos estudios en MD se utilizan algoritmos o métodos distintos, implementados por medio de software; se cuenta incluso con algoritmos distintos para realizar análisis de un mismo tipo.

En este apartado vamos a enumerar y explicar dichos tipos de estudios, en la sección "4.4 Algoritmos de Minería de Datos", página 33, se analizarán algunos de los algoritmos por medio de los cuales se responde a estas necesidades de información[20].

4.3.1 Clasificación (Aprendizaje Supervisado)

Una forma natural de comprender para el cerebro humano es la clasificación de los objetos según sus categorías. Esto se puede hacer por medio de estudio de alguna(s) variable(s) o atributo(s) de interés en el objeto en cuestión, al observar el valor de dichos atributos se clasifica el objeto en una forma excluyente. El valor de esta variable es conocido, por este motivo en MD a este tipo de estudio se le llama "**Aprendizaje Supervisado**".

Un ejemplo puede aclarar el concepto:

En un banco interesa determinar si los clientes de tarjeta de crédito tienden a "pagar al día" o si bien acostumbran a "retrasar su pago". El banco cuenta con la información histórica de las transacciones de los clientes, en otras palabras, el banco sabe quiénes tienden a retrasar sus pagos y quiénes no. Con esta situación, se puede plantear la "cantidad de retrasos en el año" como la variable objetivo del estudio. No obstante el banco no sabe por qué se da o qué explica este comportamiento, precisamente esto es lo que se desea determinar por medio del estudio de Clasificación. Con este objetivo, se puede construir una base de datos con la información histórica de los clientes que incluya un conjunto de variables que puedan estar relacionadas con este comportamiento[21]; las siguientes son algunas de las que se podrían investigar:

- *Retrasos en el año* Variable Objetivo
- *Edad* Variable funcional o Atributo
- *Estado Civil* Variable funcional o Atributo
- *Cantidad de dependientes* Variable funcional o Atributo
- *Monto o límite del crédito* Variable funcional o Atributo
- *Utilización del crédito* Variable funcional o Atributo
- *Profesión* Variable funcional o Atributo
- *Salario Mensual* Variable funcional o Atributo
- *Cuota mensual* Variable funcional o Atributo

[20] El objetivo aquí es caracterizar la tecnología de la MD. No se pretende construir un manual de MD o realizar un estudio técnico, para lo cual existe literatura especializada (se recomienda revisar la bibliografía).
[21] Igualmente, por medio de análisis de covariancias, regresiones, econometría, etc., se puede determinar el grado de correlación entre las variables.

La variable objetivo es "Retrasos en el año", se define un criterio tal que con 2 o más atrasos se dice que el cliente "tiende a retrasarse" mientras que con 1 o menos "no tiende a retrasarse". Para los créditos actuales esta información es conocida, lo que interesa es determinar si existe un patrón(es) en las otras variables que puedan determinar o explicar si un cliente es propenso a retrasarse o si no lo es. Con esta información el banco va a contar con un "perfil" de sus clientes que puede utilizar en el futuro en una campaña de reducción de la morosidad.

En este tipo de análisis se requiere de conocimiento sobre el negocio para determinar cuáles variables se pueden relacionar. Igualmente se pueden aplicar diversas técnicas estadísticas para encontrar las correlaciones entre las variables y excluir aquellas que no tienen relaciones causales entre sí (tales que su relación es más bien aleatoria)[22].

4.3.2 Predicción

En predicción se busca proyectar el valor o rango de valores que una muestra de datos podría alcanzar, dadas sus características. Una de las técnicas más utilizadas para la predicción son los Análisis de Regresión. Obsérvese que por medio de estudios de clasificación se pueden predecir valores discretos mientras que con regresiones se predicen valores continuos, o al menos "muy densos".

En la Economía y Estadística se utilizan distintos tipos de estudios de Predicción, como la Econometría y/o el Análisis de Series de Tiempo.

La econometría se basa en complejos sistemas de regresiones con diversos modelos matemáticos. En los estudios econométricos se contrastan variables explicativas y variables objetivo a partir de series de datos históricos, se aplican además un conjunto de técnicas para eliminar ruido y/o datos extremos; finalmente, el objetivo, más que predecir resultados futuros en algunas variables, es explicar su comportamiento y las relaciones de causalidad con otras variables. Con los resultados de los modelos econométricos se cuenta con información para plantear política económica, o bien, para defender teorías económicas.

El análisis de series de tiempo permite estudiar tendencias en los datos, como ciclos, estacionalidad, comportamiento atípico, etc. Se han desarrollado técnicas matemáticas avanzadas, que permiten realizar diversos tipos de análisis que complementan la intuición humana o permiten visualizar el comportamiento de grandes volúmenes de datos[23]. Entre las aplicaciones recientes en nuestro medio se encuentran:

- Detección de fraudes en tarjetas de crédito y cuentas de ahorro (análisis de comportamiento atípico)
- Predicción del efectivo en sucursales y cajeros automáticos (análisis de series de tiempo).

[22] Si bien la correlación no implica causalidad.

[23] En Costa Rica, la Escuela de Matemática de la Universidad de Costa Rica ha investigado estos temas y otros relacionados (ver adelante) por casi 20 años. Recientemente un grupo de académicos estableció una empresa de software y modelaje matemático especializada en estos temas: Predisoft.

4.3.3 Análisis de Agrupamiento[24] (Aprendizaje no Supervisado)

En los estudios de Agrupamiento, a diferencia de la Clasificación, no se a conoce priori el grupo o categoría a la que pertenece el objeto de estudio. Precisamente el objetivo de estos análisis es encontrar o determinar a qué grupo pertenecen nuestros objetos de estudio. En muchos casos existen agrupamientos en los datos que no se conocen previos al estudio, entonces, en el análisis primero se debe determinar o definir los grupos y luego clasificar a nuestros objetos dentro de ellos. En otros casos conocemos de alguna manera los agrupamientos pero no sabemos "a cuál pertenece" nuestro objeto.

Una diferencia importante con el análisis de Clasificación, es que en los estudios de Agrupamiento no se persigue explicar las relaciones de causalidad entre los atributos (variables de entrada vs. variables objetivo), es decir no se explica el por qué pertenece el objeto x al grupo y, simplemente éste se agrupa.

Dado que no sabemos a cuáles grupos pertenecen los objetos de estudio en MD, a los estudios de Agrupamiento se les conoce como **"Aprendizaje no supervisado"**.

Un conjunto de objetos físicos o abstractos pertenecen a un grupo en particular porque son similares entre sí, al tiempo que son distintos de los de los otros grupos. De esta manera, en muchos contextos los grupos pueden ser tratados como objetos independientes. Es decir, que un grupo puede ser precisamente el objeto de estudios posteriores que expliquen sus conductas y/o características.

Las aplicaciones de los estudios de Agrupamiento son muy amplias. En mercadeo y/o CRM se pueden utilizar para distinguir características de los clientes con respecto de sus patrones de consumo. En demografía se pueden agrupar los distintos tipos de casas en los vecindarios, según sus clases sociales, precios, ubicaciones, etc. En banca se pueden agrupar los clientes según sus características de clase social, estado civil, edades, etc., de manera que se puedan realizar políticas de mercadeo financiero personalizadas.

4.3.4 Asociación (Market Basket Analysis)

La Asociación o *Market Basket Analysis* se aplica a las compras de productos y/o servicios que se realizan en las tiendas y/o supermercados. Lo que se busca es encontrar patrones de asociación entre productos comprados. Interesa determinar asociaciones tales como:

- "Cuando un cliente compra un producto A entonces éste tiende a comprar también el producto B en X% de las veces" [Groth 2000].

El Market Basket Analysis tiene muchas aplicaciones: mercadeo cruzado de productos, diseño de catálogos, determinación de precios de productos y promociones, incluso la manera de acomodar los artículos en supermercados puede ser sugerida por un estudio de este tipo.

[24] En Estadística son muy conocidos estos estudios por su nombre en inglés: ***Clustering Analisys***

Un ejemplo de asociación: supongamos que en un supermercado se ha observado que cuando los clientes compran frijoles, el 25% de las veces también compran cerveza.

Este comportamiento puede ser simplemente aleatorio si en el 25% de las ventas del supermercado se vende siempre cerveza, con o sin frijoles.

Pero si la situación fuera otra, digamos que solo en el 10% de las ventas del supermercado se vende cerveza, mientras que la venta de cerveza aumenta a un 25% cuando se incluyen los frijoles. En este caso se observa una alta asociación o atracción de los productos. Esta información se puede utilizar para mejorar las ventas de cerveza o de frijoles por medio de alguna oferta o simplemente colocando los frijoles junto a la cerveza.

Ahora bien, si la relación fuera que en la totalidad de las ventas se incluye la cerveza en el 60% de las veces, pero cuando se compran frijoles solo se compra cerveza en un 25% de las ocasiones, en este caso se dice que los productos se repelen entre sí. Si este es el caso, se deben aplicar políticas de mercadeo y/o ventas distintas.

4.4 Algoritmos de Minería de Datos

En MD se integran técnicas, algoritmos y métodos de estudio de diversas áreas de especialización; para efectos prácticos llamaremos a todos estos tipos de análisis **Algoritmos de la Minería de Datos.** Se trata básicamente de: redes neuronales, aprendizaje mecánico (*machine learning)*, visualización de datos, reconocimiento de patrones, métodos de la estadística inferencial y descriptiva, métodos matemáticos, etc. Cada uno de estos algoritmos se aplica a distintos tipos de estudio. A continuación se hace una breve referencia a algunas de estas técnicas y en qué casos y/o circunstancias se pueden utilizar.

4.4.1 Clasificación por medio de Árboles de Decisión

Un árbol de decisión se representa por medio de un gráfico compuesto de nodos, ramas y hojas: los nodos representan pruebas (preguntas) en atributos específicos y las ramas representan las respuestas a dichas pruebas, finalmente las hojas son las categorías finales o distribución de los datos.

Por medio de un árbol de decisión se puede mapear un conjunto de atributos para un objeto en particular, con respecto de un resultado específico. Un ejemplo puede aclarar estas ideas (ver ilustración 5.4.1.1):

En una tienda de ropa interesa determinar cuáles de sus clientes son propensos a comprar trajes enteros de caballero. Ésta es la pregunta que se desea responder para cada cliente: "¿Es propenso a comprar trajes enteros?" Para esta pregunta el conjunto de respuestas es: { sí, no }

Con la información de los clientes se van haciendo un conjunto de pruebas, y para cada resultado de éstos se examina el subconjunto de clientes que cumplen con él y se determina si estos

cumplen o no con la pregunta "¿compra trajes enteros?". Cada prueba es un nodo dentro de árbol.

Empezamos por el sexo y se determina que tanto hombres como mujeres compran trajes. Por lo tanto la prueba no es determinante.

Dentro del grupo de las mujeres, se pregunta por el estado civil y se descubre que las casadas tienden a comprar trajes mientras que las solteras no. Con esto se construyen las primeras hojas del árbol:

- Sexo: Femenino, Estado civil: Casada
  Resultado: Si compran trajes enteros

- Sexo: Femenino, Estado civil: Soltera
  Resultado: No compran trajes enteros

Dentro de los hombres se pregunta por las edades y se determina que todos los hombres entre 31 y 40 años compran trajes, esto nos lleva a determinar otra hoja:

- Sexo: Masculino, Edad: entre 31 y 40
  Resultado: Sí compran trajes enteros

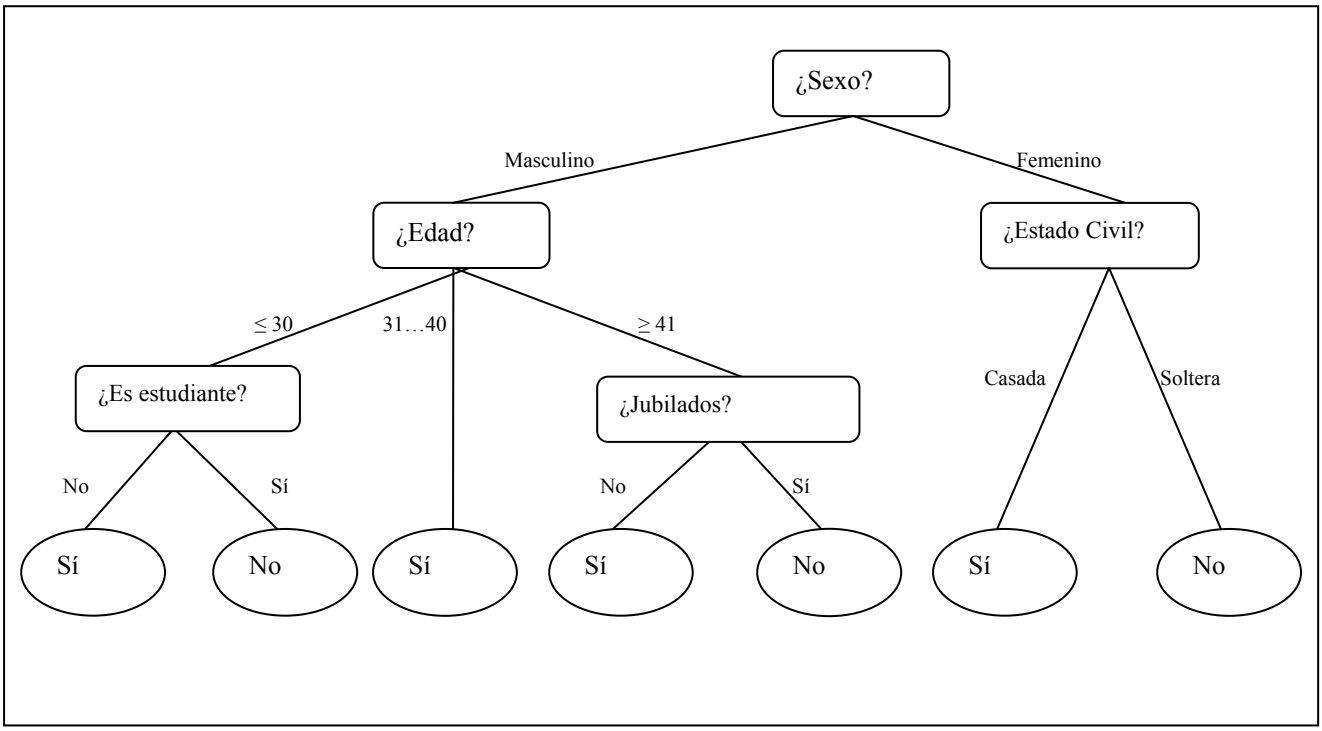
No obstante, los menores a 31 y los mayores a 40 no tienen un comportamiento fijo, algunos compran y otros no. Para los menores a 30 se hace la pregunta: "¿Es estudiante?", se logra determinar que todos los estudiantes no compran trajes mientras que todos los no estudiantes sí compran. Con esto se construyen dos hojas más:

- Sexo: Masculino, Edad: menor o igual a 30, Estudiante: Sí
  Resultado: No compran trajes enteros

- Sexo: Masculino, Edad: menor o igual a 30, Estudiante: No
  Resultado: Sí compran trajes enteros

De esta menara se van a determinar todo el conjunto de reglas, tales que se pueden aplicar para poner en práctica ofertas según las características particulares de los clientes.

Ilustración 5.4.1.1. Ejemplo de un Árbol de Decisión. Se trata de determinar quienes son propensos a comprar trajes enteros.



El árbol del ejemplo es muy sencillo, la idea es simplemente comprender la inducción básica del algoritmo. Se trata de una versión simplificada del algoritmo conocido como ID3, éste utiliza una técnica "top-down recursiva" con una estrategia "divide y vencerás".

Obsérvese que para cada nodo las preguntas significativas pueden ser distintas: no se pregunta por la edad de las mujeres, mientras que en los hombres no se pregunta por estado civil. Esto es porque cada grupo de clientes tiene características distintas, por lo que sus patrones de conducta deben responder a sus perfiles específicos. Para determinar los atributos y medidas de los test, así como su orden, existen un conjunto de pruebas estadísticas. En éstas se miden aspectos tales como "la bondad de la separación", o sea, qué tanto ayuda a explicar un atributo un comportamiento específico.

Un algoritmo más avanzado de Árboles de Decisión es el C4.5, el cual es una versión mejorada del ID3. El algoritmo básico requiere que todos los atributos sean discretos o bien que sean "discretizados" (por medio de rangos) la versión mejorada permite entre otras cosas la utilización de valores continuos, los cuales se pueden agrupar y desagrupar posteriormente. Se puede profundizar más en este tema en [Quinlan 86].

4.4.2 Redes Neuronales

Las primeras Redes Neuronales Artificiales que se desarrollaron simulaban el comportamiento del cerebro humano, en el cual las neuronas están interconectadas y reaccionan ante los estímulos que reciben unas de otras.

En el cerebro humano, las neuronas y las conexiones entre ellas (denominadas sinapsis) constituyen la clave para el procesamiento de la información. Una Red Neuronal Artificial está compuesta por un conjunto de elementos de entrada y salida que se conectan entre sí. En cada conexión reciben una ponderación o peso que determina los resultados o respuestas del nodo ante estímulos que recibe de los otros nodos. Durante los procesos de aprendizaje o entrenamiento de la red, las ponderaciones se van modificando o ajustando para que logren explicar la clase a la que pertenece un objeto. En el proceso de entrenamiento de los datos, estos se pre-clasifican, esta caracterización posteriormente se va depurando hasta llegar a un modelo satisfactorio.

Existen dos tipos de redes neuronales:

1. *Las biológicas*: emulan el comportamiento de las neuronas del cerebro.
2. *Orientadas a aplicaciones*: responden a requerimientos o características específicos de la aplicación en la que se va a utilizar.

Se utilizan largos períodos de entrenamiento, por lo que este tipo de estudios solo se utiliza en aplicaciones y/o circunstancias en las que se permiten estas características de tiempo y complejidad. Esta es una desventaja de las redes neuronales.

En un estudio o análisis de algún fenómeno, realizado por medio de Redes Neuronales existen dos etapas básicas: la de entrenamiento o aprendizaje y la de pruebas. En el entrenamiento de la red se establecen las ponderaciones entre las neuronas utilizando un conjunto de datos o patrones de entrenamiento. En la etapa de pruebas se "aplica" el modelo definido por medio de las entradas y se revisan los resultados generados por la red.

Una característica de las Redes Neuronales es su capacidad de aprender. Aprenden por la actualización o cambio de las ponderaciones de las conexiones al ir ingresando nuevas muestras de datos en el proceso de entrenamiento. En [Andina 2001] se afirma que "… normalmente, los pesos óptimos se obtienen optimizando (minimizando o maximizando) alguna 'función de energía'. Por ejemplo, un criterio popular en el entrenamiento supervisado es minimizar el ***error cuadrático medio*** entre el valor del maestro[25] y el valor de salida real."

Las redes neuronales han sido criticadas por ser muy compleja la interpretación de sus resultados pues es difícil comprender el significado simbólico de las ponderaciones antes mencionadas. No obstante, las redes neuronales tienen ventajas, como su tolerancia ante datos extremos (lo que se conoce como ruido) y su capacidad para clasificar patrones. Además, se han desarrollado algoritmos para extraer reglas a partir de una red neuronal entrenada.

[25] En los conjuntos de datos entrenados, se conoce como "maestro" al resultado generado por el modelo, el cual se compara con los resultados reales.

4.4.2.1 Cómo funcionan las redes neuronales

Los elementos básicos de las redes neuronales se denominan **Elementos de Procesamiento (EP),** se trata de unidades de procesamiento de datos que reciben entradas de información (*inputs*) las cuales se procesan matemáticamente para generar resultados (*outputs*).

Por ejemplo, en una venta de automóviles, se establece que para una marca y modelo de vehículo específico, el precio de los vehículos similares de otras marcas, forma parte de las variables que determinan el comportamiento de los clientes **cuando éstos deciden comprar un vehículo nuevo**: el precio de vehículos similares de otras marcas afecta el hecho de que los clientes renuevan el vehículo (compran el mismo modelo y marca pero de un año mas nuevo) o bien que estos decidan cambiar de marca (compran el modelo similar de otra marca). En este ejemplo se supone que se trata de clientes que ya poseen vehículos de esta marca y modelo específicos.

Para un modelo en particular, se establece que si el precio promedio de un conjunto vehículos similares de otras marcas es de un 101% o más con respecto al precio del vehículo en cuestión, el 75% los clientes viejos de este modelo tienden a renovarlo (compran el mismo modelo pero de un año más reciente), pero si la relación es de entre un 90% y 100%, el 50% de los clientes viejos no renuevan (el otro 50% tienen a comprar modelos similares de otras marcas), …, así se pueden ir estableciendo un conjunto de reglas para otros rangos de precios. Finalmente con esta información se construye un EP que formará parte de una red neuronal para el análisis de las ventas y renovaciones del modelo de vehículo en cuestión.

Entonces, un EP procesa los datos resumiéndolos y transformándolos por medio de funciones matemáticas. Los EP se conectan entre sí por medio de sus entradas y salidas, conformando de esta manera una Red Neuronal. Un EP por sí solo está limitado en cuanto a su capacidad de brindar información, pero al interactuar con los demás EP esta situación cambia[26].

Como se ha mencionado, en el proceso de entrenamiento de la red, se van modificando las ponderaciones de las conexiones según los resultados que se observan. La fuerza de una conexión depende de su peso. Este entrenamiento utiliza un método matemático que se denomina Regla de Aprendizaje.

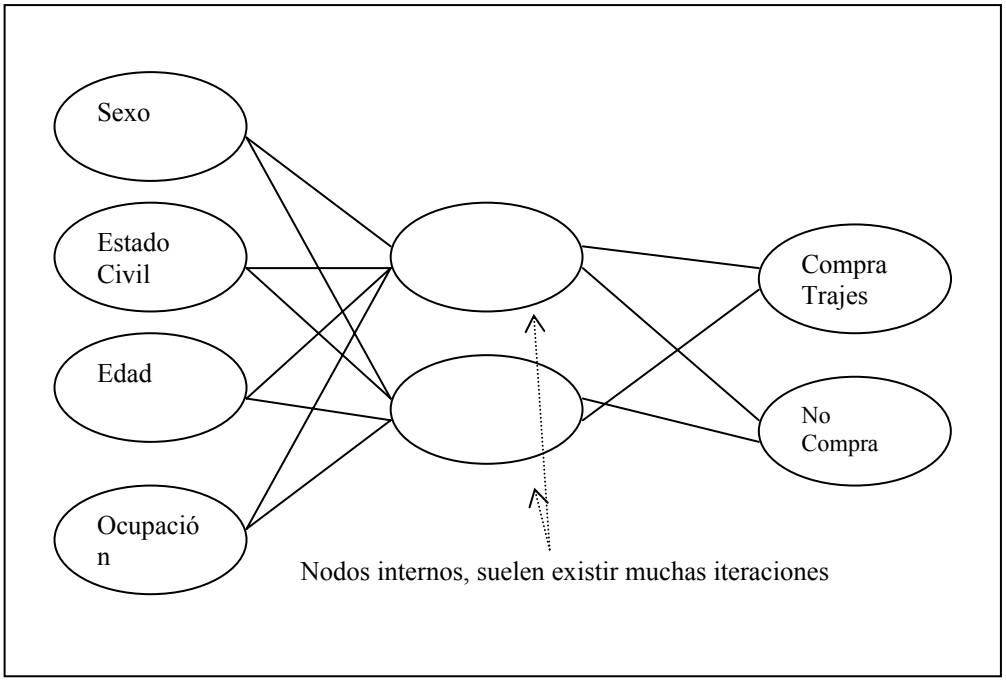
Las redes neuronales se entrenan en forma iterativa con muchas muestras históricas de los datos, para cada una de las cuales se miden sus entradas, salidas y sus ponderaciones respectivas. Esto genera un modelo que se puede utilizar con efectos predictivos y/o de clasificación. El entrenamiento va a continuar hasta que la Red produzca resultados similares a los observados empíricamente, es decir, que continúa hasta que la Red logre explicar los datos observados.

En la Ilustración 5.4.2.1 se muestra un ejemplo de una Red Neuronal para el problema similar al presentando en el apartado anterior. En este ejemplo, al igual que el anterior, se quiere determinar (clasificar) si los clientes son propensos a comprar (o no) trajes, pero se analizan menos variables que en el ejemplo anterior. En el procesamiento se van analizando los valores de los atributos de las muestras de datos y sus interacciones, para finalmente plantear un modelo

[26] Existirán muchas otras variables que afectan el comportamiento de los clientes, por ejemplo: la satisfacción de los clientes con sus vehículos, las características de estos (las "extras"), la cilindrada, etc. En modelos completos, las variables de interés tendrán ponderaciones distintas, o sea, distintos pesos.

que clasifica a los clientes entre Compradores y No Compradores de trajes. Las ponderaciones se dan entre las conexiones entre los EP, cada valor de cada atributo (salida) implica una entrada ponderada hacia otro EP.

Ilustración 5.4.2.1. Ejemplo de una Red Neuronal para clasificar a los clientes entre Compradores y No Compradores de Trajes.



Existen distintos tipos de Redes Neuronales, en el ejemplo anterior podemos hablar de una "red con alimentación hacia adelante" (*feed-forward*), que es muy común en los **aprendizajes supervisados**, dado que ya se conoce el valor final de las variables objetivos (compra: sí/no). La regla de entrenamiento conocida como *Back-Propagation* se utiliza en este tipo de redes.

En una Red Neuronal también se pueden hacer estudios del tipo de Aprendizaje no Supervisado (para realizar Análisis de Agrupamiento). El proceso es muy similar, la diferencia es que en este caso no se conoce el valor final de la variable objetivo. En otras palabras, no tenemos un "valor final deseado". En este caso, la red va a dividir u organizar las muestras de datos en grupos dependiendo de las similitudes entre las muestras.

4.4.3 Redes de Confianza Bayesianas

Por medio de una 'Red de Confianza Bayesiana' también se pueden realizar estudios de clasificación. Este tipo de estudio se basa en el Teorema de Bayes (el cual se comenta más adelante). La idea básica de este método es que, para un objeto de estudio particular, si conocemos el valor de alguno(s) de sus atributo, podríamos medir la probabilidad de que el objeto pertenezca a algún grupo o segmento específico.

El método Bayesiano se basa en las probabilidades condicionales de que un objeto de análisis

pertenezca a un grupo, dado que se conocen las probabilidades de que algunos atributos tengan valores particulares. En otros términos, con este método se calcula la probabilidad de que un objeto pertenezca a un grupo en particular a partir de la probabilidad de que algunos de sus atributos tengan algún un valor específico. En el proceso de estudio se generarán un conjunto de probabilidades para cada una de las posibles combinaciones de valores de los atributos y los grupos pre-establecidos.

Obsérvese que aquí no se determinan o establecen los grupos o segmentos dentro del universo de los datos. Es decir, por medio de éste método no se establecen estos grupos, sino que ellos deben estar predefinidos. Para establecer o descubrir grupos se utilizan los métodos de Análisis de Agrupamiento (*Clustering Analysis)*, dentro de lo que se denomina Aprendizaje no Supervisado (ver sección "Análisis de Agrupamiento (Aprendizaje no Supervisado)" página 32).

4.4.3.1 Teorema de Bayes

Como se explicó antes, las Redes de Confianza Bayesianas se basan el Teorema de Bayes, para explicarlo se supone lo siguiente:

1. Sea X una muestra de un objeto tal que no se conoce el grupo al que pertenece.
2. Sea a_i el valor observado de un atributo A de X.
3. Sea H alguna hipótesis de que el objeto X pertenece a una clase o grupo C.

Considerando los puntos anteriores vamos a definir $P(H/a_i)$ como la probabilidad de que se cumpla H (o sea, que X pertenece a la clase C) una vez que se haya observado el valor a_i en el atributo A. Esta es una *probabilidad a posteriori* dado que se deriva del hecho de que se conoce previamente el valor de A.

En forma inversa, podemos establecer que $P(a_i/H)$ como la probabilidad de que el atributo A tenga el valor a_i dado que se conoce que el objeto X pertenece a la clase C.

El teorema de Bayes establece que se puede calcular la *probabilidad a posteriori* $P(H/a_i)$ a partir de las probabilidades P(H), $P(a_i)$ y $P(a_i/H)$:

$$P\left(H/a_i\right) = \frac{P\left(a_i/H\right)P(H)}{P(a_i)}$$

4.4.3.2 Clasificación Bayesiana Naive

El siguiente ejemplo muestra el método Naive (o Simple) de Clasificación Bayesiana:

En un banco se quiere clasificar a los clientes de tarjeta de crédito en tres grupos:

1. Aquellos que se atrasan mas de una vez en el año (a1+)
2. Lo que se atrasan solo una vez (a1)
3. Lo que no se atrasan (a0)

En los cuadros siguientes se muestran los datos históricos de los clientes.

Cuadro 1: distribución de clientes por sexo

	a1	a1+	a0	*Totales*
Hombres	3.750	4.000	7.250	*15.000*
Mujeres	1.000	500	8.500	*10.000*
Totales	*4.750*	*4.500*	*15.750*	*25.000*

Cuadro 2: probabilidades de atrasos

P(a1)	P(a1+)	P(a0)
0,19	0,18	0,63

Cuadro 3: probabilidades de sexo

P(Hombre)	P(Mujer)
0,6	0,4

Cuadro 4: probabilidades condicionales de sexo, dados los atrasos

	a1	a1+	a0
P(Hombre/H)	0,79	0,89	0,46
P(Mujer/H)	0,21	0,11	0,54

Cuadro 5: por medio del teorema de Bayes se calculan las probabilidades condicionales de los atrasos según el sexo

	a1	a1+	a0
P(H/Hombre)	0,25	0,27	0,48
P(H/Mujer)	0,1	0,05	0,85

Nota: recuérdese que H es la hipótesis de que el objeto X pertenece a una clase o grupo C

Aplicando el teorema de Bayes se calculan las probabilidades del cuadro 5, el cual se puede interpretar de esta manera:

- Los hombres tienen una probabilidad de 0,25 de atrasarse 1 vez al año.
- Las mujeres tienen una probabilidad de 0,1 de atrasarse 1 vez al año.
- Los hombres tienen una probabilidad de 0,27 de atrasarse más de 1 vez al año.
- Los hombres tienen una probabilidad de 0,48 no de atrasarse en el año.
- …

Es interesante observar que, si bien se cuenta con menos mujeres que hombres, éstas tienden a ser mejores tarjeta-habientes.

El ejemplo es muy sencillo, la idea es tan solo facilitar la exposición. No obstante, se puede

aplicar este análisis a un conjunto mucho mayor de atributos que lleven al banco a establecer reglas de decisión que ayuden en la aprobación de tarjetas crédito y límites de crédito, así como a mejorar los planes de mercadeo para este producto.

5 Otras tecnologías para información gerencial

Hasta ahora en este informe se ha hecho una revisión de dos grandes grupos de tecnologías de información que se enfocan en la generación y análisis de información de dos niveles de información estratégica de alto nivel:

- Data Warehouse con resultados históricos y resumidos de datos estratégicos-gerenciales
- Data Mining para realizar estudios estadísticos y matemáticos de los datos

Además de estos dos niveles de información existen otras áreas de interés, enfoques y/o sistemas gerenciales que requieren información, la cual se va encontrar en los sistemas expuestos solo de manera parcial, o no se encontrará del todo, por ello, existen otras tecnologías de IN que cubren estos otros niveles de información.

Hay autores, fabricantes de software y proveedores de servicios de consultoría que consideran estas tecnologías como parte de la IN y las integran dentro de sus suites de productos y/o servicios. Otros las ven como áreas de interés separadas e independientes. Esta discusión no afecta en nada el enfoque que se sigue en este informe. Precisamente si se cuenta con una estrategia y metodología de IN adecuada, esto representará una oportunidad para implementar otros sistemas estratégicos y hacer que el valor agregado del área de TI sea aún mayor. Objetivo éste que de otra manera sería muy difícil alcanzar, haciendo tal trabajo además muy ineficiente.

A continuación se revisarán tres áreas de interés:

- Balanced Scorecard (Cuadro de Mando Integral)
- Sistemas de *Customer Relationship Management* (Administración de las Relaciones con los Clientes)
- Visualización de los datos

5.1 Balanced Scorecard (Cuadros de Mando Integral)

La medición de los resultados en cualquier actividad es la manera natural para determinar si se está siendo exitoso o si es necesario realizar cambios y/o ajustes en las estrategias que se llevan a cabo. La frase “no importa el juego, el concepto siempre es el mismo: si no llevas control del puntaje, cómo puedes saber si vas ganando” [Colteryahn 1988] es muy clara y resulta un tanto obvia. Cómo se podría administrar una empresa sin llevar control de todas las variables de interés,…, simplemente esto sería imposible. Además este proceso de medición y valoración debe realizarse en forma ordenada y coherente y además debe ser completa, es decir, no se deben vigilar o controlar unos aspectos y descuidar otros, sino que se debe llevar control de todo.

La teoría del área de la Administración de Empresas conocida como Balanced Scorecard (BS) fue desarrollada por Robert Kaplan y Edward Norton, a partir de investigaciones que llevaron a cabo en la Universidad de Harvard en los años 90, en la cuales buscaba generar nuevos enfoques y modelos de gestión dado que los vigentes (Calidad Total, Reingeniería, Administración por Objetivos, Justo a Tiempo, etc.) atacaban aspectos puntuales de la gestión y no conformaban un sistema global (al menos así lo veían Kaplan y Norton). En su libro “The Balanced Scorecard” [Kaplan 1996] exponen su trabajo completo, el cual plantea un sistema de gestión global y no de

aspectos específicos.

Kaplan y Norton plantean el BS como un sistema de administración, el cual se basa en el análisis de medidas, tanto financieras como no financieras, que resumen la gestión de la empresa en todas sus áreas y en todo momento. Si bien muchas empresas utilizan sistemas de medición, no lo hacen con un enfoque de gestión estratégica, sino que mas bien los utilizan para obtener "una retroalimentación táctica y para controlar operaciones de corto plazo" [Kaplan 1996]. El BS va mucho más allá de esta simple visión para convertirse en todo un sistema de administración de empresas a partir de un marco conceptual y un conjunto de medidas organizadas por áreas específicas.

En el BS las medidas financieras y las no financieras se organizan en dos niveles básicos:

- **Medidas de los resultados pasados (efecto)**
- **Medidas que generan estos resultados (causa)**

En ambos casos se debe dar seguimiento y tener control de ellas. Estas medidas se generan a partir de la Visión, Misión y Estrategia de la empresa. Esto además se aplica para cada una de las unidades de negocio (es decir, es para todos los niveles de la empresa y no solo para la alta gerencia).

Los objetivos del BS comprenden más que generar una colección de medidas ad-hoc, el BS debe traducir la misión y estrategia de las unidades de negocios en medidas y objetivos tangibles. Las medidas **deben tener un balance** entre las variables generadoras de resultados y las variables que miden los resultados: no es posible alcanzar un objetivo si los elementos que ayudan a generarlo están fallando.

Esta concepción de "causa y efecto en un control constante" es la que diferencia al BS de otros sistemas de mediciones. La idea básica es que se debe llevar un control continuo de las medidas y no simplemente "sentarse a contemplar" resultados pasados para luego tomar una decisión cuando ya quizá sea demasiado tarde.

Por otra parte, debe ser claro que el BS no es excluyente de otras áreas o fuentes de información como la generada en los sistemas de Minería de Datos, Data Warehouse, Depósitos de Datos Operativos, etc. Desde el punto de vista de la información, se trata de otro nivel de datos y necesidades, que es más bien complementario.

El BS se divide en cuatro grandes áreas:

- **Medidas financieras:** Miden las consecuencias económicas de acciones ejecutadas en el pasado. Se trata de la perspectiva financiera de los datos. Normalmente se miden variables relacionadas con la generación de utilidades, por ejemplo: ingresos y retorno del capital. También se mide otro tipo de indicadores financieros como el crecimiento de las ventas, déficit o superávit de flujo de caja, etc.

- **Medidas de los clientes:** En la perspectiva de los clientes, se debe definir el segmento del mercado en el que se compite y las medidas de los resultados. Lo que interesa medir normalmente son aspectos tales como la retención de los clientes, su satisfacción, su rentabilidad, etc.
- **Medidas de los procesos internos de negocios:** Se deben identificar los procesos internos críticos que tienen impacto en la satisfacción de los clientes y las finanzas de la empresa. Se miden el diseño y fabricación de productos, procesos de pre venta y post venta, etc.
- **Medias de crecimiento y aprendizaje:** Tiene que ver con las personas, sistemas y la cultura organizacional de la empresa. Esta perspectiva del BS identifica la infraestructura que la organización debe construir para alcanzar el crecimiento y las mayores ganancias en el futuro. Se miden aspectos tales como: si el personal tiene los conocimientos y destrezas necesarias para realizar su trabajo, el aprovechamiento de las capacitaciones, la integración y capacidad de trabajar en equipo, etc.

Un proyecto de BS consiste básicamente en tres etapas:

1. Determinar las medidas que se van a incluir dentro de cada una de las 4 áreas.
2. La generación de datos para el cálculo de estas medidas.
3. La presentación de la información.

Las medidas que van a conformar el BS de una empresa dependen de la estrategia y objetivos de la empresa, y dentro de éstas, de la estrategia y objetivos de cada unidad de negocios. El BS enlaza las relaciones de causa y efecto en la empresa al medir tanto las variables generadoras de resultados como los propios resultados.

Cuadro 6.1.1. Algunas medidas de un sistema de BS hipotético.

		Medidas	
Área	**Estrategia**	**Generadores de resultados**	**Resultados finales**
Financiera	Alcanzar metas de utilidad	Tasa de crecimiento de ventas, reducción de costos	Tasas de utilidad diarias, semanales, …
	Minimizar los recursos financieros ociosos	Tasas de vencimiento de obligaciones y de activos financieros	Porcentajes diarios de recursos ociosos en los Flujos de Caja
Clientes	Capturar nuevos clientes	Medidas de cobertura de los vendedores y o tiendas	Tasa de crecimiento de clientes nuevos
	Aumentar la lealtad de los clientes	Porcentajes de productos defectuosos devueltos	Tasa de retención de clientes
Procesos Internos	Alcanzar mercados meta nuevos	Cantidad de puntos de venta en el mercado meta	Volúmenes de ventas en el mercado meta
	Aumentar productividad	Tasas de producción por hora/día/…	Productividad por empleado
Crecimiento y aprendizaje	Mejorar la competitividad del personal	Capacitación en los sistemas de información	Utilización los sistemas informáticos
		Disponibilidad de información	

Otro factor muy importante es el hecho de que se debe concebir el BS como todo un sistema de administración que debe formar parte de la cultura empresarial. Todas las áreas de la empresa deben "alinearse y balancearse" dentro del BS y contar con un subsistema de información.

En este sentido se debe ser claro en el hecho de que un BS forma parte del mundo de la Administración de Empresas y no de la Computación. No obstante, viéndolo desde una perspectiva de TI, se pueden desarrollar aplicaciones y bases de datos para facilitar aspectos del BS tales como:

- Migración de datos y generación de la información.
- Visualización y acceso a los datos del BS
- Consulta del sistema en las PC's de los usuarios

Una opción muy sencilla y útil es la de construir un Portal Web para visualizar las medidas del BS. El acceso a la información del portal se puede definir por áreas y/o unidades de negocio. Es común que se muestren las medidas por medio de colores, los cuales determinan si estas están por debajo o por encima de las metas:

- Rojo si la medida está por debajo de un rango que se considera aceptable.
- Amarillo si la medida está dentro de los rangos aceptables.
- Verde si la medida está por encima de la meta.

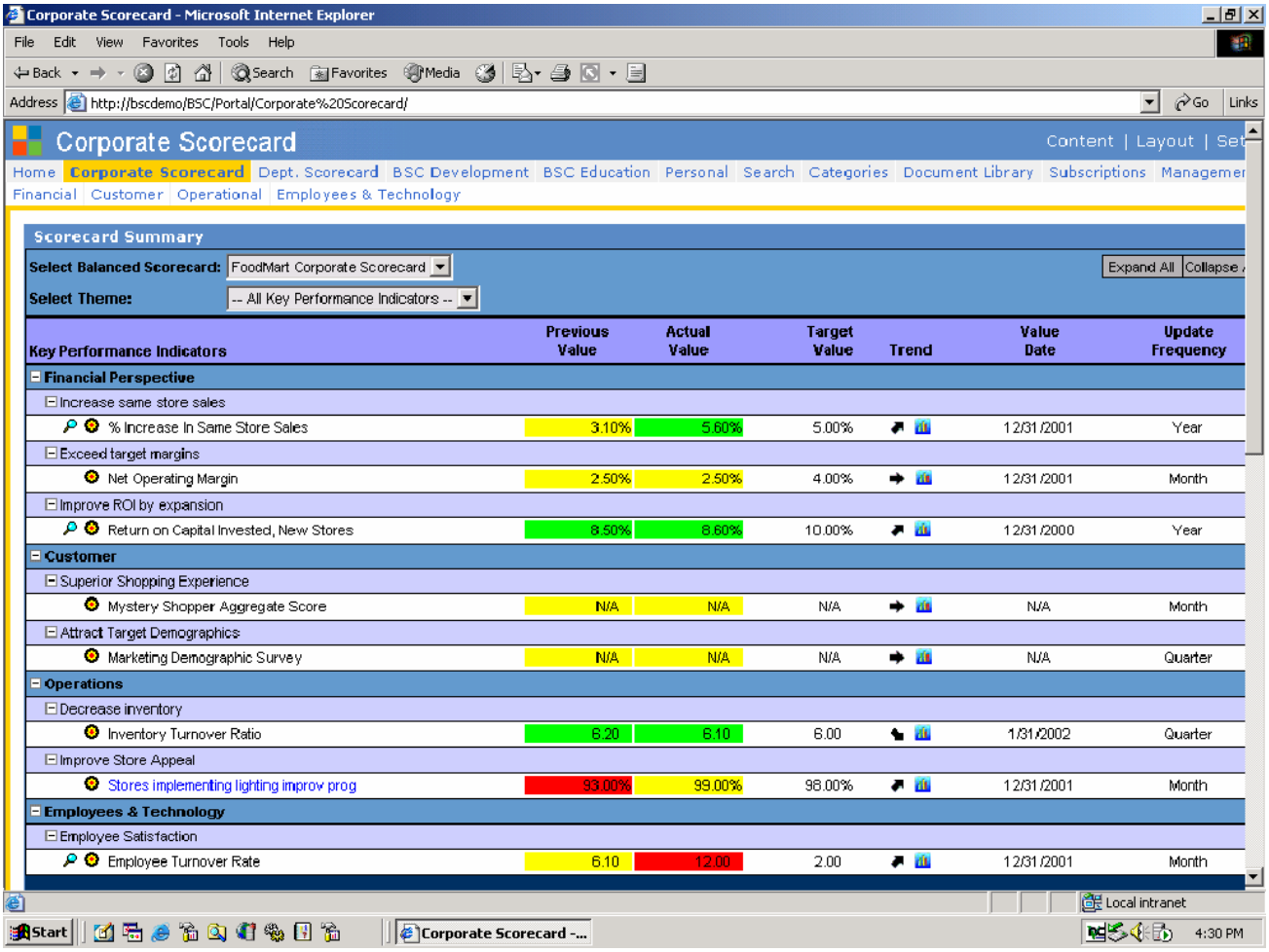
En la figura 6.1.1 se muestra un ejemplo de un portal Web, se observan los valor actual y anterior de las medidas de los diferentes niveles del BS. Esta es una herramienta de usuario final, detrás de la cual existe un sistema de bases de datos, y de migración y generación de medidas. El portal que se muestra en la figura 6.1.1. es un ejemplo construido por medio del Balanced Scorecard Framework de la empresa Microsoft, el cual se compone de un conjunto de aplicaciones, metodología y programación.

Ahora bien, la integración del BS con una estrategia de IN debe ser natural. Por ejemplo, podríamos establecer reportes históricos del Data Warehouse para analizar medidas que formen parte del BS. El BS nos brindará los valores actuales de dichas medidas mientras que en el DW obtenemos la información analítica e histórica de estas.

Podemos ser aún mas ambiciosos: suena muy tentador el que las bases de datos del DW alimenten al BS. Por lo tanto, la integración estratégica de la IN no solo resulta natural sino que además es muy eficiente.

Tal y como hemos venido insistiendo a lo largo del Informe, lo ideal es contar con un Data Warehouse integrado y validado (éste puede incluir un Depósito de Datos Operativo), a partir del cual se pueda extraer la información del BS y mostrarla por medio de alguna aplicación de usuario final. Es decir, la idea es alcanzar la integración de la tecnología a partir de la integración de una estrategia global de IN.

Figura 6.1.1. Portal que muestra los datos de un BS, ésta es una herramienta de usuario final que permite el despliegue del BS para toda la organización, dado que este se puede publicar como parte del portal corporativo, en una Intranet o bien en la Internet.



5.2 Customer Relationship Management (Administración de las relaciones con los clientes)

Los ***Customer Relationship Management*** **(CRM)** son sistemas que almacenan información histórica y detallada de cada uno de los clientes de la empresa, así como de las transacciones que éstos llevan a cabo. El CRM va a involucrar un conjunto de modelos de decisión y algoritmos que permiten o ayudan a establecer los perfiles y patrones de conducta de cada cliente, con esta información el sistema genera alertas y/o semáforos, reportes, gráficos, etc. Toda esta información finalmente ayudará en la elaboración de estrategias de ventas y mercadeo específicas, y en general ayudará a alcanzar las metas y objetivos que se hayan definido.

Muchos de los algoritmos de los CRM provienen de la Minería de Datos, pero se alimentan de modelos conceptúales y/o teóricos propios de las áreas de las ventas y del mercadeo. Se podría decir que un CRM es un caso muy específico de un sistema de Administración del Conocimiento.

Los CRM son sistemas de información que ayudan a identificar, adquirir y mantener clientes. Las empresas que mantienen un CRM adecuado, tienen mayores posibilidades de tener clientes satisfechos y leales, de tener bajos costos de adquisición de clientes, lo cual se traduce en una mejor situación financiera y en mayores posibilidades de ventas.

Los CRM proveen medios para administrar y coordinar la interacción con los clientes. Esta tecnología puede ayudar a maximizar el valor de cada cliente para la empresa al ofrecer información específica de cada uno de ellos, de manera tal que la información se pueda aprovechar por medio de campañas de mercadeo directo y ventas personalizadas y específicas para cada uno de los clientes.

En la actualidad los clientes interactúan con las empresas por medio de muchos canales: Internet, teléfono, agentes de ventas, puntos de venta, etc. Asimismo las empresas tienen múltiples líneas de negocios que interactúan con los clientes. La meta es lograr que los clientes hagan sus negocios de una manera rápida y sencilla, y que sea tal y como ellos lo requieren: en cualquier momento, lugar, por cualquier vía o lenguaje, de manera tal que los clientes se sientan con confianza en la empresa. Por ello, los canales de venta son una información de gran valor para el CRM.

Por otra parte, mucha de la información del CRM se puede aprovechar para alimentar el DW y para hacer análisis de minería de datos, o bien, del DW se puede obtener mucha información para el CRM. Por otra parte, si la empresa cuenta con un Depósito de Datos Operativo, éste puede ser la fuente de información tanto para el CRM como para el DW, así como para cualquier otro sistema de IN. De esta manera, es muy recomendable el contar con una estrategia de IN amplia que abarque todos los proyectos de este tipo en la empresa.

Debe observarse que los datos de los CRM se encuentran en un nivel más detallado que la información del DW.

5.3 Visualización de la información

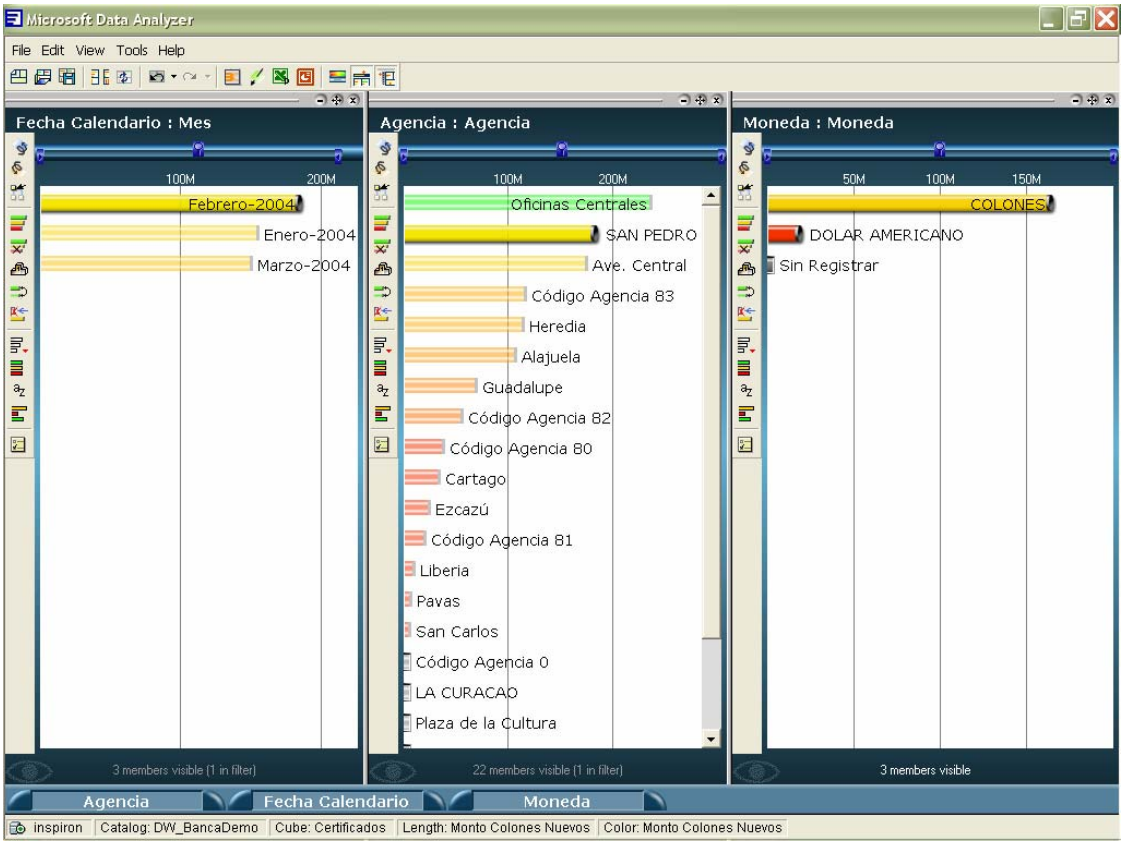
Las herramientas de visualización son las que permiten a los usuarios interactuar con los sistemas de IN de una manera inteligible. En el mercado existen *herramientas* de todos los precios y características, incluso en Internet se pueden conseguir muchas herramientas gratuitas para cada una de las áreas de IN. Para mostrar los datos por medio de gráficos y/o cuadros la creatividad de los programadores es muy importante, además existe todo un conjunto de ideas y teorías en las áreas de la estadística, matemática, administración de empresas, y psicología cognitiva, sobre formas correctas o adecuadas de mostrar los datos.

También se pueden desarrollar aplicaciones para la Internet o la Intranet, las cuales son complementarias a otras aplicaciones de escritorio. La principal utilidad de este tipo de tecnología es que permite desplegar los sistemas de IN llevándolos a muchos usuarios (dentro y fuera de la empresa) y facilita aspectos como el licenciamiento de aplicaciones. Básicamente se trata de la construcción de un portal y de utilizar dentro de él aplicaciones de visualización de datos (principalmente gráficos y cuadros). Éste portal reside en un servidor de aplicaciones y se tiene acceso a él por medio de un navegador Web.

Para los sistemas OLAP fundamentados en las bases de datos multidimensionales, se utilizan reporteadores y graficadores que muestran las medidas y dimensiones de los cubos en una forma intuitiva y sencilla para el usuario. También existen muchas herramientas que generan código SQL para hacer reportes de tablas y no de cubos. Estas herramientas deben ser gráficas o visuales, es decir, el usuario debe estar en capacidad de generar reportes sin la necesidad de escribir líneas de código (se trata de herramientas para los usuarios y no para los técnicos).

El la Figura 6.3.1. se muestra la herramienta Data Analyzer de Microsoft, ésta genera gráficos a partir de Cubos de Datos Multidimensionales. En el ejemplo se muestran las nuevas aperturas de certificados de inversión, por mes, moneda y agencia de apertura.

Figura 6.3.1: Reporte de apertura de Inversiones a Plazo en un Banco. Se observan las nuevas aperturas y estas se desagregan por mes, sexo y moneda.

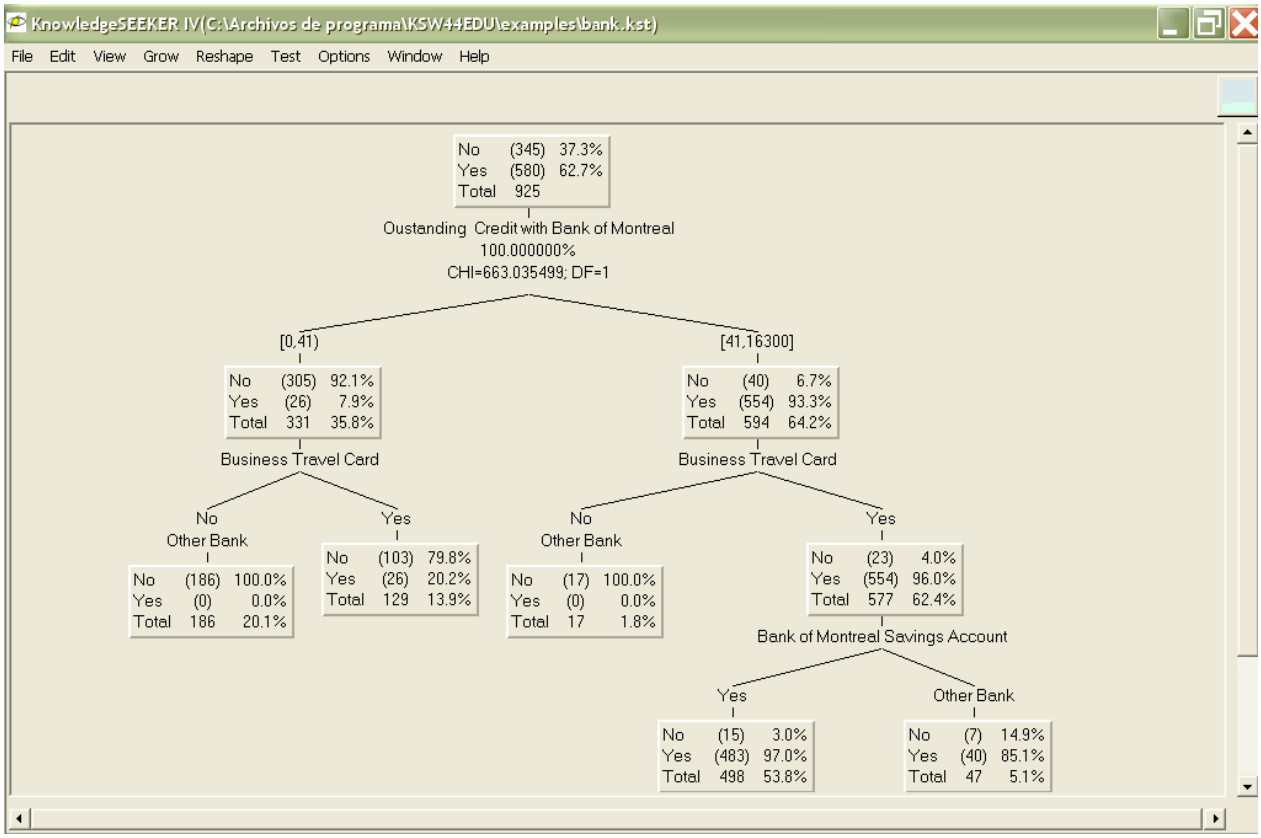


También existen herramientas enfocadas al personal técnico, que le facilitan la labor de la generación de código (MDX para cubos, SQL para tablas, XML, etc.) ya sea para aplicaciones y/o para ser utilizado en programas.

Como hemos mostrado antes, un BS se puede visualizar en el portal Web y de esta manera se integra con el DW. También existen herramientas de escritorio (no Web).

Las herramientas de MD normalmente son utilizadas por unos cuantos expertos en la empresa; estas herramientas tienen sus propias opciones de visualización, las cuales interactúan o tienen capacidad de exportar tablas y/o resultados a archivos de texto, hojas electrónicas, etc. Los resultados de los estudios de MD se pueden publicar en el portal corporativo de la empresa, junto con el DW y el BS.

Figura 6.3.1. Se muestra un árbol de decisiones para una aplicación bancaria. Se utiliza la aplicación KnowledgeSEEKER, la cual entrena un conjunto de datos y posteriormente desarrolla el árbol. El fabricante de este software es Angoss Software Corporation, para más información véase: http://www.angoss.com/.



En fin, en materia de herramientas de usuario final, basta con navegar en la Web para encontrase un sinnúmero de opciones, con todo tipo de características, precios, opciones de licenciamiento, etc. No está de más decir que si bien esta área o segmento de la estrategia de IN es fundamental, no se debe confundir jamás las herramientas de visualización con el propio proyecto de IN. De hecho, el trabajo más arduo reside en la construcción de los procesos de migración de datos, en la construcción de los repositorios de información y en la administración del proyecto. Con la visualización de los datos llegamos al punto final del proyecto, al denominado 'front-end'. Una característica deseable del sistema es que sea compatible con diferentes conjuntos de herramientas y no estar 'casados' con una tecnología particular.

6 Conclusiones y recomendaciones

La Inteligencia de Negocios es un área de especialización en la TI, cuyo fin es brindar información estratégica e histórica a las gerencias y juntas directivas de las empresas, para apoyarlos en la toma de decisiones de alto nivel. Con la Inteligencia de Negocios, el área de TI le da mayor valor agregado a las empresas al generar y poner a disposición de los usuarios, información estratégica e histórica, utilizando métodos y técnicas especializadas.

La IN se diferencia radicalmente de la tecnología informática transaccional; las diferencias se dan en todos los niveles o etapas de los sistemas de información:

1. Se utilizan bases de datos orientadas a la generación de reportes ad-hoc, en donde, en vez de registrar transacciones, se almacena en forma histórica la información sobre los principales resultados de los procesos de negocios.
2. La información que brindan los sistemas de IN es resumida y/o agregada. La información de los sistemas transaccionales es muy detallada y, generalmente, no es histórica (los objetivos son operativos y no gerenciales).
3. Otra diferencia importante se da en las aplicaciones de los usuarios finales, en la IN éstas se concentran en la presentación y visualización de los datos, en forma tabular y/o gráfica, con técnicas y métodos de visualización desarrollados específicamente para la IN. Las aplicaciones de los sistemas transaccionales se enfocan, sobre todo, en la actualización de la información por parte de los usuarios.
4. En las técnicas y metodologías de desarrollo y administración de proyectos, en la IN se tienen enfoques distintos, dado que se trabaja con información y objetivos diferentes de los transaccionales, por tal motivo se han desarrollado métodos de trabajo específicos.

Por lo tanto, con la tecnología de IN nos encontramos ante un área de especialización dentro de la TI, con sus propios métodos, objetivos y técnicas.

El la IN coexisten muchas tecnologías, algunas de éstas son:

- Data Warehouse y Data Marts: sistemas de información histórica, resumida y que integra a todas las áreas de negocios de la empresa.
- Minería de Datos: sistemas que utilizan algoritmos matemáticos y/o estadísticos para procesar datos.
- Customer Relationship Management (CRM): se trata de herramientas para analizar el comportamiento de los clientes.
- Balanced Scorecard: se utilizan para llevar un control en forma continua de los resultados de metas específicas de negocios (por medio de medidas), se “vigilan” estadísticas, razones financieras, etc.
- Visualización de la información. Se incluyen aquí las herramientas de visualización de datos para usuarios finales.

En el informe se recomienda que las iniciativas de desarrollo de los sistemas de IN se integren dentro de una "Estrategia de IN". En ésta, el desarrollo inicial del DW es el punto de partida, que es un proyecto complejo y de grandes proporciones, y constituye un producto de gran complejidad y de un gran valor agregado[27]. Adicionalmente, el DW y sus bases de datos "intermedias", pueden constituirse en la fuente de información para otros sistemas y/o aplicaciones de IN. Por ejemplo, se puede construir un Depósito de Datos Operativo, el cual va a alimentar al DW, a los sistemas de Minería de Datos, y al Balanced Scorecard; de esta manera no se incurre en la repetición de tareas o esfuerzos, más bien, se aprovechan al máximo las bases de datos construidas. El DW va a contar con información validada y depurada, almacenada en un repositorio de datos único, que integra datos de varias fuentes y/o sistemas de información. Todo esto implica grandes ahorros de recursos ya que este repositorio se va a re-utilizar. Por lo tanto, la integración de los sistemas dentro de una estrategia de IN, a partir de un sistema de DW, es muy eficiente, reduce costos y en el mediano plazo acelera el desarrollo de las aplicaciones.

Una vez que los sistemas de IN se liberan, es decir, se pasan a producción, reciben un mantenimiento mínimo, el cual muchas veces es preventivo. Los programas de extracción y migración de datos se pueden calendarizar para que se ejecuten en forma automática y vayan alimentando a las bases de datos. En algunos casos las aplicaciones usuario final pueden requerir de un mayor mantenimiento técnico, tal es el caso de los portales Web, en donde los usuarios, normalmente no pueden hacer modificaciones a la aplicación, esto lo realizan los administradores del sistema. Y en casos tales como la Minería de Datos, los usuarios son personas altamente versadas en análisis matemático y/o estadístico, que utilizan aplicaciones de un gran nivel de especialización.

La naturaleza de la IN nos lleva a concluir que estos sistemas deben siempre desarrollarse a la medida de la empresa, independientemente de las aplicaciones usuario final que se utilicen. No obstante en el mercado existen, para todas las tecnologías de IN, sistemas pre-construídos. Se debe tener especial atención en la implementación de estos sistemas dado que, en su proceso de instalación se debe incluir una etapa de adaptación del sistema a la realidad de la empresa. En el otro extremo existe la posibilidad de desarrollar el sistema a la medida, para lo cual se debe contar con experiencia y metodologías de trabajo adecuadas y probadas. Dado el estado actual de las herramientas de desarrollo en IN, el tiempo de desarrollo de estos sistemas es relativamente corto y en la práctica no es mas rápido implementar y adaptar un sistema preconstruido que desarrollar uno nuevo a la media.

Se debe prestar especial atención en la evaluación del equipo de trabajo y en la tecnología que se vaya a utilizar en el desarrollo/implementación de los sistemas de IN. Independientemente de que se haga por medio de un *outsourcing* o por medio de un desarrollo interno, es preciso evaluar las opciones. Muchos proyectos desastrosos de IN se pudieron evitar con una simple evaluación del equipo de trabajo y de la tecnología.

[27] El DW es probablemente el proyecto de IN más ambicioso y complejo, su desarrollo puede darse a lo largo de varios meses o incluso años. El DW además es dinámico, es decir, que el mismo va a sufrir constantemente de cambios de requerimientos (el DW debe cambiar al cambiar los negocios de la empresa).

6.1 ¿Cómo 'vender' un proyecto de IN?

Finalmente, nos referimos a algunas recomendaciones sobre cómo vender o convencer a las empresas de iniciar un proyecto de IN. Existen distintos grupos de personas que toman decisiones, con necesidades y enfoques diferentes; por esto, lo que le interesa a unos puede ser irrelevante para otros.

Una buena estrategia de convencimiento es empezar por el área de TI. La iniciativa de un proyecto de IN puede surgir de esta área de la empresa: un gerente de TI puede tener conocimiento y/o experiencia con IN y tratar de convencer a la empresa de destinar recursos para IN. Él tendrá un papel de 'evangelizador' en la materia y además cuenta con un conocimiento claro de las limitaciones de la información estratégica de su sistema actual. Otra opción es convencer a este gerente de asumir este papel, haciéndole ver los potenciales de la IN. Para el área de TI un proyecto de IN exitoso es una forma de promoción interna; esta puede ser una manera palpable en que TI dé mayor valor agregado a la compañía.

Los gerentes de otras áreas no relacionadas con la TI, es decir los usuarios, se pueden interesar en la IN a partir de sus necesidades y de las limitaciones de la información con que la cuentan. Por medio de conferencias, presentaciones y/o publicaciones sobre la materia, estos tomadores de decisiones pueden conocer los alcances y limitaciones de la tecnología de IN.

Es muy importante el hecho de que "hay que ver para creer", es decir, los tomadores de decisiones requieren ver un sistema "en acción" para entender lo que es realmente un sistema de IN, ya que muchas veces no se tienen claras la diferencias entre distintos tipos de tecnología.

Ralph Kimball [Kimball 1998] recomienda el desarrollo de prototipos, como parte de su metodología, para alcanzar los objetivos de desarrollo e implementación del sistema de Data Warehouse. Los prototipos además ayudarán a vender el proyecto, ya que los gerentes van a conocer la tecnología y alcances del IN por medio de un mini sistema con sus propios datos y requerimientos, y atendiendo a sus propios problemas de información. Un prototipo puede ser la primera versión, resumida y sencilla de algún Data Mart[28]

[28] Las empresas de consultoría acostumbran hacer las denominadas 'pruebas de concepto', que no son ni más ni menos que el desarrollo de un prototipo, sencillo y rápido, para un Data Mart. La elección del área por desarrollar es muy importante: debe elegirse un Data Mart para un área importante y que comprenda necesidades de información no cubiertas.

7 Bibliografía

[Andina 2001] Diego Andina y Antonio Vega Corona, Tutorial de Redes Neuronales. http://www.gc.ssr.upm.es/inves/neural/ann2/anntutorial.html Universidad Politécnica de Madrid-UPM (España) y Universidad de Guanajuato (México). 2001 (sitio visitado el 15-04-2004).

[Bloomfield 2002] Charles Bloomfield. Bringing the Balanced Scorecard to Life: The Microsoft Balanced Scorecard Framework. White Paper. Microsoft, May 2002. (http://www.microsoft.com/business/bi/, sitio visitado el 16-6-2004)

[Colteryanhn 1998] Karen Colteryahn and Patty Davis. The Balanced Scorecard. Does it answer the though questions? Whitepaper. Development Dimensions International, Inc., 1998, MKTCPWPO3-08980MB (www.ddiworld.com, sitio visitado el 12-12-2003)

[Groth 2000] Robert Groth. Data mining, Building Competitive Adavntege. Prentice Hall. 2000.

[Han 2001] Jiawei Han y Micheline Kamber. Data Mining, Concepts and techniques. Morgan Kauffman Publishers. 2001.

[Inmon 1996] W. H. Inmon. Building the Data Warehouse. New York: John Wiley & Sons, 1996.

[Jiménez 1998] Beatriz Jiménez y Rafael Ávalos. Depósitos de Datos. Club de Investigación Tecnológica. San José, Costa Rica, 1998.

[Kaplan 1996] Robert S. Kaplan y David P. Norton. The Balanced Scorecard. Translating Strategy Into Action. Harvard Business School Press, September 1996.

[Kaplan 2001] Robert S. Kaplan y David P. Norton. The Strategy Focused Organization. Harvard Business School Press, 2001.

[Kimball 1998] Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite. The Data Warehouse Lifecycle Toolkit, Expert Methods for Designing, Developing and Deploying Data Warehouses. New York: John Wiley & Sons, 1998.

[Moss 2003] Larissa T. Moss, Shaku Atre. Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Addison-Wesley Pub Co, 2003.

[Muñoz 2000] Lilia Muñoz. Calidad de los datos: Un enfoque conceptual. Club de Investigación Tecnológica. San José, Costa Rica, 2000.

[Peterson 2000] Timothy Peterson and James Pikelman. Microsoft OLAP Unleashed. SAMS, 2000.

[Quinlan 1986] J. R. Quinlan. Induction of decision trees. Machine Learling, 1:81-106, 1986.

[Quirós 2000] Franco Quirós. Medición de calidad de datos: Un enfoque práctico. Club de Investigación Tecnológica. San José, Costa Rica, 2000.

[Vitt 2002] Elizabeth Vitt, Michael Luckevich, Stacia Misner. Business Intelligence. Microsoft Press; 2002.