UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Facultad de Contaduría y Administración

DATA WAREHOUSING COMO FACTOR COMPETITIVO EN LA TOMA DE DECISIONES

Seminario de Investigación Informática

QUE PARA OBTENER EL TITULO DE

LICENCIADO EN INFORMATICA

P R E S E N T A N:

MARY KARINA RUIZ TORRES

NEY GALICIA ARROCENA

ASESOR DEL SEMINARIO:

Actuario Francisco David Mejía Rodríguez

México, D. F.

1998

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

Agradecimientos:

A Dios, por el maravilloso regalo de vivir para llegar a la culminación de esta meta tan importante.

A la Universidad Nacional Autónoma de México, por los conocimientos, profesores y amigos que han contribuido en cada paso de mi formación profesional.

Dedicatorias:

A mi madre, por que nadie en este mundo podría brindarme un apoyo y amor tan grande como tú lo has hecho.

A mis hermanos, Tavo y Mimi, por escuchar y comprender.

Karina

A Dios

"Libradme de la sabiduría que no llora,
de la filosofía que no ríe
y del orgullo que no inclina la cabeza ante un niño"

A mis padres y hermanos

A Rocío

Ney

ÍNDICE

INTRODUCCIÓN

En la evolución de nuestra sociedad, han aparecido una cantidad impresionante de elementos que se han sumado a nuestro acervo tecnológico y que han servido de base para desarrollos cada vez más sofisticados, esta constante búsqueda del hombre se ha reflejado también en la forma en que ha estructurado a sus organizaciones y en nuestros días encontramos organizaciones altamente tecnificadas junto a otras con recursos apenas suficientes para seguir realizando sus operaciones. Esta desigualdad se ha extendido a otros aspectos de la vida de los pueblos, y es nuestro deber como miembros de esta sociedad contribuir de manera significativa a reducir el impacto delas limitaciones que sufren nuestras organizaciones para poder ampliar los beneficios de su buen funcionamiento a cuantos ámbitos de nuestra vida en sociedad sea posible.

El conocimiento que logren las organizaciones de las diversas tecnologías existentes y nuestra colaboración como profesionales permitirá dar enfoques más justos a su uso y quizá en un futuro permitirá el desarrollo de nuevos avances más adecuados a la realidad imperante en las diversas regiones y culturas de nuestro mundo.

Cada vez más organizaciones se percatan del valor de los datos contenidos en las bases de datos corporativas, por lo que su deseo de información crece exponencialmente llevándolas a la búsqueda de medios que le permitan contar con la información adecuada de manera oportuna, y la tecnología de data warehousing, tema central de nuestro estudio, es una respuesta a esta necesidad.

En el primer capítulo hacemos mención de los aspectos que han definido nuestro marco teórico a lo largo de nuestra investigación.

Los capítulos 2, 3 y 4 describen algunos temas fundamentales para lograr una mejor comprensión del tema de Data Warehousing, estos son: Antecedentes de la tecnología de bases de datos, Sistemas manejadores de bases de datos y Modelado de sistemas de información respectivamente.

Los capítulos 5, 6, 7, 8 y 9 incluyen información que gradualmente nos introduce en el tema del Data Warehouse y nos permite acercarnos de manera concreta a una de las tecnologías que más ha impactado al mundo de los sistemas de información y que cada vez cobra mayor fuerza entre las comunidades involucradas en el desarrollo, distribución y uso de estos sistemas y, principalmente a interesado a la comunidad de negocios en sus esferas de toma de decisiones, pues han visto en el Data Warehousing un medio que promete revolucionar la forma en que se toman las decisiones dentro de las organizaciones de nuestros días.

El capítulo 10 presenta el análisis que hemos realizado con base en los datos recogidos de varias fuentes, las cuales se mencionan en las referencias bibliográficas y electrónicas que se encuentran al final de este documento. Este análisis muestra la conveniencia de contar con un data warehouse en la organización y la manera en que permite mejorar el proceso de toma de decisiones, los diversos factores que intervienen en el éxito o fracaso de un proyecto de esta magnitud, la forma en que se integran cada uno de los elementos del data warehouse y algunas propuestas de uso que a manera de ejemplo, permiten tener una visión acerca de las ventajas de esta tecnología.

Finalmente se encuentran las conclusiones a que hemos llegado y un breve glosario de los términos cuyo significado consideramos conveniente aclarar para una mejor comprensión del tema.

1. MARCO TEÓRICO

1.1 IDENTIFICACIÓN DEL PROBLEMA

En el mundo actual, la velocidad con que una organización pueda tomar decisiones es uno de los factores que define el éxito o fracaso del negocio, pues es sabido que en nuestros días los avances tecnológicos han abierto posibilidades antes inimaginables en las diversas áreas de actividad empresarial. Si bien en un principio la tecnología se enfocó a los sistemas operacionales (en línea), al paso del tiempo y con la ayuda de otras disciplinas como son la inteligencia artificial, la ingeniería, etc., ha sido posible desarrollar sistemas cada vez más complejos que ahora nos pueden ayudar a realizar actividades que han sido y serán puramente humanas, así, vemos ahora sistemas de apoyo a la toma de decisiones como es el data warehousing los cuales en ningún momento reemplazarán a los líderes y directores de la empresa, por el contrario, estos líderes cuentan ahora con una herramienta que les ofrece una ventaja competitiva en la realización de sus funciones y que de ser utilizada en la forma adecuada puede redituar a la organización grandes beneficios, no sólo permitiéndole cumplir sus objetivos, sino también ampliar sus horizontes de oportunidades en un mercado en el cual la competencia se muestra cada vez más agresiva y en que las alianzas estratégicas se suceden una tras otra procurando encontrar las soluciones más convenientes a los problemas que enfrentan.

México, no ha escapado a esta vorágine tecnológica, y en nuestras organizaciones enfrentamos cada vez con más urgencia la necesidad de aprovechar con mayor eficiencia los escasos y cada vez más costosos recursos de que se dispone.

Quizá uno de los recursos más costosos para las organizaciones, es la información, pues día a día se generan grandes volúmenes de datos que requieren cada vez mayor espacio de almacenamiento, herramientas más poderosas para su manipulación y personal capacitado; y que desgraciadamente en muchos casos, estos datos no se utilizan más que para llenar los archivos muertos, siendo que de un análisis adecuado de esos datos, se pueden conocer tendencias, hacer proyecciones, segmentar mercados, conocer patrones de consumo y un sinnúmero de factores que son fundamentales en la supervivencia de un negocio.

De ahí surge la necesidad de, en primer lugar, difundir el conocimiento sobre la tecnología del data warehousing; y segundo, resaltar los aspectos que determinan el uso de esta tecnología.

1.2 HIPÓTESIS

El data warehousing es una tecnología que podemos considerar como factor de competitividad debido a que permite a la organización explotar información concentrada en beneficio de sus procesos de toma de decisiones, facilitándole el análisis y comprensión de dicha información y la proyección de eventos o comportamientos futuros en las distintas actividades que afectan el desarrollo de la organización.

1.3 OBJETIVO

Destacar la importancia del uso de las bases de datos como parte esencial en la toma de decisiones mediante la aplicación de data warehousing.

1.4 METAS

a) Identificar las necesidades de información que intervienen en la toma de decisiones de la organización.
b) Identificar la evolución de la tecnología de bases de datos hasta llegar al data warehouse.
c) Destacar el data warehouse como apoyo a las herramientas de bases de datos para los sistemas ejecutivos de información.
d) Conocer las ventajas y desventajas de la implantación de un data warehouse en la organización.

2. ANTECEDENTES DE LA TECNOLOGÍA DE BASES DE DATOS

2.1 BASE DE DATOS

Una base de datos es una colección de datos relacionados acerca de una empresa con múltiples usos [Shakuntala Atre].

Datos interrelacionados almacenados en conjunto sin redundancias perjudiciales o innecesarias, su finalidad es la de servir a una o más aplicaciones de la mejor manera posible, los datos se almacenan de modo que resulten independientes de los programas que los usan; se emplean métodos bien determinados para incluir datos nuevos y para modificar y extraer los datos almacenados.

2.2 HISTORIA

Tradicionalmente, las generaciones de desarrollo computacional, son definidas en relación con la tecnología de procesamiento (transistores, circuitos integrados, integración a gran escala, etc.). Para nuestros propósitos es mejor definir las generaciones en términos de tecnologías de bases de datos significativas las cuales se presentan a continuación:

2.2.1 *Las Primeras Dos Generaciones*

- Cronológicamente el periodo puede ser considerado desde finales de los 1940's y principios de los 1960's.
- Las computadoras de los 50's se basaron en transistores con memoria de centro magnético.
- La cinta magnética fue el principal medio de almacenamiento masivo, restringiendo a los sistemas a un procesamiento serial.

- Los sistemas de información fueron limitados en alcance y tamaño por la inmadurez de la tecnología de almacenamiento y del software.
- Programas generadores de reportes (por ejemplo Mark I, desarrollado en 1956), los cuales permiten la producción de reportes sin un esfuerzo de programación significativo.
- Mantenimiento de archivos generalizado y sistemas de reportes (por ejemplo 9PAC, desarrollado en 1959), los cuales resultan del perfeccionamiento de los programas generadores de reportes.
- Sistemas de archivos con formato (por ejemplo IRS, desarrollado en 1958), el cual fue desarrollado en la mayor parte para uso de las agencias militares y de inteligencia en Estados Unidos.
- Otro logro importante durante este periodo fue el lenguaje de definición de datos COMPOOL, desarrollado en MIT como un mecanismo de definición de atributos del SAGE (Air Defense System); este fue probablemente la primera concepción de una definición de datos global.
- Creación de CODASYL (COnference on Data Systems and Languages), formada por representantes de las áreas de defensa y gobierno de Estados Unidos, así como representantes mundiales de las áreas de negocios, con el objetivo inicial de proponer un lenguaje de programación de alto nivel para usar en el desarrollo de programas aplicables a negocios. Este objetivo inicial fue alcanzado con la publicación del primer COBOL.
- El reconocimiento de las necesidades para mecanismos generalizados en el soporte de desarrollo de sistemas de información, y
- El surgimiento de prototipos de algunas de las facilidades requeridas de los DBMSs integrados.

En retrospectiva, probablemente la investigación más significativa en el campo de sistemas de información fue el trabajo de General Electric. Este trabajo resultó en el primer DBMS integrado comercial, IDS (Integrated Data Store), y tiene un profundo efecto en los sistemas de información de la siguiente generación.

2.2.2 *Tercera Generación*

- Abarca desde principios de los años 60's hasta mediados de los 70's.
- Las máquinas de este periodo fueron representadas por la serie IBM 360 e IBM 370, con circuitos integrados y discos de almacenamiento magnético.
- Por el lado del software, los elementos dominantes fueron sistemas operativos multiusuarios que ofrecían métodos de acceso de datos, lenguajes de programación de alto nivel y DBMSs nacientes.
- Quizá el producto más importante fue Information Management System (IMS) de IBM, desarrollado a mediados de los 60's por IBM y North American Aviation (ahora Rockwell International) posteriormente IBM lo desarrollo como el primer sistema de administración de comunicación de datos y bases de datos generalizadas a gran escala.
- En 1965 se integró el CODASYL List Processing Task Force. Este grupo fue renombrado Database Task Group (DBTG) en 1967, y en 1968 propuso la extensión al lenguaje COBOL para permitir a los programas escritos en COBOL manipular bases de datos navegacionales. En 1969 el DBTG publicó de manera semiformal recomendaciones para los lenguajes de definición y manipulación de bases de datos. Hay tres puntos destacados de estas recomendaciones:

  a) Las proposiciones incluían, además de un lenguaje de esquematización (o lenguaje de definición de datos) un lenguaje de subesquematización para la definición de vistas de usuarios de las bases de datos.
  b) Asumían que el acceso a las bases de datos es sólo a través de un programa de aplicación escrito en un lenguaje de programación convencional de alto nivel.
  c) Las propuestas fueron hechas tomando en cuenta un bajo nivel de abstracción de las consideraciones de acceso y almacenamiento. En la práctica esto significa que los programas escritos para una base de datos en particular están influenciados por cualquier cambio que se genere en la implantación de esa base de datos.

- Las principales causas que motivaron el cambio durante este periodo pueden ser agrupadas bajo tres grupos:

  a) El creciente problema de la aplicación de backlog: la capacidad de atención era superada por la demanda en el campo de los sistemas de información.
  b) Desarrollos en tecnologías relacionadas ofrecían poder de cómputo a los usuarios que demandaban cada vez más facilidades de bases de datos interactivas.
  c) Había una creciente preocupación por la necesidad de bases teóricas para el trabajo en bases de datos.

En resumen, se puede decir que la tercera generación se caracterizó por:

- Las bases de datos se convirtieron en parte importante del desarrollo de los sistemas de información.
- Creciente preocupación por la inconveniencia de la tecnología disponible en términos de productividad, facilidades interactivas, y bases teóricas, y
- Surgimiento de propuestas por un nuevo enfoque de administración de bases de datos, basadas en un formalismo abstracto.

2.2.3 *La Cuarta Generación*

- Este periodo comienza a mediados de 1970. En términos de sistemas de información es un periodo dominado por el desarrollo de sistemas manejadores de bases de datos (DBMSs) relacionales.
- IBM desarrollo el proyecto System R como demostración de la factibilidad del modelo relacional como base de los DBMSs de ambientes comerciales. Este proyecto hizo dos importantes contribuciones:

a) El lenguaje de bases de datos SQL (Su nombre original fue SEQUEL), el cual se ha convertido en el lenguaje relacional estándar de ISO y de facto .
b) Varios productos entre los que destacaron SQL/DS y DB2.

- Es importante destacar el volumen de recursos invertidos durante esta época en la investigación de DBMSs, y los productos que surgieron fueron modificados con los desarrollos respectivos para adecuarse al equipo y software deseados.

En la actualidad existen tres áreas principales que están apoyando el desarrollo de una nueva generación de tecnología de bases de datos:

a) La primera es la creciente demanda de los usuarios por un mejor desempeño.
b) La segunda es la introducción de mayor "inteligencia" al DBMS, y
c) La tercera es la distribución.

3. SISTEMAS MANEJADORES DE BASES DE DATOS (DBMS)

3.1 DEFINICIÓN

Un sistema manejador de bases de datos (DBMS -Data Base Management System-), es un conjunto de rutinas, funciones, métodos de acceso, áreas de trabajo, almacenamiento y control requeridas para el tratamiento del manejo de información bajo el concepto de base de datos [Martin J.].

Un DBMS es probablemente mejor definido como una pieza sofisticada de software, la cual soporta la creación, manipulación y administración de sistemas de bases de datos.

Los elementos de un DBMS son:

- Lenguaje de Definición de Datos (DDL -Data Definition Language).
- Diccionario de Datos (DD -Data Dictionary-).
- Lenguaje de Manipulación de Datos (DML -Data Manipulation Language-).

Un sistema de bases de datos consta de dos partes: el *Sistema Manejador de Bases de Datos* (DBMS), el cual es el programa que organiza y mantiene listas de información, y la *Aplicación de Bases de Datos*, la cual es un programa que nos permite recuperar, consultar y modificar la información almacenada por el DBMS.

Un DBMS provee los siguientes servicios:

- *Definición de datos:* Provee un método de definición y almacenamiento de una cierta cantidad de datos.
- *Mantenimiento de datos:* Mantiene los datos utilizando un registro para cada elemento, los campos contienen información particular que describe cada elemento.

- *Manipulación de datos:* Provee servicios que permiten al usuario insertar, modificar, borrar y ordenar los datos de la base de datos.
- *Integridad de datos:* Provee uno o más métodos para asegurarse que los datos son correctos.
- *Despliegue de datos:* Provee algún método para presentarle los datos al usuario.

3.2 CARACTERÍSTICAS

Un DBMS requiere hardware para el almacenamiento de bases de datos y software para accesar a la base de datos por contenido más que por localización. El software DBMS soporta distintas características a diferencia del sistema de archivos:

- *Interfaces:* Tales como los lenguajes de consulta y las formas de llenado. Los usuarios utilizan estas interfaces para salvar y modificar registros.
- *Mecanismos de seguridad:* El DBMS asigna derechos a los usuarios para el acceso a datos.
- *Reglas de negocio:* Las reglas de negocio son condiciones específicas de la empresa acerca de los valores de los datos en la base de datos. El DBMS verifica que los valores introducidos o modificados por los usuarios sean conforme a las reglas mencionadas.
- *Transparencia en concurrencia:* Muchos usuarios pueden intentar modificar el mismo dato en el mismo tiempo. Cuando esto ocurre, los cambios introducidos por un usuario pueden ser sobrescritos por un segundo usuario sin tomar en cuenta los cambios introducidos por el primero. Múltiples usuarios necesitan algún tipo de control de concurrencia para que ellos no ignoren los cambios hechos por otros. Un DBMS soporta transparencia en la concurrencia, múltiples usuarios pueden utilizar el DBMS al mismo tiempo sin tener conocimiento que otros están utilizándolo.
- *Procesamiento de transacciones:* Una transacción es una secuencia de operaciones de base de datos que deben ser ejecutadas como una unidad. Las transacciones son deseables debido a que garantizan que los datos satisfagan las reglas del negocio. Si una de las operaciones de la transacción falla, entonces la transacción completa es abortada y algunos cambios hechos a la base de datos son deshechos inmediatamente.

- *Respaldos y mecanismos de recuperación:* Estas facilidades habilitan a los usuarios para hacer copias de sus bases de datos y recuperar una base de datos de una copia si la base de datos está dañada. El DBMS puede crear un *log* de los cambios hechos a la base de datos.

Por otro lado, una aplicación de bases de datos es un programa que permite a los usuarios ingresar, modificar, borrar y obtener reportes de los datos que se encuentran en la base de datos. Los lenguajes utilizados para crear aplicaciones de bases de datos pueden ser agrupados en tres categorías:

- *Lenguajes procedurales:* La gran mayoría de los lenguajes de programación pueden ser descritos como procedurales. El código de la aplicación es escrito como una serie de procedimientos, cada uno de ellos hace el trabajo de una porción de la aplicación, tal como una consulta a la base de datos o un procedimiento para modificar los datos contenidos en la base de datos. Como ejemplo de este tipo de lenguajes encontramos: Pascal, COBOL, BASIC y C. Los ejemplos más comunes de lenguajes procedurales para bases de datos específicas son: dBASE, PAL (Paradox Application Language) y el R/BASIC Language (utilizado por Advanced Revelation).
- *Lenguajes de Consulta Estructurados (SQL -Structured Query Language-):* Fue diseñado como un lenguaje para accesar explícitamente a DBMSs basados en el modelo relacional. SQL se describe más propiamente como un sublenguaje, ya que no contiene facilidades de manejo de pantalla o de entrada/salida. Su objetivo principal es proveer un método estándar para accesar a las bases de datos, independientemente del lenguaje en que se encuentre escrita el resto de la aplicación.
- *Otros Lenguajes:* Este grupo abarca todos aquellos lenguajes que no caen completamente dentro de las dos categorías previas. Los lenguajes más comunes de este tipo son los *Lenguajes de Programación Orientados a Objetos* (OOP) tales como Modula-2 o C++, estos lenguajes representan un enfoque de programación completamente diferente, donde las acciones son definidas sobre "objetos", en vez de ser tomadas como series de procedimientos.

Otro tipo de lenguaje que es utilizado con bases de datos es un lenguaje *macro* (o script). Tienen una lista del conjunto de teclas que el usuario introduce manualmente en la aplicación para automatizar ciertas tareas. Es sumamente específico para una aplicación particular.

Finalmente, *Query-By-Example* (QBE); el cual no es estrictamente un lenguaje; es una interfaz que presenta al usuario con una o mas tablas en blanco que corresponden a las tablas en la base de datos. El usuario elige las columnas que serán incluidas en la consulta a través de una combinación de teclas, y define las condiciones de la consulta llenando las condiciones dentro de las columnas apropiadas. El DBMS traslada el QBE en las acciones necesarias para cumplir con el requerimiento del usuario.

3.3 FACILIDADES

Los DBMSs modernos incluyen una amplia variedad de facilidades, básicamente son las siguientes[2]:

- *Administración del almacenamiento secundario:* El objetivo de los DBMSs es la administración de grandes cantidades de datos compartidos. Por grande queremos decir que la información es demasiada para ser manejada en la memoria principal.
- *Persistencia:* Los datos deben ser persistentes; ésta es una diferencia muy importante contra la programación tradicional, en la cual las estructuras deben ser codificadas en un archivo para existir después de la ejecución de una aplicación. En nuestros días se utilizan lenguajes de programación persistentes de reciente aparición.
- *Control de concurrencia:* Los datos son compartidos. El sistema debe soportar acceso simultáneo a la información compartida en un ambiente de armonía que controle los conflictos de acceso y presente a cada usuario una base de datos coherente.

- *Protección de los datos:* Las bases de datos son una fuente de información invaluable que debe ser protegida contra errores humanos, de otros programas, fallas en el equipo de cómputo y mal uso. Los mecanismos de verificación de la integridad se enfocan a prevenir inconsistencias en los datos almacenados. Los protocolos de recuperación y respaldo de las bases de datos la protegen de fallas en el hardware. Finalmente, los mecanismos de control de la seguridad previenen el acceso de usuarios no autorizados y contra la modificación de la información relevante.
- *Interfaz hombre-máquina:* Involucra una gran cantidad de características, generalmente giran en torno a la representación lógica de los datos.
- *Distribución:* En muchas aplicaciones, la información reside en distintos lugares. Aún en una misma empresa, es común encontrar información interrelacionada dispersa en varias bases de datos, ya sea por razones históricas o para mantener las bases de datos de un tamaño más compacto. Estas bases de datos pueden ser manejadas por diferentes sistemas (interoperabilidad) y basadas en distintos modelos (heterogeneidad).
- *Compilación y optimización:* La traducción de las peticiones contra los niveles externo y lógico en programas ejecutables. Esto por lo general involucra varios pasos de compilación y de optimización de manera que el desempeño no se vea degradado por la conveniencia de utilizar interfaces más amigables.

Algunas de estas características se encuentran relacionadas básicamente con el nivel físico de los datos: control de concurrencia, recuperación y administración del almacenamiento secundario. Otras, como la optimización, se pueden ubicar en los tres niveles.

3.4 MODELOS

3.4.1 Sistema Manejador de Archivos (FMS)

Es el *único modelo que describe cómo son* almacenados los datos en el disco. En este modelo, cada campo o dato es almacenado secuencialmente sobre el disco en un gran archivo. Fue el primer método usado para almacenar datos en una base de datos computarizada y la simplicidad es su única ventaja. Los productos existentes actualmente sobre este modelo son de bajo nivel. Sus desventajas son claras. Primero, no hay otra indicación de la relación entre los elementos más que la secuencia de almacenamiento. El programador, y algunas veces el usuario, tiene que conocer exactamente como son almacenados los datos en el archivo para poder manipularlos.

El FMS crea problemas con respecto a la integridad de datos, todos los valores de los campos tienen que ser verificados por el programa de aplicación antes de ser almacenados. No hay forma de encontrar rápidamente un registro particular, cada búsqueda inicia desde el principio del archivo examinando cada registro.

Por otra parte, la única manera de ordenar los datos es leyendo el archivo completo y sobreescribiéndolo en un nuevo orden.

Finalmente su gran desventaja, es que no permite realizar cambios fácilmente a la estructura de la base de datos.

3.4.2 Sistema de Base de Datos Jerárquica

En este modelo los datos son organizados en una estructura de árbol que se origina desde una raíz. Cada clase de dato es localizada en diferentes niveles de una rama particular que proviene de la raíz. La estructura de datos en cada nivel de clase es llamado *nodo;* si no nacen de él nuevas ramas el último nodo en las series es considerado una *hoja.*

Permiten definir las relaciones uno a muchos. Además la estructura jerárquica hace fácil y rápida la búsqueda de datos.

La estructura física de los datos en el disco no tiene importancia en el modelo jerárquico; el DBMS puede almacenar los datos como una lista ligada de campos, con apuntadores que van del padre al hijo y de rama a rama, finalizando en un valor nulo o apuntador terminal en la última hoja.

Este diseño facilita la adición de nuevos campos en cualquier nivel, ya que el DBMS únicamente tiene que modificar el apuntador terminal al siguiente nodo de la rama en la lista. Por conveniencia, podemos definir un registro como un padre y todos sus hijos.

El primer problema radica en la estructura inicial de la base de datos, la cual es arbitraria y debe ser definida por el programador cuando la base de datos es creada. La relación padre-hijo no puede ser modificada sin rediseñar la estructura. Otro problema creado por la rigidez de la estructura jerárquica, es la dificultad para modificar la definición de los niveles de clases, ya que se tiene que redefinir la estructura.

La desventaja más significativa de este modelo es que no provee un método de definición sencillo para el uso de relaciones muchos a muchos; una solución a este problema es el almacenamiento de múltiples copias del mismo dato en múltiples niveles. Otro enfoque de solución al problema de la relación muchos a muchos es ir aumentando relaciones padre-hijo secundarias y apuntadores en la estructura jerárquica.

Este método crea numerosas relaciones circulares, las cuales se vuelven más complejas, la arquitectura de base de datos se convierte gradualmente en el modelo de red.

3.4.3 Sistema de Base de Datos de Red

El término "red" no tiene relación con el medio físico en el que actualmente corren las bases de datos, el modelo de red define conceptualmente las bases de datos en las cuales existen las relaciones muchos a muchos. Las relaciones entre los diferentes datos son referidas comúnmente como conjuntos que los distinguen estrictamente de las relaciones padre-hijo definidas en el modelo jerárquico.

Un sistema de base de datos de red se identifica por líneas o apuntadores cíclicos para mapear las relaciones entre los diferentes elementos de datos.

Las interrelaciones entre los diferentes conjuntos pueden convertirse en un modelo cada vez más complejo y difícil de mapear. Tal como las bases de datos jerárquicas, las bases de datos de red pueden ser muy rápidas, especialmente mediante el uso de índices de apuntadores que permiten la ubicación directa en el primer elemento de un conjunto en una búsqueda.

El diseño inicial de la base de datos es arbitrario, una vez que este es instalado cualquier cambio requiere crear una nueva estructura. El modelo de red permite adicionar nuevos datos o modificaciones a los ya existentes de manera simple, ya que sólo se tiene que definir un nuevo conjunto de relaciones propias con el resto del conjunto de datos.

3.4.4 Modelo de Bases de Datos Relacionales

En el modelo relacional el dato es organizado en conjuntos lógicos matemáticos dentro de una estructura tabular. En un RDBMS, cada campo se convierte en una columna dentro de una tabla, y cada registro se convierte en un renglón. Diferentes relaciones entre varias tablas, son definidas a través del uso de funciones matemáticas, tales como el *JOIN* y *UNION*.

Cada tabla tiene una o más columnas con el mismo nombre que se encuentra en otra tabla; son estos nombres de columnas comunes los que son utilizados para relacionar diferentes tablas. Sín embargo, los nombres de columnas no tienen que ser idénticos en el modelo relacional.

El modelo relacional tiene distintas ventajas sobre los modelos jerárquico y de red, la más importante de las cuales es su completa flexibilidad en la descripción de las relaciones entre los diferentes elementos de datos. Modificar la estructura de la base de datos es tan simple como aumentar o borrar columnas de una tabla, lo cual no afecta a otra tabla de ningún otro modo. Pueden ser creadas nuevas tablas como proyecciones (subconjuntos) de tablas existentes, y otras tablas pueden ser removidas.

No se tiene que reconstruir la estructura de la base de datos completa para hacer cambios, esto representa un incremento en la preservación de la integridad de datos.

La mayor decisión para un diseñador de una base de datos relacional es la definición de tablas. El proceso de descomponer los datos a ser almacenados dentro de subconjuntos en tablas, es llamado normalización. El modelo de bases de datos relacional define cinco niveles de normalización, en los cuales cada nivel reduce el monto de datos duplicados en la base de datos.

En un DBMS propiamente relacional, la información sobre las estructuras que comprende la base de datos se mantiene en un conjunto separado de tablas, comúnmente llamado sistema de tablas o diccionario de base de datos. Esta información consta de elementos de datos tales como nombres de las tablas, nombres de las columnas en dichas tablas, y el tipo de datos almacenado en cada columna.

Mientras el modelo relacional (como los modelos jerárquico y de red) no especifica cómo son almacenados los datos en disco, la preservación de la integridad implica que los datos deban ser almacenados en un formato que controle el acceso al DBMS que lo creó.

El énfasis sobre la integridad de datos hace al modelo relacional ideal para sistemas de procesamiento de transacciones, y por lo tanto para bases de datos cliente/servidor. En los otros modelos de bases de datos, los cambios tienen que ser realizados directamente a los datos mismos, lo cual puede causar conflictos cuando múltiples usuarios están modificando los mismos registros.

3.4.5 Bases de Datos Orientadas a Objetos

El paradigma de objetos para la construcción de software se basa en la simple premisa de que los objetos modulares autocontenidos son más fáciles de mantener, ampliar y reutilizar que el enfoque tradicional "orientado a la acción" donde el software se divide en procedimientos. Esto representa un cambio en el enfoque tradicional top-down. En vez de concentrarse en las funciones necesarias que se deben ejecutar, el enfoque se cambia a las entidades sobre las cuales se deben aplicar las funciones.

Los componentes fundamentales de un modelo de bases de datos orientado a objetos son:

- Los objetos son entidades atómicas o abstractas que corresponden a cosas en el ambiente de la aplicación siendo representadas en la base de datos, y pueden encontrarse a varios niveles de abstracción y en varias modalidades (media).

- Las relaciones entre los objetos describen asociaciones entre objetos. Tales relaciones son modeladas como atributos de los objetos, así como por asociación de objetos.
- Las clasificaciones de objetos agrupan a aquellos objetos que poseen características comunes.

3.5 ARQUITECTURAS DE SISTEMAS DBMS

El tipo de sistemas de computadoras en que las bases de datos pueden correr, pueden ser clasificados en cuatro categorías o plataformas principales: centralizada, PC, Cliente/Servidor y distribuidas. La arquitectura del DBMS mismo no determina necesariamente el tipo de sistema computacional en que la base de datos tiene que correr.

3.5.1 Plataforma Centralizada

En un sistema centralizado, todos los programas se ejecutan sobre una computadora principal, incluyendo el DBMS, las aplicaciones que accesan a la base de datos y las facilidades de comunicación que envían y reciben datos de las terminales de usuarios.

Todo el procesamiento de datos en un sistema centralizado toma lugar en el host, y el DBMS debe estar corriendo antes de que las aplicaciones puedan accesar la base de datos. El DBMS es responsable de mover los datos hacia y desde los sistemas de almacenamiento (disco), utilizando los servicios que ofrece el sistema operativo.

Las aplicaciones se comunican con los usuarios en las terminales y con el DBMS; el DBMS se comunica con los dispositivos de almacenamiento (los cuales no se limitan únicamente a discos duros) y con las aplicaciones.

Las principales ventajas de un sistema centralizado es su seguridad y la habilidad de manejar enormes montos de datos en dispositivos de almacenamiento. Además, soportan un gran número de usuarios simultáneos. Las desventajas se refieren generalmente a los costos de instalación y mantenimiento. Grandes sistemas de mainframes y de minicomputadoras requieren soporte especializado, como es en sistemas de enfriamiento y sistemas de control climático.

Finalmente, el precio de instalación del hardware para grandes sistemas centralizados, generalmente asciende a millones de dólares, y sus costos de mantenimiento también son altos.

3.5.2 *Sistemas de Computadoras Personales*

Un DBMS al correr en una PC, la PC actúa como host y terminal. A diferencia de los grandes sistemas, las funciones del DBMS y de la aplicación son combinadas dentro de una aplicación. Las aplicaciones de bases de datos sobre una PC manejan las entradas de los usuarios, la salida hacia pantalla y el acceso a los datos en el disco. Combinando estas funciones dentro de una unidad, el DBMS adquiere poder, flexibilidad y velocidad, a cambio del decremento en la seguridad de datos e integridad.

Varios sistemas de bases de datos multiusuario basados en PCs manejan el mismo número de usuarios que los pequeños sistemas centralizados. Sin embargo, los problemas de manejar transacciones múltiples simultáneamente, el incremento en el tráfico de la red y los límites en el poder de procesamiento de las PCs que están corriendo el DBMS, causan que se incremente en gran medida la complejidad y la disminución en el nivel de ejecución. La solución que fue desarrollada para corregir estas limitantes es el sistema de bases de datos cliente/servidor.

3.5.3 Sistemas de Bases de Datos Cliente/Servidor

De manera sencilla, una base de datos Cliente/Servidor divide el procesamiento de la base de datos entre dos sistemas: el cliente (el cual ejecuta la aplicación de la base de datos) y el servidor (el cual ejecuta todo o parte del DBMS actual). El servidor de archivos LAN provee recursos compartidos, tales como espacio en disco para las aplicaciones, e impresoras. El servidor de bases de datos puede encontrarse corriendo en la misma PC como el servidor de archivos.

La aplicación del cliente, se identifica como *front-end*, maneja todas las pantallas y el procesamiento de entrada y salida del usuario. El *back-end* del servidor de bases de datos maneja el procesamiento de datos y el acceso a disco.

La división del procesamiento entre dos sistemas reduce la cantidad de tráfico de datos en la red. Mientras los sistemas cliente corren generalmente sobre PCs, el servidor de bases de datos puede correr sobre otra PC o mainframe.

La desventaja de los sistemas de bases de datos descritos anteriormente es que requieren que los datos se encuentren almacenados en un sistema único. Esto puede ser un problema cuando se tiene la necesidad de atender a usuarios dispersos en un área geográfica, o bien, que necesitan compartir porciones de sus bases de datos departamentales con otros departamentos o un host central.

3.5.4 Sistemas de Procesamiento Distribuido

El procesamiento distribuido ha existido en su forma simple desde hace varios años; de una manera muy limitada, los datos son compartidos entre varios hosts mediante el envío a través de conexiones directas (sobre la misma red) o a través de conexiones remotas vía telefónica o líneas de datos dedicadas. Una aplicación que es ejecutada sobre uno o más hosts extrae la porción de datos que ha sido modificada durante un periodo definido por el programador, y después se transmiten los datos a un host centralizado u otros hosts que se encuentran distribuidos.

Posteriormente, las otras bases de datos son actualizadas, de manera que todos los sistemas están en sincronía unos con otros. Este tipo de procesamiento distribuido ocurre generalmente entre computadoras departamentales o redes LAN y sistemas propietarios.

Mientras este sistema resulta ideal para compartir porciones de datos entre diferentes hosts, éste no controla el acceso de usuarios a los datos que no se encuentran almacenados en su host local. Los usuarios deben cambiar sus conexiones a los diferentes hosts para accesar a diferentes bases de datos, recordando en dónde se encuentra cada base de datos. La combinación de los datos que se obtienen de distintas bases de datos que existen en diferentes hosts, significa que tanto los usuarios como los programadores deben realizar varios cambios importantes. Además, surge la duplicación de datos; aunque el almacenamiento en discos ha disminuido su costo, almacenar los mismos datos en distintos discos resultaría en un momento dado demasiado costoso, y por otra parte, mantener todos estos conjuntos de datos duplicados aumenta el grado de complejidad del sistema.

La solución a estos problemas es la tecnología denominada de *procesamiento distribuido.*

Bajo un sistema de este tipo, un usuario solicita datos del host local; si éste no encuentra los datos solicitados, realiza una búsqueda a través de la red y al encontrarlos,

envía de regreso al usuario los datos solicitados sin que él sepa que dichos datos han sido obtenidos de un sistema diferente, a excepción quizá, de un pequeño retraso en la obtención de los datos.

Un DBMS distribuido provee distintos grados de optimización, la siguiente lista describe tres de ellos:

1. *No-Optimización de la consulta:* El usuario debe localizar cada archivo a ser accesado, formular una subconsulta para accesar a cada archivo y fusionar los resultados de la subconsulta.
2. *Transparencia en replicación y localización:* Un DBMS soporta transparencia en la replica y localización de datos, sin que el usuario tenga conocimiento de que existe más de una copia.
3. *Transparencia en fragmentación:* Un DBMS distribuido soporta transparencia en fragmentación; en estos sistemas, el administrador de la base de datos puede particionar o fragmentar la tabla en distintas partes o fragmentos y almacenar cada fragmento en un lugar diferente. Un DBMS distribuido soporta la transparencia en fragmentación si el usuario no tiene conocimiento que una tabla ha sido fragmentada, cuando esto ocurre un optimizador de software determina cómo reconstruir una tabla de sus fragmentos distribuidos.

3.6 TECNOLOGÍA CLIENTE/SERVIDOR

Actualmente el enfoque principal es sobre aplicaciones de bases de datos. La razón de ello es que una gran cantidad de empresas están reduciendo sus costos de cómputo mediante "downsizing" de sus bases de datos. La idea del downsizing es simple, mover las bases de datos corporativas de sistemas grandes y centralizados, a sistemas pequeños y menos costosos que no requieren de soporte y mantenimiento extensivos. La división en el poder de procesamiento, que es la base de la arquitectura cliente/servidor, hace posible que los sistemas pequeños manipulen los datos.

Los vendedores pueden proveer versiones para un solo usuario de su software cliente/servidor, pero un verdadero sistema cliente/servidor necesita algún tipo de red con una o más estaciones de trabajo interconectadas y servidores. Las estaciones de trabajo deben tener además un tipo de CPU.

3.6.1 Capacidades

La manera en que los sistemas cliente/servidor son implementados, depende de las plataformas en las que corren el front-end y el back-end, y el grado en que el procesamiento será dividido. Hasta el momento, no hay una manera estándar de clasificar los diferentes niveles o implementaciones de los sistemas cliente/servidor.

A continuación se presenta una tabla en la que se clasifican los sistemas cliente/servidor de acuerdo a su implementación, de la más completa a la menos completa.

CLASIFICACIÓN	DESCRIPCIÓN	COMENTARIOS
Clase 1: Completamente Procesamiento Distribuido	• Los datos residen en múltiples sistemas y/o plataformas. • El acceso del usuario es transparente: los usuarios se conectan a un servidor, el cual accesa a los otros sistemas. • Los usuarios no pueden accesar a los datos fuera del DBMS que se encuentra corriendo en el servidor. • Múltiples front-ends proveen consultas, modificaciones a los datos y servicios de reportes.	La implementación a datos es muy limitada.

Clase 2: Completamente Cliente/Servidor	• Los datos residen en uno o más servidores. • El usuario o la aplicación hacen conexiones explícitas a cada servidor. • El servidor ejecuta todo el procesamiento del DBMS. • Los usuarios pueden accesar a los datos únicamente a través del DBMS que se encuentra corriendo en el servidor. • Múltiples front-ends proveen consultas, modificaciones a datos y servicios de reportes.	Es el tipo de sistema cliente/servidor más común actualmente.
Clase 3: Cliente/Servidor Puenteado	• Los sistemas gateway y aplicaciones, crean un puente entre la aplicación front-end del usuario y el DBMS que se encuentra corriendo en un sistema no cliente/servidor. • Los sistemas gateway trasladan las consultas, modificaciones de datos, etc. a procedimientos y llamada al sistema de bases de datos que los puede procesar. • El gateway soporta múltiples front-ends.	Es utilizado generalmente en sistemas basados en PCs y DBMSs que se encuentran corriendo sobre un mainframe o minicomputadora, o como una liga entre un sistema de Clase 2 y un mainframe o mini.

Clase 4: Cliente/Servidor Limitado	• El servidor provee algunas funciones del DBMS, generalmente mediante el almacenamiento de datos y funciones de indexación. • El servidor no siempre previene a los usuarios para accesar a los datos fuera del DBMS corriendo en el servidor. • Gran parte del procesamiento toma lugar en el cliente. • Múltiples front-end son soportados, aunque no tantos como los que soporta la Clase 2.	Los sistemas de este tipo incrementan la funcionalidad del servidor de bases de datos basadas en PCs, utilizando formatos de archivos comunes tales como dBASEs con extensión .DBF
Clase 5: Cliente/Servidor Propietario	• Se requiere una plataforma de hardware propietaria y un sistema operativo. • Los datos pueden ser accesados únicamente a través del front-end que proporciona el proveedor del DBMS.	Utilizado generalmente a principios de los 80´s; recientemente involucrados en sistemas abiertos.

3.6.2 Ventajas

Las principales ventajas de un sistema cliente/servidor surgen de la división del procesamiento entre el sistema cliente y el servidor de bases de datos. Desde que el procesamiento de la base de datos se lleva a cabo en el back-end, la velocidad del DBMS no se encuentra ligada con la velocidad de la estación de trabajo. Como resultado, la estación de trabajo necesita estar habilitada para correr el software front-end.

Esta división del trabajo reduce además la carga de trabajo sobre la red al conectar las estaciones de trabajo. En vez de enviar el archivo de la base de datos completo, el tráfico de la red se reduce gracias a las consultas y las respuestas del servidor de bases de datos. Algunos servidores pueden almacenar y ejecutar procedimientos y consultas en el servidor mismo.

Otro beneficio que se obtiene, es la independencia entre estaciones de trabajo; los usuarios no se encuentran limitados a un solo tipo de sistema o plataforma. En un sistema cliente/servidor, las estaciones de trabajo pueden ser PCs compatibles con IBM, Macintosh, estaciones de trabajo UNIX, o una combinación de las anteriores, y pueden correr múltiples sistemas operativos, tales como MS/PC-DOS, MS Windows, IBM OS/2, o System 7 de Apple.

Otra ventaja es la preservación de la integridad de datos. Actualmente, varios servidores de bases de datos corren un DBMS con base en el modelo relacional, de modo que los usuarios no pueden accesar a los datos fuera del DBMS. Además, el DBMS puede proveer servicios de protección de datos, tales como almacenamiento de archivos encriptados; respaldos en cinta en tiempo real, lo cual ocurre mientras la base de datos está siendo accesada; generación de discos espejo, en el cual los datos son escritos automáticamente en una base de datos duplicada sobre un disco duro distinto.

El procesamiento de transacciones es un método por el cual el DBMS mantiene un log (registro) de todas las modificaciones hechas a la base de datos en un periodo determinado de tiempo.

3.6.3 Desventajas

La mayor desventaja de los sistemas cliente/servidor es el incremento en los costos de administración y personal de apoyo que mantienen el servidor de bases de datos.

Además, debe considerarse la complejidad; debido a la gran cantidad de partes que conforman el sistema cliente/servidor. Por otra parte, cuando se tienen múltiples front-ends, se incrementa la necesidad de dar mantenimiento a los programas de aplicación.

Los cambios a la estructura de la base de datos llevan consigo distintos efectos a través de los distintos front-ends. Esto requiere de un proceso largo y complejo para llevar a cabo las modificaciones necesarias a las aplicaciones de front-end, y esto dificulta el mantenimiento de todas ellas sin causar la ruptura en el acceso a la base de datos por parte de los usuarios.

4. MODELADO DE SISTEMAS DE INFORMACIÓN

Un sistema de información puede ser visto como un modelo de una parte de la realidad de la organización, los hechos que existen en la organización y las actividades que tienen lugar en ella. Por lo tanto, el problema de desarrollar un sistema de información puede ser considerado como un problema de descripción del modelo. Estos modelos son desarrollados como partes de, ampliamente hablando, dos actividades mayores: los requerimientos de ingeniería y las actividades de diseño de la ingeniería. Los requerimientos de ingeniería involucran investigar los problemas y requerimientos de la comunidad usuaria, y desarrollar una especificación de los sistemas de información deseados. La ingeniería de diseño utiliza estas especificaciones y sobre la base de las diferentes restricciones de diseño impuestas por requerimientos no funcionales y por las características de los sistemas objetivo.

La creciente demanda de sistemas de información de tamaño, alcance y complejidad cada vez mayores ha provocado la introducción de varios lenguajes de modelado de alto nivel a través de los cuales los requerimientos funcionales de las aplicaciones y los componentes de los sistemas de información pueden ser modelados a nivel conceptual. Las contribuciones al área del modelado conceptual provienen de las investigaciones realizadas en la Inteligencia Artificial (en particular de la representación del conocimiento), lenguajes de programación y bases de datos. En los últimos años, el interés en esta área ha sido mostrado por lingüistas, psicólogos e investigadores en administración y dirección de negocios.

Desarrollar un sistema de información es una tarea de diseño en la cual el contenido de la especificación final no puede ser conocido de antemano. En particular el área del modelado de requerimientos y el análisis se caracterizan por su informalidad e incertidumbre.

La calidad de un esquema conceptual y finalmente la del sistema de información dependen en forma importante de la habilidad del desarrollador para extraer y comprender el conocimiento acerca del domino modelado, el Universo de Discusión (UoD) y del sistema de información en sí mismo. Este conocimiento es parcialmente poseído por una comunidad variada de usuarios finales y es en parte incorporado en estructuras formales de sistemas de información existentes.

Debido a la naturaleza de la tarea, los desarrolladores se ven forzados a recolectar una cantidad impresionantemente amplia de conocimientos acerca de la empresa, la cual es después abstraída en una especificación formal. Esta especificación debe servir de punto de referencia para cualquier aspecto de desarrollo o procedimiento de mantenimiento de un sistema de información; su importancia jamas podrá ser, sin embargo, resaltada en forma suficiente. El desarrollo de tal especificación se lleva a cabo utilizando lo que se conoce como Modelado Conceptual.

4.1 ASPECTOS METODOLÓGICOS

El área de los sistemas de información se encuentra dominada por referencias al mundo real y se ha dicho que los problemas que se encuentran en esta área son una combinación de problemas empíricos, formales y de ingeniería[13].

Los problemas empíricos se relacionan con el hecho de que al desarrollar los sistemas de información uno se encuentra constantemente obligado a observar los fenómenos del mundo real. Los problemas formales se refieren a la abstracción, estructura y representación del conocimiento de manera que sea posible razonar este conocimiento. Los problemas de ingeniería aparecen cuando se intenta llevar a cabo la construcción establecida por los principios de formalidad adoptados.

Un sistema de información es una descripción formal de un modelo abstracto de una parte de la realidad (UoD). Esta descripción puede cambiar con el tiempo de acuerdo a los cambios en el propio UoD.

Dado que un sistema de información es un artefacto hecho por el hombre muchos de los conceptos que se encuentran en estos sistemas son simplificaciones de conceptos que se encuentran en el mundo real.

Esta visión reconoce la importancia de desarrollar en primer lugar modelos que estén orientados a comprender el dominio de la aplicación y, posteriormente, a través de una serie de transformaciones, desarrollar modelos más formales guiados por las consideraciones de diseño e implantación.

La tarea de transformar los requerimientos de los usuarios en una especificación detallando estos requerimientos de manera formal involucra modelado y análisis. El propósito del modelado es obtener conocimiento acerca del UoD y representarlo de manera que se permita a un desarrollador razonar este conocimiento, comunicar su comprensión a los usuarios finales para su comprobación y modificar el modelo de acuerdo a ello. El análisis se refiere a las actividades involucradas en comprender el conocimiento obtenido.

Dentro del proceso de desarrollo de un sistema de información, la tarea del modelado conceptual puede (y debería) llevarse a cabo en varias etapas complementarias:

- *Análisis de la empresa:* Este involucra el estudio del área del negocio, su misión, planes, problemas, actividades, etc. Uno de los objetivos de esta tarea es delimitar el área de interés y relevancia.
- *Formación conceptual:* Este tipo de tarea implica la construcción de una concepción inicial común y el consenso sobre los objetos principales con relevancia en el domino del problema. En otras palabras, esta tarea contribuye a desarrollar un diccionario corporativo de conceptos, su significado, sus relaciones, sus reglas y restricciones.
- *Estudio de los sistemas de información existentes:* Incluye el estudio y valoración de las bases de datos existentes en la empresa y su contenido de información, así como la calidad de la información, su apoyo al negocio y su potencial para proveer de información de apoyo adecuada.

- *Diseño de sistemas de información mejorados:* Esta es la tarea de desarrollar componentes y sistemas de información nuevos o modificados de modo que correspondan a las necesidades de apoyo a las actividades del negocio.
- *Validación de los requerimientos funcionales y no funcionales:* Esta tarea compuesta tiene el propósito de asegurar que los requerimientos corresponden a las necesidades del negocio y de sus usuarios. Esta tarea incluye procesos tales como la negociación (en caso de requerimientos en conflicto), validación y seguimiento. Para obtener análisis y modelos adecuados del domino de una aplicación es necesario contar con el apoyo de lenguajes, guías y herramientas adecuadas.

Los lenguajes de modelado conceptual son usados con el fin de apoyar la comunicación entre los analistas y los usuarios finales durante las fases de obtención de datos y verificación de la especificación. Obtener y verificar los requerimientos son labores que necesitan un adecuado intercambio de información entre aquellos que comprenden el domino del problema y aquellos que necesitan modelar el dominio del problema[13].

Tradicionalmente, los modelos conceptuales han puesto mayor atención a los aspectos estructurales de la aplicación (Abrial, 1974; Chen, 1976; Codd, 1979; Hammer & McLeod, 1981; Shipman, 1981; Su, 1983) dando origen a los llamados modelos semánticos de datos. En menor escala, otros enfoques han sido desarrollados orientándose hacia el modelado de los aspectos de comportamiento de la aplicación (DeMarco, 1978; Rolland & Richard, 1982; Jackson, 1983).

4.2 PROPIEDADES DE UN ESQUEMA CONCEPTUAL

La parte funcional de una especificación de requerimientos usualmente toma la forma de un esquema conceptual definido de acuerdo a algún modelo, incorporando propiedades estáticas y dinámicas (de comportamiento) y reglas del dominio de la aplicación.

Desde una perspectiva de bases de datos, un esquema conceptual puede ser visto como un conjunto de reglas que describen la información que puede ingresarse y residir en la base de datos. El cumplimiento de estas reglas es realizado por lo que puede ser llamado un procesador conceptual de información (una mezcla de la base de datos, programas de aplicación y procedimientos de interacción humano-computadora) cuya función es mantener la población de la base de datos de acuerdo con el esquema conceptual.

Un esquema conceptual debe contener:

- Todas las definiciones de los tipos de hechos permitidos en la población de la base de datos.
- Todas las restricciones que se refieren a los posibles estados de la base de datos así como las transiciones permitidas en la población, y
- Todas las reglas que se relacionan con los hechos que se pueden obtener de la base de datos utilizando estas reglas.

De acuerdo con ISO (Van Griethuysen, 1982) un esquema conceptual se define como [13]:

"La descripción de los posibles estados de los aspectos del UoD incluyendo las clasificaciones, reglas, leyes, etc., del UoD."

5. SITUACIÓN ORGANIZACIONAL

5.1 PROBLEMÁTICA ACTUAL

La información vital corre en una empresa a través de múltiples y aisladas islas de información en donde los usuarios tienen que localizar, accesar e integrar datos desde múltiples bases de datos y archivos de diferentes computadoras localizadas en distintos puntos geográficos a los que no es posible accesar en muchas ocasiones, esto se debe a tres razones principales:

1. *Múltiples computadoras*: El número de computadoras en una organización está proliferando, las empresas poseen varios mainframes, estaciones de trabajo y computadoras personales; los usuarios generan y almacenan esencialmente archivos privados y bases de datos en estas máquinas. De tal modo que cada computadora contiene múltiples bases de datos y archivos que necesitan ser accesados, así que cada computadora es una isla.
2. *Múltiples sistemas de manejo de bases de datos*: Casi todas las computadoras tienen un sistema de archivos para el almacenamiento de datos, cada sistema de archivos está organizado en una estructura jerárquica. Además, varias computadoras soportan uno o más sistemas manejadores de bases de datos (DBMS) que contienen distintas estructuras de archivos interrelacionadas. Cada archivo y cada DBMS contienen información separada. Los usuarios de un DBMS no están habilitados para accesar a archivos, y los usuarios de un sistema de archivos no puede accesar a un DBMS.
3. *Distancias geográficas entre computadoras*: Las computadoras pueden residir físicamente en distintas habitaciones, edificios o ciudades alrededor del mundo. Esta distancia geográfica hace difícil para un usuario accesar a los datos de otros usuarios.

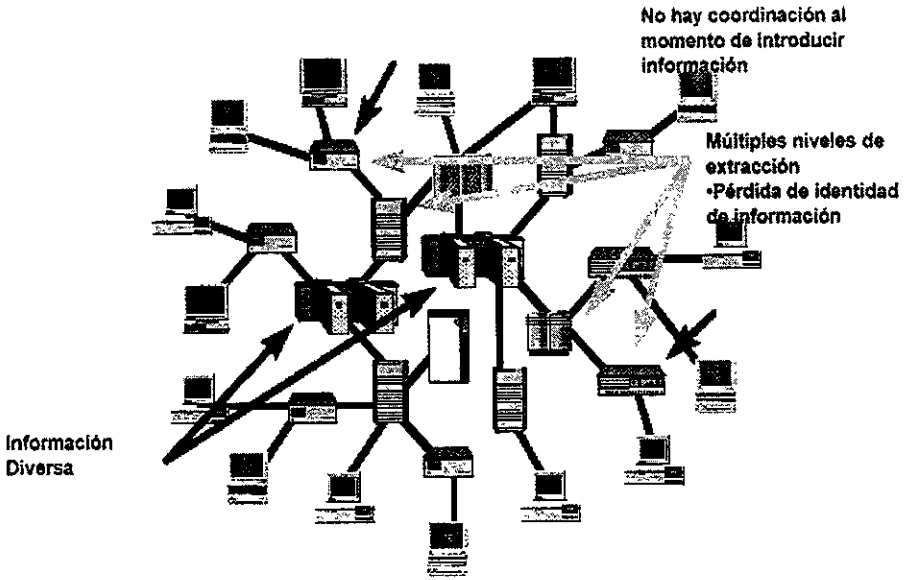


Fig 1 - Problemática organizacional

Por lo anterior, los usuarios deben ejecutar tres acciones esenciales para obtener información significativa de distintos lugares:

- *Localizar:* Determinar si los datos relevantes se encuentran disponibles en algún lado, y de ser así, encontrar el lugar exacto que contiene esos datos.
- *Accesar* Los usuarios deben estar habilitados para formular peticiones por separado a cada uno de los lugares (islas) que contienen la información.
- *Integrar:* Coordinar y consolidar los resultados de sus peticiones dentro de un formato integrado que puedan revisar y utilizar para tomar decisiones.

5.2 MEDICIÓN DE LOS SISTEMAS DE INFORMACIÓN

En las difíciles economías actuales, las compañías se están reorganizando y moviéndose a donde les sea posible para ahorrar dinero y atraer clientes. Esta nueva corporación hace énfasis en un mejor rendimiento de la empresa; por ejemplo, la productividad y calidad en bienes y servicios obligan a todas las partes de la organización, incluyendo los sistemas de información, a hacer esfuerzos por mejorar lo que hacen y la forma como operan. Esto es más fácil de decir que de hacer y muchas organizaciones enfrentan dificultades para identificar en dónde comenzar, qué hacer y cuáles deberían ser sus metas. El aprendizaje organizacional y el cambio son las características destacables de los negocios de los años 90 y el punto de inicio para este esfuerzo es la obtención de información adecuada y exacta, así como su análisis apropiado[18].

No es sorprendente pues que la comparación, medición y evaluación se hayan convertido en algunas de las principales preocupaciones de muchos negocios en años recientes. Se considera que existen tres componentes de la información efectiva y del análisis que dirigen el esfuerzo de mejora de la compañía:

1. Alcance y administración del desempeño de la información,
2. Obtención, análisis y uso del nivel de datos de la compañía, incluyendo datos de clientes y datos de las operaciones, y relacionando datos del desempeño con los datos de rendimientos financieros.
3. Comparaciones competitivas y "benchmarking".

Estos criterios también pueden ser utilizados por las subunidades de la corporación para determinar la forma como ellas están contribuyendo al desempeño y calidad globales.

Una de las áreas que debe recibir un considerable interés es la función de los sistemas de información. Muchas compañías están considerando a la tecnología de la información para ayudarlos a apoyar la reestructuración de las empresas, y hacerlo con una menor cantidad de recursos.

Desafortunadamente, los sistemas de información han sido una de las áreas más difíciles de medir y evaluar. Los sistemas de información presentan además, un problema de credibilidad en varias organizaciones, pues los ejecutivos consideran que los sistemas de información contribuyen de forma mínima o nula al logro de los objetivos de la organización. Es por ello que los presupuestos destinados a sistemas de información, se están haciendo cada vez más pequeños, de ahí la necesidad de medirlos para demostrar su contribución en el desempeño de la organización, tanto a través de servicios y sistemas de calidad como de su rentabilidad.

Cada organización cuenta con su propio conjunto de métricas y parámetros de comparación de los sistemas de información.

Un pre-requisito para el éxito de cualquier programa, interno o externo, de medición y evaluación es contar con datos adecuados y exactos[18]. Las prácticas actuales de medición de los sistemas de información son problemáticas en tres aspectos:

1. *Lagunas de información.* La mayoría de las organizaciones no dan seguimiento al avance en una o más de las siguientes áreas fundamentales para monitorear mejoras en el desempeño operacional y la calidad:

    - Desempeño de los productos y servicios (sistemas y operaciones)
    - Operaciones internas y desempeño,
    - Desempeño de los proveedores,
    - Información financiera y de costos

2. *Falta de estándares.* La medición de los sistemas de información sufre de falta de estándares en casi todas las áreas. Las actuales prácticas de medición se caracterizan por el desacuerdo existente sobre lo que se debe considerar como datos adecuados. Como consecuencia, no es posible hacer comparaciones de las estadísticas de una organización a otra.
3. *Dificultad en la recolección de datos.* En muchos casos la recolección de datos en las organizaciones con sistemas de información impone una seria carga para el personal que, por lo general ya se encuentra bajo demasiada presión para completar su trabajo. Como consecuencia la exactitud de los datos es frecuentemente obstaculizada.

El desempeño de los sistemas de información debe ser medido por 3 razones:

1. La información que proporcionan puede ser de ayuda para determinar si los administradores están desempeñando su actividad ejecutiva en forma eficiente y si se está mejorando continuamente.
2. Ayuda a incrementar la productividad mostrando tendencias en el tiempo e identificando áreas donde es necesario mejorar.
3. Finalmente, pueden ser el catalizador del cambio organizacional apoyando las revisiones de funciones de la compañía y de la planeación.

La clave para lograr cada uno de estos puntos es demostrar el vínculo existente entre los datos recolectados y el desempeño funcional y corporativo. El análisis de estas métricas podría arrojar información valiosa acerca de la función de los sistemas de información y su impacto en el desempeño de la organización.

Incluso las más simples métricas descriptivas pueden ser difíciles de interpretar si no se cuenta con la liga hacia el desempeño. Las métricas solo son útiles si son claramente asociadas con las mediciones del impacto organizacional.

Vincular las métricas con los sistemas de información con contextos más amplios del desempeño funcional y corporativo es un componente esencial de las mediciones efectivas. Por el momento, aunque se logran recolectar grandes cantidades de datos, los esfuerzos para relacionarlos a contratos más amplios son aun insuficientes.

5.2.1 Comparaciones Competitivas y Benchmarking Profesional

Benchmarking o la comparación de una organización con otra es una estrategia para proporcionar dirección a muchos aspectos del desempeño organizacional. Como mínimo, el benchmarking puede proporcionar a los directores o administradores la certeza de que el costo de sus sistemas de información no se encuentra fuera de lo considerado por otros organismos. Pero el beneficio real del benchmarking es que desafía a los administradores a actualizarse con las mejores prácticas de negocio y a mejorar su productividad continuamente.

Aunque muchas compañías se comparan más con otras, esto se hace por lo general de manera informal. Algunos de los métodos actualmente usados son:

- *Preguntar:* Los administradores de los sistemas de información asisten a conferencias, ferias, etc., sobre una base regular. Parte del beneficio de esto es aprender de los participantes de modo informal, acerca de lo que se está haciendo y de cómo llegaron al éxito.
- *Visitando otras compañías.*
- *Proyectos conjuntos:* En algunas industrias se tiene la oportunidad de colaborar en proyectos conjuntos, lo que permite a los participantes ver de manera directa la forma como trabajan otras compañías.
- *Organizaciones industriales.*
- *Servicios de benchmarking.*

6. SISTEMAS DE INFORMACIÓN GERENCIAL (SIG)

Cuando se observa el proceso de diseño de un Sistema de Información Gerencial en la práctica, nos encontramos que cada problema es usualmente iniciado de la nada. En el mejor de los casos, un analista experimentado y con conocimientos en sistemas similares que ha desarrollado con anterioridad trata de "copiar" lo que considera relevante para el caso en que se encuentra.

Se muestra aquí un intento por definir marcos de trabajo que ejemplifiquen alternativas generales de diseño. El concepto de descomposición funcional y diseño ha proporcionado el enfoque básico de ordenamiento y unificación.

El marco de trabajo esta basado en las siguientes ideas.

Primero identificamos la estructura general del Sistema de Información Gerencial, es decir las funciones generales que el sistema debe contener y sus relaciones. Aquí el Sistema de Información Gerencial y la función están definidos en términos muy generales, de modo que incluyen actividades de administración y procesamiento de datos.

Las instancias especificas de la estructura general del Sistema de Información Gerencial hacen posible la definición de los tipos básicos de Sistemas de Información Gerencial. El grado de mecanización y automatización (utilización de la computadora) en la implantación de las funciones generales es el criterio que nos conduce a estos tipos. Tales tipos proporcionan alternativas de diseño y estructura de donde podemos seleccionar cuando nos enfrentamos a un problema.

De este modo, se establecen distintos grados de integración entre las actividades de administración y procesamiento de datos dentro del Sistema de Información Gerencial.

6.1 ORGANIZACIÓN Y ESTRUCTURA

Nuestra visión del procesamiento de la información de la organización distingue los procesos y los sistemas de información gerencial. Los procesos son actividades por medio de las cuales las entradas físicas -materia prima, dinero, bienes de capital, y trabajo- son transformados en bienes o servicios finales.

El sistema de información gerencial -en un sentido muy general- es el conjunto de actividades que regulan (dirigen, planean, controlan, deciden) los procesos (en particular los flujos).

La entrada y salida de estas actividades es información. La información que sale (políticas, planes, programas, reglas, instrucciones, etc.) es el significado que establece la regulación.

Además, podemos definir un ambiente, como todo lo que se encuentra fuera del Sistema de Información Gerencial y que puede afectarlo.

En consecuencia, de acuerdo a nuestra definición, el Sistema de Información Gerencial incluye componentes de administración y de información y se encuentra orientado a toda la organización. Por supuesto, instancias especificas del Sistema de Información Gerencial pueden existir para niveles y decisiones especificas, pero todos ellos son parte, con diferentes grados de integración, del Sistema de Información Gerencial de toda la organización.

Desde un punto de vista de descomposición funcional los componentes de información y de administración de un Sistema de Información Gerencial pueden ser divididos en: Funciones de Administración (MF -Management Functions-) o actividades especificas de decisión que deben realizarse para regular los procesos, y las Funciones de Procesamiento de Datos (DPF -Data Processing Functions-) o la recolección de datos y su transformación en información necesaria para tomar decisiones.

Las funciones de administración pueden ser divididas en funciones de:

a) *Generación de planes estratégicos:* O generación de las políticas de mas alto nivel y planes que regulan el comportamiento de la organización como un todo.
b) *Desarrollo de planes tácticos:* Proceso de convertir los objetivos estratégicos en planes detallados y obtener y ubicar los recursos necesarios para realizar esos planes.
c) *Operaciones de control:* Aseguramiento de que las tareas especificas de cada día que se han implementado en los planes tácticos son llevadas a cabo en forma efectiva y eficiente.

En cuanto a las funciones de procesamiento de datos, podemos distinguir los siguientes tipos:

a) *Funciones de procesamiento de datos básicas:* Por ejemplo la simple recolección de datos, almacenamiento y tareas de transformación realizadas en cualquier organización. Estas se pueden clasificar en:

I. Obtener: conseguir o recolectar los datos elementales necesarios para soportar -posiblemente después de una transformación- una funciones de administración dada.
II. Mantener: almacenar, actualizar, seleccionar y extraer los datos de archivos computarizados o manuales.
III. Calcular: realizar las transformaciones de los datos -posiblemente provenientes de (I) o (II)- a través de operaciones simples como ordenamientos, comparaciones, adiciones, multiplicaciones, agregación, etc.
IV. Proveer: imprimir o desplegar los datos -provenientes de (I), (II) y (III)- necesarios para ejecutar una función de administración dada.

b) *Funciones de procesamiento de datos analíticas:* O transformaciones complejas de los datos orientadas a obtener elementos especialmente significativos. Lo más importante de estas funciones de procesamiento de datos es modelar el comportamiento a través de los datos, por lo general utilizando algún tipo de procesamiento estadístico.

Las funciones detalladas involucradas son:

I. Agregación: Mantener los datos (usualmente) consolidados y clasificados como históricos organizados (utilizando en la mayoría de los casos el tiempo como índice). En los aspectos mecánicos esta función es similar al mantenimiento, pero su objetivo es completamente diferente. La acumulación esta orientada a descubrir patrones y relaciones dentro y entre los datos; mientras que el mantenimiento es utilizado sólo para alimentar un cómputo específico.

II. Análisis: Realizar transformaciones sobre históricos acumulados para delinear patrones subyacentes, relaciones y de modo general aquello que pueda servir para aseverar eventos actuales o futuros. Las transformaciones típicas son estimaciones en series de tiempo y predicciones. Por supuesto existe siempre una interacción de esta función con una función de administración correspondiente que proporciona modelos de comportamiento tentativos y finalmente evalúa la veracidad de los modelos derivados.

III. Cálculo: Ejecuta modelos que permiten cuantificar o predecir consecuencias de funciones de administración propuestas o en ejecución. Los modelos pueden ser simples como las relaciones contables (por ejemplo una explosión del ensamblaje de un producto para cuantificar el número de componentes en un plan de producción o la expresión del costo de los bienes vendidos como un porcentaje de las ventas netas) o tan complejos como un modelo econométrico (relación de ventas y precios, publicidad, etc.). Estos modelos pueden provenir de (II) o de una función de administración.

IV. Informar: Proporcionar las conclusiones derivadas de la ejecución de los modelos en una rutina, excepción o como bases planteadas.

Las funciones del Sistema de Información Gerencial y de otros elementos que se han definido pueden ser ligadas a través de los flujos típicos de información y flujos físicos que existen entre ellos, como se muestra en la figura 11 (Modelo gráfico de un Sistema de Información Gerencial).

Este modelo puede ser considerado también como la representación de la estructura general -componentes generales y sus relaciones- de un Sistema de Información Gerencial.

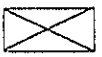
Lo que distingue la estructura de un Sistema de Información Gerencial en especifico en la practica es la forma en que las diferentes funciones y flujos de información son implantados, -para un problema de una área dada- el grado de participación de computo en su desempeño y el grado de integración con otras áreas. De esta manera se genera una clase de estructura por medio de la mecanización (computarización) de las funciones de procesamiento de datos Básicas y Analíticas. Otra clase aparece de la total o parcial automatización (computarización) de funciones de administración (en cualquier nivel). Por supuesto mecanizar un funciones de procesamiento de datos dado o automatizar un funciones de administración implica cambiar la forma en que se realizan las cosas. Por ejemplo, mecanizar una función de procesamiento de datos analítica implica la posibilidad de soportar decisiones en formas que no serían posibles sin una computadora, tales como análisis estadístico de series de tiempo para predecir el comportamiento de una variable dada. De la misma manera, la automatización de una función de administración significa desarrollar modelos o reglas de decisión para generar las decisiones (recomendadas) a través de una computadora.

Por otro lado, un conjunto dado de funciones de procesamiento de datos y funciones de administración puede ser implantado para un problema muy específico de una área o para un conjunto de áreas dando origen a otras clases de estructuras, dependiendo del grado de integración que tenemos sobre las funciones.

Estas clases de estructuras pueden ser también consideradas como clases de diseño a partir de las cuales se debe seleccionar un diseño específico, de entre un número de alternativas disponibles en cada clase, cuando se desarrolla un sistema en la práctica.

6.2 TIPOS

<table>
<tr><td rowspan="2">Nivel de las Funciones de Administración</td><td colspan="3">Computarización de:</td></tr>
<tr><td>Funciones de procesamiento de datos</td><td>Funciones Gerenciales (Parcial)</td><td>Funciones Gerenciales (Total)</td></tr>
<tr><td>Operacional</td><td>Procesamiento de datos tradicional (PD Básico)</td><td>Operación Semiautomática</td><td>Operación Automática</td></tr>
<tr><td rowspan="2">Táctico</td><td>Procesamiento de datos tradicional (PD Básico)</td><td rowspan="2">Planeación Dirigida por Modelos</td><td rowspan="2"></td></tr>
<tr><td>Decisiones apoyadas por Computadora (PD Analítico)</td></tr>
<tr><td>Estratégico</td><td>Decisiones apoyadas por Computadora (PD Analítico)</td><td>Planeación Dirigida por Modelos</td><td></td></tr>
</table>



No es posible

6.2.1 Procesamiento de Datos Tradicional

El procesamiento de datos tradicional proviene de la mecanización estricta de las funciones de procesamiento de datos básicas. Tal mecanización puede ser llevada a cabo a varios niveles de sofisticación e integración. Esta va desde procesamiento de nóminas hasta sofisticados controles de producción con obtención de datos en línea, en tiempo real y distribuido y facilidades de búsquedas sobre esos datos. Puede ir desde aplicaciones aisladas de procesos batch (procesamiento de facturas) hasta procesamiento integrado utilizando tecnología de Sistemas Administradores de Bases de Datos. (DBMS).

Por lo general la mayor parte de los Sistemas de Información Gerencial que se encuentran en la práctica corresponden a: procesamiento contable y financiero, procesamiento de nóminas y personal, y control de inventarios y de producción. Como ejemplo presentamos un modelo de la aplicación simplificada de pago de salarios donde las funciones de procesamiento de datos básicas han sido computarizadas.

Los recientes avances tecnológicos tienden a facilitar e integrar el procesamiento de datos tradicional. Este es el caso de las Bases de Datos, procesamiento interactivo y distribuido y las herramientas de automatización. Además, dado que las aplicaciones de este tipo están bien definidas, existe una tendencia creciente a integrar la parte computarizada de estos Sistemas de Información Gerencial en software generalizado.

6.2.2 Operaciones Automatizadas

Los Sistemas de Información Gerencial del tipo de operaciones automatizadas corresponden a la automatización parcial o total de algunos funciones de administración de nivel operacional. Tal automatización esta basada en modelos derivados o reglas de decisión heurísticas. Existen niveles de sofisticación bastante distintos. Podemos tener desde reglas basadas en la practica hasta complejas calendarizaciones y ruteo de vehículos. Otros ejemplos típicos de estos Sistemas de Información Gerencial son los de requisiciones para inventarios por computadora, autorizaciones de crédito, determinaciones del tamaño de los lotes de producción y decisiones de distribución y calendarización. En la figura 13 se muestra un modelo simplificado de sistemas de requisiciones para inventarios en forma computarizada.

6.2.3 Decisiones Apoyadas por Computadora

La principal característica de las decisiones apoyadas por computadora es la computarización de las funciones analíticas de procesamiento de datos. Se orientan principalmente a:

a) Prever el futuro basándose en datos históricos de la organización y otros datos exógenos.
b) Contestar preguntas del tipo "Que pasaría si...?" en términos de consecuencias especificas para la organización.
c) Esto significa construir modelos con el comportamiento de las variables de estado relevantes como una función del control y de las variables exógenas para el problema que se enfrenta. La idea subyacente en estos modelos es permitir simular el curso futuro de la organización para un conjunto dado de condiciones especificadas por decisiones y variables exógenas. Actualmente estos modelos no generan decisiones pero predice las consecuencias estimadas de las acciones dadas.

Uno de los ejemplos más relevantes de este tipo de sistemas es Corporate Financial Planning System. Un modelo simplificado de este sistema se muestra en la figura 14. La principal idea de este sistema es modelar el comportamiento del mercado y de los costos financieros de producción y otros costos que pudieran ser incurridos para satisfacer una proyección de ventas dada. Los resultados del modelo son las proyecciones de ventas y los resultados financieros, usualmente en el esquema de un estado financiero proforma, para escenarios dados del ambiente económico, actitud de los competidores, políticas de mercadotecnia y otras variables exógenas. Actualmente existen algunas herramientas de este tipo como son: SIMPLAN, IBM Trend Analysis/370 (procesador de series de tiempo), TSP (paquete econométrico), IFPS (modelado de sistemas financieros).

6.3 INTEGRACIÓN DEL SIG

La integración del Sistema de Información Gerencial puede alcanzarse en dos formas. Primeramente, las actividades relacionadas pueden ser integradas en un nivel dado del Sistema de Información Gerencial. Esto es llamado integración horizontal. Por ejemplo, los sistemas de procesamiento de datos básicos relacionados con asuntos del personal -nómina, historial del personal, desarrollo del personal, entrenamiento, etc.- pueden ser integrados en un sistema único de bases de datos.

En segundo lugar, distintos niveles para el mismo tipo de actividad pueden ser combinados en un sistema dado. Esto es llamado integración vertical del Sistema de Información Gerencial. Por ejemplo, un sistema integrado de diseño, planeación, calendarización y control de la producción, donde un solo conjunto funciones proporciona la información necesaria para estas funciones.

Una faceta importante de los sistemas de información es la separación entre la base de datos y el esquema conceptual. Esta distinción es en primer lugar una distinción entre la extensión y el propósito. Al intentar desarrollar un esquema conceptual, uno se debe interesar más en los propósitos del UoD. De este modo, un desarrollador analiza algunos aspectos del mundo con el propósito de determinar las relaciones entre sus objetivos.

7. MINADO DE DATOS

7.1 DEFINICIÓN

El minado de datos es uno de los temas de mayor novedad en el mundo de la tecnología de información. Existen quizá tantas definiciones de minado de datos como vendedores de software de análisis de datos. Al igual que con otros términos los proveedores y algunos analistas utilizan "minado de datos" de manera indiscriminada. El resultado de esto es un conglomerado de definiciones que incluyen a todas las herramientas utilizadas para ayudar a los usuarios a analizar y entender sus datos. Minado de datos es un conjunto de técnicas utilizadas en una forma automatizada para explorar de forma exhaustiva y descubrir relaciones complejas a partir de grandes conjuntos de datos. Es importante hacer notar que estas técnicas pueden ser aplicadas a varias formas de representar los datos, incluyendo dominios basados en texto, dominios de multimedia, y dominios de datos representados de forma tabular, teniendo en cuenta la tecnología de bases de datos relacionales[14].

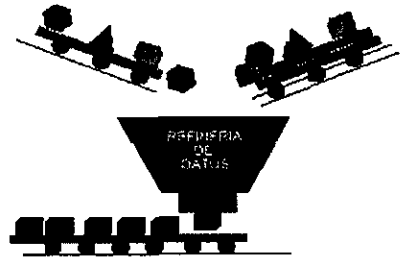


Fig. 2.- Minado de datos

Una diferencia importante entre el minado de datos y otras herramientas de análisis de datos es el enfoque que se utiliza para explorar las interrelaciones entre los datos. Muchas de las herramientas de análisis disponibles soportan un enfoque basado en verificación, en el cual el usuario hace hipótesis acerca de relaciones específicas y una vez formuladas utiliza las herramientas para verificar o refutar tales hipótesis.

Estas herramientas se apoyan en la intuición del analista para establecer la hipótesis y refinar el análisis basándose en los resultados de consultas complejas a la base de datos. La efectividad de estas herramientas de análisis se encuentra limitada por diversos factores, incluyendo la habilidad que posea el analista para hacer preguntas adecuadas y mostrar resultados, administrar la complejidad del espacio de atributos y hacer consideraciones "fuera del domino" del problema.

El minado de datos en contraste, utiliza un enfoque basado en el descubrimiento, en el cual se utilizan algoritmos de reconocimiento de patrones para determinar las relaciones entre los datos. Los algoritmos de minado de datos pueden identificar distintas relaciones entre datos multidimensionales de manera concurrente, resaltando aquellas que resulten dominantes o excepcionales.

Muchas de las técnicas que se utilizan en el minado de datos han existido por varios años y han tenido su origen en el mundo de la inteligencia artificial. Pero en nuestros días han comenzado a utilizarse en los grandes sistemas de bases de datos.

Durante los últimos 15 o 20 años, las computadoras han sido utilizadas para capturar información detallada de transacciones en una gran variedad de empresas, incluyendo empresas de telecomunicaciones, bancos y otras con un alto volumen de operaciones transaccionales.

Estos sistemas transaccionales están diseñados para capturar información detallada acerca de cualquier aspecto del negocio y tienen capacidad para realizar una gran cantidad de operaciones por minuto gracias a los avances de la tecnología de bases de datos y de arquitectura de computadoras. La necesidad de contar con información ha provocado la proliferación de data warehouses que integran información de sistemas de operación múltiples y distintos para apoyar la toma de decisiones. Además, por lo general incluyen datos de fuentes externas tales como datos demográficos de clientes y proveedores.

Recientemente, se ha generalizado la adopción de sistemas escalables y abiertos. Esto incluye sistemas de administración de bases de datos, herramientas de análisis y, más recientemente, intercambio de información y publicación entre distintos servicios. Estos factores han propiciado la existencia de una gran presión en la "cadena de valor" de la información. Del lado del origen, la cantidad de información almacenada en los data warehouse corporativos esta creciendo rápidamente. El "espacio de decisión" es demasiado complejo, existen demasiados datos y complejidad que puede ser relevantes a un cierto problema. En el lado final de la cadena, el conocimiento requerido por los tomadores de decisiones para definir el curso del negocio genera una gran tensión sobre los sistemas tradicionales de apoyo a la toma de decisiones.

7.2 TÉCNICAS Y HERRAMIENTAS PARA MINADO DE DATOS

Las aplicaciones para minado de datos pueden ser descritos en términos de una arquitectura de tres niveles. Estos niveles son: aplicaciones, enfoques y algoritmos y modelos. Los tres niveles se encuentran asentados sobre un repositorio de datos[14] .

7.2.1 Aplicaciones

Las aplicaciones de minado de datos pueden ser clasificadas en conjuntos de problemas que tienen características similares entre diferentes dominios de aplicación. La parametrización de la aplicación es distinta entre las distintas industrias y aplicaciones. Los mismos enfoques y modelos subyacentes utilizados para dar capacidad de detección de fraudes a un banco puede ser utilizado para desarrollar aplicaciones de detección de fraudes en seguros médicos. La diferencia se encuentra en la forma como los modelos son parametrizados, esto es cuales atributos del dominio específico se recuperan del repositorio y se utilizan en el análisis y la forma como son utilizados.

7.2.2 Enfoques

Cada clase de aplicaciones de minado de datos es apoyada por un conjunto de enfoques de algoritmos utilizados para descubrir las relaciones relevantes entre los datos y pueden ser: asociación, análisis en secuencia, "clustering", clasificación, y estimación.

Asociación: Estos enfoques se centran en una clase de problemas caracterizada por análisis de mercado - canasta. El análisis clásico de mercado-canasta trata la compra de un número de artículos (contenidos en una "canasta de compras") como una transacción única. La meta es encontrar tendencias a través de un gran número de transacciones que pueden ser utilizadas para comprender y explotar los patrones naturales de compra. Esta información puede utilizarse para ajustar inventarios, modificar la distribución de anaqueles o pisos, o introducir campañas promocionales dirigidas a ciertos sectores de consumidores. Estos enfoques tienen su origen en las empresas de ventas a menudeo, pero se pueden aplicar a empresas que desarrollan campañas publicitarias, sectores financieros, etc.

Análisis en secuencia: El análisis tradicional de mercado-canasta trata una colección de artículos como parte de una transacción en un momento dado. Una variante de este problema ocurre cuando existe información adicional para relacionar una secuencia de compras en una serie de tiempo. En este caso no sólo puede ser importante la coexistencia de los artículos dentro de la transacción, sino también el orden en el cuál dichos artículos aparecen dentro delas transacciones y la cantidad de tiempo entre cada una de ellas.

Clustering: Los enfoques de clustering se enfocan a problemas de segmentación. Estos enfoques asignan registros con una gran cantidad de atributos en grupos o segmentos relativamente pequeños. Este proceso de asignación es realizado de manera automática por algoritmos especializados que identifican las características sobresalientes del conjunto de datos y dividen el espacio n-dimensional definido para el conjunto de atributos dentro de ciertos límites. No es necesario identificar los grupos deseados o los atributos que se han de utilizar para realizar la segmentación de los datos.

El clustering es uno de los primeros pasos para el minado de datos. Identifica grupos o registros relacionados que pueden utilizarse como puntos de inicio para la exploración de otras relaciones. Esta técnica soporta el desarrollo de modelos de segmentación de poblaciones. Análisis adicionales, utilizando análisis estándar u otras técnicas de minado de datos pueden identificar características especiales de estos segmentos con respecto a una entrada deseada.

Clasificación: Esta es quizá la técnica de minado de datos más utilizada, en ella se emplea un conjunto de ejemplos preclasificados para desarrollar un modelo que puede abarcar a la población total de registros. Este tipo de análisis es particularmente adecuado para aplicaciones de análisis de riesgo y detección de fraudes. Este enfoque utiliza algoritmos de clasificación basados en arboles de decisión o en redes neuronales. El uso de los algoritmos de clasificación comienza con un conjunto preclasificado de transacciones de entrenamiento, el algoritmo de entrenamiento define los parámetros requeridos para una adecuada discriminación de las transacciones. El algoritmo codifica estos parámetros en un modelo llamado clasificador.

Una vez que se ha desarrollado un clasificador adecuado, se le puede utilizar en un modelo predictivo para clasificar nuevos registros dentro de las clases predefinidas.

Estimación: Una variación del problema de clasificación involucra la generación de marcadores entre varias dimensiones en los datos. Más que el uso de un clasificador binario este enfoque genera un marcador de aceptación basado en un conjunto preclasificado de entrenamiento.

Otras técnicas: Existen otros enfoques que son utilizados de manera conjunta con las técnicas anteriores incluyendo razonamiento de casos, lógica difusa, algoritmos genéticos. Cada una de estas técnicas tiene sus propias fortalezas y debilidades en función de su orientación, capacidad de discriminación, desempeño y necesidad de entrenamiento.

7.2.3 Algoritmos y Modelos

Este aspecto del minado de datos es atractivo para ejecutivos y profesionales de la informática quienes buscan obtener mayor rendimiento sobre grandes y complejos volúmenes de datos. Las herramientas actuales de minado de datos han evolucionado a partir de las investigaciones de inteligencia artificial en reconocimiento de patrones, pero en su mayoría presentan problemas ya sea en las interfaces de usuario, controles de ejecución y parametrización de modelos. Esta situación implica una cantidad importante de retos que pueden colocar a los usuarios en un conjunto de herramientas fragmentadas, lo que ocasiona que sea necesario realizar actividades de pre y post procesamiento de los datos para obtener resultados efectivos del minado de datos. Las actividades de pre-procesamiento incluyen la selección del subconjunto adecuado de datos por razones de desempeño y consistencia. El post-procesamiento involucra la subselección de resultados masivos y la aplicación de técnicas de visualización para proporcionar comprensión adicional del problema. Estas actividades son críticas para la efectividad en situaciones como:

Posibilidad de datos "sucios". Las herramientas de minado de datos no cuentan con un modelo de alto nivel de los datos sobre los cuales operan, no tienen estructuras orientadas a la aplicación (semántica) y, en consecuencia, simplemente recuperan datos que consideran correctos y determinan las conclusiones a partir de ellos. Los usuarios deben tomar las precauciones necesarias para asegurarse de que los datos que se están analizando están "limpios". Esto puede requerir de un análisis significativo de los valores de los atributos que son alimentados en las herramientas de minado. Sin embargo, si la empresa cuenta con un adecuado proceso de depuración de datos las herramientas de minado de datos pueden generar mayores beneficios derivados al analizar los datos.

Incapacidad de "explicar" los resultados en términos humanos: Muchas de las herramientas empleadas en el minado de datos utilizan algoritmos matemáticos de gran complejidad que no pueden ser fácilmente representados en términos humanos; incluso con métodos como árboles de decisión o reglas de inducción, el volumen y formato de la información puede ser de tal magnitud que se requiera de procesamiento adicional y/o visualización.

Representación de la información: La mayoría de las fuentes de datos para las aplicaciones actuales de minado de datos residen en grandes sistemas de bases de datos relacionales paralelos. La información típicamente se encuentra normalizada y los atributos que se utilizan en las herramientas de minado de datos pueden localizarse en varias tablas. Las herramientas de minado de datos operan sobre un conjunto de atributos presentados a través de un archivo plano, por lo general Unix, y se debe utilizar código condicional para proporcionar la representación no normalizada que estas herramientas necesitan.

Muchas de las herramientas de minado están restringidas por el tipo de datos con los cuales pueden operar. Otra limitación es la impuesta por Unix a los archivos planos de un máximo de 2GB, aunque este problema se puede superar con los sistemas operativos de 64 bits, pero en el caso de los 2 GB las herramientas de minado de datos se ven limitadas en el tamaño de los conjuntos de datos que pueden analizar.

Los sistemas paralelos de bases de datos relacionales almacenan los datos en múltiples discos y los accesan a través de múltiples CPUs. Las arquitecturas de las bases de datos actuales hacen que los conjuntos de resultados generados por el motor de la base de datos sean eventualmente ruteados por un proceso coordinador de una sola consulta. Esto puede ocasionar un cuello de botella en la utilización de bases de datos paralelas. Al extraer grandes conjuntos de datos, el procesarlos puede requerir de una gran capacidad de cómputo. Aunque muchas de las herramientas de minado de datos están diseñadas para operar sobre datos provenientes de sistemas de bases de datos paralelas, muchas de ellas no son paralelas en sí mismas. Este aspecto de desempeño es reducido mediante un muestreo del conjunto de datos de entrada.

Los usuarios deben asegurarse de que están obteniendo los datos representativos que les permitan aplicar los algoritmos de descubrimiento. Debido a que los propios algoritmos determinan cuales son los atributos importantes en el reconocimiento de patrones esto genera un escenario que puede requerir una solución iterativa.

En el caso de los algoritmos que requieren conjuntos de datos para entrenamiento, estos conjuntos deben cubrir de manera adecuada al total de la población. Una vez más, los usuarios se deben asegurar de utilizar los conjuntos de datos de entrenamiento adecuados que cubran al total de la población.

En nuestros días las herramientas de minado son fuertemente algorítmicas, pero requieren de una gran experiencia para ser implantadas de manera efectiva. Conforme maduren todas estas herramientas, y se logren más avances en la conectividad de los servidores, desarrollo de modelos de negocio y se mejoren las interfaces de usuario se colocará al minado de datos como una de las principales herramientas para la toma de decisiones.

La mayoría del software de extracción de datos actualmente trabaja en un mainframe o en las bases de datos basadas en archivos.

"La extracción de datos es el próximo progreso lógico en la cadena de bases de datos de soporte de decisiones y de análisis. En el low end usted tiene queries ad hoc y en el high end existe la extracción de datos" [Karen Rubenstruck, vicepresidenta en Meta Group].

Las herramientas de consulta tradicionales le permiten a los usuarios arrancar un conjunto de datos específico, aunque el software de extracción de datos usa un híbrido de análisis estadístico, redes nerviosas y lógica para reconocer patrones.

Por ejemplo, un analista de mercados puede registrar ciertos parámetros de búsqueda para identificar las tendencias del mercado, patrones de compra de los consumidores, o demandas de seguros fraudulentos.

La utilización de Data Warehousing combinada con la extracción de datos elimina la necesidad de limpiar y cerrar los datos antes de utilizar la Base de Datos.

Las técnicas de extracción de datos, a diferencia de los queries SQL ordinarios, son altamente sensitivas a los datos faltantes o inconsistentes. Los Data Warehouses, como el Red Brick Warehouse basado en bases de datos relacionales, normalmente importan los datos y los reformatean, o los limpian para ofrecer almacenes de datos consistentes y consolidados.

Pilot Software introducirá una aplicación integrada de extracción de datos que combina OLAP con técnicas de extracción de datos. El Pilot Discovery Server permite que los profesionales de ventas y mercadotecnia usen y analicen la información de sus clientes dentro de su medio Data Warehouse sin tener que extraer y reformatear la información.

El Pilot Discovery Server corre en el Microsoft SQL Server en Windows NT y en las bases de datos Oracle 7 en HP-UX.

7.4 VISUALIZACIÓN

Muchas compañías se han percatado de que los datos de transacciones que han recolectado durante la operación diaria de su negocio tiene un valor estratégico importante. Estos datos contienen información descriptiva acerca de varios eventos y temas que son la fuerza vital de la empresa. A través de la observación de patrones significativos y asociaciones, una compañía puede descubrir información con la cual puede emprender acciones para incrementar sus ganancias, reducir costos, o expandir sus mercados. Las nuevas correlaciones pueden conducir hacia nuevas estrategias de producto s que darán a la firma una ventaja competitiva.

De acuerdo al Gartner Group, cerca del 10 % de las grandes empresas existentes en EUA tienen proyectos de minado de datos en progreso. Las primeras empresas en adoptar técnicas de minado de datos han sido firmas del sector financiero, ventas y mercadotecnia, y mercadotecnia de bases de datos. Para estas firmas, el análisis automático, dirigido por reglas de los datos ofrece una alternativa más rápida a los métodos tradicionales de análisis estadístico.

Los avances en la tecnología están promoviendo también el desarrollo del minado de datos. Los avances en hardware y software apoyan esta tendencia haciendo más fácil y económico capturar y almacenar grandes cantidades de datos transaccionales.

El término minado de datos ha sido utilizado para referirse a una variedad de herramientas, en su más amplio sentido , minado de datos significa buscar patrones dentro de una colección de hechos u observaciones. El minado de datos automatizado es el descubrimiento de conocimientos utilizando una mezcla sofisticada de técnicas que van desde el análisis estadístico tradicional, la inteligencia artificial y los gráficos por computadora.

La materia prima para le minado de datos puede ser de aspectos como ventas, observaciones científicas, datos demográficos, datos de transacciones, etc. Dado que el tiempo es un aspecto importante, los datos históricos contenidos en los data warehouses es una fuente ideal de datos.

Los objetivos del minado de datos caen en tres categorías principales:

- *Explicativos:* Para explicar algunos eventos o mediciones observados.
- *Confirmativo:* Para confirmar una hipótesis.
- *Exploratorios:* para analizar relaciones inesperadas o nuevas entre los datos.

La construcción de modelos estadísticos a partir de los datos observados ha sido por mucho tiempo la técnica más utilizada para realizar predicciones útiles. Sin embargo, el análisis estadístico tradicional esta limitado por distintos aspectos. El análisis se vuelve cada vez más difícil a medida que crece el número de variables que deben ser incluidas. Por lo general se requiere limitar el numero de casos que se utilizarán en el análisis, en consecuencia, sólo una muestra de la gran base de datos esta disponible para un análisis exhaustivo. Finalmente, cuando no hay factores que interactúen dentro de los datos o cuando no existe una relación lineal entre los datos es difícil aplicar el método estadístico tradicional.

Algunas organizaciones comerciales tienen personal con el suficiente conocimiento de técnicas estadísticas para estructurar y realizar estos análisis e interpretar sus resultados. E incluso con expertos en estadística puede llevarse meses el diseño y construcción de modelos adecuados. Con la gran cantidad de datos y la velocidad a la que cambian es necesario tomar un nuevo enfoque.

Hasta hace poco tiempo, el minado de datos sólo era manejado por especialistas dentro de las organizaciones o dentro de firmas de consultoría. Las técnicas que se utilizan en las herramientas de minado de datos incluyen varias de las siguientes:

- *Razonamiento de casos:* Con esta técnica se derivan reglas a partir del análisis de situaciones o casos.
- *Descubrimiento de reglas:* El descubrimiento automático de reglas implica la ejecución de algoritmos de análisis de datos que recorren grandes cantidades de datos en busca de patrones o de correlaciones a partir de las cuales se pueden formular reglas. Esta búsqueda puede ser directa (para localizar datos que apoyen una regla ya propuesta) o no (permitiendo que los patrones que se detecten sugieran posibles reglas). IDIS es un producto de Information Discovery Inc, que permite descubrir relaciones entre datos y generar reportes en describiendo los resultados y sus niveles de confianza asociados. IDIS también puede descubrir anomalías en los datos que pueden ser el resultado de errores accidentales o de usos indebidos.

- *Marcadores:* Para utilizar la técnica de marcadores, los datos históricos deben ser analizados y se debe construir un árbol de decisión basado en un cierto conjunto de valores.
- *Procesamiento de señales:* Las técnicas de procesamiento de señales pueden identificar grupos de observaciones con características similares. Un producto llamado DataEngine (de la compañía alemana MIT GmbH) incorpora módulos que utilizan estas técnicas, así como lógica difusa y redes neuronales. Algunas organizaciones han utilizado DataEnigine para aplicaciones de control de calidad, proyecciones y segmentación de clientes.
- *Redes neuronales:* Las redes neuronales son modelos predictivos basados en principios similares a aquellos que rigen el cerebro humano. En una red de nodos (neuronas), cada nodo recibe una entrada y envía una salida a los nodos subsecuentes basándose en lo que recibió como entrada. La red es "entrenada" utilizando una muestra de datos para determinar los "pesos" adecuados para cada nodo. Se producen entonces valores específicos para los datos siguientes, separándolos de acuerdo a las categorías establecidas. Una vez que la red neuronal ha sido validada, puede ayudar a analizar y predecir eventos a partir de entradas de datos nuevos.

La visualización de datos se refiere a la representación de los datos en un formato gráfico[19]. Algunos beneficios de la visualización de datos son:

a) *Identificación de grupos/segmentos*: Las técnicas de minado de datos para el análisis de segmentos presenta sus resultados de manera numérica de acuerdo a mediciones de distancias y coeficientes de correlación. El ojo humano puede identificar grupos de forma más rápida cuando se encuentran representados de manera gráfica que cuando aparecen representados estadísticamente.
b) *Observación de patrones/tendencias*: De igual forma, para la detección de patrones o de tendencias -en especial cuando se involucran múltiples variables- las gráficas son superiores a las tablas de números. Las gráficas permiten una rápida identificación de áreas de interés o relevancia dentro de grandes bases de datos, que son estudiadas por los analistas utilizando técnicas de minado de datos o estadísticas.

c) *Comprensión de resultados*: A través de las gráficas, el significado de los datos se vuelve más evidente. Una representación gráfica realista explica más fácilmente el significado e impacto de los resultados del análisis a una audiencia conocimientos o experiencia en el área.

Algunas herramientas de visualización de datos representan a los datos en formatos que facilitan su despliegue y manipulación. dbExpress de Computer Concepts utiliza un método que permite a los usuarios seleccionar y desplegar datos, crear vistas de múltiples fuentes, etc. NetMap de Software AG utiliza un método similar que permite a los usuarios desplegar grandes cantidades de datos gráficamente y analizar sus relaciones de manera interactiva.

Algunos proveedores se han enfocado a algún tipo de aplicación específica en sus herramientas de visualización. Nielsen Spotligth de AC Nielsen se enfoca a la búsqueda de condiciones excepcionales en mercadotecnia y ventas para dar a los administradores una visión del desempeño de los productos. Visible Decisions describe a riskDiscovery y a marketDiscovery como la herramienta para mostrar el rendimiento de acciones a través del tiempo.

Advanced Visual Systems (AVS) provee una aplicación visual para la administración de riesgo en empresas de comercio de valores.

Conforme crece el tamaño de las bases de datos y se reconoce el valor potencial del conocimiento y las ventajas del análisis de datos, más organizaciones están buscando adquirir herramientas más sofisticadas de análisis de datos. La mayoría de las herramientas en esta área proporcionan soluciones especializadas para minado de datos, visualización, o ambas. La aplicación de este conjunto de herramientas es complejo y requiere de una dirección especializada., Sin embargo, las herramientas se están volviendo más fáciles de usar, automatizadas y aplicables a empresas tanto con aplicaciones OLTP como OLAP.

8. DATA WAREHOUSING

8.1 ANTECEDENTES

En 1987, las grandes bases de datos (VLDB -Very Large Data Bases-) existían casi exclusivamente sobre sistemas propietarios. De hecho, el mundo "abierto" comenzaba a tomar auge en los sistemas grandes. Casi todas las grandes bases de datos corrían en mainframes IBM con MVS y los productos compatibles de Amdhal y National Semiconductor (ahora Hitachi). La única alternativa real era Teradata, la cual ingresó al mercado en 1984 y para 1987 tenía aproximadamente 50 sistemas instalados. Teradata corría sobre una máquina dedicada a bases de datos empleando una arquitectura basada en microprocesadores paralelos e implementando el modelo relacional.

El recorrido de las grandes tablas podía ser realizado con mucha mayor velocidad en paralelo con Teradata que con la mayoría de los más poderosos mainframes. Esa es una de las causas por las cuales las bases de datos más grandes en 1987 eran de Teradata.

Aunque Teradata tuvo un mercado pequeño en 1987, fue el comienzo de un cambio importante. Por más de una década los mainframes de IBM habían sido el único lugar en el cual se podían colocar las grandes bases de datos comerciales. El haber logrado proporcionar desempeño interactivo en búsquedas complejas sobre grandes volúmenes de datos fue uno de los factores que influyó en el éxito de Teradata, pues ningún sistema relacional sobre un mainframe de IBM podía hacerlo.

El hecho de que comenzara a aparecer el término relacional en las grandes bases de datos fue otro fenómeno importante. La mayor parte del mercado de bases de datos para mainframe perteneció a IMS de IBM y a los cinco grandes productos de bases de datos independientes: Datacomm/DB de Applied Data Research, Model 204 de Computer Corp. of América, Total de Cincom, IDMS de Cullinet y ADABAS de Software AG.

Aunque algunas de estas compañías contaban con productos SQL ninguno había sido capaz de proporcionar un producto SQL de alto desempeño en grandes bases de datos.

En los mainframes de 1987, las grandes bases de datos para soporte a las decisiones estaban en Model 204 de CCA y las grandes bases de datos transaccionales en IMS.

Hace aproximadamente diez años, la tecnología relacional abarcó a las bases de datos pequeñas, pero su impacto aún era mínimo en las grandes bases de datos (VLDB). Para fines de 1987, la tecnología relacional comenzó a tomar importancia en el campo de las VLDB.

En primer lugar, Tandem Computers introdujo Nonstop SQL, una base de datos relacional diseñada para proporcionar un alto desempeño y procesamiento transaccional tolerante a fallas. Esto fue un desarrollo de gran relevancia, pues hasta entonces se había considerado que el modelo relacional sólo era útil para el soporte a decisiones. Tandem recalcó que muchos compradores habían asumido que necesitaban las estructuras de red de un IMS o un IDMS para un procesamiento transaccional eficiente, sin embargo Tandem demostró lo contrario.

Para entonces IBM lanzó DB2/2 y demostró los grandes avances que había logrado en su desempeño. El mundo de los sistemas a gran escala se estaba volviendo relacional; gran escala quiere decir el mundo de los mainframes.

De este modo para el final de 1987 la tecnología relacional había cubierto el mundo de las VLDB. DB2 había tomado el mercado de los mainframes, Teradata mostraba que el modelo relacional podía ser usado en bases de datos de apoyo a las decisiones y procesamiento paralelo, y Tandem demostraba que el modelo se podía utilizar en paralelo en grandes sistemas de procesamiento transaccional tolerantes a fallas.

En 1987 sucedieron cambios revolucionarios en el mundo de las bases de datos en sistemas a gran escala. Algunas empresas comenzaron a modificar sus esquemas para adaptarse a tales cambios. La versión 6 de Oracle fue diseñada para soportar un límite teórico de 32 terabytes, pero para mediados de los 80´s la compañía se movía hacia el manejo de bases de datos mucho mayores.

Mientras tanto, la relación entre nCUBE y Oracle se fue desarrollando para implantar una versión de MPP en Oracle. En 1991, Oracle y nCUBE mostraron el resultado de su benchmark de 1000 TPS (transacciones por segundo). Oracle 7 incluyó soporte a MPP.

A finales de los 80´s Sequent formó una alianza con Ingres para explotar su arquitectura SMP para aplicaciones de bases de datos. En 1991, Informix decidió reescribir su motor y estableció la estrategia para dar servicio al mercado de usuarios finales en Unix con una arquitectura paralela. Posteriormente Ingres fue adquirido por ASK y se separo de su sociedad con Informix. Informix introdujo Online Dynamic Server con Parallel Data Query en 1993. El MPP fue incluido en Informix XPS y distribuido comercialmente en 1996.

Estos cambios, combinados con el gran éxito del hardware para sistemas abiertos en paralelo condujo a una nueva generación en el mundo de las soluciones de grandes bases de datos. Con Informix y Oracle encabezando los cambios, Unix se convirtió en el ambiente preferido para las bases de datos medias a grandes. Hoy en día, tanto en data warehouse como en OLTP, las bases de datos más grandes se encuentran construidas sobre Unix con bases de datos relacionales sobre arquitecturas paralelas.

Otros grandes del mundo de las VLDB también realizaron cambios importantes. Sybase se enfocó más en la arquitectura del producto, eficiencia del procesamiento y escalabilidad en la carga de trabajo que en el tamaño de la base de datos. Posteriormente en sociedad con NCR, Sybase desarrolló su Navigation Server (ahora Sybase MPP) para plataformas MPP.

En inicio limitado a NCR 3600 Navigation Server comenzó a despuntar. Reintroducido con mejoras para las plataformas NCR WorldMark 5100 y la IBM RS/6000 SP - mejorado con avances en su desempaño en la línea de productos Sybase (como System 11 y Sybase IQ) - Sybase ha demostrado lo que es posible lograr con esta tecnología.

Teradata, pionera en las bases de datos paralelas se estableció como líder en plataformas de soporte a decisiones a gran escala para fines de los 80's. Para 1992 WalMart implemento la primera base de datos de 1 terabyte en Teradata. Para 1996 esta base de datos alcanzó los 4 TB, y NCR hizo demostraciones con una base de datos de 10 TB sobre Teradata.

En el mundo de los mainframes, DB2 se ha convertido en un DBMS de propósito general y en la principal plataforma para el desarrollo de data warehouses basados en mainframes. En el mundo de las VLDB el mainframe sigue siendo un ambiente fundamental para las implantaciones de gran escala. Se considera que hasta el momento aproximadamente el 25% de los data warehouses construidos se encuentran en mainframes y aproximadamente cerca del 70 u 80% de los datos comerciales aún residen en mainframes.

"Después de tres meses de utilizar un data warehouse, usted se dará cuenta que de todos los criterios que lo ayudaron a decidirse por la implantación de un data warehouse, el más importante es el desempeño". [Alan Paller, Director del Data Warehouse Institute]

8.2 DEFINICIÓN

Existen cuatro elementos importantes en el data warehousing que son:

1. La adquisición, transformación e integración de datos operacionales.
2. El sistema manejador de bases de datos.
3. Los sistemas de soporte a las decisiones del cliente.
4. Los sistemas de almacenamiento.

El almacenamiento de la información y la recuperación de la misma, puede ser el elemento más importante. Un sistema de almacenamiento y recuperación de información pobremente diseñado puede causar problemas de ejecución, costo, expansión y escalabilidad. Y por otra parte, un sistema bien diseñado puede ser considerado como inútil si los tiempos de respuesta son muy altos o no permite una futura expansión.

El objetivo principal de un proyecto de data warehousing es investir a los usuarios de toda aquella información que ha sido previamente inasequible. Un diseño de data warehouse común, es creado utilizando como base un modelo de datos de la empresa (EDM -Enterprise Data Model-). Sin embargo, un EDM sólo representa la mitad de lo que es necesario. El diseñador debe preguntar a los usuarios acerca de la información que requieren que se encuentre en el sistema, es decir, realizar un *análisis de requerimientos del negocio.*

"El objetivo principal de data warehousing, es la creación de una vista única de datos que pueden residir en bases de datos ubicadas físicamente en distintos lugares. Esto permite a los desarrolladores y directivos trabajar con un modelo de datos único." [The Butler Group].

El párrafo anterior define el objetivo del data warehouse, pero no define lo qué es un data warehouse. La definición de lo que es un data warehouse no resulta fácil, debido a que un data warehouse puede ser diferentes cosas para mucha gente; por lo que en ocasiones resulta más sencillo definir un data warehouse en términos de las propiedades que todo data warehouse comparte.

Podemos decir que un data warehouse es una base de datos cuyos datos han sido previamente seleccionados y limpiados del ambiente operacional. Mientras que el ambiente operacional ha sido optimizado para llevar a cabo el *procesamiento de transacciones*, el ambiente del data warehouse ha sido optimizado para el *procesamiento de consultas*. Un data warehouse manipula los datos en modo de sólo lectura, es decir, no se llevan a cabo modificaciones o borrados en él.

8.3 PROPIEDADES

De acuerdo a las propiedades del data warehouse, podemos definirlo como: Una colección de datos orientados al sujeto, integrados, variantes en el tiempo y no-volátiles; que apoyan el proceso de toma de decisiones.

8.3.1 Orientado al Sujeto

Esta primer característica del data warehouse significa que está orientado al mayor número de usuarios de la empresa.

Esto significa que un data warehouse almacena información acerca de los elementos que son importantes para la organización día a día. Esto está en contraste con los procesos clásicos orientados a sistemas, los cuales son desarrollados para mantener los tipos de datos de las transacciones diarias. Un data warehouse almacena información acerca de los elementos que son definidos implícitamente por el proceso orientado a datos.

En los procesos orientados a sistemas, los diseñadores se enfocan al diseño de la base de datos y al diseño de los procesos. Data warehousing se enfoca al modelado de datos y al diseño de la base de datos.

Los datos a los que los data warehouses y los sistemas orientados a procesos están enfocados son diferentes (especialmente a nivel de detalle). Un data warehouse únicamente necesita contener información que es importante para el proceso de apoyo a la toma de decisiones. Un sistema orientado a procesos necesita información que no es importante para la toma de decisiones. Otra diferencia importante entre los datos operacionales y los contenidos en el data warehouse se basa en la relación entre datos; los datos operacionales mantienen una relación prolongada entre dos o más tablas basadas en la regla del negocio que las afecta. El data warehouse abarca un determinado espacio de tiempo y las relaciones que se encuentran en el son muchas.

Varias reglas del negocio (y correspondientemente, varias relaciones) son representadas en el data warehouse entre dos o más tablas.

El ambiente operacional en una institución financiera, está diseñado con base en aplicaciones y funciones tales como préstamos, ahorros, tarjetas bancarias y créditos. El ambiente de data warehouse está organizado con base en sujetos de mayor importancia como son los clientes, proveedores, productos y actividades. La clasificación entre áreas de sujetos afecta el diseño y la implementación de los datos encontrados en el data warehouse.

8.3.2 Variante en el Tiempo

Los datos en el data warehouse son considerados conforme a un determinado momento en el tiempo, a diferencia de los datos operacionales que son tomados en el momento de acceso, es decir, en el ambiente operacional cuando se accesa a una unidad de datos, se espera que éste refleje los valores tomados en el momento del acceso. Los datos de un data warehouse son una toma instantánea en algún momento de los datos operacionales. Esto implica que los datos contenidos en el data warehouse no pueden ser modificados.

Debido a que los datos en el data warehouse son tomados con base en un determinado momento (no "justo en el momento"), se dice que estos datos son "variantes en el tiempo". La siguiente figura ilustra lo anterior:

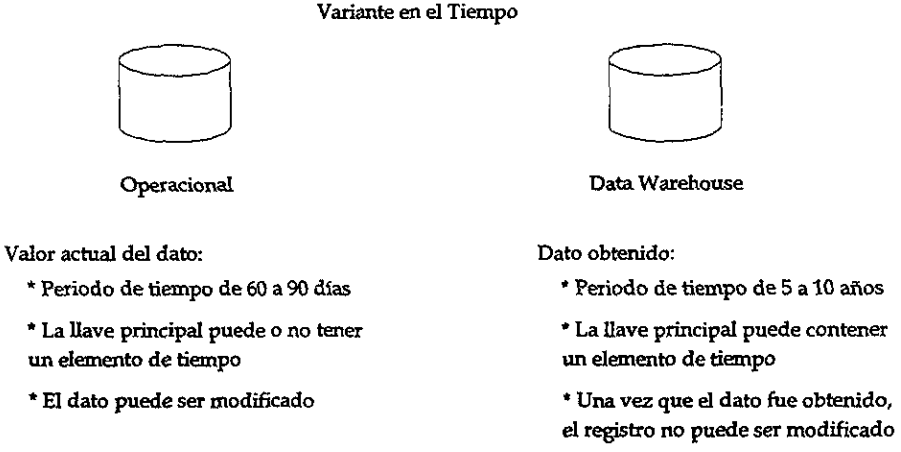


Fig. 3 Propiedad de variación en el tiempo de un data warehouse

La variación en el tiempo de los datos de un data warehouse se identifica de diferentes formas, la más simple de ellas es que los datos de un data warehouse representan datos de un largo periodo de tiempo (de cinco a diez años; el periodo de tiempo representado en el ambiente operacional es mucho más corto).

Las aplicaciones que deben ejecutarse y que deben estar disponibles para el procesamiento de transacciones debe llevar la cantidad mínima de datos para así tener un mayor grado de flexibilidad. Por lo tanto, las aplicaciones operacionales manejan periodos de tiempo cortos.

La segunda forma de identificar la variación en el tiempo, es a través de la estructura principal. Cada estructura principal en el data warehouse contiene (implícita o explícitamente) un elemento de tiempo, tal como día, semana, mes, etc. El elemento de tiempo se encuentra casi siempre al principio de la llave que se encuentra en el data warehouse. En ocasiones, el elemento de tiempo existirá implícitamente, tal como es el caso en que un archivo completo está duplicado al final del mes, o a la mitad.

La tercer forma en que la variación en el tiempo aparece, es que una vez que el dato del data warehouse ha sido correctamente grabado, no puede ser modificado. El dato contenido en el data warehouse es, para propósitos prácticos, una gran serie de obtenciones. Si el dato obtenido ha sido tomado incorrectamente, entonces la obtención puede ser cambiada. Ahora, si la obtención ha sido realizada correctamente, no puede ser alterada. En algunos casos una modificación puede ser ilegal o no ética. Los datos operacionales pueden ser modificados.

8.3.3 Integrado

Todos los datos en un data warehouse tienen un alto grado de integración. Esta integración se muestra de distintas maneras - principalmente en la consistencia de los elementos en el data warehouse. Por ejemplo, nombres, mediciones de variables, convenciones, atributos físicos, etc. -

Uno de los aspectos más importantes del ambiente del data warehouse es que los datos que se encuentran dentro de él están integrados. Siempre. Sin excepciones.

En contraste con las integración que se encuentra en el data warehouse está la falta de integración que se encuentra en el ambiente de aplicaciones, y las diferencias son severas, tal como se muestra en la figura 4.

A través de los años, los diseñadores de diferentes aplicaciones han tomado un gran número de decisiones individuales acerca de cómo debe ser construida una aplicación. Las decisiones acerca del estilo y el diseño individualizado del diseñador se producen en diferentes formas; diferencias en codificación, en estructuras, en características físicas, en convenciones para el nombramiento de variables, etc. La habilidad colectiva de varios diseñadores para crear aplicaciones consistentes es legendaria. La figura anterior muestra algunas de las diferencias más importantes en las que una aplicación puede ser diseñada.

Codificación: Los diseñadores pueden codificar el campo GÉNERO de diferentes maneras. Un diseñador representa GÉNERO como una "M" y una "F". Otro, diseñador lo representa con un "1" y un "0"; otro, como una "x" y una "y". Y así, otro diseñador lo representa como "masc" y "fem". La forma en que GÉNERO llegue al data warehouse no tiene importancia. "M" y "F" son tan buenos como cualquier otra representación. Lo que importa, es que no importa de dónde provenga GÉNERO, éste debe llegar al data warehouse en un estado consistente e integrado.

Por consiguiente, cuando GÉNERO es cargado al data warehouse desde alguna aplicación donde éste ha sido representado en un formato diferente a "M" y "F", el dato debe ser convertido al formato del data warehouse.

Medición de atributos: Los diseñadores han elegido medir pipeline de varias formas a través del tiempo. Un diseñador, almacena el dato pipeline en centímetros; otro diseñador lo almacena en términos de pulgadas, un tercer diseñador lo almacena en pies; y un último diseñador, almacena la información de pipeline en yardas. Cualquiera que sea la fuente, cuando la información de pipeline llega al data warehouse, éste necesita que la medida sea la misma.

Tal y como se muestra en la figura 4, los elementos de integración afectan a casi todos los aspectos de diseño (las características físicas de los datos, el dilema que surge al tener más de una fuente de datos, la inconsistencia en el nombramiento de estándares, la inconsistencia en el formato de fechas, etc.).

Cualquiera que sea el diseño, el resultado es el mismo; los datos necesitan ser almacenados en el data warehouse de manera singular, formalmente aceptable siempre y cuando los sistemas operacionales almacenen los datos de manera diferente.

Al analizar el data warehouse, el analista debe enfocarse al uso que tendrán los datos en el data warehouse más que preguntarse acerca de la credibilidad o consistencia de los datos.

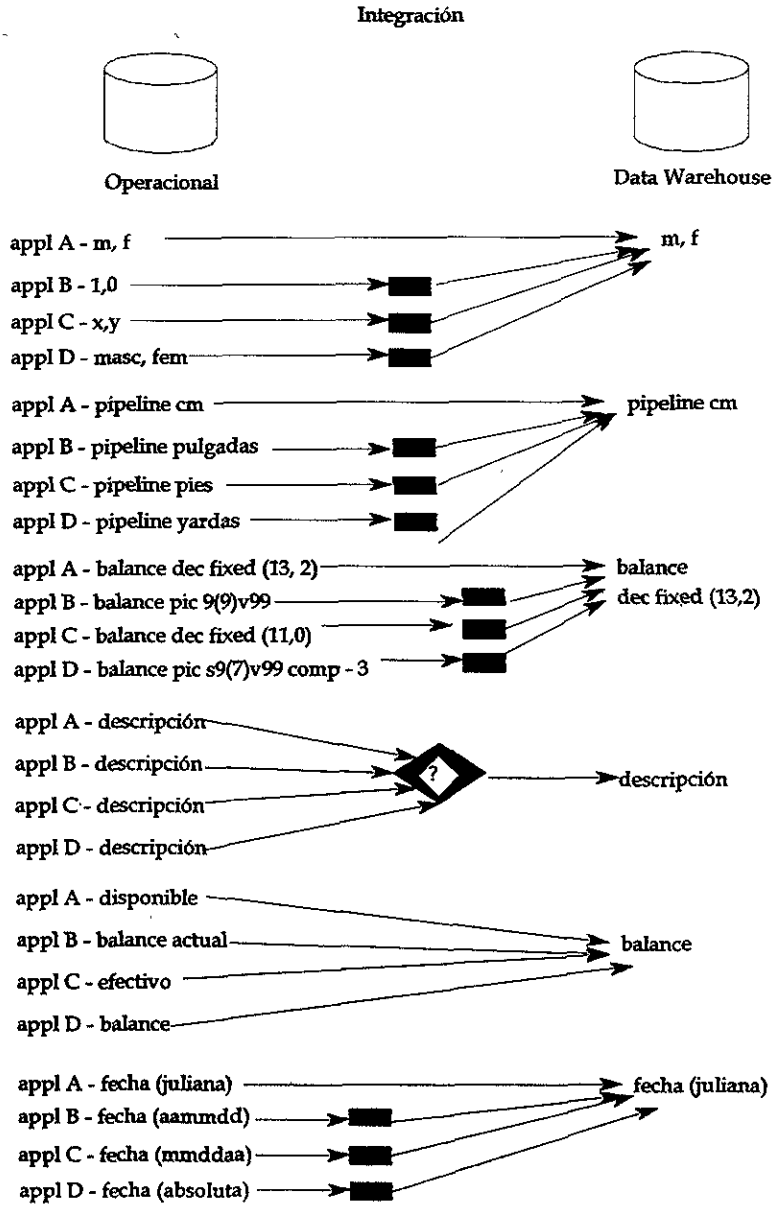


Fig. 4.- Propiedad de integración de un data warehouse

8.3.4 No-Volátil

Los data warehouses sólo permiten dos operaciones básicas: la carga inicial de los datos y el acceso a esos datos en modo de sólo lectura (una vez cargados). Esto significa que la funcionalidad de un data warehouse es totalmente diferente de los sistemas operacionales y por consiguiente los requerimientos del DBMS para esos dos tipos de sistemas son diferentes. Un data warehouse no necesita preocuparse por elementos tales como un deadlock o modificaciones registro por registro, por nombrar algunos.

Casi todos los datos de un data warehouse provienen del ambiente operacional. Los datos no se obtienen directamente, son filtrados y modificados de modo que se cubran las necesidades del data warehouse. Dichos datos se mantienen en el data warehouse hasta que se decida que no tienen relevancia o que deben ser modificados.

La siguiente figura ilustra esta característica del data warehouse:

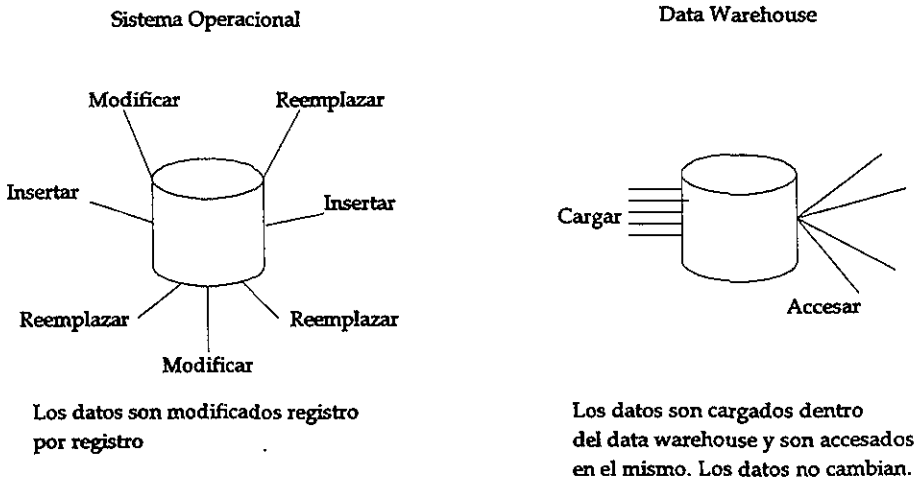


Fig. 5.- Propiedad de no volatilidad de un data warehouse

Como se puede observar en la figura anterior, las modificaciones (inserciones, borrados y reemplazos) se realizan regularmente en el ambiente operacional registro por registro. Sin embargo, la manipulación básica de datos que ocurre en el data warehouse es mucho más simple, ya que como se ha mencionado, en él sólo existen dos tipos de operaciones. Existen algunas consecuencias poderosas de esta diferencia básica entre el procesamiento operacional y el procesamiento del data warehouse. En el nivel de diseño, la necesidad de ser cuidadoso con la existencia de modificaciones no existe en el data warehouse, ya que las modificaciones no están permitidas. Esto significa que en el nivel de diseño físico, se tiene libertad para optimizar el acceso a los datos, particularmente tratándose con los elementos de normalización.

Otra consecuencia de la simplicidad de las operaciones del data warehouse se encuentra en la tecnología utilizada para ejecutar el ambiente del data warehouse. Tener que soportar las modificaciones registro por registro en línea (como es a menudo el procesamiento operacional) requiere que la tecnología tenga un fundamento muy complejo bajo una fachada de simplicidad. La tecnología debe soportar respaldos y recuperación, integridad de datos y transacciones, y la detección y corrección de deadlocks. Lo cual es innecesario para en procesamiento del data warehouse.

Las características del data warehouse (orientado al sujeto, integración de datos dentro del data warehouse, variante en el tiempo, y la simplicidad en la administración de datos) conducen a un ambiente que es muy diferente al clásico ambiente operacional.

La fuente de casi todos los datos del data warehouse, es el ambiente operacional. Esto da a pensar que se trata de una redundancia masiva de datos entre los dos ambientes. De hecho, existe un mínimo de redundancia entre los datos del ambiente operacional y del ambiente del data warehouse.

Hay que tener en consideración que:

- Los datos son filtrados del ambiente operacional al data warehouse. Muchos datos nunca pasan fuera del ambiente operacional. Únicamente los datos que son necesarios en el procesamiento de toma de decisiones se encuentran en el ambiente del data warehouse.
- El periodo de tiempo de los datos, es muy diferente de un ambiente a otro. Los datos en el ambiente operacional están muy frescos; los datos en el data warehouse son mucho más viejos. Desde la perspectiva de tiempos, existe una pequeña franja entre el ambiente operacional y el de data warehouse.
- El data warehouse contiene un resumen de datos que nunca se encuentra en el ambiente operacional.
- Los datos padecen una transformación fundamental al pasar dentro del data warehouse. La figura 5, muestra que los datos son alterados significativamente al ser seleccionados para formar parte del data warehouse. No es el mismo dato el que reside en el ambiente operacional al que se tiene desde el punto de vista de integración.

De acuerdo a estos factores, la redundancia de datos entre ambos ambientes es una rara ocurrencia, resultando menor a 1% de redundancia entre ambos.

8.4 ESTRUCTURA

El data warehouse consta de cinco tipos de datos (como puede observarse en la siguiente figura).

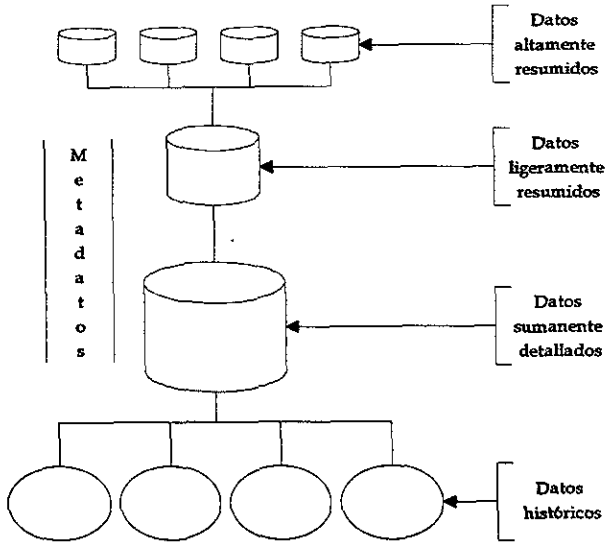


Fig. 6.- Estructura de un data warehouse

Metadato:

a) Puede ser utilizado como un indicador de lo que está siendo almacenado en la base de datos y de cómo los datos han sido modificados para ser incluidos en el data warehouse.

b) Contiene los algoritmos utilizados (la estructura de los datos) en el proceso de simplificación de datos, es decir, el mapeo del ambiente operacional al data warehouse.

Datos sumamente detallados:

a) Generalmente almacenados en dispositivos de almacenamiento de gran velocidad.

b) Refleja lo sucedido recientemente.

c) Contiene una gran cantidad de datos que son almacenados en un bajo nivel de granularidad.

Datos históricos:

a) No son accesados frecuentemente, por lo que generalmente son almacenados en algún dispositivo de almacenamiento masivo.

b) En cierto momento, los datos que se encuentran detallados, son movidos a los datos históricos.

Datos ligeramente resumidos:

a) Generalmente se almacenan en un disco.

b) Son obtenidos de los datos detallados.

Datos altamente resumidos:

a) Generalmente se almacenan en un disco.

b) Son fácilmente accesibles y compactos.

c) Son obtenidos de los datos detallados actuales y de los datos ligeramente resumidos.

8.5 ARQUITECTURA E INFRAESTRUCTURA DE UN DATA WAREHOUSE

Al momento de decidir el inicio de un proyecto de data warehouse lo primero que debemos hacer para aclarar algunos de los errores de comprensión asociados a la construcción de un data warehouse es definirlo en términos de su arquitectura y de su infraestructura.

Una arquitectura es un conjunto de reglas o estructuras que proporcionan un marco de referencia para el diseño completo de un sistema o producto. Uno de los primeros componentes de la arquitectura de un Data warehouse es una base de datos de sólo lectura utilizada para el soporte a las decisiones.

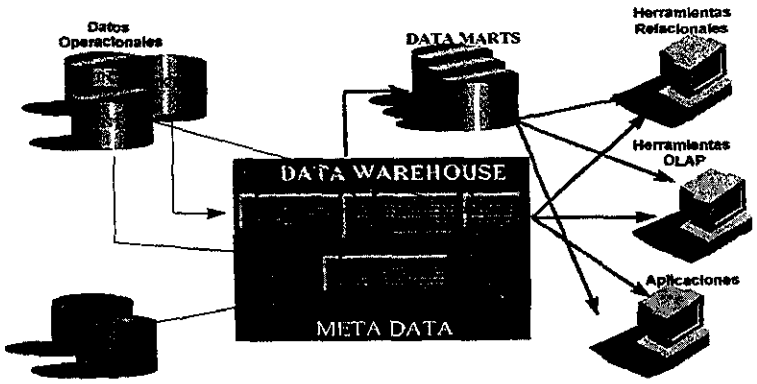


Fig. 8.- Arquitectura de un data warehouse

De la figura anterior podemos distinguir las siguientes características:

1. *Los datos son extraídos de sistemas, bases de datos o archivos fuente:* En la mayoría de las compañías los sistemas propios son la fuente dominante de datos. Otras fuentes de datos pueden ser aquellas compradas a compañías que se especializan en proporcionar datos. Estas fuentes pueden estar en diferentes formatos o diferentes medios de modo que puede o no ser necesaria la extracción selectiva de los datos.

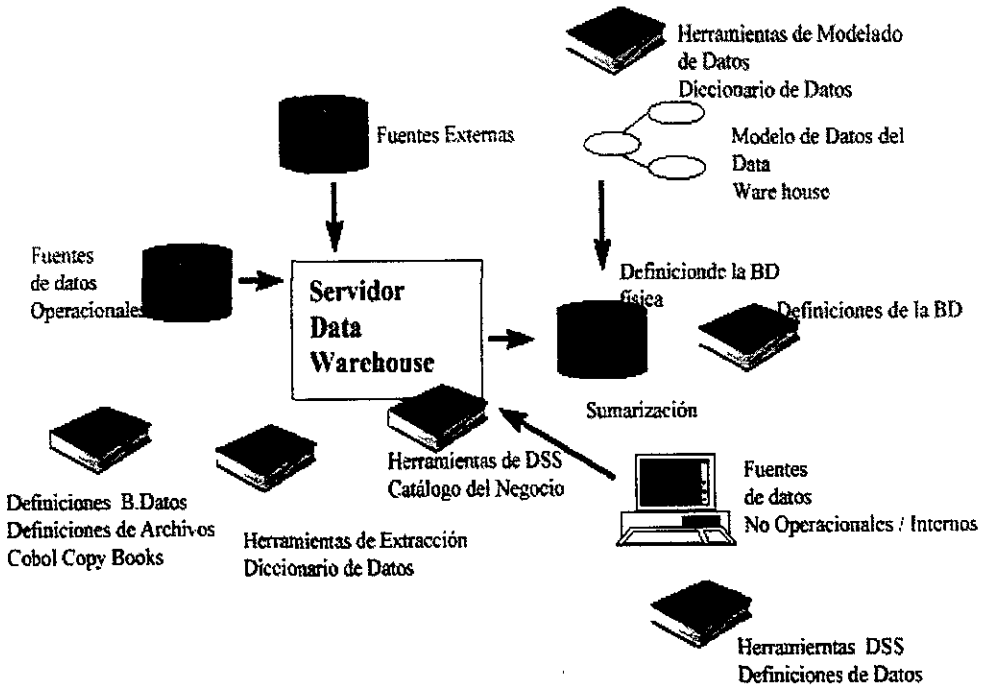


Fig. 7.- Arquitectura de un data warehouse

2. *Los datos provenientes de las fuentes son integrados y transformados antes de ser cargados en el data warehouse:* Esta característica es muchas veces subestimada. Si los datos provienen de múltiples sistemas, bases de datos y plataformas, alguna forma de integración o transformación será necesaria.
3. *Los datos para el soporte de decisiones residen en una base de datos separada y de sólo lectura:* Inherente a la idea del data warehouse existe la idea de que el procesamiento operativo y el procesamiento para toma de decisiones son fundamentalmente diferentes. El primero se refiere a la ejecución diaria del negocio, estos sistemas por lo general son transaccionales que reciben, actualizan y almacenan los datos generados por los sistemas del negocio. El objetivo del procesamiento para toma de decisiones es proporcionar información analítica para apoyar en la formulación de decisiones de negocio estratégicas y tácticas. El proceso de toma de decisiones requiere un rango de datos históricos para análisis comparativo de manera que los usuarios sean capaces de reconocer tendencias y patrones de información en el tiempo. Las diferencias fundamentales entre la funcionalidad de los sistemas operativos y de toma de decisiones requieren de estilos substancialmente diferentes de diseño de las bases de datos.
4. *El acceso de los usuarios al Data warehouse se realiza a través de un front-end o alguna aplicación:* Con la mayoría de la Data warehouses la línea frontal del sistema es el ambiente de acceso a los datos.

Separando estas características básicas de la arquitectura de un Data warehouse podemos obtener una mayor comprensión del papel que cada una jugará en el sistema completo. Esta descripción cubre los elementos "genéricos" o básicos del Data warehouse. Debido a que no existe una "forma correcta" para implantar un data warehouse, los elementos genéricos deben ser siempre parte de su arquitectura para funcionar de manera correcta.

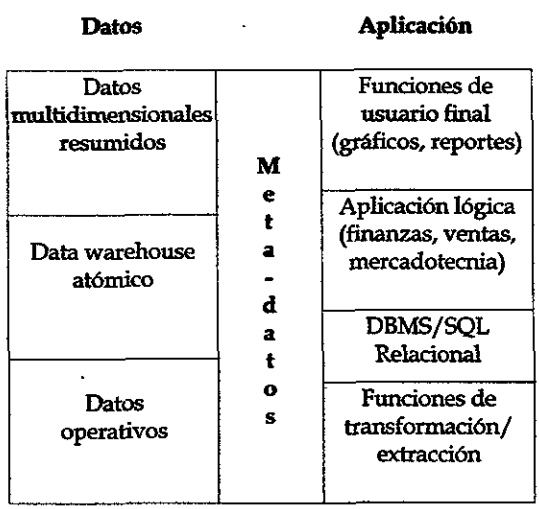
Parte de la tarea de construir una data warehouse es combinar sus elementos genéricos con la arquitectura de los sistemas actuales. Sin embargo, integrar el data warehouse con la arquitectura de los sistemas actuales puede resultar más complicado de lo que parece. En muchas compañías las arquitecturas de los actuales sistemas de procesamiento de datos son en extremo complejas y sofisticadas. Estos factores nos conducen a consideraciones técnicas, de interoperabilidad, estrategia y política.

Construir un data warehouse es un proceso para encontrar la solución técnica correcta de acuerdo a las necesidades de apoyo a las decisiones y la creación de una arquitectura sólida dentro de los parámetros con los que se debe trabajar. Mientras que algunas arquitecturas pueden parecer mejores que otras, también algunas serán más difíciles de implantar que otras. Sin embargo, la arquitectura de data warehouse seleccionada debe ser la solución técnica más apropiada de acuerdo a las metas de la organización, restricciones de arquitectura y requerimientos de apoyo a las decisiones.

La infraestructura se refiere a las plataformas, bases de datos, gateways, redes, herramientas y otros componentes necesarios para hacer que la arquitectura funcione. La capacitación en estas tecnologías también debe ser considerado parte de la infraestructura.

Mientras que la arquitectura y la infraestructura están íntimamente relacionadas, una arquitectura actualmente puede requerir de diferentes infraestructuras, dependiendo del ambiente particular de cada organización. La forma como las organizaciones establezcan su infraestructura puede se diferente dependiendo de tiempo, presupuestos y planes estratégicos, sin embargo un data warehouse requiere que la infraestructura se encuentre en su lugar.

Para colocar a las herramientas disponibles bajo observación, primero debemos aclarar algunos aspectos del ambiente del data warehouse, en el cual reside la mayor parte de los datos multidimensionales. Actualmente, el warehouse tiene dos arquitecturas paralelas: la arquitectura de los datos y la arquitectura de la aplicación.



El nivel base de la arquitectura de datos esta constituido por los datos de operación a partir de los cuales las instancias detalladas de los datos son recopiladas. Por lo general este nivel se forma con aplicaciones heredadas que se utilizan como proveedores del data warehouse. Después tenemos el nivel de detalle más bajo del data warehouse - llamado en ocasiones almacén de datos "atómicos" o de datos de operación (ODS -Operational Data Store-). Este nivel representa datos extraídos de los sistemas en producción, posiblemente recodificados o transformados de alguna forma para garantizar su exactitud y consistencia, y almacenados en un formato relacional. Los niveles siguientes de la arquitectura de datos son resúmenes multidimensionales de vistas, las cuales permiten que los datos sean desplegados y procesados de acuerdo a diversas dimensiones de descripción. Los metadatos - o información acerca de las entidades, relaciones y dimensiones descriptivas - se requieren para apoyar la operación del data warehouse en todos sus niveles.

La arquitectura de la aplicación comienza con una base en la que se encuentran funciones de transformación utilizadas para convertir los datos operativos en datos atómicos dentro del data warehouse.

Después vienen las funciones de búsqueda y de administración tradicionales de bases de datos relacionales. Dado que el data warehouse es actualizado en forma periódica, en todo o en parte, el énfasis en este nivel se hace en la adecuada selección del RDBMS. El siguiente nivel de aplicación consiste en funciones de soporte específicas del negocio, tales como herramientas estadísticas, de análisis financiero, ventas y mercadotecnia, presupuestación y proyección, etc.

El último nivel de aplicación se constituye de aquellas funciones más relacionadas con las actividades de los usuarios finales: selección específica para búsquedas, despliegue de texto e imágenes, etcétera.

8.6 DATA WAREHOUSE Y TECNOLOGÍA PARALELA

En la práctica, se ha detectado que existen empresas que están planeando establecer data warehouses, la mayoría de ellas comenzarían con data warehouses de aproximadamente 50Gb., sin embargo esperan que esta capacidad sea sobrepasada a gran velocidad.

El reto técnico de los data warehouse va más allá del manejo de grandes volúmenes de datos. También tiene que ver con el incremento en el número de usuarios, quienes esperan tiempos de respuesta razonables en búsquedas que pueden incluir algunos renglones o millones de ellos. Se espera que los data warehouses residan en múltiples plataformas distribuidas en distintos sistemas manejadores de bases de datos.

Estas características de los data warehouses han hecho que las organizaciones busquen plataformas de hardware y software que sean escalables y capaces de proporcionar un rendimiento y un tiempo de respuesta aceptables. Agregado a estos factores se encuentran las consideraciones típicas de precio/rendimiento, viabilidad del proveedor, soporte técnico, incremento de la productividad, paquetes de aplicación y coexistencia con el ambiente existente.

Escalabilidad es:

- La habilidad de agregar poder de procesamiento, memoria, discos, y otros componentes críticos de hardware de acuerdo a como cambien las necesidades de la organización, mientras se mantenga la misma, o una mejor, relación precio/desempeño.
- Una solución que proporciona una relación lineal entre recursos consumidos/requeridos y una creciente carga de trabajo (de usuarios, búsquedas y volumen de datos), siempre y cuando se mantengan tiempos de respuesta consistentes.
- La habilidad de administrar un ambiente cada vez mas complejo con un número mínimo de personal especializado y con herramientas que manipulen recursos cada vez mayores en diversidad y tamaño.

El fenómeno del data warehouse ha traído nueva vida al uso del procesamiento paralelo masivo (MPP -Massively Parallel Processing-) para propósitos comerciales. Muchos vendedores de esta tecnología, como Kendall Square Research, Thinking Machines, y nCUBE han despuntado y abierto oportunidades para productos como IBM RS/6000 SP, ICL Goldrush MegaServer, Tandem Himalaya, Pyramid Reliant RM1000, Unisys OPUS, AT&T 5100M, Maspar Decision Series, y White Cross WX 9000 Series.

Algunas de estas maquinas, tales como Tandem Himalaya, Maspar Decision Series, y White Cross WX 9000 Series, soportan sistemas manejadores de bases de datos de búsquedas paralelas propietarios, mientras que otros soportan sistemas manejadores de bases de datos de otros fabricantes como Oracle, Sybase, IBM e Informix, quienes están vendiendo extensiones a sus DBMS relacionales para explotar la tecnología MPP.

La situación actual de la tecnología de soporte a los DBMS relacionales para manejo de MPP es como sigue:

Oracle 7.1 esta disponible en distintas plataformas de MPP, incluyendo IBM RS/6000 SP, nCUBE, Maiko Computing Surface, AT&T GIS 3600 e ICL Goldrush MegaServer. y Oracle ofreció que su versión 7.3 tendría procesadores con afinidad para explotar arquitecturas de MPP.

IBM DB2 PE (Parallel Edition) solo esta disponible para la plataforma IBM RS/6000 SP.

Sybase MPP esta disponible para plataformas AT&T GIS 3600, además de versiones en pruebas para IBM RS/6000 SP y plataformas HP y SUN.

Informix versión 8.0 XPS se prueba para plataformas IBM RS/600 SP, ICL Goldrush MegaServer y AT&T GIS 3600, además de otras plataformas proyectadas entre las que se encuentran Sequent y Pyramid.

8.7 DISEÑO

8.7.1 Modelo Multidimensional

Conforme madura la arquitectura de los data warehouses se está desarrollando un modelo básico para el diseño de bases de datos de soporte a la toma de decisiones, este modelo se conoce como modelo multidimensional.

El modelo multidimensional representa a los datos como un arreglo en el cual cada dimensión es un aspecto del negocio que se esta analizando. El tiempo es siempre una de las dimensiones, las otras dependen del problema de negocio que se estudie.

En un arreglo multidimensional, cada celda representa la intersección de sus dimensiones, lo cual define el alcance del análisis.

Los seguidores del modelo consideran que las bases de datos diseñadas a través de él son más fáciles de utilizar por el usuario final. Esta apreciación es correcta si se considera que el modelo relacional fue desarrollado cuando se hacía mayor énfasis en las necesidades de los sistemas de operación de procesamiento de transacciones en línea (OLTP -On Line Transactional Processing-) donde la norma es cumplir con un requerimiento de tiempo de respuesta para transacciones volátiles.

En una aplicación OLTP, las pantallas y reportes predefinidos mantienen los detalles de la base de datos normalizada ocultos para el usuario final. Nunca se permite que el usuario final tenga acceso a la base de datos de manera directa.

Sin embargo, en un ambiente de data warehouse, el objetivo principal es proporcionar acceso directo a los usuarios finales. Dada esta situación, un data warehouse no puede ser actualizado por los usuarios finales. El data warehouse es una colección de datos orientado a la tarea, integrado, variante en el tiempo y no volátil para el soporte a la toma de decisiones operando en un ambiente de procesamiento analítico en línea (OLAP -On Line Analytical Processing-).

Desarrollar un modelo normalizado es un paso esencial en el diseño de un data warehouse eficiente. Un modelo no normalizado es difícil de mantener cuando el sistema crece, además reduce la flexibilidad necesaria para soportar múltiples perspectivas.

Algunos autores han sugerido que para el desarrollo de un data warehouse eficiente se debe seguir una metodología que involucre:

- Construir el modelo lógico de la información disponible en la aplicación fuente.
- Trabajar con los directivos y analistas del negocio para determinar que porción de la información es útil en el data warehouse. Esta porción es la información del negocio necesaria para hacerlo funcionar.
- Construir un modelo que identifique las dimensiones.

- Trabajar con los analistas del negocio que proporcionen reportes administrativos a los directivos para determinar el nivel de agregación requerido, la frecuencia de carga de datos y el número de ciclos de mantenimiento del data warehouse.
- Construir el data warehouse piloto y determinar su efectividad en el cumplimiento de los requerimientos.

La evolución de un data warehouse es de naturaleza iterativa. Los analistas de negocios y los directivos se dan cuenta de que necesitan nueva información conforme se familiarizan con el uso de las facilidades del data warehouse. Los requerimientos para el data warehouse pueden impactar los sistemas de operación, conforme los usuarios determinan que requieren información que no se encuentra actualmente en tales sistemas.

No sólo los administradores de datos tendrán que adaptarse a los requerimientos del ambiente de soporte a las decisiones. La naturaleza iterativa del desarrollo del ciclo de vida del data warehouse establecerá demandas en la organización de la tecnología de información para establecer mejoras a los sistemas en un grado que no se había visto anteriormente. La tecnología de información deberá establecer un proceso de control de cambios que permita que las mejoras al data warehouse sean llevadas a cabo con mayor rapidez que con los procedimientos tradicionales de control de cambios.

Tanto el modelo relacional como el modelo multidimensional tienen un papel importante dentro de la administración de los datos corporativos. El modelo lógico proporciona los fundamentos para el diseño de las bases de datos; el modelo multidimensional, el cual representa una vista del modelo conceptual, es optimizado para facilitar el uso del data warehouse por parte de los usuarios finales.

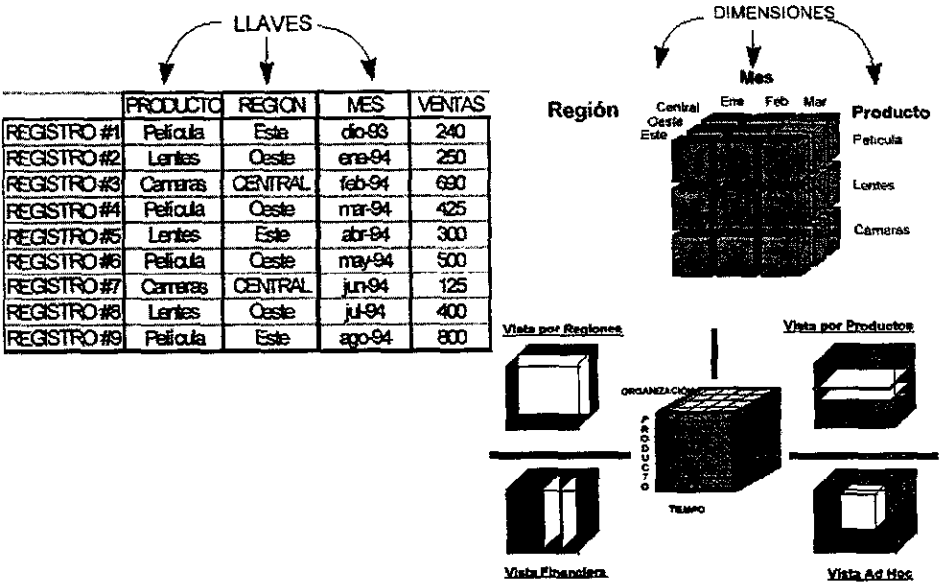
8.7.2 Procesamiento Analítico en Línea (OLAP -On Line Analitical Processing-)

Hablar de procesamiento analítico en línea por lo general nos hace pensar en agregación, análisis y reorganización de vistas. Estas nociones se encuentran contenidas en las herramientas OLAP, ya sea que operen separadas del almacén de datos o como un nivel encima de una base de datos SQL. Las herramientas OLAP son por lo general referenciadas dentro del contexto de los data warehouses como una parte de acceso a los datos o data marts.

El problema de pensar en las herramientas OLAP aisladas de otras aplicaciones analíticas es que OLAP debería proporcionar el núcleo estructural de los datos -llamado estructurado dimensional- que es vital en cualquier análisis. El análisis de series de tiempo, regresiones, proyecciones, descubrimiento de conocimiento basado en reglas, y otros análisis se pueden ver beneficiados por el estructurado dimensional. OLAP permitiría mejorar las aplicaciones analíticas existentes. Para lograrlo, nociones analíticas como tipo, variable, e independencia entre dimensiones deben ser definidas con claridad dentro del ambiente de OLAP.

Cualquier cosa que se pueda rastrear dentro de una situación es una variable o medida. El conjunto de todas las variables que se están rastreando para una situación en particular es llamado dimensión de las variables. A menos que las instancias de las variables tengan etiquetas que las identifiquen, resultan completamente inútiles.

Es sabido que nunca se daría seguimiento a una serie de variables sin tener una forma de identificarlas en un momento dado. Cada conjunto de factores de identificación es una dimensión de identificadores de una situación. Una dimensión de identificadores es un conjunto de factores de identificación de una variable de un mismo tipo. En contraste con las variables, las dimensiones del identificador tienen valores que ya son conocidos.



	PRODUCTO	REGION	MES	VENTAS
REGISTRO #1	Película	Este	dic-93	240
REGISTRO #2	Lentes	Oeste	ene-94	250
REGISTRO #3	Cámaras	CENTRAL	feb-94	690
REGISTRO #4	Película	Oeste	mar-94	425
REGISTRO #5	Lentes	Este	abr-94	300
REGISTRO #6	Película	Oeste	may-94	500
REGISTRO #7	Cámaras	CENTRAL	jun-94	125
REGISTRO #8	Lentes	Oeste	jul-94	400
REGISTRO #9	Película	Este	ago-94	800

Fig.9.- Organización de un data warehouse

8.8 ARQUITECTURA DE REFERENCIA

La *construcción de un data warehouse se encuentra representada por una relación* entre bloques (adquisición de datos, almacenamiento de datos, distribución de datos y acceso a datos) y capas (indicadores del entorno y del negocio, administración del metadato, transporte e infraestructura y herramientas, tecnologías y roles). Los bloques hacen referencia a la funcionalidad del data warehouse. Las capas representan el ambiente necesario para implementar los bloques.

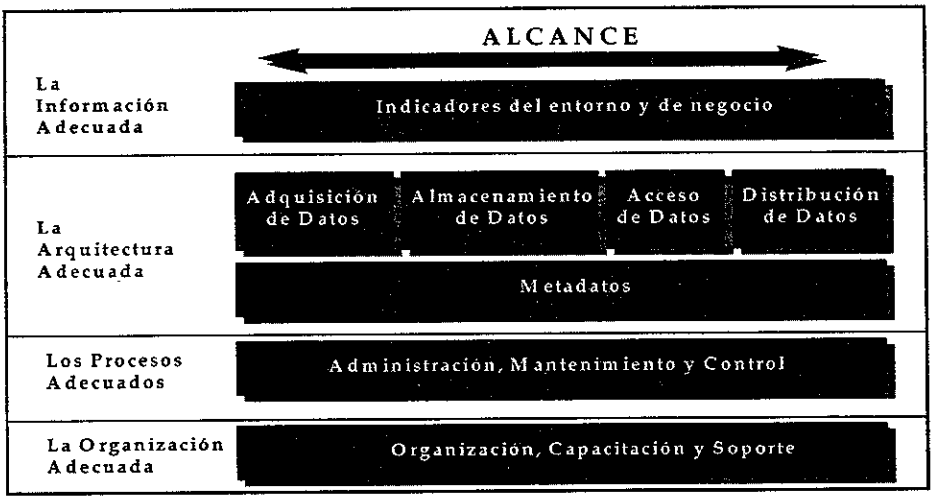


Fig. 10.- Arquitectura de referencia

8.8.1 Adquisición de Datos (Fuentes de Datos)

Las fuentes de datos que alimentan un data warehouse pertenecen a las siguientes categorías:

- Producción de datos: Se refiere a las bases de datos operacionales que mantienen la información obtenida de aplicaciones operacionales. Las bases de datos operacionales pueden encontrarse en una gran variedad de tecnologías como puede ser relacional, no-relacional o basada en archivos.
- Legado de datos: Son datos que se encuentran almacenados de manera aislada ya que no proporcionan apoyo a las aplicaciones operacionales. Sin embargo, tienen un valor histórico significativo para el análisis de tendencias, y deben ser incluidos en el data warehouse con la aplicación de las cifras correctas. Estos datos, también se utilizan con propósitos de minería en el data warehousing.
- Sistemas internos de oficina: Estas son fuentes de datos que no son almacenados en una base de datos operacional o que no son utilizados por una aplicación operacional. Este dato puede ser *no-estructurado* (como puede ser en formas no electrónicas), *estructurado* (como puede ser en reportes, gráficos, hojas de cálculo o procesadores de palabras) o

semi-estructurado (como en un reporte anual). Estos datos son utilizados para apoyar el análisis entre departamentos.

- Fuentes externas: Estos datos provienen de fuentes que no son controladas, operadas o propietarias de la empresa. Estas fuentes pueden ser *electrónicas* (Dow Jones u otras firmas de análisis de mercados) o *no-electrónicas* (como son los reportes de consultores, artículos de revistas o periódicos).

8.8.2 Almacenamiento de Datos (Construcción del Data Warehouse)

Es el componente más grande de esta arquitectura que consta de tres componentes:

1. *Refinamiento*: Este componente es responsable de estandarizar, limpiar, filtrar, comparar y capturar en el tiempo la información extraída de las fuentes de datos seleccionadas. En este componente, el *metadato* es mapeado a los nombres y definiciones de datos estándar.

2. *Reingeniería*: Este componente es responsable de la adaptación de los datos para conocer las necesidades de análisis del usuario. La reingeniería en el contexto del data warehouse difiere de la reingeniería de las aplicaciones o procesos del negocio. La reingeniería involucra:

- La integración de diferentes tipos de datos desde múltiples sistemas para crear un nuevo dato.
- Particionar los datos en series de tiempo para su análisis.
- Trasladar y formatear.
- Transformar y remapear los datos almacenados a las fuentes de datos originales para habilitar la actualización prolongada de los datos derivados, creados y transformados.

3. *Data warehouse*: Este componente es responsable del modelado de datos del data warehouse, la reducción de grandes volúmenes de datos a paquetes de datos manejables, y el metadato del data warehouse lógico y físico. El manejo de datos para

reducir los grandes volúmenes de datos que entran y los datos almacenados previamente utilizando técnicas de agregación y resumen, es una parte de este componente junto con la responsabilidad de crear datos altamente resumidos a partir de datos poco resumidos almacenados al mismo tiempo en el data warehouse. El grado de sumarización se refiere al nivel de agregación de los datos. Por ejemplo, los resúmenes semanales son menos resumidos que los resúmenes quincenales.

8.8.3 Distribución de Datos (Construcción del Data Mart)

Este es el segundo gran componente de la arquitectura. Este elemento es utilizado principalmente para crear el data mart desde los componentes del data warehouse. Los elementos que se requieren para construir el data mart son:

De Refinamiento y Reingeniería:

- Filtrado y comparación.
- Integración y partición.
- Resumen y agregación.
- Precalcular y derivar.
- Estampar la fuente de datos en una dimensión del tiempo.
- Extracción y creación del metadato.

De Creación del data mart:

- Modelado.
- Resumen.
- Agregación.
- Reconciliación y validación.
- Construir consultas.
- Creación del glosario y navegación del metadato.

El Data Mart tiene componentes similares al data warehouse. La principal diferencia no es tanto en el tipo o componentes de la arquitectura como en el objetivo del

usuario final. Los componentes del Data Mart aplican un conjunto de pasos diferentes de reingeniería y refinamiento junto con los objetivos del negocio y sus usuarios finales.

8.8.4 Acceso de Datos (Acceso y Uso)

Este bloque de la arquitectura consta de dos componentes: *acceso y recuperación,* y *análisis y reportes.* Este bloque permite la recuperación de la inversión y justifica el valor de la implementación completa del data warehouse. A continuación se muestran sus componentes:

De acceso y recuperación:

- Acceso directo al data warehouse.
- Acceso al Data Mart.
- Reingeniería.
- Transformación de la estructura multidimensional.
- Crear un almacén local.
- Administrar el metadato del data warehouse y Navegar en el metadato.

De análisis y reportes:

- Herramientas de reportes.
- Herramientas de análisis y del DSS.
- Herramientas del modelado de negocios.
- Herramientas de minado de datos.
- Nueva producción de aplicaciones (OLAP).
- Administración y reportes del metadato.

8.8.5 Indicadores del Entorno y del Negocio (Administración de Datos)

Las tareas de extracción, carga, actualización, seguridad, archivar y recuperar el data warehouse desde archivos son apoyadas por la capa de administración de datos, cuyos componentes son:

- Extracción y administración de nuevos requerimientos de datos/consultas.
- Carga, almacenamiento, actualización (refresh) y modificación de sistemas.
- Sistemas de seguridad y autorización.
- Sistemas de archivación, recuperación y depuración.

Esta capa de administración de datos incorpora las políticas, procedimientos, programas y operaciones de seguridad, autorizaciones de acceso, archivación y recuperación, y depuración de los datos. Un reto particular es presentado por el tamaño potencial del data warehouse. El tamaño del data warehouse tiene impacto en el manejo de la consolidación de datos de múltiples índices, el lugar físico de los datos e índices, y en la recuperación rápida de datos desde múltiples medios. Los aspectos del procesamiento paralelo de consultas, y el uso de procesadores paralelos para el acceso y recuperación de datos, también son manejados en esta capa. La administración de estas actividades es responsabilidad de la capa de administración del metadato.

8.8.6 Administración del Metadato

Esta capa es responsable de la administración del metadato que es utilizado por el data warehouse como una descripción completa de los datos almacenados en él. El metadato provee además guías y punteros a los datos localizados en el data warehouse. Sus componentes principales son:

- Esquema y glosario del data warehouse y del data mart.
- Extracción del metadato, creación, almacenamiento y manejo de modificaciones.

- Consultas predefinidas, reportes, administración de índices.
- Administración de la actualización y replicación.
- Administración de entradas, archivación, recuperación y depuración.

Los modelos lógicos y físicos del data warehouse y del data mart, así como su esquema y el glosario de negocios y técnico, son manejados en esta capa. El enorme reto de la administración de grandes bases de datos con su gran complejidad en las áreas de múltiples índices, consolidación de datos, llaves compuestas, y versiones de datos, es dirigido y manejado en esta capa.

8.8.7 Administración, Mantenimiento y Control (Transporte)

Esta capa utiliza la tecnología de actualización y replicación, transferencia de datos y redes, así como componentes middleware. Además provee seguridad y autenticación para los requerimientos de transporte. Los componentes de esta capa son:

- Transferencia de datos y redes de distribución.
- Agentes cliente/servidor y herramientas middleware.
- Sistemas de replicación.
- Sistemas de seguridad y autenticación.

La capa de transporte direcciona los puentes de comunicación necesarios entre las plataformas hardware/software que son separadas como un resultado del particionamiento de plataformas de los distintos bloques.

La transferencia de datos y redes de distribución, contienen los siguientes tipos de sistemas:

- Protocolos de red (TCP/IP, SNA/APPN, IPX).
- Estructuras de administración de red tales como OpenView de Hewlett Packard, NetView de IBM, SunNet Manager de SunSoft's.

- Sistemas operativos de red.
- Tipos de red tales como Ethernet, Token Ring, FDDI.

Los agentes cliente/servidor y el middleware, contiene los siguientes tipos de sistemas:

- Gateways de bases de datos tales como Builders´EDA/SQL, Sybase Enterprise Connect, DRDA/DDCS de IBM.
- Middleware orientado a los mensajes tal como MQSeries de IBM.
- Agente de requerimientos (ORBs -Object Request Broker) tales como SOM de IBM, DSOM, ORB Plus de Hewlett Packard y Object Broker de DEC.

Los componentes de los sistemas de replicación contienen los siguientes tipos de sistemas:

- Sistemas de propagación y replicación tales como Data Propagator Relational (DPropR) y Non-Relational (DPropNR) de IBM , Sybase Replicator Server, Symmetric Replication de Oracle.
- Productos específicos de data warehouse tales como Prism Solutions´Warehouse y Change Manager.

8.8.8 Organización, Capacitación y Soporte (Infraestructura)

Los componentes de esta capa son:

- Sistemas de administración.
- Administración del flujo de trabajo.
- Sistemas de almacenamiento.
- Sistemas de procesamiento.

Los sistemas de administración proveen de funciones, facilidades y servicios para invocar, manejar y terminar herramientas y aplicaciones, basadas en eventos o bajo la dirección de constructores de sistemas y el usuario.

El componente de administración del flujo de trabajo apoya los procesos de integración y administración para coordinar la ejecución ordenada y especificada de herramientas, aplicaciones y actividades que completan correctamente la extracción, actualización, replicación, modificación, agregación y sumarización, y otras tareas de mantenimiento y administración de sistemas del Data Warehouse y del Data Mart. Es la automatización de las pequeñas y complejas tareas requeridas para mantener y actualizar el data warehouse y el data mart, así como proveer reportes predefinidos y resultados de consultas, que incrementan la eficiencia y la productividad de los constructores de sistemas y de los usuarios.

Los sistemas de almacenamiento proveen los servicios de administración de bases de datos y de archivos para las fuentes de datos, los catálogos de bases de datos del data warehouse y del data mart, y el almacenamiento multidimensional y local para su acceso y uso.

Los sistemas de procesamiento son los ambientes de operación de los bloques: Fuentes de datos, el data warehouse y el data mart, las herramientas de acceso y uso, el middleware y otros componentes de infraestructura discutidos previamente.

Otros sistemas importantes de la capa de infraestructura son:

- Administradores de configuración.
- Administradores de almacenamiento.
- Administradores de seguridad.
- Administradores de Distribución de Software.
- Administradores de licencias.
- Monitores de desempeño.
- Analizadores de capacidad.

8.9 DATA WAREHOUSING Y SISTEMAS DE APOYO A LA TOMA DE DECISIONES

Hasta este momento, contamos con información acerca de la aplicación de las bases de datos en el transcurso del tiempo, podemos identificar que desde su origen han representado un papel importante en el desempeño de las funciones que se realizan en las organizaciones; inicialmente a nivel operativo, convirtiéndose gradualmente en una herramienta de apoyo esencial en la toma de decisiones.

Podemos destacar que dependiendo del nivel jerárquico en que nos encontremos dentro de la estructura organizacional, la necesidad de información varía; es decir, así como cada persona tiene una función específica que realizar en la organización, la información que requiere también es específica acerca de algún tema en particular, por ejemplo, el director del departamento de ventas, requiere información acerca de las ventas que se han llevado a cabo ya sea por departamento, por región, por planta, por periodo de tiempo y/o por producto.

Asimismo podemos notar que gracias a los avances que presentan en la tecnología de la información, las organizaciones pueden acumular mayores cantidades de información, incrementándose a su vez la necesidad de manipular y mantener estos grandes volúmenes de información organizados de manera tal, que permita la obtención de datos específicos en un momento determinado.

Ahora bien, sabemos que en el mercado actual de tecnología de información, se cuenta con una gran variedad de herramientas y tecnologías que pueden apoyar y contribuir en gran medida el logro de objetivos organizacionales, y que en nuestro caso de estudio es la toma de decisiones.

Antes que nada, debemos destacar que existen dos tipos de sistemas de negocios: unos apoyan las funciones operacionales y otros reportan dichas funciones. Lo cual muestra la diferencia entre los sistemas orientados a las entradas de datos y los que se encuentran orientados a la salida de datos. Sin embargo, ambos pretenden lograr lo siguiente: acceso a datos, disponibilidad, integración y análisis.

Los sistemas operacionales (que corren el negocio) son: Sistemas de Procesamiento de Transacciones (Transaction Processing System -TPS-), Sistemas de Reingeniería de Procesamiento del Negocio (Business Process Reengineered -BPR-), Sistemas de Intercambio Electrónico de Datos (Electronic Data Interchange -EDI-) y los Sistemas de Integración tradicionales (Systems Integration -SI-). Todos estos sistemas generan una gran cantidad de datos que tanto los analistas de negocios, especialistas de mercadotecnia, ejecutivos y operativos quieren tener en sus manos, por lo que impidiendo el acceso a esta información también limitan así su efectividad. Liberando estos datos, proporcionan oportunidades en el ahorro de costos y nuevas formas de obtención de ingresos.

Por otro lado encontramos a los sistemas diseñados para el análisis y reporte de los negocios, estos incluyen: los Sistemas de Apoyo a las Decisiones (Decision Support Systems -DSS-), Sistemas Ejecutivos de Información (Executive Information Systems -EIS-) y las Aplicaciones CASE Basadas en el Razonamiento (CASE-based reasoning -CBR-) . En general, estos sistemas tienen una profunda necesidad de información. La conversión de datos a información involucra la reorganización de datos, la derivación de nuevos datos, su integración y presentación a los usuarios finales. Por lo anterior, podemos observar que el acceso a los datos operacionales en una consideración importante.

A continuación se describe cómo el uso de un data warehouse en substitución o bien en combinación con otro tipo de sistema, contribuye a la aplicación eficiente y óptima de la información en el proceso de toma de decisiones.

8.9.1 Sistemas de Reingeniería de Procesamiento del Negocio

Este tipo de sistemas son una opción para responder a los cambios estratégicos operacionales y organizacionales. Reingeniería implica cambio en la organización, que para los ejecutivos es un aspecto de riesgo, ya que están acostumbrados a realizar sus actividades de un modo determinado. La cuestión es, de qué manera se le puede proporcionar la información necesaria a los ejecutivos para ayudarles a decidir cómo llevar a cabo la reingeniería del negocio efectivamente. Para ello, se requiere tener acceso, disponibilidad, integración y análisis de la información; todo lo cual puede ser obtenido vía data warehouse.

Es decir, un proyecto de data warehouse apoya el diseño de un gran sistema de reingeniería de procesos, ya que la iniciativa misma de la reingeniería tiene como requerimiento principal mejorar el acceso a datos de los usuarios finales. Ahora bien, conociendo los problemas del pasado en el intento de proveer a los usuarios finales datos operacionales, el uso de un data warehouse permitirá correr fácilmente y de manera ininterrumpida sistemas de negocios con acceso ilimitado a la información [5E].

8.9.2 Sistemas Basados en el Conocimiento

Los sistemas basados en el conocimiento son otro tipo de sistemas utilizados en la operación del negocio, conocidos también como sistemas de inteligencia artificial operacional (AI). Generalmente se encuentran integrados con otro sistema de información; este tipo de aplicaciones piden datos en vez de crearlos. La limitada popularidad de los sistemas basados en el conocimiento se debe, en gran parte, a su incapacidad de accesar a los datos transaccionales en los volúmenes requeridos con el desempeño aceptable. Sus aplicaciones, leen directamente las bases operacionales, por lo que actúan deficientemente. Los diseñadores se ven forzados a hacer concesiones que limitan la inteligencia resultante y la utilidad de estos sistemas [5E].

Se considera que el procesamiento inteligente con el alto desempeño y el acceso propiamente organizado del data warehouse, puede proporcionarle a la inteligencia artificial una oportunidad de desarrollo.

8.9.3 Sistemas de Procesamiento de Transacciones

El mayor problema al que se enfrentan las organizaciones con este tipo de sistemas, es el ciclo de mantenimiento. Los analistas de negocios y otros usuarios finales de datos, ven al los sistemas de procesamiento de transacciones (TPS) como la raíz de todos los datos [5E], es decir, un sistema que debe recopilar y crear todos los datos que ellos necesitan para analizar, pero que sin embargo, incrementan la dificultad en su mantenimiento. Por ejemplo, un sistema de mercadotecnia enfocado principalmente a las ventas de un producto, extiende su enfoque hacia la obtención de registros de servicio, además de canales de distribución, ventas industriales y una gran cantidad de información histórica adicional. Con un data warehouse, muchos de los nuevos requerimientos de datos pueden ser fácilmente localizados junto con los datos pertenecientes a un TPS al que pertenecen.

8.9.4 Intercambio Electrónico de Datos

Los sistemas de intercambio electrónico de datos (EDI) son creadores de grandes cantidades de datos, sin embargo, los datos deben ser formateados y las bases de datos deben ser organizadas para un procesamiento muy específico, lo cual prohibe grandemente la introducción de datos auxiliares utilizados por otra aplicación. El EDI y los datos auxiliares pueden coexistir en una base de datos integrada y estar disponible tanto para usuarios finales como para otros sistemas de negocios.

8.9.5 Integración de Sistemas

Los proyectos de integración de sistemas involucran la coordinación del procesamiento de dos o más funciones del negocios (por ejemplo, ventas y manufactura). Actualmente estos proyectos intentan facilitar la integración de unidades de negocio para fusionar o adquirir compañías [5E]. Sin embargo, el problema que se plantea es el mismo que en la mayoría de los sistemas, es decir, ¿cómo se pueden integrar o emigrar los sistemas sin afectar su operación actual?, bien, la respuesta es muy a menudo el data warehouse, ya que permite diseñar una base de datos integrada que sirva a las funciones de ambos sistemas.

8.9.6 Sistemas de Apoyo a las Decisiones

La palabra apoyo, puede ser un término limitado para la familia de aplicaciones que apoyan a los trabajadores del conocimiento en la organización. Algunas personas piensan que se trata de aplicaciones que apoyan a la filosofía popular del "libro abierto" [5E] en la que se permite a los empleados de todos los niveles accesar a información financiera y propia del negocio, lo cual puede tener un impacto directo en la organización.

Ahora bien, los información que el DSS presenta al usuario es fácil de entender y aplica reglas del negocio para derivar nueva información. Estos requerimientos requerimientos de información provocan:

- La decodificación de los datos, de modo que sean presentados a los usuarios con información descriptiva, no encriptada.
- La organización física por tema (sujeto) de los datos para una recuperación rápida sin consultas complejas u operaciones de unión (joins) masivas.
- Aplicación de fórmulas y análisis comunes para proveer información consistente y simplificada.

En un data warehouse estas funciones son aplicadas fácilmente, y los usuarios finales del sistema de apoyo a las decisiones aseguran la integridad, disponibilidad y comprensibilidad de los datos.

8.9.7 Sistemas de Información Ejecutivos

Este tipo de sistemas permiten gran flexibilidad en la división de datos, lo cual incrementa su funcionalidad, aunque sus requerimientos son similares a los de DSS, los EIS necesitan que sus datos sean mucho más integrados, resumidos, históricos y no-volátiles. Hay que considerar que para integrar datos no solo se requiere reunir los datos de diferentes fuentes en la misma base de datos, las tablas deben ser combinadas físicamente a través de técnicas de normalización orientadas al tema (sujeto).

Los datos utilizados en la operación de los negocios incluyen frecuentemente el estado actual de los negocios. Una vez que los datos son sobreescritos, su valor previo no puede ser recuperado (sin ocupar los sistemas de archivos). Lo anterior no es aceptable para los sistemas que reportan los negocios.

En las aplicaciones EIS se requiere de datos históricos no-volátiles que le permitan presentar información apropiada para su análisis, esto es debido a que se requiere un conteo de ciertos elementos que han sido afectados en un periodo de tiempo.

El data warehousing provee un lugar y un proceso para el almacenamiento de información integrada, resumida, reorganizada, histórica y no-volátil que permita una gran variedad de reportes y muchas otras capacidades simplificadas. La necesidad de una nueva manera de representar físicamente los datos para EIS ha dirigido al data warehousing a técnicas de modelado tales como el esquema de estrella.

8.9.8 Herramientas CASE Basadas en el Razonamiento

Sistemas tales como CBR, minería de datos y aplicaciones no estructuradas de recuperación de datos abren nuevos caminos a las organizaciones para explorar sus datos. Por otra parte, este tipo de aplicaciones enfocan la medida de su éxito a la cantidad de datos a la que pueden accesar. Las aplicaciones CBR apoyan la toma de decisiones debido a que consideran los factores de un evento actual, buscando un evento similar ocurrido previamente y extrapolando/interpolando estos datos para obtener una respuesta óptima[5E]. Esta respuesta es tan óptima de acuerdo a la cantidad de casos que han sido planteados.

En las aplicaciones de recuperación de información no estructuradas los usuarios buscan información en archivos grandes. Mientras mayor sea la cantidad de datos requeridos, más específica debe ser la consulta.

Ninguno de estos almacenes de datos son soportados por las bases de datos comunes, solo un data warehouse puede asegurar realmente el éxito de este tipo de aplicación.

8.10 CASOS DE ESTUDIO

Muchos administradores de sistemas de información en servicios financieros, cuidado de la salud y telecomunicaciones iniciaron sus data warehouses y data marts hace algunos años. Han dominado el modelado y los esquemas de estrella, y han enfrentado distintos elementos políticos que surgen cuando desarrollan su primer sistema de soporte a las decisiones.

Los problemas a los que nos enfrentamos son los siguientes: consultas inesperadas que desafían aún a los índices más sofisticados; warehouses que atraen a los hackers y el abuso de usuarios novatos; sobrecarga de redes LAN y procesadores; problemas de administración, que varían desde help desks que no están bien fundados hasta el manejo que realiza el departamento financiero en cuanto a la recuperación de la inversión (Return on Investment -ROI-) . A continuación se presentan los nombres y comentarios de algunas empresas que han implementado un data warehouse.

8.10.1 Telus Comunications, Edmonton, Alberta, Canadá.

Desarrollaron su warehouse en 1994, contiene 600Gb., de datos de ventas y de usuarios. Su inversión inicial fue de 4 millones de dólares que podrían ser recuperados en un año y medio debido a los reducidos costos de almacenamiento y a su habilidad de capacitar al personal de tecnología de información. Esa inversión fue para la base de datos Oracle, el servidor HP/Unix, así como el HP Intelligent Warehouse iniciado con 200Gb de datos.

Aunque la recuperación de su inversión se llevó a cabo en un tiempo mayor del previsto, se considera lo siguiente: se tenía estimado un número de 100 usuarios de su warehouse, y en el primer año han tenido 200. Se estimaba tener 2,000 consultas al mes, y actualmente se tienen 6,000 consultas y este número se encuentra en crecimiento.

El problema con el outsourcing es que, el data warehouse requiere de una investigación a largo plazo, por lo que se requiere tener un equipo dedicado a ello [15].

valor para el cliente. Si consideramos los patrones contractuales, observamos que estén llevando a cabo sus mejores estrategias de compra y se les provee de información en un ambiente muy fácil de manipular, tal como un OLAP, nos encontraremos con que se están creando nuevos negocios [15]. Y esto es muy mensurable en términos de recuperación de la inversión. Podemos empezar a ver en la retención de clientes, nuevos clientes o nuevas líneas de producto. En ese momento el warehouse se está pagando a si mismo.

8.10.5 Blue Cross/Blue Shield Association, Washington.

Desarrollaron su warehouse el año pasado para servir como componente en una arquitectura de tres niveles. Optaron por una base de datos de Red Brick Systems, y herramientas de consulta y reportes de Brio Technology.

8.10.6 Siemens Business Communication Systems Inc., Santa Clara, California.

Desarrollaron su warehouse basado en mainframe para el Web, de manera que pueda ser accesado por 6,000 de sus empleados. Están utilizando PeopleSoft, la cual es una aplicación cliente/servidor. Como su servidor es un mainframe DB2, toman los datos fuera de las tablas DB2 y los ponen dentro del warehouse. Es una herramienta gráfica, fácil de usar, se puede mantener y administrar sin requerir de un sistema de información. Estar en una plataforma DB2 hace muy difícil encontrar herramientas.

9. HERRAMIENTAS PARA DATA WAREHOUSING

Un aspecto fundamental de las nuevas herramientas para data warehousing es el manejo de tecnología de índices de mapas de bits, los cuales mejoran de manera significativa el tiempo de respuesta sobre los métodos de índices tradicionales, a través de una reducción significativa del número de operaciones de lectura de los datos. Además permite a más usuarios el acceso al warehouse de manera simultánea, los índices de mapas de bits también simplifican a los usuarios el envío de series de consultas para analizar los datos. Aún más, se puede alcanzar un nivel de respuesta aceptable con un gasto menor en hardware que con los métodos de índices tradicionales y sin la necesidad de diseñar bases de datos especializadas como es el caso de los esquemas de estrella, los cuales requieren de computadoras de procesamiento paralelo masivo o de DBMS especializados dedicados al análisis multidimensional.

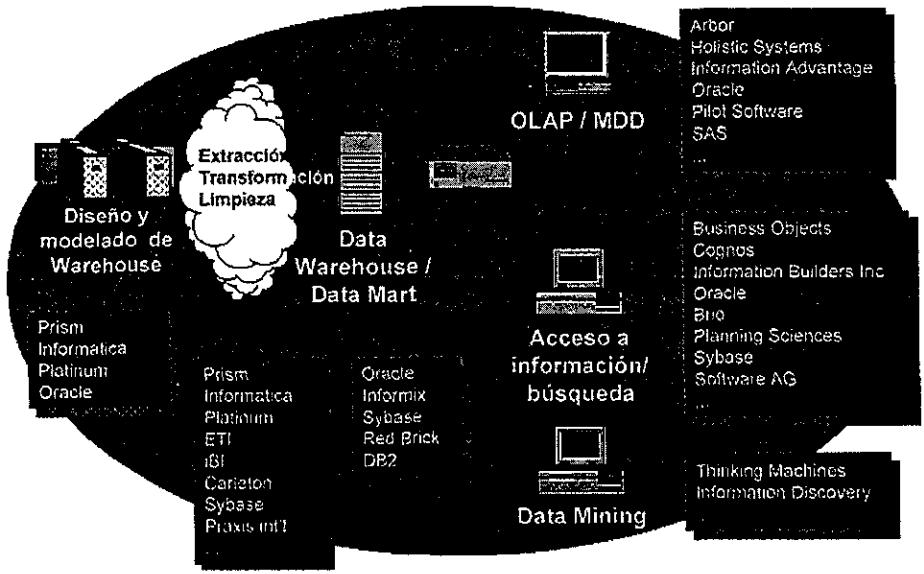


Fig. 11.- Proveedores de herramientas de data warehousing

9.1 ÁRBOLES B Y HASHING

La mayoría de los sistemas de administración de bases de datos proporcionan un amplia variedad de formas de acceso a los datos. La mezcla de tareas que se tratan de realizar y la naturaleza de los datos determinan el método de acceso o métodos que pueden proporcionar el mejor desempeño. Prácticamente todas las bases de datos relacionales soportan tanto el acceso secuencial con el acceso a través de índices como formas fundamentales de localizar los datos. El acceso secuencial se refiere simplemente a leer los datos desde el inicio del archivo hasta que se localice el renglón o registro deseado, momento en el cual dicho renglón o registro puede ser actualizado, modificado, consultado o borrado. Este puede ser un método efectivo para tablas pequeñas donde todos los renglones se encuentran juntos, o en aplicaciones que necesitan desempeñar alguna acción sobre todos los registros de la base de datos[4E].

Pero en el caso de tener demasiado s datos almacenados en la base, una búsqueda secuencial implicaría el uso de una considerable cantidad de tiempo. En estos casos es mejor construir un índice en un árbol B y almacenar la información en una jerarquía (árbol) de páginas. Cada página contendría muchos índices así como punteros a la siguiente página de índices. La ventaja de los árboles B sobre las búsquedas secuenciales es que el DBMS necesita revisar sólo una pequeña cantidad de páginas y recuperar sólo aquellas que contienen al índice.

En casos de alta cardinalidad, algunos DBMS como Oracle o CA/Ingres proporcionan métodos de acceso de hash. En vez de construir un índice, los métodos de hashing utilizan una función matemática para calcular directamente la página en la cual el renglón se encuentra almacenado. Considerando que la función distribuye los registros de manera equitativa entre las páginas de la base de datos, sólo se requiere una operación de lectura a los datos para recuperar el renglón deseado.

En el otro extremo, en los cases de baja cardinalidad, la construcción de un árbol B no proporciona mayor ventaja en el número de operaciones de lectura ni en la velocidad de acceso, pues para buscar renglones con los valores deseados se recuperan una gran cantidad de páginas.

Los índices en árboles B o hash han presentado problemas en consultas que tienen un conjunto complejo de condiciones, en otras palabras, en el tipo de consultas por las cuales se han desarrollado los warehouse.

Los mapas de bits resultan atractivos para las aplicaciones de warehousing porque, cuando son implantados de manera adecuada como parte de un esquema de índices, pueden mejorar de manera significativa el tiempo de respuesta. Esto se debe a que las búsquedas complejas de muchas condiciones, consolidaciones, cálculos y uniones pueden ser realizadas por completo en los índices sin leer los datos. Incluso el uso limitado de los mapas de bits puede reducir la cantidad de operaciones de lectura a la base de datos y mejorar el desempeño en algunas consultas.

La metodología más común de los proveedores de bases de datos es el uso de índices complejos basados en combinaciones de árboles B, mapas de bits y listas de registros ID.

La complejidad y variedad de las estructuras de índice revelan dos hechos. Primero, no existe una solución sencilla para un adecuado desempeño de las consultas en grandes bases de datos. Se requiere además una gran variedad de estructuras integradas y de alto desempeño para dar a las consultas la flexibilidad que requieren en tiempo de respuesta. Segundo, las empresas necesitan un diseñador de bases de datos que comprenda lo que se encuentra en la base de datos y la forma como se utiliza. Un buen diseño físico de la base de datos es un pre-requisito para su buen desempeño.

9.2 SYBASE IQ

Sybase IQ esta diseñado para manejar data warehouses de tamaño moderado que pueden incluir hasta varios cientos de gigabytes sin requerir de hardware demasiado costoso. Es descendiente de Expressway, adquirido por Sybase a fines de 1994. IQ difiere de Expressway en que se encuentra integrado al catálogo de bases de datos Server SQL de Sybase, puede realizar uniones (joins) y tiene soporte al SQL tradicional.

Sybase también incluyó en IQ soporte para computadoras de procesamiento simétrico con carga paralela y construcción de índices, procesamiento de mapas de bits y realización de ordenamientos.

Los diseñadores de IQ se concentraron en las operaciones de alto desempeño en velocidad y cálculos sobre cualquier tipo de datos de cualquier cardinalidad. Esto llevó a Sybase a desarrollar una variedad de estructuras de índice para los cuales los mapas de bits son un componente fundamental.

En IQ el usuario ve a los datos como una base de datos relacional, aunque de hecho los datos no se encuentran almacenados como tales; todos los datos se encuentran en estructuras de índices. IQ utiliza compresión par mantener un tamaño de base de datos menor al espacio ocupado por los datos de manera natural.

Debido a que IQ esta diseñado para data warehouses de sólo lectura, no permite actualizaciones interactivas de datos, las cuales resultarían especialmente demandantes de tiempo de procesamiento. En vez de ello, todas las actualizaciones deben ser realizadas de manera periódica en modo batch. Sybase también diseñó sus estructuras de índices de modo que los grande cambios de tamaño de los índices no requiera de su reconstrucción.

Características	Beneficios
Almacenamiento de datos vertical y compresión inteligente	• Reducción en más del 98% en las entradas y salidas del disco en tiempo de respuesta rápido
	• Elimina búsquedas en la tabla y tiempos de respuesta impredecibles
Diseño autoajustado	• Elimina los esfuerzos de afinación basados en consultas del DBA.
Mapas de bits en baja cardinalidad	• Cuenta rápida de registros
Indices bit-wise en alta cardinalidad	• Consolidación dinámica rápida y rangos de búsqueda de datos relacionales
Optimizador de consultas	• Selección automática del método de acceso más rápido para resolver una consulta.
Interfaces estándar abiertas	• Soporte a un amplio rango de herramientas de consulta.

9.3 HP DATAMART MANAGER

HP ofrece HP DataMart Manager, un software administrador de búsquedas en un data warehouse para mejorar el desempeño de data marts y data warehouses. Sus características son:

1. *Incremento del desempeño:* Este software de administración de consultas incluye las tecnologías de administración de tablas resumen utilizadas en algunos de los ambientes de data warehousing más complejos y grandes del mundo llevándolas a ambientes de data warehouse de alcance más limitado. HP DataMart Manager está diseñado para manejar aspectos que incluyen afinación del desempeño, administración de consultas y facilidad de uso para usuarios finales.

2. *Afinación del desempeño con administración de tablas resumen:* El desempeño es un aspecto crítico de la administración de los ambientes de warehouse. Los data warehouses utilizan tablas resumen para precalcular respuestas a las consultas más complejas y comunes de los usuarios. Los datos pueden ser presentados de manera más rápida en un formato fácil de comprender para usuarios no capacitados. HP DataMart Manager es un middleware que selecciona de manera dinámica las tablas resumen más adecuadas para cada consulta, proporcionando tiempos de respuesta medidos en segundos. Adicionalmente, HP DataMart Manager avisa al administrador a través de reportes y despliegues gráficos en cuál de esas tablas deben ser creadas o borradas de acuerdo a los patrones de uso. Dado que HP DataMart Manager es un middleware residente entre el servidor de la base de datos y el cliente, la administración de las tablas resumen no afecta a las herramientas del usuario final o a sus aplicaciones. El resultado es un warehouse de alto desempeño que puede ser administrado fácil y eficientemente.

3. *Administración de consultas:* La mayoría de los data warehouses son accesados por usuarios con múltiples accesos a los datos y herramientas de reporte. HP DataMart Manager soporta casi todas las herramientas conocidas de acceso a los datos, reporte y herramientas OLAP y combinaciones de herramientas que corren en UNIX, MS Windows, NT, Macintosh y OS/2. Además HP DataMart Manager soporta una gran variedad de aplicaciones de apoyo a la toma de decisiones basadas en navegadores de red. HP DataMart Manager simplifica la administración de consultas en ambientes con múltiples herramientas de acceso a datos, reportes y OLAP. Proporciona un nivel medio, simple y comprensible para manejar interfaces con herramientas y aplicaciones.

   Debido a que HP DataMart Manager registra todas las consultas, permite al administrador generar reportes para:

   a) Conocer los tiempos de respuesta a los usuarios
   b) Mejorar la administración de consultas que sobrecargan al sistema
   c) Agregar usuarios para el uso del data mart

3. *Facilidad de uso para usuarios finales:* La complejidad del data warehouse puede intimidar a los usuarios casuales de la red y a usuarios no técnicos. Si no son asesorados adecuadamente, los usuarios no capacitados pueden saturar la red con consultas inútiles, limitando la disponibilidad para los demás. Y si no se conocen las expectativas de desempeño de los usuarios el warehouse puede ser desaprovechado por los tomadores de decisiones limitando su efectividad dentro de la organización. La facilidad de uso es posible con casi cualquier herramienta de acceso a datos compatible con ODBC o aplicaciones de apoyo a la toma de decisiones, incluyendo Microsoft Access, Microsoft Excel, Brioquery, Business Objects, Cognos Impromptu, y otras.

4. *Acceso al data warehouse a través del web:* Las capacidades de administración de consultas de HP DataMart Manager pueden ser especialmente efectivas cuando el warehouse es accesado por una gran cantidad de usuarios a través de navegadores web. La solución HP OpenWarehouse Web está basada en un servidor de seguridad HP combinado con un administrador HP DataMart y una variedad de aplicaciones de apoyo a las decisiones basadas en el Web.

9.4 RED BRICK WAREHOUSE

Red Brick 4.0, posee una columna índice, llamada TargetIndex, este tipo de índice esta basado en mapas de bits y diseñado para selecciones rápidas de una tabla dada. Esto complementa los árboles B, índices de patrones y los StarIndex ya existentes en Red Brick. Las consultas pueden ser resueltas utilizando una combinación de todos los tipos de índices.

Aunque los usuarios no actualizan los datos en el data warehouse, Red Brick considera este aspecto importante para mantener la limpieza de los datos y optimizar la administración de los mismos.

Para manejar los problemas derivados de los datos con alta cardinalidad, Red Brick posee tres estructuras de índices al construir el TargetIndex: muy baja cardinalidad (2 a 16 valores), cardinalidad media (8 a 256 valores) y alta cardinalidad (64 a 128 000 valores).

El índice de baja cardinalidad esta implantado como un árbol B con mapas de bits en las hojas y el valor del mapa de bits en la cabecera. Estos índices típicamente toman un 20% de tiempo para construirse como índices de arboles B puros. Red Brick no comprime los mapas de bits, debido a que sus mediciones mostraron que la compresión para datos de baja cardinalidad no proporcionaban una ganancia importante en el desempeño.

Conforme se incrementa la cardinalidad, el número de ocurrencias de cada valor decrece. De este modo, en los datos con cardinalidad media, Red Brick proporciona árboles B con una lista comprimida de RID en las hojas, y para los datos de alta cardinalidad se utilizan árboles B con listas no comprimidas de RID en las hojas.

Para obtener un buen desempeño en la actualización de los datos con mapas de bits, Red Brick establece apuntadores dentro de los mapas de bits o de las listas comprimidas de RID para hacer más rápida la localización de los bits o RID que se necesita modificar. Además administra las listas RID en los índices de alta cardinalidad para mantener el desempeño en las actualizaciones.

Red Brick ha puesto especial atención en el optimizador de búsquedas para obtener el máximo desempeño de estos índices. El StarIndex de Red Brick, le permite realizar uniones de datos con un esquema estrella. Este esquema se caracteriza por una tabla central de hechos que esta rodeada de tablas de dimensiones en las cuales se basan los cálculos.

9.5 ORACLE 7.3

Oracle ha agregado índices de mapas de bits para adecuarse a sus índices de hash y de árboles B. Los mapas de bits de Oracle están considerados para datos de baja cardinalidad, mientras que los árboles B se utilizan para los datos de alta cardinalidad.

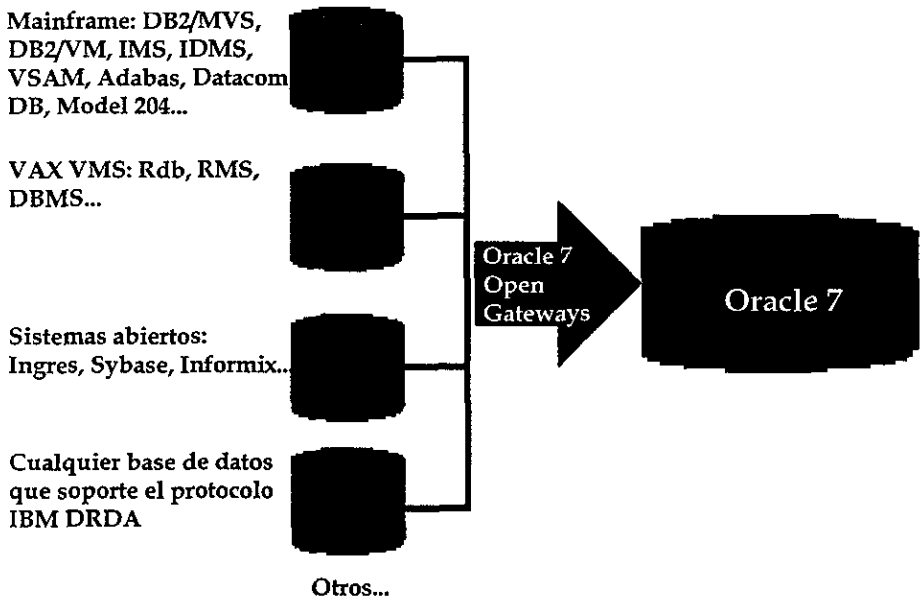


Fig. 12.- Oracle 7

El enfoque de Oracle esta orientado a mapas de bits en las hojas de los árboles B. Los mapas de bits están comprimidos utilizando un método adoptado del servidor de texto de Oracle. Una característica importante de este método es que permite la manipulación lógica de los mapas de bits en su estado comprimido.

Oracle pretende que los mapas de bits sean utilizados como parte de un mezcla de métodos regulares de acceso y se permita que los mapas de bits sean actualizados en tiempo real.

9.6 TERADATA

NCR y Brio Tecnology establecieron un acuerdo global para proveer soluciones de negocios que cumplan con las necesidades de sus clientes para que la información del data warehouse esté disponible de manera segura y a un costo adecuado para toda la empresa y a través de Internet.

Conforme a este acuerdo, los productos de software de Brio Technology -Brio Query, Brio Query.Sever, Brio Query.Insight y Brio.QuickView- estarán incluidas como parte de la solución completa de data warehouse de NCR para mejorar el alcance del motor de data warehouse del RDBMS Teradata de NCR.

Brio Technology considera que en el ambiente de negocios actual el principal reto es dar capacidad a los tomadores de decisiones para contestar el "porqué", en vez de el "qué", proporcionándole un rápido desempeño en análisis profundos y reportes. Como parte del data warehouse de NCR, el conjunto Brio Enterprise puede proporcionar a los clientes de información de la compañía datos útiles en el momento preciso vía Internet. Brio Enterprise cuenta con consultas integradas de plataforma cruzada, análisis multidimensional, diagramas interactivos y capacidad de reporte.

Los beneficios de, y el retorno de la inversión del data warehouse se multiplican cuando los administradores y usuarios en la empresa tienen acceso a la información vital de la corporación. En particular, la combinación NCR/Brio es conveniente en empresas de telecomunicaciones, menudeo y financieras, en donde NCR ha permanecido durante mucho tiempo.

Brio Technology desarrolló la familia de soluciones de apoyo a decisiones Brio Enterprise para facilitar a los administradores, vendedores y profesionistas en mercadotecnia tomar decisiones basadas en el conocimiento del negocio.

La familia de productos Brio Enterprise incluye:

- Brio Query: Es una herramienta para usuarios del negocio que requieren acceso directo al data warehouse, permite realizar consultas, análisis, gráficos en 3-D y SmartReports que permite integración dinámica de metadatos y cuenta con un motor OLAP. Con el nuevo Open Metadata Interpreter, Brio Query tiene la capacidad de leer metadatos existentes eliminando la dependencia sobre metadatos propietarios. Esta disponible para todas las plataformas cliente/servidor, incluyendo Windows NT y Unix.
- Brio.Insight: Permite realizar OLAP interactivo a través de un navegador. Está diseñado para consumidores que necesitan ir más allá de los reportes estadísticos para realizar análisis interactivo en ambientes conectados a la red.
- Brio.QuickView: Posee un portafolio de reportes dentro de un navegador similar a ReportViewer, que da a los usuarios mayor flexibilidad y una forma eficiente de visualizar la información.
- Brio Query.Server: Integra conocimiento del negocio en toda la empresa estableciendo ligas entre el data warehouse de la organización y los usuarios que utilizan cliente/servidor y web. Brio Query. Server maximiza los recursos del sistema a través del procesamiento de consultas y la producción de reportes fuera de las hora pico y distribuye los resultados de manera automática vía correo electrónico, el web, servidores de archivos o impresoras.

9.7 INFORMIX

La familia de productos Informix Metacube cuenta con metadatos para informar acerca de los datos que deben ser incluidos en el data warehouse. Metacube Warehouse Manager permite crear y administrar un modelo de metadatos que incluye la representación lógica y multidimensional del data warehouse. Esta representación incluye la descripción de varias medidas de las estructuras de datos fuente, las dimensiones y niveles de dimensiones utilizadas para la consolidación, los atributos de los elementos de dato, etc.

Para facilitar la administración del modelo de metadatos Informix Metacube Warehouse Manager proporciona un editor de jerarquías gráfico que facilita especificar las relaciones entre los elementos en una jerarquía dimensional. Además se puede incluir ayuda en línea para cada uno de los atributos en el modelo de metadatos.

El modelo de metadatos también proporciona facilidades para especificar todas las características físicas de la base de datos, tales como tablas, uniones y columnas. Con Metacube Warehouse Manager se puede mapear el modelo lógico con la representación física del modelo en la base de datos. El modelo de metadatos también incluye la información física relevante sobre consolidación, partición y los tamaños relativos para las optimizaciones de desempeño basadas en el costo.

Metacube Warehouse Manager permite definir cómo desplegarán la información que se encuentra dentro del data warehouse las aplicaciones de apoyo a las decisiones. Cualquier cambio físico a la base de datos puede significar que la representación en el modelo lógico difiera del modelo físico. Metacube Warehouse Manager ofrece un modelo de sincronización que realiza comparaciones entre el modelo de metadatos lógico y las tablas físicas actuales y permite reajustar el modelo lógico cuando sea necesario. Metacube Warehouse Manager permite el acceso a todas sus funciones a través de una interface gráfica.

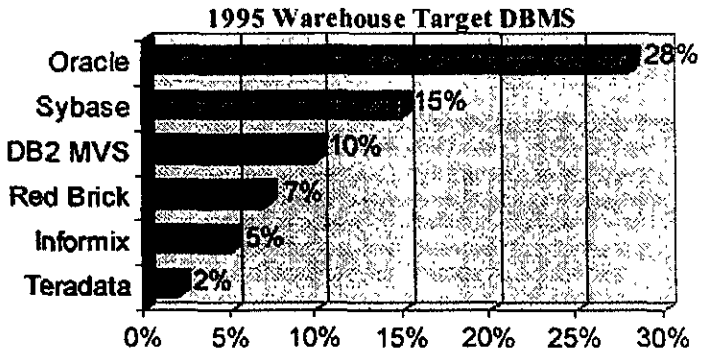
Informix ha desarrollado una estrategia enfocada en la tecnología, de servicio superior al cliente, y fuertemente cooperativa para ambientes de cómputo corporativo que van desde grupos de trabajo en OLTP y aplicaciones de data warehouse.

9.8 VISUAL WAREHOUSE DE IBM

Visual Warehouse es utilizada para obtener datos de fuentes de datos corporativos o de un warehouse centralizado al ambiente de los grupos de trabajo. Visual Warehouse ayuda a los usuarios con un mínimo conocimiento técnico a crear y mantener un data warehouse a un nivel de unidad de negocio o de departamento.

Soporta múltiples bases de datos como son: DB2 para OS/2, para AIX y para OS/400; y permite extraer datos de la familia DB2, IMS, archivos planos, VSAM, ORACLE y SYBASE.

Visual Warehouse contiene una solución completa para el ambiente LAN, satisfaciendo las necesidades de un departamento de accesar los datos del negocio. Esta versión (v1.3) posee adaptadores para warehouses no-relacionales, lo cual permite a Visual Warehouse extraer datos de archivos planos IMS, VSAM y MVS, así como VM. Proporciona además las actividades necesarias de diseño e implantación de un data warehouse robusto.



Fuente: Meta Group ADS Data Warehousing 1995

Fig. 13.- Mercado del data warehousing

10. ANÁLISIS: DATA WAREHOUSING COMO FACTOR COMPETITIVO EN LA TOMA DE DECISIONES

El oficio de la gerencia es difícil ya que no hay fórmulas que uno pueda seguir. Dentro de las restricciones sociales y legales, se tienen que alcanzar resultados económicamente aceptables, de cualquier modo que éstos se definan. No basta diseñar la estructura organizacional y supervisar la construcción de la misma, para no limitarse a sólo reaccionar a las medidas de los competidores o a las fluctuaciones del mercado, la gerencia tiene la responsabilidad de desarrollar una estrategia eficaz para la organización. El papel del estratega es particularmente exigente puesto que requiere comprender la esencia del negocio y cómo está cambiando, esto significa que la dirección debe entender la tecnología aplicable y cómo ésta modifica los productos y procesos, evaluar a los clientes y desarrollar capacidad organizacional medida frente a la competencia.

Lo particularmente difícil de una estrategia es que casi siempre se requiere que los miembros de la organización cambien en formas fundamentales. Esto perturba las relaciones sociales entre individuos y grupos, y a menudo se requiere que la gente desarrolle nuevas habilidades.

Para formular una estrategia corporativa eficaz, se requiere tener visión y creatividad, ayudados por un análisis cuidadoso de lo que está ocurriendo en la industria de uno y de la forma en que la eficiencia corporativa se puede traducir en ventaja competitiva.

Podemos decir que las tareas clave que constituyen los fundamentos del trabajo de la dirección son: modelar el ambiente de trabajo, fijar la estrategia, asignar recursos, formar gerentes, crear la organización y supervisar las operaciones.

El ambiente de trabajo de una compañía está definido por tres elementos principales:

- Las normas de rendimiento que imperan y que fijan el ritmo y la calidad de los esfuerzos de la gente,
- Los conceptos mercantiles que definen cómo es la compañía y cómo opera, y
- Los conceptos sobre las personas y los valores que imperan y que definen cómo se trabaja ahí.

De estos tres elementos, las normas de rendimiento constituyen el más importante porque, hablando en términos de dirección, definen la calidad del esfuerzo que hace la organización. Las altas normas provienen , desde luego, de algo más que metas exigentes.

La visión estratégica de la dirección que toma en cuenta la industria, los clientes y el ambiente específico, lleva a una innovación dirigida a una posición competitiva particular. Eso es lo que distingue una visión útil de las generalidades sin sentido que en ocasiones se usan para describir una estrategia.

Identificar los puntos de competitividad en producto, en características, en servicio es una prioridad y está implícito que en esta realización se debe entender en detalle cómo se comparan sus productos y servicios con los de los competidores. Demasiados directivos basan sus estrategias en supuestos no comprobados y en varias ilusiones sobre su rendimiento corporativo. Hoy no se puede hablar sobre estrategia sin hablar de dar a los clientes mejor valor que los competidores.

Reconociendo que es difícil generar ventajas competitivas duraderas, los directivos deben basarse en las capacidades existentes mientras buscan al mismo tiempo nuevas fuentes de ventaja.

Otra función fundamental de los directivos es supervisar las operaciones y la ejecución. Eso significa dirigir el negocio día por día, produciendo planes sensatos, descubriendo temprano problemas y oportunidades y respondiendo vigorosamente a ellos. Los grandes directivos utilizan información para descubrir temprano los problemas e identificar ventajas potenciales. No es cuestión de disponer de más información, simplemente de usar mejor la información; las cifras y los hechos tienen significado para aquellos que conocen a los clientes, los productos y los competidores, y nunca dejan de estudiar los hechos y cifras en busca de indicios de una ventaja en el mercado.

Ante todo, es necesario mantenerse informados acerca de una amplia gama de decisiones operativas que se toman a diversos niveles de la compañía. En estas circunstancias, es posible aplazar un a decisión, darle una nueva dirección o incluso impedir que se siga adelante.

Otra habilidad importante es saber economizar energía y horas para dedicarlas a cuestiones, decisiones o problemas especiales que requieren de atención personal. Las condiciones de los negocios cambian de manera continua y rápida y hay que revisar la estrategia corporativa para tomar en cuenta el cambio.

Las relaciones entre diferentes propuestas ofrecen oportunidades de combinación y reestructuración, es necesario reconocer que el directivo tiene un amplio campo para amplios intereses y curiosidad. Cuantas más cosas sepa, más oportunidades tendrá de descubrir partes que se relacionan entre sí. La contribución más significativa de un directivo puede ser ver relaciones que nadie había visto.

Para planificar estrategias eficaces, los gerentes generales tienen que entender los puntos fuertes y los puntos débiles de su compañía, la naturaleza de su industria y las características de sus competidores.

La esencia de la formulación de una estrategia puede incluir:

- Posicionar a la compañía de manera que sus capacidades proporcionen la mejor defensa contra la fuerza competitiva,
- Influir en el equilibrio de fuerzas mediante medidas estratégicas, y mejorar así la posición de la compañía, y
- Anticiparse a los cambios y a los factores subyacentes de esas fuerzas y responder a ellos con la esperanza de explotar el cambio escogiendo una estrategia apropiada para el nuevo equilibrio competitivo antes de que los opositores se den cuenta de ello.

Al momento de implantar una nueva tecnología, es necesario considerar algunos aspectos que pueden influir en el éxito o fracaso de dicha tecnología dentro de una organización.

Los cambios tecnológicos requieren que muchas partes de la empresa se adapten a ese cambio. Para que esta transición sea menos impactante podemos tomar en cuenta los siguientes puntos.

- Involucrar a los usuarios en cada fase del diseño. Si una organización es pequeña o está organizada en una jerarquía estricta, es posible diseñar los sistemas de acuerdo a las especificaciones de los altos mandos. Pero si la organización mantiene un clima menos estricto será necesario involucrar a un mayor número de usuarios en cada fase del diseño. Dado que los sistemas de apoyo a la toma de decisiones necesitan información de las bases de datos corporativas, es necesario incluir al personal del área de sistemas de información.
- Limitar el alcance. La cantidad de datos que serán transportados a los sistemas determina los costos del hardware y los parámetros de selección de las herramientas. Algunas formas de reducir la cantidad de datos dentro de los sistemas son: acordar con el usuario que se incremente la cantidad de datos con cada versión del sistema, identificar los datos que pueden mantenerse fuera de línea y aquellos que son necesarios en línea.

- Especificaciones de diseño. Es necesario recordar que las especificaciones para un sistema de apoyo a la toma de decisiones son muy diferentes de las de un sistema operativo de la empresa.

En cuanto al costo, es necesario considerar que un sistema de información no es un fin en sí mismo, sino un medio para llegar a un objetivo. No tiene en consecuencia un valor intrínseco. Su valor depende de los usos a los que sea sometido. La tecnología de la información no resuelve los problemas que una organización enfrenta, pero le permite a la organización enfrentar los problemas de una mejor manera. El ambiente en que una organización se desarrolla hoy en día es muy diferente al de hace algunos años, los límites espaciales y temporales se han reducido y como resultado tenemos ahora más competidores en cada subregión del mercado. Las organizaciones se desplazan de un ambiente de "hacer y vender" a un ambiente de "sentir y responder".

Las compañías se están viendo obligadas a modificar sus actitudes y ellas mismas se han denominado como "orientadas al mercado" por un lado y como "orientadas al cliente" por otro. Este es un proceso de cambio radical que no puede realizarse sin el apoyo de la tecnología de la información. En las nuevas organizaciones es básica la habilidad de la gente de comunicar, pues es necesario desarrollar e intercambiar conocimiento. El nivel de complejidad de estas nuevas organizaciones en consecuencia es mayor al observado en esquemas anteriores de administración.

En un mundo de negocios cambiante, es necesario que los sistemas de información apoyen los cambios a gran velocidad. Las aplicaciones deben ser capaces de evolucionar conforme la organización requiera. Algunos autores sugieren que la información es compleja, siempre en expansión e imposible de controlar completamente. El significado de la información depende del uso que se le dé y de sus usuarios.

La tecnología de la información requiere poner información corporativa a disposición de los administradores y otros usuarios finales para incrementar la productividad y, en consecuencia, obtener una ventaja competitiva. En respuesta, los administradores de tecnología de información han incrementado su capacidad en cuanto a consultas de bases de datos, OLAP y reportes para sus usuarios finales.

Hay que enfatizar que los data warehouses son realmente una arquitectura, por lo que necesita tenerse muy presente con qué tecnologías se cuenta para que éstas sean bien integradas con la solución del proveedor.

Si el sistema es dato de baja, debido a que se están realizando mejoras o por alguna razón un disco falla, todo lo que no tenía interés ahora toma importancia debido a que esto afecta la disponibilidad del warehouse, el cual tiene una gran cantidad de usuarios dependiendo de él. Así del lado de la tecnología de la información, desde el punto de vista de la infraestructura, piensan que el warehouse no es un elemento crítico para la organización. Por ello es importante que tanto la comunidad de usuarios como el personal de infraestructura se involucren en un cambio cultural.

Nos encontramos con el problema de que existe una gran variedad de fuentes de información acerca de cómo desarrollar un data warehouse, pero no se encuentra información que indique cuáles son los problemas a los que se enfrenta el desarrollo y cuáles son las posibles soluciones.

Nos encontramos con aplicaciones que le ofrecen a la organización nuevos caminos de explotación de sus datos, los requerimientos para estas aplicaciones son comunes en un alto grado, ya que buscan lo siguiente:

- Disponibilidad de datos
- Organización de datos
- Integración de datos
- Retención de datos

Todos ellos importantes consideraciones de diseño.

La implantación de un data warehouse no es una tarea fácil de realizar en ninguna organización, cualesquiera que sean sus características: tamaño, presencia en el mercado, estructura organizacional, personal, recursos, etc. Debido a que las condiciones que afectan en general a un proyecto de data warehouse varían entre una organización y otra, quizá uno de los elementos más valiosos en estos proyectos es la experiencia que posea el grupo encargado del proyecto así como la capacidad de integración con la comunidad de usuarios a los que está destinado el data warehouse y los proveedores con los cuales se establecerán relaciones que no solo deberán ser estrictamente comerciales, sino que implicarán el establecimiento de una red de transferencia del conocimiento entre todas las partes involucradas en el proyecto de data warehouse.

Otro factor que puede determinar el éxito de un proyecto de data warehouse es el alcance que se defina desde su inicio, pues dadas las características de estos proyectos es posible comenzar con un data warehouse de tamaño moderado que considere la agregación paulatina de otras áreas o aspectos del negocio conforme evolucione el grado de aceptación y uso de esta tecnología dentro de la propia organización, es decir, se puede implantar un data warehouse en un departamento y posteriormente conforme se perciba su aceptación y rendimiento extender su alcance de manera tal que abarque completamente a la organización.

El personal involucrado en el proyecto de data warehouse no puede ser puramente interno ni externo a la organización, debemos asegurarnos que se logre una cooperación entre proveedores de conocimiento, proveedores de herramientas y personal de la empresa a fin de que todos ellos aporten los elementos necesarios para llevar a buen término el proyecto establecido.

Lo anterior se debe a la necesidad de evaluar si nuestro personal posee la experiencia necesaria para involucrarse en un proyecto de esta magnitud, situación que en muchos casos no sucederá, razón por la cual nos podemos ver forzados a recurrir a la asesoría de personal externo involucrado en el desarrollo de proyectos similares en otras empresas, aunque sin el necesario conocimiento de los procesos y condiciones propias de nuestra organización, y de proveedores de herramientas genéricas que requerirán adecuaciones para el caso particular que nos ocupe. Consideramos que organizar al personal de un proyecto en esta forma beneficiará a todos los participantes del mismo, pues aportará experiencia para los nuestros y a los externos les dará un campo de ensayo y mejora de sus productos y metodologías.

Al elegir las herramientas, es necesario observar las tecnologías y asegurarse que los vendedores se integren a la organización, es necesario trabajar continuamente con ellos, por ejemplo: Oracle para el desempeño del DBMS, QDB Solutions para medir la calidad de los datos, Platinum para la solución del metadato, Evolutionary Technologies Extract ETI para el movimiento y replicación de datos.

Se recomienda seleccionar un vendedor a quien le interese el éxito del proyecto como a la organización misma, tal y como si fuera a seleccionarse un socio, de manera que el vendedor pueda estar revisando los esquemas, haciendo sugerencias, indicando cómo adecuar las tablas para tener un mejor desempeño. Es una manera de outsourcing sin serlo tal cual, debido a que el proveedor y la organización trabajan conjuntamente hasta crear un producto fuerte. No todas las grandes compañías son los mejores socios, en varias ocasiones una pequeña compañía puede proporcionar un muy buen servicio, calidad o herramientas que contribuyen a la solución de problemas en la calidad de los datos. Hay que considerar la tecnología que están utilizando.

Desde nuestro punto de vista no consideramos adecuado recomendar una metodología específica aplicable a todos los proyectos de data warehouse, pues como se ha mencionado en varias partes del texto, las condiciones entre uno y otro varían de manera importante, razón por la cual cada empresa deberá adecuar las metodologías existentes a las necesidades de su proyecto. El hacer una recomendación en este momento sería arrogancia nuestra y un intento por limitar las posibilidades disponibles en este campo, razón por la cual únicamente mencionamos la existencia de una arquitectura de referencia adaptable a distintas situaciones que podrían en un momento dado presentarse en un proyecto de data warehouse, y permite a la organización agregar o eliminar libremente elementos o fases a esta arquitectura de referencia.

Una preocupación latente en todo proyecto de data warehouse es la afectación que sufrirán los sistemas de información actuales con que cuenta la organización al momento de implantar el nuevo sistema de data warehouse, nuestra investigación muestra que fuera de afectarlos, puede darles mayor funcionalidad.

Aunque no significa que sistemas de apoyo a las decisiones previos (hablando de sistemas como los basados en el conocimiento, de intercambio electrónico de datos, CBR, EIS, etc.) proporcionen información poco útil y tampoco que el uso de ellos apoye deficientemente a la organización en el desempeño de sus actividades, sino que con la combinación de uno de estos sistemas y la tecnología de data warehousing, la información presentada al usuario será de mucho mayor exactitud y será la más representativa acerca de lo que está buscando, llegando así a una explotación óptima de su información. El data warehouse puede reducir la necesidad de mantener un conjunto disperso de aplicaciones de apoyo a la toma de decisiones, encargándose de agrupar las funciones de varios de ellos dentro de una aplicación más completa y funcional.

La organización puede percibir los beneficios que le aporta el data warehouse conforme la comunidad de usuarios que se benefician de él en forma real se incremente y se cumpla con su objetivo primario de mejorar el proceso de toma de decisiones. Debido a que cada decisión tomada afecta los resultados de operación de la empresa, contar con información adecuada en el momento oportuno pude traducirse en beneficios para la misma. En todo caso la única desventaja de un data warehouse sería no utilizarlo.

La confiabilidad de la información que proporciona un data warehouse no depende directamente de él mismo, sino que tendríamos que profundizar en el origen y obtención de datos, pues es conveniente recordar que en cualquier sistema hay una máxima que dice "si al sistema entra basura, del sistema sale basura", por lo cual la confiabilidad de la información arrojada por el data warehouse será mayor conforme mejores sean los procesos de obtención de datos desde sus fuentes primarias.

La administración del data warehouse es una función igualmente cambiante entre una organización y otra, dependerá fundamentalmente de los requerimientos que tengan los directivos de información que apoye la toma de decisiones y de la forma como opere la empresa, ya que por ejemplo, algunas organizaciones requerirán que su información sea actualizada cada semana por lo que semanalmente se ejecutarán los procesos de carga y sumarización de infromación respectivos, mientras que otras requerirán su información mensual y procederán en consecuencia a este requerimiento.

En realidad, las tecnologías que el ambiente del data warehouse incorpora no son nuevas, son la convergencia de productos que han venido evolucionando durante la última década, como son: sistemas manejadores de bases de datos distribuidas, productos para la conversión de datos, tecnologías cliente/servidor y la integración de herramientas de análisis y reporte que sirven como base para los analistas del negocio.

El data warehouse representa una oportunidad trascendente para la administración de datos, ya que su enfoque es principalmente hacia la entrega de datos integrados de alta calidad para ser utilizados por los tomadores de decisiones de la empresa, la administración de datos aparece como la propietaria de la lógica del data warehouse. Otro aspecto relevante en el entorno del data warehouse son los metadatos, ya que podemos considerarlos como un recurso dentro de la organización. Aunque, en la mayoría de los casos sólo se mantienen los metadatos relacionados con el actual portafolio de productos en la base de datos debido a que las bases de datos operativas cambian constantemente no se conservan los primeros datos. En el ambiente de data warehouse debemos proveer acceso a datos históricos que ya han sido archivados, por lo tanto debemos ser capaces de recuperar información utilizando el formato de la base de datos que se encontraba vigente en el momento en que los datos fueron almacenados. El manejo de las versiones de los metadatos es una de las mayores tareas para la administración de datos dentro del ambiente de data warehouse.

El mercado del data warehouse esta cambiando a gran velocidad. En tanto los proveedores y los usuarios tratan de mantenerse actualizados en este cambio, tienden a enfocarse en como aplicar las nuevas tecnologías a los problemas de los negocios, en vez de buscar como resolver los problemas del negocio con las nuevas tecnologías. Al justificar un sistema de data warehouse debemos enfocarnos en los objetivos y beneficios del negocio más que en adoptar la tecnología sólo porque es lo último del mercado.

El objetivo de un sistema de data warehouse es incrementar la calidad y exactitud de la información del negocio y hacer llegar esta información a los usuarios del negocio en una forma accesible y entendible. Un data warehouse es un lugar para almacenar datos. Un sistema de data warehouse proporciona una solución completa para entregar información a los usuarios.

Un sistema de data warehouse proporciona a los usuarios conocimiento del negocio para tomar decisiones. Considerando este conocimiento del negocio como una nueva generación de tecnologías de acceso, manipulación y presentación de datos que permite a los usuarios responder a cuestiones del negocio utilizando datos internos y externos. Podemos pensar a un sistema de data warehouse como un sistema que proporciona herramientas para el conocimiento del negocio que procesan y transforman los datos almacenados en información del negocio. Los datos para el warehouse en la actualidad son tomados e integrados de múltiples sistemas operacionales internos y de algunos proveedores externos, sin embargo, se espera que los data warehouses contengan datos de otras fuentes.

Al diseñar y construir un data warehouse se debe considerar que existen diferentes tipos de decisiones y que cada tipo requiere de herramientas y datos distintos del negocio. Los sistemas de warehouse son instrumentos que apoyan la toma de decisiones tácticas relacionadas con la operación diaria del negocio y en las decisiones estratégicas que involucran planeación a largo plazo.

Las herramientas para el conocimiento del negocio soportan tres tipos de tareas principales:

- Búsqueda y reporte de hechos conocidos.
- Análisis de hechos conocidos.
- Descubrimiento de hechos desconocidos.

Las tareas de *búsqueda y reporte* hacen referencia al despliegue de la información de manera visual, las herramientas que apoyan este tipo de tareas han existido por muchos años. Estas herramientas se usan en muchos casos para dar seguimiento a las operaciones diarias del negocio. En este contexto un warehouse ofrece la ventaja de que los datos han sido seleccionados e integrados desde múltiples sistemas operacionales.

De este modo el warehouse contiene datos detallados que reflejan el estado actual de los datos en los sistemas operacionales y de este modo son conocidos como data warehouse operacionales o almacenes de datos operacionales.

Las herramientas de *análisis de datos* proporcionan capacidades de modelado multidimensional, funciones estadísticas y matemáticas y proyecciones. Son utilizadas para analizar y proyectar tendencias y medir la eficiencia del negocio a través del tiempo. Estas evaluaciones proporcionan soporte para la toma de decisiones estratégicas del negocio y para prever la forma de mejorar la eficiencia de las operaciones del negocio. Este tipo de procesamiento es conocido como procesamiento analítico en línea OLAP (Online Analytical Processing).

Las herramientas OLAP permiten al usuario analizar y dividir o bien integrar datos a través de múltiples dimensiones tales como tiempo, mercado y/o tipo de producto.

El data warehouse para OLAP no solo ofrece la ventaja de filtrar e integrar datos sino también la de disponer de datos históricos esenciales para proyecciones y de análisis de tendencias. Un warehouse que soporta OLAP puede ser considerado como un sistema de data warehouse de apoyo a las decisiones.

Actualmente existen diferencias sobre el uso de las herramientas de OLAP en análisis de datos multidimensionales (MDA -Multidimensional Data Analysis-). Los vendedores de estas herramientas ofrecen dos tipos de ellas: aquellas que accesan datos almacenados en sistemas de bases de datos multidimensionales (MDBMS) y aquellas que accesan datos almacenados en bases de datos relacionales. El debate se centra en determinar que tipo de DBMS es mejor para almacenar y dar mantenimiento a datos multidimensionales.

Un análisis de los productos de análisis de datos multidimensionales podría enfocarse a dos aspectos principalmente:

1. La forma como las herramientas de MDA satisfacen los requerimientos del usuario.
2. El desempeño y la escalabilidad.

Los MDBMS proporcionan capacidades tanto de análisis como de administración de datos. Proporcionan un desempeño aceptable al analizar y explorar múltiples niveles de datos resumidos, pero son menos adecuados para la manipulación de grandes cantidades de datos detallados. Cuando las herramientas de MDA son utilizadas con datos almacenados en bases de datos relacionales, el RDBMS proporciona la capacidad para la manipulación de datos, mientras que el front-end del cliente proporciona el mecanismo de análisis. La ventaja de esta opción es que en el RDBMS se pueden almacenar datos tanto detallados como resumidos, y los datos pueden ser compartidos con otras herramientas del negocio.

Las herramientas de análisis de datos trabajan por lo general con datos resumidos. Aunque se pueden construir sumarios durante el procesamiento analítico, es mucho mas eficiente preconstruirlos siempre que sea posible. Hacer lo anterior reduce considerablemente sobrecargas de procesamiento y simplifica el trabajo al usuario. Los sumarios son almacenados en bases de datos especiales conocidas como data marts. Los data marts son construidos a partir de datos históricos detallados almacenados en data warehouses de sistemas DSS (Decision Support Systems) y, en algunos casos, son construidos directamente a partir de las bases de datos operacionales o de los data warehouses operacionales.

Las herramientas de búsqueda, reporte y análisis de datos son utilizadas para procesar o visualizar hechos conocidos. En otras palabras, los usuarios de estas herramientas saben que tipo de información desean accesar y analizar. Sin embargo, una nueva rama de conocimiento del negocio ha comenzado a surgir y se refiere a explorar datos de hechos desconocidos, esto es, información que es desconocida para el usuario.

Este estilo de procesamiento permite a los usuarios buscar nuevas oportunidades de negocio y buscar patrones de datos antes desconocidos.

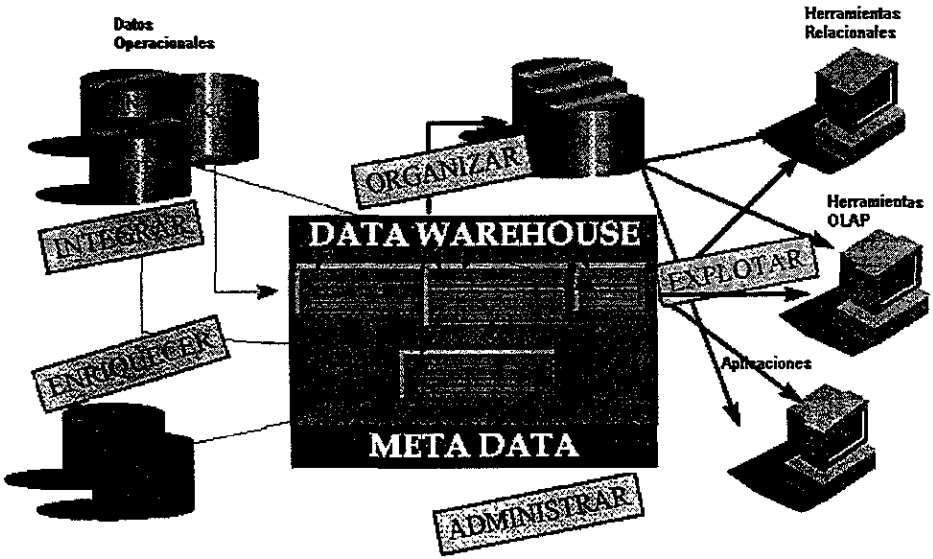


Fig. 14.- Visión integral de un data warehouse

La *exploración de datos* significa "excavar" a través de grandes cantidades de datos almacenados en la mayoría de los casos en data warehouses de sistemas DSS. Las herramientas que soportan la exploración de datos son conocidas como herramientas de minado de datos o de descubrimiento de datos. Sin embargo algunos autores las incluyen bajo el titulo de herramientas de OLAP, e incluso algunos utilizan el término minado de datos para cubrir todos los aspectos del conocimiento de un negocio.

Propuestas de uso:

En cuanto a los usos más comunes, con base en un estudio realizado por MetaGroup, encontramos que el enfoque principal es hacia la mercadotecnia, finanzas, ventas y sistemas de información de clientes.

A continuación se mencionan a manera de ejemplo, algunos posibles usos que se pueden dar a esta tecnología, los cuales se encuentran clasificados de acuerdo al giro de la empresa en la cual son aplicables:

Financieras:

- Análisis de transacciones por periodos: retiros, depósitos, cambios de moneda, aperturas de cuentas, etc.
- Análisis en cambios de tasas de interés.
- Análisis acerca de la aplicación de tarjetas de crédito.

De transporte:

- Análisis de rutas.
- Cantidad de usuarios.
- Ingresos.

De servicios:

- Necesidades básicas.
- Demanda de productos.
- Variaciones de precios.
- Competencia.
- Lealtad de clientes.

Educativas:

- *Análisis de ingresos: por colegiaturas, por donaciones, por venta de material didáctico,* etc.
- Demanda.
- Análisis de la población estudiantil: hombres, mujeres, por grado, etc.
- Recursos.

Comerciales:

- Inventarios.
- Demanda.
- Análisis de mercado.

CONCLUSIONES

El hombre durante su devenir histórico se ha provisto de medios que le permitan realizar sus tareas de mejor manera, con mayor precisión y facilidad, en nuestros días se ha alcanzado un nivel de sofisticación tal que podemos hablar de sistemas de inteligencia artificial, dispositivos para la realización automática de procesos, sistemas de apoyo a las decisiones y así progresivamente, sin embargo debemos detenernos en estos últimos, pues han sido el objeto de nuestra investigación y consideramos necesario resaltar que si bien, estos sistemas pueden detectar tendencias, realizar análisis estadísticos y matemáticos, sugiriendo cursos de acción, la decisión final es responsabilidad de seres humanos, pues aún los más intrincados sistemas de cómputo son incapaces de suplir al hombre en una función que siempre será prerrogativa suya: la toma de decisiones. La mayoría de los casos los datos no mienten, pero la interpretación que hacemos de ellos si.

Si hacemos un recorrido a lo largo de la historia del hombre, encontraremos etapas bien definidas que han evolucionado desde las antiguas sociedades nómadas hasta nuestra "sociedad de conocimiento", logrando identificar marcadas diferencias entre los factores que determinaron la riqueza de esas sociedades. De este modo presenciamos que el conocimiento se ha convertido en el elemento fundamental de ventaja competitiva en las organizaciones y encontraremos a la tecnología de la información como su soporte principal.

Durante años, las organizaciones consideraron al producto como el elemento más importante de su operación, enfocando sus esfuerzos a obtener productos cada vez "mejores"; pero se olvidaron de tomar en cuenta los gustos, necesidades y conducta de sus clientes. Sin embargo, este viejo esquema de administración ha puesto de manifiesto sus deficiencias en los últimos años, cuando nos enfrentamos a importantes cambios en la cultura y comportamiento de los mercados internacionales y en particular de los individuos.

Ante esta situación surgen nuevos modelos de negocio enfocados al cliente, en contraste con aquellos enfocados al producto, donde de forma vertiginosa la información que las empresas poseen y que han almacenado durante años cobra especial importancia, pues los tomadores de decisiones de la empresa se dan cuenta de que en esos datos, se encuentran ocultos los patrones de comportamiento de los clientes, las tendencias de los mercados, los ciclos de los productos y una infinidad de aspectos que de ser identificados de manera correcta y oportuna pueden definir el éxito o fracaso del negocio.

Pero las organizaciones se enfrentan a una competencia cada vez mayor y a mercados sumamente exigentes y cambiantes, lo que las lleva a reconocer la necesidad de contar con herramientas y aplicaciones suficientemente rápidas, confiables y eficientes para analizar los datos y mostrar los resultados de manera comprensible al usuario final. Bajo tales condiciones, las organizaciones de nuestros días apuestan cada vez con mayor vehemencia a las nuevas tecnologías en busca de la preciada solución que les permita tomar ventaja sobre sus competidores y ofrecer a sus clientes más y mejores productos y/o servicios. Por su parte, tanto las empresas fabricantes de hardware como las desarrolladoras de soluciones de software lanzan al mercado sus productos pugnando cada una por su parte ser "la empresa líder en soluciones a problemas de cómputo", donde las empresas se enfrentan a una infinidad de opciones que prometen resolver sus problemas de datos. Esta situación puede llevarnos a sobrestimar las ventajas de estas nuevas tecnologías y a caer en un uso inapropiado de las mismas, generando con ello errores que tendrán repercusiones en la inversión realizada, cambios culturales en la organización, su entorno y en la política necesaria para mantener en la organización un proyecto de este tipo.

La naturaleza del data warehouse es tal que su impacto en una organización es total: involucra y afecta a la tecnología existente, altera la forma como las diversas unidades funcionales de la organización trabajan, e impacta de manera significativa la forma como la gente, procesos y tecnología son administrados. Es claro que un data warehouse está más allá de un ejercicio trivial, por ello es necesario hacer un gran esfuerzo de planeación para realizarlo correctamente.

Las promesas del data warehousing han capturado la imaginación de los profesionales de sistemas y de sus colegas encargados de la dirección de estas áreas. Más allá de contar con la capacidad de realizar consultas que permitan soportar la toma de decisiones, el data warehouse esta siendo considerado como una forma mas completa de comprender a los clientes y sus requerimientos, conocer las condiciones financieras de productos y servicios y acelerar la introducción de nuevos productos y servicios al mercado. Y aún más - y quizá lo más arriesgado – el data warehousing se concibe como la tecnología que dará a las empresas la capacidad de generar mayores ganancias y mejorar su rendimiento.

Un aspecto que es importante considerar en un data warehouse, es su efecto sobre la gente de la organización. Al iniciar cualquier proyecto de sistemas dentro de una organización, es necesario tomar en cuenta todos los factores que los puedan influir, y uno de ellos es la gente, el éxito de un data warehouse depende en gran medida del grado de cooperación que se logre por parte de todas y cada una de las áreas funcionales de la empresa. Muchas organizaciones han invertido una gran cantidad de tiempo, dinero y prestigio en la construcción de data warehouses que no han podido utilizar. Cuando los problemas se presentan, la última fuente de ellos es la gente. Quizá no se planeó adecuadamente la comunicación sobre tiempo, costo, y recursos requeridos para construir el data warehouse; y entonces se presentan los problemas de datos incompletos, inadecuados. O quizá los intereses de la gente acerca de los cambios que generará la introducción de un data warehouse no son orientados de forma adecuada, y a menudo los participantes de un proyecto mantienen intereses ocultos que pueden detener o destruir un proyecto de data warehouse.

Entre los intereses más comunes encontramos la del síndrome de "son mis datos". Los gerentes de producto e individuos encargados de ciertas unidades funcionales desarrollan fuertes sentimientos de protección hacia los datos que son generados o utilizados por sus departamentos. Otros intereses involucran control, poder, miedo, y protección personal. Esta situación puede utilizarse para ocultar excepciones a las políticas, tratos especiales y otras condiciones inusuales.

Descubrir la existencia de intereses ocultos es la clave para eliminarlos como barreras para el progreso. Una forma de lograrlo es establecer programas y hacer que los individuos se sujeten a ellos basándonos en el razonamiento de que al atender las tareas que se deben lograr es fácil separar los aspectos reales de aquellos que son personales o de autoprotección.

Cubrir los intereses "personales o de autoprotección" puede requerir de una cantidad considerable de la capacidad de la organización. En el caso más simple, discutir y capacitar al personal acerca de la importancia de los datos como un recurso corporativo puede resolver el problema. Sin embargo, en otros casos, los ejecutivos responsables del proyecto deben proporcionar "fuerza organizacional" y resolver el problema. Establecer claramente que *los datos son un recurso corporativo y disponible para todos aquellos que* lo necesiten puede ser la mejor forma de evitar problemas y de fijar un punto de vista objetivo de la situación.

La propiedad de los datos es un problema universal. Las firmas que hacen un buen uso de los datos tienen también la capacidad de:

- Establecer una cultura en la cual los datos son propiedad del negocio para beneficio de la organización entera y de sus clientes,
- Procesar toda la información en cada transacción una sola vez, de manera completa y exacta desde el primer empleado de la organización que recibe dicha información,
- Proporcionan acceso a los datos a todos aquellos que lo necesitan,
- Extienden los procesos de información para incluir tanto a clientes como vendedores en donde sea posible.

Los intereses ocultos siempre existirán. El truco para los ejecutivos en informática es manejar adecuadamente el proceso y utilizar los "intereses ocultos" en beneficio de la organización.

Ahora bien, las organizaciones de nuestro país tienen a su disposición la posibilidad de adquirir diversas herramientas que les permiten planear, desarrollar y administrar su propio data warehouse, de modo tal, que pueden tener pleno control de su información aprovechando con ello la totalidad de sus recursos.

Los datos deben estar disponibles cuando son necesarios, los usuarios a menudo encuentran difícil accesar y analizar los datos que se encuentran en distintos lugares y probablemente no conocen qué datos están disponibles en cada base de datos. Finalmente, los usuarios se encuentran con la dificultad de integrar los datos obtenidos de las bases de datos.

Al realizar esta investigación nos hemos dado cuenta de la infinidad de información disponible acerca de cada uno de los aspectos involucrados en un proyecto de data warehousing. En este trabajo hemos logrado concretar la información necesaria y suficiente para comprender con claridad lo que esta nueva tecnología implica para las organizaciones y, sobre todo, para los profesionales en el área de la informática que deseen familiarizarse con ella. Sea pues este el precedente para todos aquellos profesionales de la informática que deseen profundizar en el campo del data warehousing y que, apoyados por la experiencia en el área logren desarrollar e implantar modelos de esta tecnología que beneficien a la sociedad, en particular a la de nuestro país cuyas características no han sido aprovechadas en su totalidad.

Una última reflexión es importante: *si bien es cierto que la carencia de información nos lleva a la ignorancia y en consecuencia, posiblemente al fracaso; el exceso de ella nos hace conformar una masa tan impenetrable como la ignorancia misma, recordemos pues que la ignorancia es una desgracia voluntaria.*

GLOSARIO

Adición: Actividad de combinar datos de diversas tablas para formar una unidad de información más compleja .

Agregar: Incorporar múltiples fuentes de datos o dimensiones para crear una dimensión nueva.

API (Application Program Interface): Es el conjunto de llamadas, subrutinas o interrupciones de software que comprenden una interface documentada de manera que un programa de alto nivel, tal como un programa de aplicación, puede hacer uso de los servicios y funciones de otra aplicación, sistema operativo, sistema operativo de red, manejador u otro programa (software) de bajo nivel.

Aplicación: Algún programa para la inserción, modificación, consulta de datos o creación de reportes que procesa datos para el usuario, esto incluye al software de productividad (hojas de cálculo, procesadores de palabras, programas de bases de datos, etc.), programas hechos a la medida y programas para manejo de nómina, inventarios y facturación.

Apuntador: En el manejo de bases de datos un apuntador es una dirección que contenida en los datos especifica la localización de datos en otro registro o archivo.

Arquitectura de red: El diseño de un sistema de comunicaciones, el cual incluye el hardware, software, métodos de acceso y protocolos a utilizar. Además define el método de control: si las computadoras pueden actuar independientemente o son controladas por otras computadoras monitoreando la red. Esto determina la flexibilidad futura y la conectividad a redes foráneas. El método de acceso en una LAN puede ser Ethernet, Token Ring y LocalTalk.

Arquitectura de software: El diseño de una aplicación o sistema de software que incorpora protocolos e interfaces para interactuar con otros programas y para flexibilidad y expansiones futuras. Un programa stand-alone puede tener programación lógica pero no una arquitectura de software.

Base de datos distribuida: Es una base de datos donde varias computadoras, conectadas a una red, pueden compartir parte de sus datos haciéndolos accesibles a otros usuarios.

Base de datos multidimensional: Base de datos diseñada alrededor de un conjunto de dimensiones; se usa en el análisis multidimensional.

Base de Datos: Es una colección de datos y ligas entre ellos, estructurados de tal manera que pueden ser accesados por un número diferente de programas de aplicación o lenguajes de consulta.

Becnchmark: Prueba de desempeño de una computadora o dispositivo periférico. El mejor benchmark es el conjunto actual de programas de aplicación y archivos de datos que utilizará la organización. Corriendo benchmarks sobre una computadora única es razonablemente efectivo; sin embargo, probar un sistema multiusuario es más complicado. A menos que el ambiente usuario pueda ser duplicado casi por completo, el benchmark puede ser inexacto. Herramientas: Linpack, Dhrystones, Whetstones, Khornerstones and SPECmark.

Campo: Unidad física de datos de uno o más bytes de tamaño. Una colección de campos forman un registro. Un campo define una unidad de datos sobre un documento fuente, pantalla o reporte. El campo es el común denominador entre el usuario y la computadora. Cuando se realiza una consulta o modificación a la base de datos, la referencia a la base es por el nombre del campo. Existen varios términos que se refieren a la misma unidad de almacenamiento como un campo.

Cliente/Servidor: Arquitectura en la cual el cliente realiza peticiones y el servidor provee de resultados. El cliente provee al usuario la interfase y lleva a cabo parte o todo el procesamiento de la aplicación. El servidor mantiene la bases de datos y procesos requeridos del cliente para extraer o modificar datos de la base de datos.

Cliente: Estación de trabajo o computadora personal en un ambiente cliente/servidor. Un final del espectro en un requerimiento entre programas.

Consulta: Petición formal claramente especificada de información.

Data mart: Implementación de data warehouse con un ámbito de datos y funciones de data warehouse más pequeño y restringido, que sirve a un departamento único o una parte de la organización.

DBMS: Es una colección de software que administra el acceso y modificación a una base de datos.

Deadlock: Estancamiento que ocurre cuando dos elementos en un proceso están esperando cada uno para que otro responda. En una red, si un usuario está trabajando con un archivo A y necesita del archivo B para continuar, y otro usuario está utilizando el archivo B y necesita del archivo A, entonces ambos se encuentran en espera hasta que se desocupe el recurso que desean utilizar. El software debe ser capaz de resolver esta situación.

Host: Computadora principal en un ambiente de procesamiento distribuido. Generalmente se refiere a un computador de tiempo compartido o un computador central que controla una red.

Join (unión): En el manejo de bases de datos, combinar un archivo con otro de acuerdo a una condición determinada creando un tercer archivo con datos de los archivos comparados.

Ligas físicas: (1) Conexión electrónica entre dos dispositivos. (2) En el manejo de datos, un apuntador en un índice o registro que se refiere a la localización física del dato en otro archivo.

Ligas lógicas: Los usuarios relacionan un dato lógicamente por el nombre del elemento, sin embargo, los campos actuales del dato se encuentran físicamente en sectores del disco.????

Log: Registro de la actividad de la computadora utilizado para propósitos estadísticos, así como recuperación y respaldo.

Middleware: Software que se encuentra entre la aplicación y el programa de control (sistema operativo, programa de control de red y DBMS). Provee una interfase de programación única para una aplicación que será escrita, y la aplicación correrá ambientes diferentes tal y como el middleware lo hace.

Minicomputadora: Computadora de escala media que funciona como una estación de trabajo única, o como un sistema multiusuario con hasta cientos de terminales.

Normalización: Proceso que identifica los datos redundantes que pueden existir en una estructura lógica, determina claves únicas necesarias para el acceso de los elementos de datos y ayuda a establecer las relaciones necesarias entre los elementos de datos.

OLAP: (OnLine Analytical Processing database) Base de datos diseñada para accesar rápidamente a datos resumidos. Utilizando técnicas especiales de indexación, procesa consultas que pertenecen a grandes cantidades de datos más rápido que una tradicional base de datos relacional.

OLTP: (OnLine Transacton Processing) Procesamiento de transacciones tal y como son recibidas por la computadora. También llamado "en línea" o "sistemas de tiempo real", los archivos maestros son modificados tan pronto como las transacciones son introducidas en las terminales o llegan sobre las líneas de comunicación.

Plataforma: Arquitectura de hardware de un modelo particular o familia de computadoras. Es el estándar con el cual los desarrolladores de software escriben sus programas. El término a menudo se refiere al sistema operativo, el cual implica una arquitectura de hardware particular.

Procesamiento distribuido: Sistema de computadoras conectado en una red, cada computadora maneja su carga de trabajo localmente, y la red ha sido diseñada para apoyar al sistema como un todo.

Query: ver consulta.

Registro: Conjunto de campos relacionados que almacenan información acerca de un elemento.

Servidor: Computadora que en una red LAN almacena los programas y archivos de datos compartidos por usuarios sobre la red.

Tipo de Dato: Conjunto de valores con una representación física.

Usuario: Persona o programa de aplicación que accesa a una base de datos.

VLDB: Acrónimo de Very Large Data Bases

REFERENCIAS BIBLIOGRÁFICAS

[1] *"The OLAP Report: Succeeding with Online analytical processing" Business intelligence*, 1995

[2] ABITEBOUL, Serge/ HULL, Richard/ VIANU, Vianu, *Foundations of databases*, Addison-Wesley, 1a. ed., 1995.

[3] BARROS, Oscar, *Information Systems. Databases: Their creation, management and utilization, "Management information systems, types and integration"*, Pergamon, Vol. 6, No. 4, Pag. 243 - 254, New York, 1981.

[4] *Building Effective Decision Support Systems*, Prentice Hall, 1982

[5] COMPUTERWORLD, *Vendedores de Data Warehouse Realizan Extracción de Datos*, Pag. E-21, Año 17, No. 477, México D.F., Julio 22-26 de 1996

[6] *Decision Support Systems: An Organizational Perspective*, Addison-Wesley, Massachusetts, 1978

[7] HARJINDER, S. Gill/ PRAKASH, C. Rao, *Data Warehousing. La integración de información para la mejor toma de decisiones*, Prentice Hall Hispanoamericana - QUE, Pp.:382, México, 1996.

[8] *HP DataMart Manager Product Brief*, Hewlett Packard, 1996

[9] IIVARI, Juhani/ HIRSCHHEIM, Rudy, *Information Systems. Databases: Their creation, management and utilization, "Analyzing information systems development"*, Pergamon, Vol. 21, No. 7, Pag. 551 - 575, Great Bretain, 1996.

[10] JENSEN, Christian S./ SNODGRASS, Richard T., *Information Systems. Databases: Their creation, management and utilization, "Semantics of time-varying information"*, Pergamon, Vol. 21, No. 4, Pag. 311 - 352, Great Bretain, 1996.

[11] LANING, Laurence J., *Database Advances "Corporate data architecture: the key to supporting management"*, 1993.

[12] LARSON, James A., *Database Directions. From Relational to Distributed, Multimedia, and Object-Oriented Database Systems*, Prentice Hall PTR, Pp.: 261, Pub.: 1995

[13] LOUCOPOULUS, Pericles/ZICARI, Roberto Zicari, *Conceptual modeling, databases and CASE. An integrated view of Information Systems Development*, John Wiley & Sons, Inc, 1ª. Ed., New York, 1992.

[14] MOXTON, Bruce, *Defining Data Mining. DBMS*, Data Warehouse Supplement, 1996.

[15] PORTER, Patrik L./ RADCLIFF, Deborah, *Data Warehousing for Grown-Ups*, Software Magazine, Junio 1997, (Casos de estudio)

[16] RYMER, J., *"Business Intelligence: The third tier" Distributed computing Monitor*, Junio 1995.

[17] SALEMI, Joe, *Guide to Client/Server Databases*, Ziff Davis, Pp.: 312, Pub.: 1995

[18] SMITH, H. A./ MCKEEN, J. D., *DataBase Advances. Measuring IS: How dows your organization rate?*, Vol. 27, No. 1, 1996.

[19] WELDON, Jay Louise, *Data Mining and Visualization. Database, programming and design.*, Mayo 1996.

[20] WHITE, Colin, *Data warehouse What's in a name? Database Programming and Design*, Vol. 9, Marzo 1996.

[21] WHITTINGTON, R. P., *Database Systems Engineering. Oxford Applied Mathematics and Computing Science Series*, Clarendon Press

[22] WINTER, Richard, *Database Programming & design VLDB vision. "Retracting our steps to a time when a gigabyte was a large database. A decade of VLDB"*, Vol. 10, No. 3, Marzo 1997.

REFERENCIAS ELECTRÓNICAS

[1E] http://www.cs.usask.ca/homepages/grads/crossman/review4.html

[2E] http://www.cait.wustl.edu.cait/papers/prism/

[3E] http://www.tekptnr.com/tpi/tdw/review/vboarl.htm

[4E] http://www.techweb.cmp.com/iw/556/56olbit.htm, EDELSTEIN, Herb, *Technology analysis: faster data warehouses* , Diciembre 4, 1995.

[5E] http://www.idwa.org/spr96/roads.htm, *All Roads Lead To The Data Warehouse*, DOUG, Laney, Consulting Manager, Prism Solutions