Universidad Autónoma de Querétaro

Facultad de Informática

Tecnologías de Bases de Datos

T E S I N A

Que para obtener el titulo de Licenciado en Informática

Presenta: Margot Verenize Rojas Reséndiz

Profesor Titular: I.S.C. Jabel Reséndiz González

Querétaro Qro. Marzo 2003

CARTA DE ACEPTACIÓN

Por este medio, se otorga constancia de aceptación de tesina para obtener el título de Licenciado en Informática, que presenta la pasante **MARGOT VERENIZE ROJAS RESÉNDIZ** con el tema denominado "*Tecnologías de Bases de datos*".

Este trabajo fue desarrollado como una investigación derivada del curso de titulación **"ADMINISTRACIÓN DE BASE DE DATOS",** dando cumplimiento a uno de los requisitos contemplados en el artículo 34 del reglamento de titulación vigente, en lo referente a la opción de titulación por realización y aprobación de cursos de actualización.

Se extiende la presente para los fines legales a que haya lugar y para su inclusión en todos los ejemplares impresos de la tesina, a los doce días del mes de marzo del dos mil tres.

ATENTAMENTE

ING. JABEL RESENDIZ GONZÁLEZ
PROFR. CURSO DE TITULACIÓN

Tecnologías de Bases de Datos

Resumen

El entorno en el que nos encontramos actualmente detonado por la globalización de mercados agresivos y competitivos demanda mucho de cada uno de nosotros, independiente si hablamos de un individuo o una organización se requiere de dar respuestas rápidas y asertivas a los requerimientos de información e incluso anticiparse y prospectar oportunidades de negocio, la razón es muy sencilla: aquellas organizaciones e individuos que no respondan a las tendencias actuales de comercialización y competitividad se vuelven obsoletos y sin remedio serán absorbidos o desaparecidos de la escena comercial. Por tal motivo se invierten en tecnología y soluciones con las cuales se puede mantener en este mundo cambiante, ahora las empresas no dependen tan solo de factores como ubicación, productos, etc. Sino también del conocimiento. Tal conocimiento basado en información comprensible, detallada y relevante es crucial para lograr y sostener ventaja competitiva. Pero las tareas de recolectar, procesar, limpiar y transformar la información necesaria para la toma de decisiones no es una tarea sencilla mas si consideramos que una empresa tiene distintas áreas que a veces se encuentran alejadas de los ejecutivos de negocios., esta tesina propone hacer un estudio de las tecnologías que descubren, procesan, administran y presentan datos convertidos en Información conocida como el principal conocimiento que sostiene al negocio además de soluciones comerciales que proponen por todos los medios minimizar el tiempo para analizar mucha información con mayor velocidad, precisión en tiempo y forma que el usuario demanda para satisfacer las metas del mismo, llamadas Datawarehouse, Data Mining,Olap, Knowledge Discovery in Databases.

Índice

Resumen

Introducción

Introducción

Hoy, la importancia e impacto de las bases de datos es incuestionable a medida que organizaciones gubernamentales, instituciones académicas, y entidades comerciales crean y mantienen importantes bases de datos que contienen toda clase de información desde documentos de texto en lenguaje natural, tablas estadísticas, datos financieros y objetos multimediales hasta datos de naturaleza técnica y científica. Muchas bases de datos están compuestas de *metadatos*, lo cual significa que los registros guardan "datos acerca de los datos" tales como información acerca del tamaño y carácter de otra base de datos en lugar de ser la fuente primaria de contenido tal como nombre y domicilio de una persona. Las tecnologías de bases de datos, incluyendo métodos de arquitectura y acceso, se están desarrollando rápidamente para mantenerse a l día con esta demanda de mecanismos de administración de la información.

Los diseñadores y administradores de bases de datos enfrentan muchos desafíos que reflejan la complejidad del floreciente entorno de la información. Las tecnologías de bases de datos deben manejar masivas cantidades de datos, extraer información útil desde estos repositorios, y tener la habilidad para reflejar las relaciones entre los datos mantenidos en diferentes bases de datos. Además de la arquitectura y sistema deben proveer integridad, recuperación, concurrencia, y seguridad. Para responder a estos desafíos, los tres modelos fundamentales de bases de datos (jerárquico, red y relacional), han servido como la base para desarrollar modelos de datos más potentes y flexibles, tales como los modelos relacional extendido y el relacional de objetos. Un esquema de datos y una arquitectura bien definida aseguran almacenamiento de datos lógico y eficiente lo cual incrementa la capacidad de la base de datos y extiende las capacidades de los lenguajes de consulta y otros métodos de acceso.

Data warehouse

Normalmente la información le llega a cada persona de una manera casi azarosa: cartas, conversaciones, artículos, e-mail o programas de radio o de TV. De forma similar, mucha información le llega así a las empresas: no desde un único canal ni de forma ordenada, sino como porciones desparramadas de información, arribando desde diferentes direcciones y que se almacenan en múltiples lugares. Este sistema sería eficaz si lo que se quiere es sólo guardar la información; pero si lo que se desea es disponer fácilmente de los datos en el momento preciso a fin de tomar las decisiones adecuadas, se hace imprescindible contar con un Data Warehouse.

Como su nombre lo indica, el Data Warehouse actúa como un área de almacenamiento central (warehouse significa almacén) para la información. Pero no sólo es eso. Es también un organizador, un "purificador" y un "visualizador" de la información: un proceso que llena los "baches" encontrados en la mayoría de las bases de datos y provee un acceso sencillo, inteligible, simplificado y organizado a los datos.

El valor real del Data Warehouse es que suministra un depósito único y centralizado, con los datos -provenientes de diferentes departamentos de una misma empresa- depurados, consolidados e integrados, de forma tal que el analista pueda entenderlos y utilizarlos en el contexto de su negocio. Los Data Warehouses pueden variar en tamaño desde pequeñas compañías con docenas de gigabytes de datos hasta multinacionales con terabytes de datos.

Existe una importante sinergia entre Data Warehouse y Data Mining, debido a que éste último resulta mucho más efectivo cuando se corre contra el primero, ya que el Data Warehouse provee acceso a datos que abarcan a todo el ámbito corporativo. Asimismo, es esta correlación de datos diversos -un tipo de información que jamás se pensaría en comparar- lo que generalmente produce los hallazgos más interesantes.

Data Mining

Es un hecho común, hasta ahora, que las empresas generen -como resultado de sus operaciones- enormes volúmenes de datos, pero, a fin de cuentas, producen muy poca información utilizable y concreta. De forma semejante a un minero que busca incansablemente dentro de un gran depósito geológico el escaso metal precioso, el trabajo incesante del Data Mining permite encontrar la minúscula parte de información útil en una montaña de informativa.

Sintéticamente, el Data Mining es el proceso de examinar exhaustiva y minuciosamente inmensas cantidades de datos a fin de identificar, extraer y descubrir nuevos conocimientos, de forma automática y en un período de tiempo relativamente corto. En otras palabras, es el proceso -asistido por computadora- de encontrar información relevante, clave y difícil de obtener (como correlaciones, tendencias, patrones, regularidades o modelos), a menudo oculta y sepultada en grandes volúmenes de datos. Al permitir analizar la información desde diferentes perspectivas y al hacerla comprensible, los analistas a menudo descubren patrones e identifican tendencias que no han visto antes, relaciones que ni siquiera saben que existen e incluso que nunca hubieran pensado que existieran.

Típicamente, el proceso de búsqueda del Data Mining es interactivo (una búsqueda para probar hipótesis), aunque también puede llevarse a cabo automáticamente por el sistema. Una vez terminado el proceso de búsqueda, el sistema de Data Mining representa sus reportes en forma de una gráfica tridimensional (o incluso con un cierto grado de multidimensionalidad) que puede ser rotada, manipulada y visualizada desde cualquier ángulo. Más tarde los analistas deberán interpretar y examinar estos resultados y tomar las acciones necesarias basadas en aquellos descubrimientos, por ejemplo, elaborando un nuevo conjunto de preguntas para reforzar la búsqueda o algún aspecto de los descubrimientos.

A fin de que su aplicación sea útil, las correlaciones encontradas deben ser tan poco obvias que parezcan ilógicas, irracionales, casi sin sentido. Por ejemplo, que "la mayoría de los

que compraron un determinado tipo de tabla de surf posiblemente veraneen este año en Nueva Zelanda", o que "el 76% de las veces que un cliente llevó gaseosa también compró detergente biodegradable", o que "tanto los desodorantes de hombre como los de mujer, se venden mejor juntos que separados, entre las 17:00 y las 19:00 del fin de semana, en las sucursales de la zona sur". Es muy poco probable que a alguna persona de marketing se le hubiera ocurrido comparar datos sobre la venta de estos productos, y éste es sólo un ejemplo de la enorme variedad de relaciones que el Data Mining es capaz de encontrar. Cuando el programa encuentra correlaciones interesantes, los traduce en gráficos simples, permitiéndoles a los gerentes tomar decisiones m ás racionales, y no sólo basadas en la intuición. No obstante, el Data Mining ayuda a confirmar un presentimiento o a desmentir una creencia: en un ejercicio netamente colaborativo, el ser humano sugiere las ideas (hipótesis) y la máquina las confirma o las rechaza según la evidencia aportada por los datos.

El Data Mining se utiliza tanto en los negocios como en la ciencia. Desde la comprensión del comportamiento de los clientes hasta el análisis de las decisiones de expertos, desde la predicción de los posibles cambios en el mercado hasta el descubrimiento de patrones en el cuidado de la salud, desde la detección de fraudes en tarjetas de crédito hasta el descubrimiento de galaxias, desde la mejora de las promociones de ventas hasta la síntesis de drogas, el Data Mining tiene una enorme gama de aplicaciones.

OLAP (On-line Analytical Processing)

Un sistema OLAP se puede entender como la generalización de un generador de informes. Las aplicaciones informáticas clásicas de consulta, orientadas a la toma de decisiones, deben ser programadas. Atendiendo a las necesidades del usuario, se crea una u otra interfaz. Sin embargo, muchos desarrolladores se dieron cuenta de que estas aplicaciones eran susceptibles de ser generalizadas y servir para casi cualquier necesidad, esto es, para casi cualquier base de datos. Los sistemas OLAP evitan la necesidad de desarrollar interfaces de consulta, y ofrecen un entorno único valido para el análisis de cualquier

información histórica, orientado a la toma de decisiones. A cambio, es necesario definir dimensiones, jerarquías y variables, organizando de esta forma los datos.

Para los desarrolladores de aplicaciones acostumbrados a trabajar con bases de datos relacionales, el diseño de una base de datos multidimensional puede ser complejo o al menos, extraño. Pero en general, nuestra experiencia nos dice que el diseño de dimensiones y variables es mucho más sencillo e intuitivo que un diseño relacional. Esto es debido a que las dimensiones y variables son reflejo directo de los informes en papel utilizados por la organización.

Una vez que se ha decidido emplear un entorno de consulta OLAP, se ha de elegir entre R-OLAP y M-OLAP. R-OLAP es la arquitectura de base de datos multidimensional en la que los datos se encuentran almacenados en una base de datos relacional, la cual tiene forma de estrella (también llamada copo de nieve o araña). En R-OLAP, en principio la base de datos sólo almacena información relativa a los datos en detalle, evitando acumulados (evitando redundancia).

En un sistema M-OLAP, en cambio, los datos se encuentran almacenados en archivos con estructura multidimensional, los cuales reservan espacio para todas las combinaciones de todos los posibles valores de todas las dimensiones de cada una de las variables, incluyendo los valores de dimensión que representan acumulados. Es decir, un sistema M-OLAP contiene precalculados (almacenados) los resultados de todas las posibles consultas a la base de datos.

M-OLAP consigue consultas muy rápidas a costa de mayores necesidades de almacenamiento, y retardos en las modificaciones (que no deberían producirse salvo excepcionalmente), y largos procesos *batch* de carga y cálculo de acumulados. En R-OLAP, al contener sólo las combinaciones de valores de dimensión que representan detalle, es decir, al no haber redundancia, el archivo de base de datos es pequeño. Los procesos *batch* de carga son rápidos (ya que no se requiere agregación), y sin embargo, las consultas pueden ser muy lentas, por lo que se aplica la solución de tener al menos algunas consultas precalculadas.

En M-OLAP, el gran tamaño de las variables multidimensionales o el retardo en los procesos *batch* puede ser un inconveniente. En este documento se proponen algunas soluciones a estos problemas, aplicables en tiempo de diseño de la base de datos.

Datamarts

Un Data Marts (Mercado de Datos) es una versión más reducida de un Data Warehouse, a menudo conteniendo información específica de algún departamento, como marketing, finanzas o mantenimiento de la red. Idealmente, el Data Marts debería ser un subconjunto del Data Warehouse, a fin de mantener consistencia en las prácticas de administración de datos corporativos y para mantener la seguridad y la integridad de la información cruda que se está usando. Para las grandes compañías, el Data Marts usualmente contiene una docena de gigabytes de datos.

Capitulo 1. Datawarehouse

El nivel competitivo alcanzado en las empresas les ha exigido desarrollar nuevas estrategias de gestión. En el pasado, las organizaciones fueron típicamente estructuradas en forma piramidal con información generada en su base fluyendo hacia lo alto; y era en el estrato de la pirámide más alto donde se tomaban decisiones a partir de la información proporcionada por la base, con un bajo aprovechamiento del potencial de esta información. Estas empresas, han reestructurado y eliminado estratos de estas pirámides y han autorizado a los usuarios de todos los niveles a tomar mayores decisiones y responsabilidades. Sin embargo, sin información sólida para influenciar y apoyar las decisiones, la autorización no tiene sentido.

Esta necesidad de obtener información para una amplia variedad de individuos es la principal razón de negocios que conduce al concepto de Datawarehouse. El énfasis no está sólo en llevar la información hacia lo alto sino que a través de la organización, para que todos los empleados que la necesiten la tengan a su disposición.

El Datawarehouse (Datawarehousing, Warehouse y Datawarehouse) convierte entonces los datos operacionales de una organización en una herramienta competitiva, por hacerlos disponibles a los empleados que lo necesiten para el análisis y toma de decisiones.

El objetivo del Datawarehouse será el de satisfacer los requerimientos de información interna de la empresa para una mejor gestión. El contenido de los datos, la organización y estructura son dirigidos a satisfacer las necesidades de información de los analistas. El Datawarehouse es el lugar donde la gente puede accesar sus datos.

El Datawarehouse no es un producto que pueda ser comprado en el mercado, sino más bien un concepto que debe ser construido. Datawarehouse es una combinación de conceptos y tecnología que cambian significativamente la manera en que es entregada la información a la gente de negocios. El objetivo principal es satisfacer los requerimientos de información internos de la empresa para una mejor gestión, con eficiencia y facilidad de acceso.

La manera tradicional hasta ahora de entregar la información es a través de emisión de reportes impresos desde los sistemas operacionales, con consultas a nivel de cliente y extracción ocasional de datos para suplir actividades basadas en papel. Los problemas con

la entrega de la información actual son muchos, incluyendo inconsistencia, inflexibilidad y carencia de integración a través de la empresa.

El Datawarehouse puede verse como una bodega donde están almacenados todos los datos necesarios para realizar las funciones de gestión de la empresa, de manera que puedan utilizarse fácilmente según se necesiten. El contenido de los datos, la organización y estructura son dirigidos a satisfacer las necesidades de información de analistas. Los sistemas transaccionales son dinámicos, en el sentido que constantemente se encuentran actualizando datos.

Analizar esta información puede presentar resultados distintos en cuestión de minutos, por lo que se deben extraer y almacenar fotografías de datos (snapshots), para estos efectos, con la implicancia de un consumo adicional de recursos de cómputo. Llevar a cabo un análisis complejo sobre un sistema transaccional, puede resultar en la degradación del sistema, con el consiguiente impacto en la operación del negocio.

El datawarehouse intenta responder a la compleja necesidad de obtención de información útil sin el sacrificio del rendimiento de las aplicaciones operacionales, debido a lo cual se ha convertido actualmente en una de las tendencias tecnológicas más significativas en la administración de información.

Los almacenes de datos (o Datawarehouse) generan bases de datos tangibles con una perspectiva histórica, utilizando datos de múltiples fuentes que se fusionan en forma congruente. Estos datos se mantienen actualizados, pero no cambian al ritmo de los sistemas transaccionales. Muchos datawarehouses se diseñan para contener un nivel de detalle hasta el nivel de transacción, con la intención de hacer disponible todo tipo de datos y características, para reportar y analizar. Así un datawarehouse resulta ser un recipiente de datos transaccionales para proporcionar consultas operativas, y la información para poder llevar a cabo análisis multidimensional. De esta forma, dentro de una almacén de datos existen dos tecnologías complementarias, una relacional para consultas y una multidimensional para análisis.

1.1 Definición

Existen muchas definiciones para el Data Warehouse voy a citar algunas, la más conocida fue propuesta por Inmon[MicroSt96] (considerado el padre de las Bases de Datos) en 1992: "Un Data Warehouse es una colección de datos orientados a temas, integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales". En 1993, Susan Osterfeldt[MicroSt96] publica una definición que sin duda acierta en la clave del Data Warehouse: "Yo considero al Data Warehouse como algo que provee dos beneficios empresariales reales: Integración y Acceso de datos. Data Warehouse elimina una gran cantidad de datos inútiles y no deseados, como también el procesamiento desde el ambiente operacional clásico"

" *Datawarehouse* es una colección de información corporativa derivada directamente de los sistemas operacionales y de algunos orígenes de datos externos"

" Es un proceso, no un producto. Es una técnica para consolidar y administrar datos de variadas fuentes con el propósito de responder preguntas de negocios y tomar decisiones, de una forma que no era posible hasta ahora. Consolidar datos desde una variedad de fuentes. Dentro del marco conceptual de Data Warehousing los agruparemos dentro del proceso de Transformación de Datos. Manejar grandes volúmenes de datos de una forma que no era posible, o no era costo efectiva."

1.2 Sistema de Información

Un **sistema de información** es un conjunto de personas, datos y procedimientos que interactúan entre si, con el propósito de alcanzar un objetivo común. Pertenece a los *Sistemas Administrativos*, es decir, los sistemas de información son subsistemas de los sistemas administrativos.

El alcance de un sistema de información es justamente la organización misma, ya que esta representa su campo de acción. En toda organización se generan flujos de información que se mueven por toda la organización. Estos flujos de datos son agrupados bajo una serie de

esquemas y forman los denominados sistemas de información que se han dividido de acuerdo al siguiente esquema (figura 1):

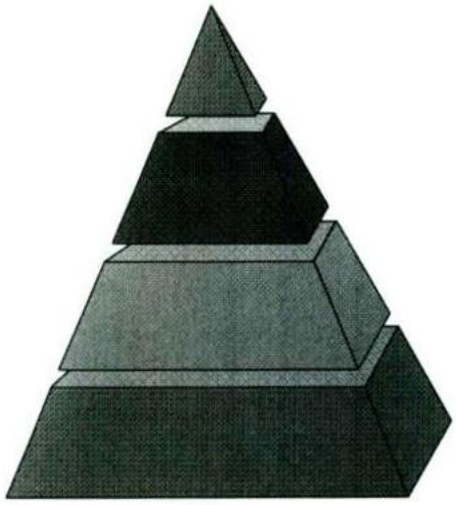


Figura 1 Sistema de Información

- **Sistemas Estratégicos**, orientados a soportar la toma de decisiones, facilitan la labor de la dirección, proporcionándole un soporte básico, en forma de mejor información, para la toma de decisiones. Se caracterizan porque son sistemas sin carga periódica de trabajo, es decir, su utilización no es predecible, al contrario de los casos anteriores, cuya utilización es periódica.

Destacan entre estos sistemas: los Sistemas de Información Gerencial (MIS), Sistemas de Información Ejecutivos (EIS), Sistemas de Información Georeferencial (GIS), Sistemas de Simulación de Negocios (BIS y que en la práctica son sistemas expertos o de Inteligencia Artificial-AI).

- **Sistemas Tácticos**, diseñados para soportar las actividades de coordinación de actividades y manejo de documentación, definidos para facilitar consultas sobre información almacenada en el sistema, proporcionar informes y, en resumen, facilitar la gestión independiente de la información por parte de los niveles intermedios de la organización.

Destacan entre ellos: los Sistemas Ofimáticos (OA), Sistemas de Transmisión de Mensajería (E-mail y Fax Server), coordinación y control de tareas (Work Flow) y tratamiento de documentos (Imagen, Trámite y Bases de Datos Documentarios).

- **Sistemas Técnico-Operativos,** que cubren el núcleo de operaciones tradicionales de captura masiva de datos (Data Entry) y servicios básicos de tratamiento de datos, con tareas predefinidas (contabilidad, facturación, almacén, presupuesto, personal y otros sistemas administrativos). Estos sistemas están evolucionando con la irrupción de censores, autómatas, sistemas multimedia, bases de datos relacionales más avanzadas y data warehousing.

- **Sistemas Interinstitucionales**, este último nivel de sistemas de información recién está surgiendo, es consecuencia del desarrollo organizacional orientado a un mercado de carácter global, el cual obliga a pensar e implementar estructuras de comunicación más estrechas entre la organización y el mercado (Empresa Extendida, Organización Inteligente e Integración Organizacional), todo esto a partir de la generalización de las redes informáticas de alcance nacional y global (INTERNET), que se convierten en vehículo de comunicación entre la organización y el mercado, no importa dónde esté la organización (INTRANET), el mercado de la institución (EXTRANET) y el mercado (Red Global).

Sin embargo, la tecnología data warehousing basa sus conceptos y diferencias entre dos tipos fundamentales de sistemas de información en todas las organizaciones: los sistemas técnico-operacionales y los sistemas de soporte de decisiones. Este último es la base de un data warehouse

1.2.1 Sistemas técnico - operacionales

Como indica su nombre, son los sistemas que ayudan a manejar la empresa con sus operaciones cotidianas. Estos son los sistemas que operan sobre el "backbone" (columna vertebral) de cualquier empresa o institución, entre las que se tiene sistemas de ingreso de órdenes, inventario, fabricación, planilla y contabilidad, entre otros.

Debido a su volumen e importancia en la organización, los sistemas operacionales siempre han sido las primeras partes de la empresa a ser computarizados. A través de los años, estos sistemas operacionales se han extendido, revisado, mejorado y mantenido al punto que hoy, ellos son completamente integrados en la organización.

Desde luego, la mayoría de las organizaciones grandes de todo el mundo, actualmente no podrían operar sin sus sistemas operacionales y los datos que estos sistemas mantienen.

1.2.2 Sistemas de Soporte de Decisiones

Por otra parte, hay otras funciones dentro de la empresa que tienen que ver con el planeamiento, previsión y administración de la organización. Estas funciones son también críticas para la supervivencia de la organización, especialmente en nuestro mundo de rápidos cambios.

Las funciones como "planificación de marketing", "planeamiento de ingeniería" y "análisis financiero", requieren, además, de sistemas de información que los soporte. Pero estas funciones son diferentes de las operacionales y los tipos de sistemas y la información requerida son también diferentes. Las funciones basadas en el conocimiento son los sistemas de soporte de decisiones.

Estos sistemas están relacionados con el análisis de los datos y la toma de decisiones, frecuentemente, decisiones importantes sobre cómo operará la empresa, ahora y en el futuro. Estos sistemas no sólo tienen un enfoque diferente al de los operacionales, sino que, por lo general, tienen un alcance diferente.

Mientras las necesidades de los datos operacionales se enfocan normalmente hacia una sola área, los datos para el soporte de decisiones, con frecuencia, toma un número de áreas diferentes y necesita cantidades grandes de datos operacionales relacionadas.

Son estos sistemas sobre los se basa la tecnología data warehousing.

1.3 Características de un Data Warehouse

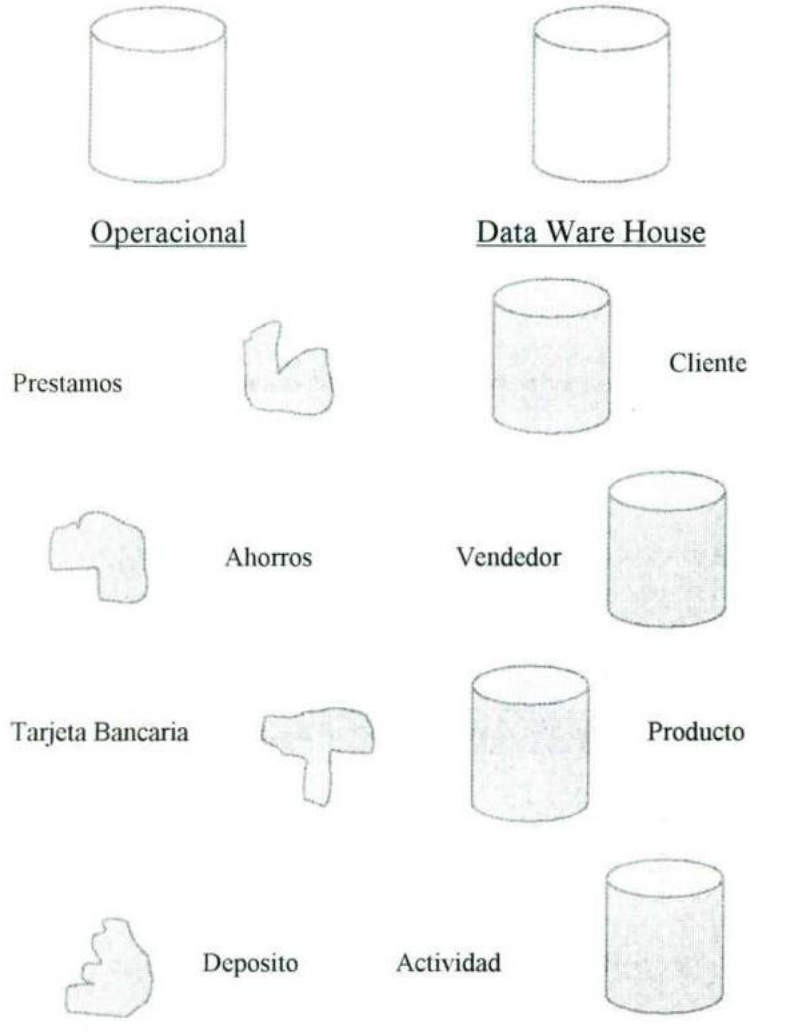
De acuerdo con Bill Inmon, autor de Building the Data Warehouse Construyendo el almacén de datos, ampliamente reconocido como el gurú creador del concepto data warehousing, existen generalmente cuatro características que describen un almacén de datos:

1.3.1. Orientado al sujeto:

Los datos se organizan de acuerdo al sujeto en vez de la aplicación, por ejemplo, una compañía de seguros usando un almacén de datos podría organizar sus datos por cliente, premios, y reclamaciones, en lugar de por diferentes productos (automóviles, vida, etc.). Los datos organizados por sujetos contienen solo la información necesaria para los procesos de soporte para la toma de decisiones. En la Figura N° 1 se muestra el contraste entre los dos tipos de orientaciones.

Orientado al Sujeto

Figura No. 1



El Data Ware House tiene una fuerte orientación al Sujeto

El ambiente operacional se diseña alrededor de las aplicaciones y funciones tales como préstamos, ahorros, tarjeta bancaria y depósitos para una institución financiera. En el ambiente data warehousing se organiza alrededor de sujetos tales como cliente, vendedor, producto y actividad. Por ejemplo, para un fabricante, éstos pueden ser clientes, productos, proveedores y vendedores. La alineación alrededor de las áreas de los temas afecta el diseño y la implementación de los datos encontrados en el data warehouse. Las principales áreas de los temas influyen en la parte más importante de la estructura clave.

Las aplicaciones están relacionadas con el diseño de la base de datos y del proceso. En data warehousing se enfoca el modelamiento de datos y el diseño de la base de datos. El diseño del proceso (en su forma clásica) no es separado de este ambiente.

Las diferencias entre la orientación de procesos y funciones de las aplicaciones y la orientación a temas, radican en el contenido de la data a nivel detallado. En el data warehouse se excluye la información que no será usada por el proceso de sistemas de soporte de decisiones, mientras que la información de las orientadas a las aplicaciones, contiene datos para satisfacer de inmediato los requerimientos funcionales y de proceso, que pueden ser usados o no por el analista de soporte de decisiones.

Otra diferencia importante está en la interrelación de la información. Los datos operacionales mantienen una relación continua entre dos o más tablas basadas en una regla comercial que está vigente. Las del data warehouse miden un espectro de tiempo y las relaciones encontradas en el data warehouse son muchas.

1.3.2 Integrados:

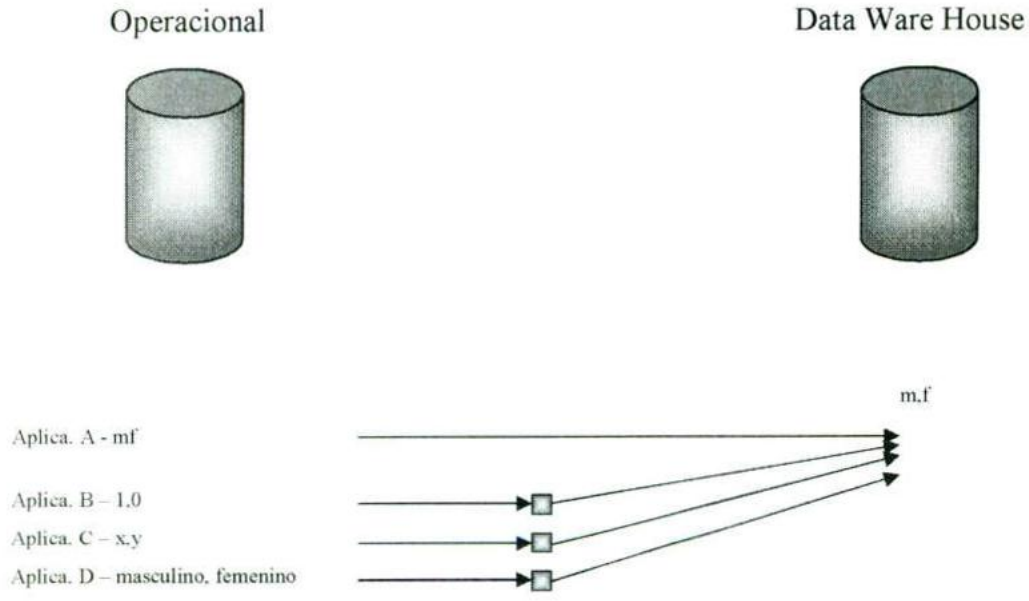
El aspecto más importante del ambiente data warehousing es que la información encontrada al interior está siempre integrada. La integración de datos se muestra de muchas maneras: en convenciones de nombres consistentes, en la medida uniforme de variables, en la codificación de estructuras consistentes, en atributos fisicos de los datos consistentes, fuentes múltiples y otros. El contraste de la integración encontrada en el data warehouse con la carencia de integración del ambiente de aplicaciones, se muestran en la Figura N° 2, con diferencias bien marcadas.

Se diferencian en la codificación, en las estructuras claves, en sus características fisicas, en las convenciones de nombramiento y otros. La Figura N° 2 mencionada, muestra algunas

de las diferencias más importantes en las formas en que se diseñan las aplicaciones.

- ***Codificación***. Los diseñadores de aplicaciones codifican el campo GENERO en varias formas. Un diseñador representa GENERO como una "M" y una "F", otros como un "1" y un "0", otros como una "X" y una "Y" e inclusive, como "masculino" y "femenino".
  No importa mucho cómo el GENERO llega al data warehouse. Probablemente "M" y "F" sean tan buenas como cualquier otra representación. Lo importante es que sea de cualquier fuente de donde venga, el GENERO debe llegar al data warehouse en un estado integrado uniforme. Por lo tanto, cuando el GENERO se carga en el data warehouse desde una aplicación, donde ha sido representado en formato "M" y "F", los datos deben convertirse al formato del data warehouse.

- ***Medida de atributos***. Los diseñadores de aplicaciones miden las unidades de medida de las tuberías en una variedad de formas. Un diseñador almacena los datos de tuberías en centímetros, otros en pulgadas, otros en millones de pies cúbicos por segundo y otros en yardas.



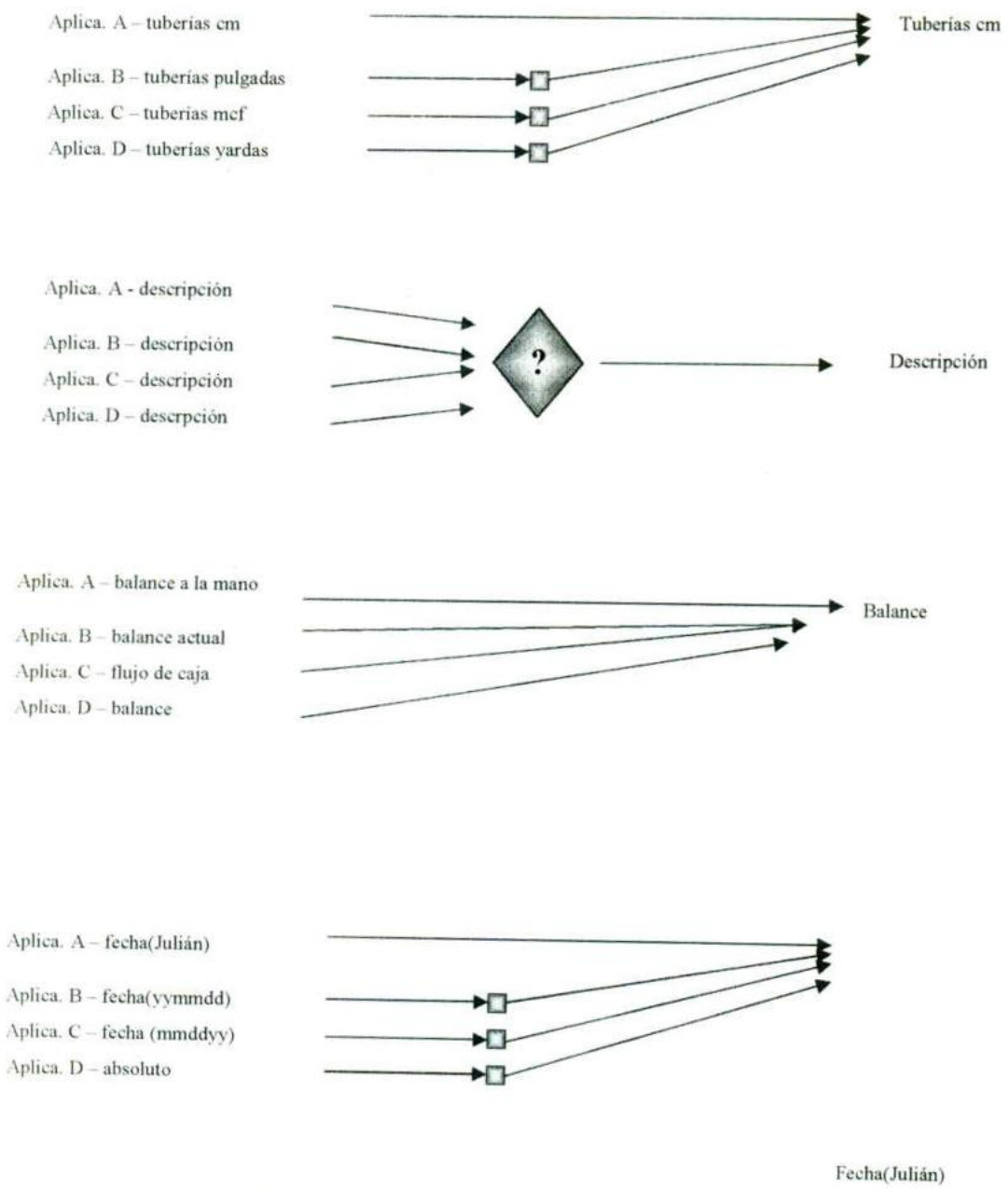


Figura No. 2

Fig. 2. Cuando los datos se mueven al data warehouse desde las aplicaciones orientas al ambiente operacional, los datos se integran antes de entrar al deposito

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

Cualquiera que sea la fuente, cuando la información de la tubería llegue al data warehouse necesitará ser medida de la misma manera.

- ***Convenciones de Nombramiento.-*** El mismo elemento es frecuentemente referido por nombres diferentes en las diversas aplicaciones. El proceso de transformación asegura que se use preferentemente el nombre de usuario.

- ***Fuentes Múltiples.-*** El mismo elemento puede derivarse desde fuentes múltiples. En este caso, el proceso de transformación debe asegurar que la fuente apropiada sea usada, documentada y movida al depósito.

Tal como se muestra en la figura, los puntos de integración afectan casi todos los aspectos de diseño - las características físicas de los datos, la disyuntiva de tener más de una de fuente de datos, el problema de estándares de denominación inconsistentes, formatos de fecha inconsistentes y otros.

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

Cuando el analista de sistema de soporte de decisiones observe el data warehouse, su enfoque deberá estar en el uso de los datos que se encuentre en el depósito, antes que preguntarse sobre la confiabilidad o consistencia de los datos.

1.3.3 De Tiempo Variante

Toda la información del data warehouse es requerida en algún momento. Esta característica básica de los datos en un depósito, es muy diferente de la información encontrada en el ambiente operacional. En éstos, la información se requiere al momento de accesar. En otras palabras, en el ambiente operacional, cuando usted accesa a una unidad de información, usted espera que los valores requeridos se obtengan a partir del momento de acceso.

Como la información en el data warehouse es solicitada en cualquier momento (es decir, no "ahora mismo"), los datos encontrados en el depósito se llaman de "tiempo variante".

Los datos históricos son de poco uso en el procesamiento operacional. La información del depósito por el contraste, debe incluir los datos históricos para usarse en la identificación y evaluación de tendencias. Ver Figura 3:

De tiempo variante

Figura No. 3

Operacional	Data Warehouse
Valor actual de los datos: • Horizonte de tiempo 60-90 días • Clave puede, como no, tener un elemento clave	Datos Instantáneos • Horizonte de tiempo 5-10 años • La clave contiene un elemento de tiempo • Una vez que el snapshot se realice el

El tiempo variante se muestra de varias maneras:

1. La más simple es que la información representa los datos sobre un horizonte largo de tiempo - desde cinco a diez años. El horizonte de tiempo representado para el ambiente operacional es mucho más corto - desde valores actuales hasta sesenta a noventa días.

Las aplicaciones que tienen un buen rendimiento y están disponibles para el procesamiento de transacciones, deben llevar una cantidad mínima de datos si tienen cualquier grado de flexibilidad. Por ello, las aplicaciones operacionales tienen un corto horizonte de tiempo, debido al diseño de aplicaciones rígidas.

2. La segunda manera en la que se muestra el tiempo variante en el data warehouse está en la estructura clave. Cada estructura clave en el data warehouse contiene, implícita o explícitamente, un elemento de tiempo como día, semana, mes, etc.

El elemento de tiempo está casi siempre al pie de la clave concatenada, encontrada en el data warehouse. En ocasiones, el elemento de tiempo existirá implícitamente, como el caso en que un archivo completo se duplica al final del mes, o al cuarto.

3. La tercera manera en que aparece el tiempo variante es cuando la información del data warehouse, una vez registrada correctamente, no puede ser actualizada. La información del data warehouse es, para todos los propósitos prácticos, una serie larga de "snapshots" (vistas instantáneas).

Por supuesto, si los snapshots de los datos se han tomado incorrectamente, entonces pueden ser cambiados. Asumiendo que los snapshots se han tomado adecuadamente, ellos no son alterados una vez hechos. En algunos casos puede ser no ético, e incluso ilegal, alterar los snapshots en el data warehouse. Los datos operacionales, siendo requeridos a partir del momento de acceso, pueden actualizarse de acuerdo a la necesidad.

1.3.4 No Volátil

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

En la Figura N° 4 se muestra que la actualización (insertar, borrar y modificar), se hace regularmente en el ambiente operacional sobre una base de registro por registro. Pero la manipulación básica de los datos que ocurre en el data warehouse es mucho más simple. Hay dos únicos tipos de operaciones: la carga inicial de datos y el acceso a los mismos. No hay actualización de datos (en el sentido general de actualización) en el depósito, como una parte normal de procesamiento.

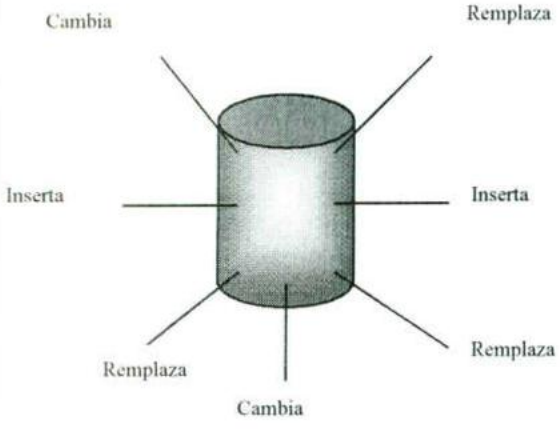
Hay algunas consecuencias muy importantes de esta diferencia básica, entre el procesamiento operacional y del data warehouse. En el nivel de diseño, la necesidad de ser precavido para actualizar las anomalías no es un factor en el data warehouse, ya que no se

hace la actualización de datos. Esto significa que en el nivel físico de diseño, se pueden tomar libertades para optimizar el acceso a los datos, particularmente al usar la normalización y des normalización física.

Otra consecuencia de la simplicidad de la operación del data warehouse está en la tecnología subyacente, utilizada para correr los datos en el depósito. Teniendo que soportar la actualización de registro por registro en modo on - line (como es frecuente en el caso del procesamiento operacional) requiere que la tecnología tenga un fundamento muy complejo debajo de una fachada de simplicidad.

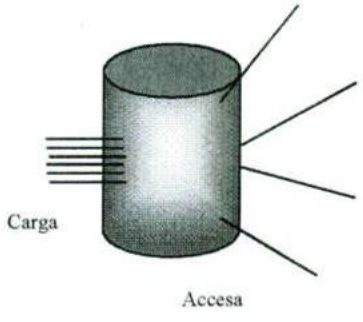
No Volátil

Figura No. 4



Operacional

Normalmente el Data se actualiza registro por



Data warehouse

El Data es cargada en el deposito de datos y ahí es accesada pero una vez que el snapshot esta hecho los datos en el deposito no cambian.

La tecnología permite realizar backup y recuperación, transacciones e integridad de los datos y la detección y solución al estancamiento que es más complejo. En el data warehouse no es necesario el procesamiento.

La fuente de casi toda la información del data warehouse es el ambiente operacional. A simple vista, se puede pensar que hay redundancia masiva de datos entre los dos ambientes. Desde luego, la primera impresión de muchas personas se centra en la gran redundancia de datos, entre el ambiente operacional y el ambiente de data warehouse. Dicho razonamiento es superficial y demuestra una carencia de entendimiento con respecto a qué ocurre en el data warehouse. De hecho, hay una mínima redundancia de datos entre ambos ambientes.
Se debe considerar lo siguiente:

- Los datos se filtran cuando pasan desde el ambiente operacional al de depósito. Existe mucha data que nunca sale del ambiente operacional. Sólo los datos que realmente se necesitan ingresarán al ambiente de data warehouse.
- El horizonte de tiempo de los datos es muy diferente de un ambiente al otro. La información en el ambiente operacional es más reciente con respecto a la del data warehouse. Desde la perspectiva de los horizontes de tiempo únicos, hay poca superposición entre los ambientes operacional y de data warehouse.
- El data warehouse contiene un resumen de la información que no se encuentra en el ambiente operacional.
- Los datos experimentan una transformación fundamental cuando pasa al data warehouse. La mayor parte de los datos se alteran significativamente al ser seleccionados y movidos al data warehouse. Dicho de otra manera, la mayoría de los datos se alteran física y radicalmente cuando se mueven al depósito. No es la misma data que reside en el ambiente operacional desde el punto de vista de integración.

En vista de estos factores, la redundancia de datos entre los dos ambientes es una ocurrencia rara, que resulta en menos de 1%.

1.4 Estructura Del Data Warehouse

Los data warehouses tienen una estructura distinta. Hay niveles diferentes de esquematización y detalle que delimitan el data warehouse. La estructura de un data warehouse se muestra en la Figura N° 5.

En la figura, se muestran los diferentes componentes del data warehouse y son:

- ***Detalle de datos actuales.-*** En gran parte, el interés más importante radica en el detalle de los datos actuales, debido a que:
- Refleja las ocurrencias más recientes, las cuales son de gran interés .
- Es voluminoso, ya que se almacena al más bajo nivel de granularidad.
- Casi siempre se almacena en disco, el cual es de fácil acceso, aunque su administración sea costosa y compleja.

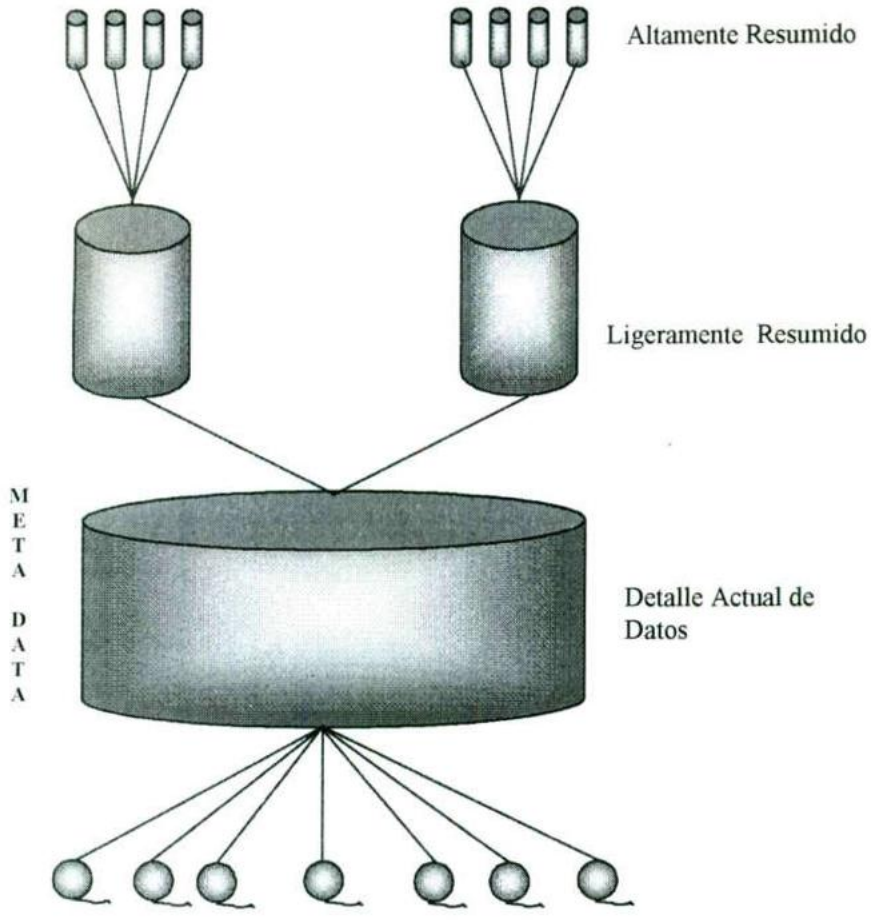
- ***Detalle de datos antiguos.-*** La data antigua es aquella que se almacena sobre alguna forma de almacenamiento masivo. No es frecuentemente accesada y se almacena a un nivel de detalle, consistente con los datos detallados actuales. Mientras no sea prioritario el almacenamiento en un medio de almacenaje alterno, a causa del gran volumen de datos unido al acceso no frecuente de los mismos, es poco usual utilizar el disco como medio de almacenamiento.

- ***Datos ligeramente resumidos.-*** La data ligeramente resumida es aquella que proviene desde un bajo nivel de detalle encontrado al nivel de detalle actual. Este nivel del data warehouse casi siempre se almacena en disco. Los puntos en los que se basa el diseñador para construirlo son:
- Que la unidad de tiempo se encuentre sobre la esquematización hecha.
- Qué contenidos (atributos) tendrá la data ligeramente resumida.

- ***Datos completamente resumidos.-*** El siguiente nivel de datos encontrado en el data warehouse es el de los datos completamente resumidos. Estos datos son compactos y fácilmente accesibles.

Estructura de los datos en un Data Warehouse

Figura No. 5



- ***Metadata.-*** El componente final del data warehouse es el de la metadata. De muchas maneras la metadata se sitúa en una dimensión diferente al de otros datos del data warehouse, debido a que su contenido no es tomado directamente desde el ambiente operacional.

La metadata juega un rol especial y muy importante en el data warehouse y es usada como:

- Un directorio para ayudar al analista a ubicar los contenidos del data warehouse.
- Una guía para el mapping de datos de cómo se transforma, del ambiente operacional al de data warehouse.
- Una guía de los algoritmos usados para la esquematización entre el detalle de datos actual, con los datos ligeramente resumidos y éstos, con los datos completamente resumidos, etc.

La metadata juega un papel mucho más importante en un ambiente data warehousing que en un operacional clásico.

A fin de recordar los diferentes niveles de los datos encontrados en el data warehouse, considere el ejemplo mostrado en la Figura N° 6.

El detalle de ventas antiguas son las que se encuentran antes de 1992. Todos los detalles de ventas desde 1982 (o cuando el diseñador inició la colección de los archivos) son almacenados en el nivel de detalle de datos más antiguo.

El detalle actual contiene información desde 1992 a 1993 (suponiendo que 1993 es el año actual). En general, el detalle de ventas no se ubica en el nivel de detalle actual hasta que haya pasado, por lo menos, veinticuatro horas desde que la información de ventas llegue a estar disponible en el ambiente operacional.

1.5 Arquitectura Datawarehouse

El término Datawarehouse se utiliza indistintamente para hablar de la arquitectura en sí como también para uno de los componentes que la conforman, específicamente el que tiene relación con el almacenamiento físico de los datos.

La estructura básica de la arquitectura de un Datawarehouse incluye:

- Base de datos operacional / Nivel de base de datos externo
- Nivel de acceso a la información
- Nivel de acceso a los datos
- Nivel de directorio de datos (Metadata)
- Nivel de gestión de proceso
- Nivel de mensaje de la aplicación
- Nivel de data warehouse

- Nivel de organización de datos

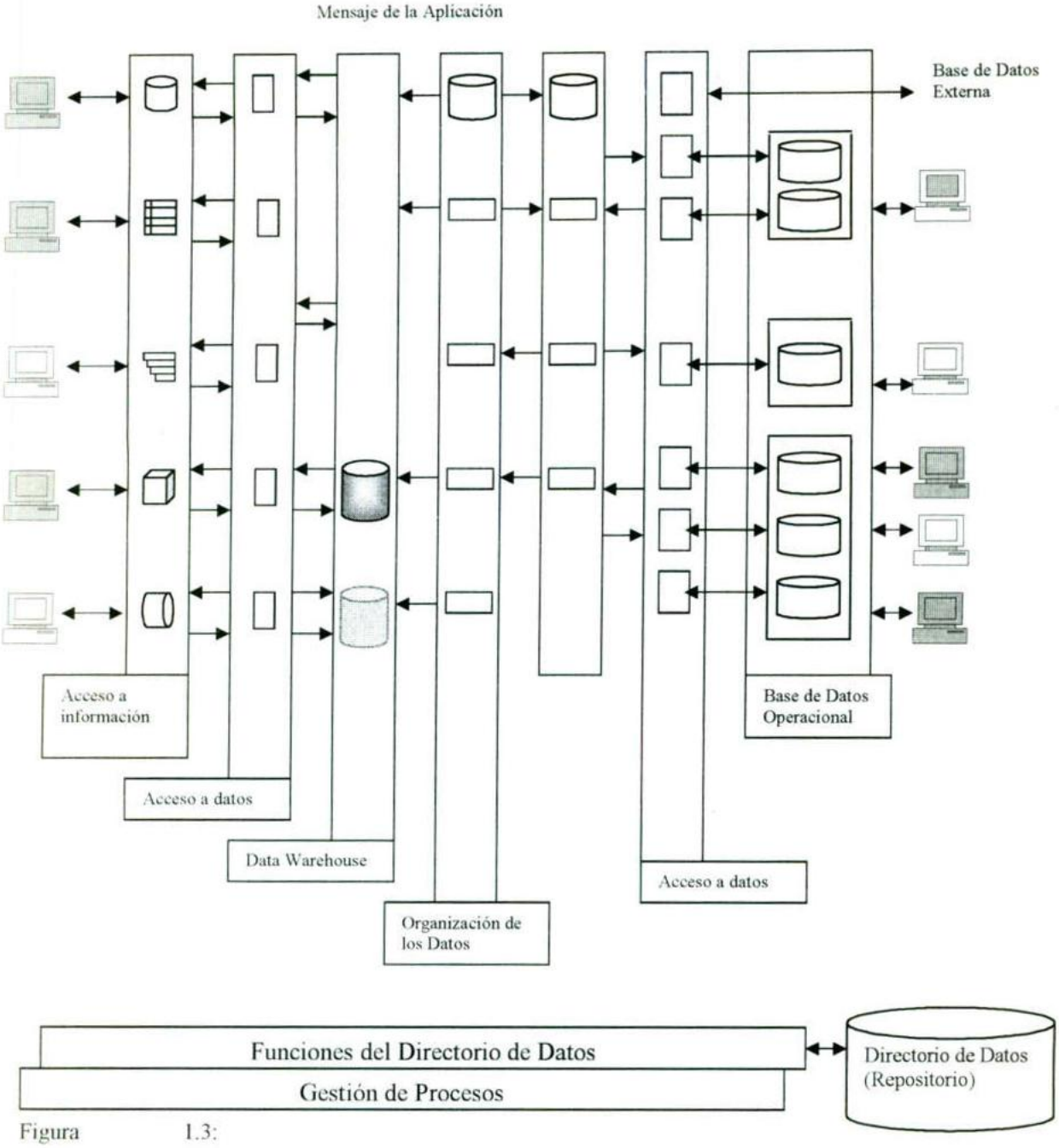


Figura 1.3:

Estructura básica de la arquitectura de un Datawarehouse

- **Base de datos operacional / Nivel de base de datos externo**

Los sistemas operacionales procesan datos para apoyar las necesidades operacionales críticas. Para hacer eso, se han creado las bases de datos operacionales históricas que proveen una estructura de procesamiento eficiente, para un número relativamente pequeño de transacciones comerciales bien definidas.

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

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

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

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

- **Nivel de acceso a la información**

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

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

Actualmente, existen herramientas más y más sofisticadas para manipular, analizar y presentar los datos, sin embargo, hay problemas significativos al tratar de convertir los datos tal como han sido recolectados y que se encuentran contenidos en los sistemas operacionales en información fácil y transparente para las herramientas de los usuarios finales. Una de las claves para esto es encontrar un lenguaje de datos común que puede usarse a través de toda la empresa.

- **Nivel de acceso a los datos**

El nivel de acceso a los datos de la arquitectura data warehouse está involucrado con el nivel de acceso a la información para conversar en el nivel operacional. En la red mundial de hoy, el lenguaje de datos común que ha surgido es SQL. Originalmente, SQL fue desarrollado por IBM como un lenguaje de consulta, pero en los últimos veinte años ha llegado a ser el estándar para el intercambio de datos.

Uno de los adelantos claves de los últimos años ha sido el desarrollo de una serie de "filtros" de acceso a datos, tales como EDA/SQL para accesar a casi todo los Sistemas de Gestión de Base de Datos (Data Base Management Systems - DBMSs) y sistemas de archivos de datos, relacionales o no. Estos filtros permiten a las herramientas de acceso a la información, accesar también a la data almacenada en sistemas de gestión de base de datos que tienen veinte años de antigüedad.

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

El acceso a los datos universales significa que, teóricamente por lo menos, los usuarios finales sin tener en cuenta la herramienta de acceso a la información o ubicación, deberían ser capaces de accesar a cualquier o todos los datos en la empresa que es necesaria para ellos, para hacer su trabajo.

El nivel de acceso a los datos entonces es responsable de la interfase entre las herramientas de acceso a la información y las bases de datos operacionales. En algunos casos, esto es

todo lo que un usuario final necesita. Sin embargo, en general, las organizaciones desarrollan un plan mucho más sofisticado para el soporte del data warehousing.

- **Nivel de Directorio de Datos (Metadata)**

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

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

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

- **Nivel de Gestión de Procesos**

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

- **Nivel de Mensaje de la Aplicación**

El nivel de mensaje de la aplicación tiene que ver con el transporte de información alrededor de la red de la empresa. El mensaje de aplicación se refiere también como "subproducto", pero puede involucrar sólo protocolos de red. Puede usarse por ejemplo, para aislar aplicaciones operacionales o estratégicas a partir del formato de datos exacto, recolectar transacciones o los mensajes y entregarlos a una ubicación segura en un tiempo seguro.

- **Nivel Data Warehouse (Físico)**

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

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

- **Nivel de Organización de Datos**

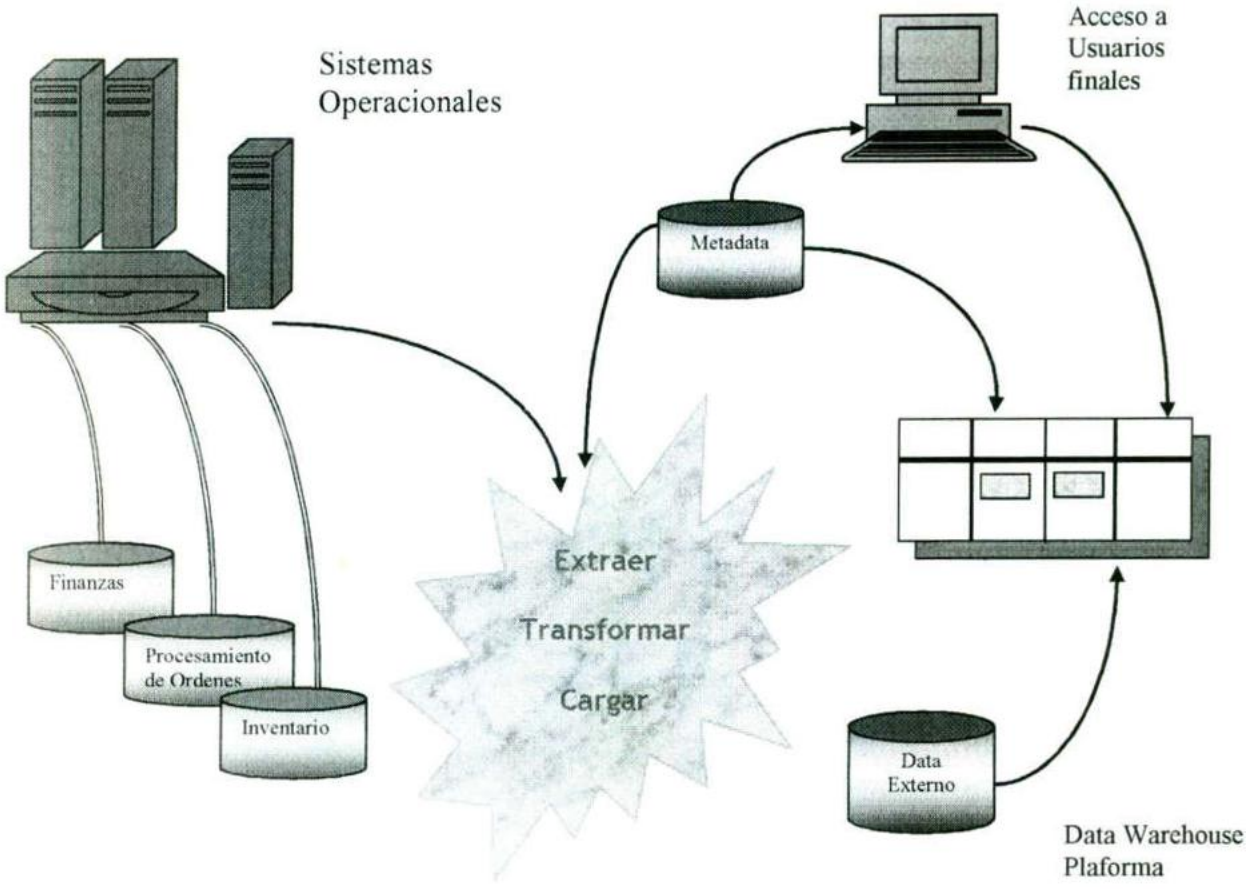
El componente final de la arquitectura data warehouse es la organización de los datos. Se llama también gestión de copia o réplica, pero de hecho, incluye todos los procesos necesarios como seleccionar, editar, resumir, combinar y cargar datos en el depósito y accesar a la información desde bases de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

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

1.5.1 Operaciones en un Data Warehouse

En la Figura N° 7 se muestra algunos de los tipos de operaciones que se efectúan dentro de un ambiente data warehousing.

Figura No. 7 Operaciones en un Data Warehouse



a) Sistemas Operacionales

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

Las bases de datos operacionales se organizan como archivos indexados (UFAS, VSAM), bases de datos de redes/jerárquicas (I-D-S/II, IMS, IDMS) o sistemas de base de datos relacionales (DB2, Oracle, Informix, etc.). Según las encuestas, aproximadamente del 70% a 80% de las bases de datos de las empresas se organizan usando DBMSs no relacional.

b) Extracción, Transformación y Carga de los Datos

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

Tomar los datos desde varias bases de datos operacionales y transformarlos en datos requeridos para el depósito, se refiere a la transformación o a la integración de datos. Las bases de datos operacionales, diseñadas para el soporte de varias aplicaciones de producción, frecuentemente difieren en el formato.

Los mismos elementos de datos, si son usados por aplicaciones diferentes o administrados por diferentes software DBMS, pueden definirse al usar nombres de elementos inconsistentes, que tienen formatos inconsistentes y/o ser codificados de manera diferente. Todas estas inconsistencias deben resolverse antes que los elementos de datos sean almacenados en el data warehouse.

c) Metadata

Otro paso necesario es crear la metadata. La metadata (es decir, datos acerca de datos) describe los contenidos del data warehouse. La metadata consiste de definiciones de los elementos de datos en el depósito, sistema(s) del (os) elemento(s) fuente. Como la data, se integra y transforma antes de ser almacenada en información similar.

d) Acceso de usuario final

Los usuarios accesan al data warehouse por medio de herramientas de productividad basadas en GUI (Graphical User Interface - Interfase gráfica de usuario). Pueden proveerse a los usuarios del data warehouse muchos de estos tipos de herramientas.

Estos pueden incluir software de consultas, generadores de reportes, procesamiento analítico en línea, herramientas data/visual mining, etc., dependiendo de los tipos de usuarios y sus requerimientos particulares. Sin embargo, una sola herramienta no satisface todos los requerimientos, por lo que es necesaria la integración de una serie de herramientas.

e) Plataforma del data warehouse

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

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

Muchas de las organizaciones quieran o no escogen una plataforma por diversas razones: el Sistema X es nuestro sistema elegido o el Sistema Y está ya disponible sobre un sistema UNIX que nosotros ya tenemos. Uno de los errores más grandes que las organizaciones cometen al seleccionar la plataforma, es que ellos presumen que el sistema (hardware y/o DBMS) escalará con los datos.

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

f) Datos Externos

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

1. **Evolución del Depósito**

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

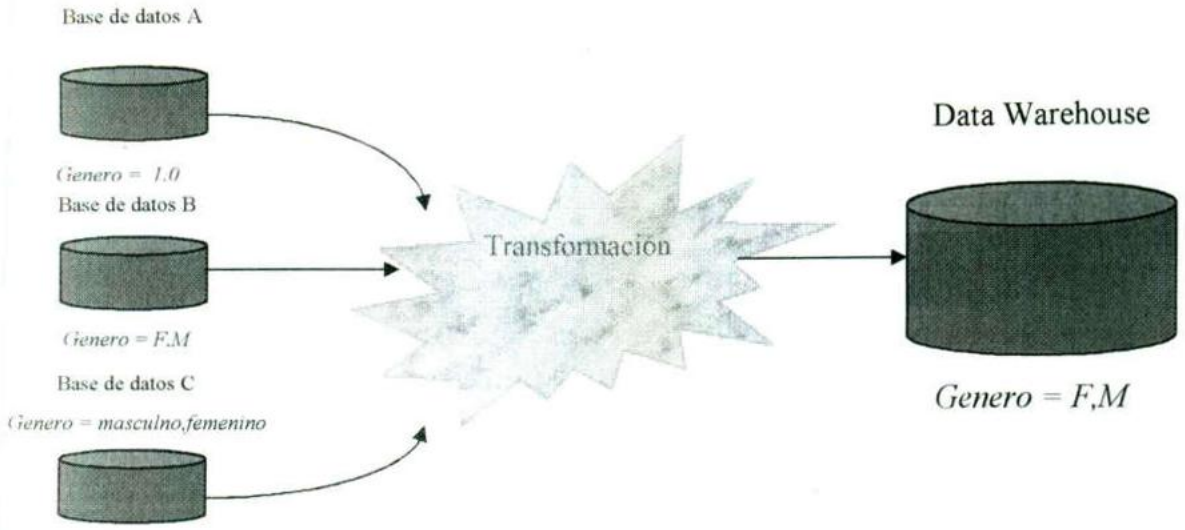


Figura 8: forma de inconsistencia

La transformación de datos también se encarga de las inconsistencias en el contenido de datos. Una vez que se toma la decisión sobre que reglas de transformación serán establecidas, deben crearse e incluirse las definiciones en las rutinas de transformación.

Se requiere una planificación cuidadosa y detallada para transformar datos inconsistentes en conjuntos de datos conciliables y consistentes para cargarlos en el data warehouse.

1.6.2 Metadata

Otro aspecto de la arquitectura de data warehouse es crear soporte a la metadata. Metadata es la información sobre los datos que se alimenta, se transforma y existe en el data warehouse. Metadata es un concepto genérico, pero cada implementación de la metadata usa técnicas y métodos específicos.

Estos métodos y técnicas son dependientes de los requerimientos de cada organización, de las capacidades existentes y de los requerimientos de interfase de usuario. Hasta ahora, no hay normas para la metadata, por lo que la metadata debe definirse desde el punto de vista del software data warehousing, seleccionado para una implementación específica.

Típicamente, la metadata incluye los siguientes ítems:

- Las estructuras de datos que dan una visión de los datos al administrador de datos.

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

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

Uno de los desafíos de mantener un data warehouse, es idear métodos para identificar datos nuevos o modificados en las bases de datos operacionales. Algunas maneras para identificar estos datos incluyen insertar fecha/tiempo en los registros de base de datos y entonces crear copias de registros actualizados y copiar información de los registros de transacción y/o base de datos diarias.
Estos elementos de datos nuevos y/o modificados son extraídos, integrados, transformados y agregados al data warehouse en pasos periódicos programados. Como se añaden las nuevas ocurrencias de datos, los datos antiguos son eliminados. Por ejemplo, si los detalles de un sujeto particular se mantienen por 5 años, como se agregó la última semana, la semana anterior es eliminada.

1.6 TRANSFORMACION DE DATOS Y METADATA

1.6.1 Transformación de Datos

Uno de los desafíos de cualquier implementación de data warehouse, es el problema de transformar los datos. La transformación se encarga de las inconsistencias en los formatos de datos y la codificación, que pueden existir dentro de una base de datos única y que casi siempre existen cuando múltiples bases de datos contribuyen al data warehouse.
En la Figura N° 8 se ilustra una forma de inconsistencia, en la cual el género se codifica de manera diferente en tres bases de datos diferentes. Los procesos de transformación de datos se desarrollan para direccionar estas inconsistencias.

- Las definiciones del sistema de registro desde el cual se construye el data warehouse.
- Las especificaciones de transformaciones de datos que ocurren tal como la fuente de datos se replica al data warehouse.
- El modelo de datos del data warehouse (es decir, los elementos de datos y sus relaciones).
- Un registro de cuando los nuevos elementos de datos se agregan al data warehouse y cuando los elementos de datos antiguos se eliminan o se resumen.
- Los niveles de sumarización, el método de sumarización y las tablas de registros de su data warehouse.

Algunas implementaciones de la metadata también incluyen definiciones de la(s) vista(s) presentada(s) a los usuarios del data warehouse. Típicamente, se definen vistas múltiples para favorecer las preferencias variadas de diversos grupos de usuarios. En otras implementaciones, estas descripciones se almacenan en un Catálogo de Información.

Los esquemas y subesquemas para bases de datos operacionales, forman una fuente óptima de entrada cuando se crea la metadata. Hacer uso de la documentación existente, especialmente cuando está disponible en forma electrónica, puede acelerar el proceso de definición de la metadata del ambiente data warehousing.

La metadata sirve, en un sentido, como el corazón del ambiente data warehousing. Crear definiciones de metadata completa y efectiva puede ser un proceso que consuma tiempo, pero lo mejor de las definiciones y si usted usa herramientas de gestión de software integrado, son los esfuerzos que darán como resultado el mantenimiento del data warehouse.

1.7 Data warehouse frente a sistemas operacionales.

Hasta ahora, el propósito principal de los primeros sistemas de base de datos era cumplir las necesidades de los sistemas operacionales, que son, característicamente de naturaleza transaccional. Ejemplos clásicos de sistemas operacionales podrían ser:

- Libro general de cuantas
- Pagos de cuentas
- Gestión financiera
- Procesamiento de órdenes
- Entrada de órdenes
- Inventario

Los sistemas operacionales, por naturaleza, están principalmente ligados al manejo de una única transacción. El típico sistema operacional trata con una orden, un pedido, un elemento del inventario, etc. Un sistema operacional, por lo general, trata con eventos predefinidos y, debido a la naturaleza de tales eventos, necesita un acceso muy rápido. Cada transacción, normalmente, trabaja con pequeñas cantidades de datos.

La mayor parte del tiempo, el negocio necesita que el sistema operacional no cambie en exceso. La aplicación que graba la información, así como la que controla su acceso a ella, o sea, los informes de un negocio bancario, no cambia mucho con el tiempo. En este tipo de sistemas, la información necesaria cuando un cliente inicia una transacción debe ser corriente. Antes de que un banco conceda una retirada, debe estar seguro del balance actual del cliente. Por otro lado, una aplicación de data warehouse tiene una naturaleza distinta. Una transacción típica maneja grandes cantidades de datos. Una aplicación de data warehouse responde preguntas como:

- ¿Cuál es el depósito medio por sucursal?
- ¿Qué día de la semana es el más ocupado?
- ¿Que clientes con un promedio alto en balances no están participando en la actualidad de una cuanta checking – plus?

Dado que se manejan preguntas del tipo <<¿que pasaría si?>>, cada petición es única. El interfaz que soporta al usuario debe tener un diseño flexible. Hay muchas aplicaciones diferentes que acceden a la misma información, cada una de ellas de forma particular.

Dos aproximaciones al proceso de construcción:

- Una aproximación consiste en gastar un tiempo extra en construir primero una núcleo de data ware house, para luego, usarlo como base para realizar rápidamente muchos datamarts. La construcción inicial de esta aproximación dura más, ya que se ha empleado tiempo en analizar las necesidades de datos del warehouse completo, identificando los elementos de información que se usarán en una gran cantidad de mercados. La ventaja consiste en que, una vez construido el datamart, ya tenemos los planos del warehouse.

- La otra aproximación consiste en construir primero un data mart especifico para trabajo en grupo. Esta aproximación pone rápidamente los datos en manos de los usuarios, pero el trabajo para poner la información en el datamart puede no ser rentalizable cuando se mueven dichos datos a una warehouse o cuando se intenta usar datos similares en un datamart diferente. Se gana velocidad, pero no portabilidad.

Los datos que se requieren para soportar el procesamiento analítico. De hecho, los datos se almacenan fisicamente de manera distinta. Un sistema operacional se optimiza para actualizaciones transaccionales, mientras que un data warehouse se optimiza para grandes queries con amplios conjuntos de datos. Estas diferencias se hacen aparentes cuando se realiza un seguimiento del uso de la unidad central de proceso (CPU) en una computadora con datawarehouse y otra con sistema operacional.

1.8 Costos v/s Valor De Datawarehouse

En todo proyecto es importante e inevitable realizar un análisis desde la perspectiva Costo/Valor.

A grandes rasgos, los costos asociados a un proyecto Datawarehouse incluyen el costo de construcción y, la manutención y operación una vez que está construido. En cuanto al valor,

éste considera, el valor de mejorar la entrega de información, el valor de mejorar el proceso de toma de decisiones y el valor agregado para los procesos empresariales.

Costos De Un Datawarehouse

Costos de Construcción:

Los costos de construir un Datawarehouse son similares para cualquier proyecto de tecnología de información. Estos pueden ser clasificados en tres grandes categorías:

- *Recursos Humanos*: la gente necesita contar con un enfoque fuerte sobre el conocimiento del área de la empresa y de los procesos empresariales. Además es muy importante considerar las cualidades de la gente, ya que el desarrollo del Datawarehouse requiere participación de la gente de negocios como de los especialistas tecnológicos; estos dos grupos de gente deben trabajar juntos, compartiendo su conocimiento y destrezas en un espíritu de equipo de trabajo, para enfrentar los desafíos de desarrollo del Datawarehouse.

- *Tiempo:* Se debe establecer el tiempo no tan solo para la construcción y entrega de resultados del Datawarehouse, sino también para la planeación del proyecto y la definición de la arquitectura. La planeación y la arquitectura, establecen un marco de referencia y un conjunto de estándares que son críticos para la eficacia del Datawarehouse.

- *Tecnología:* Muchas tecnologías nuevas son introducidas por el Datawarehouse. El costo de la nueva tecnología puede ser tan sólo la inversión inicial del proyecto.

Costos De Operación

Una vez que está construido y entregado un Datawarehouse debe ser soportado para que tenga valor empresarial. Son justamente estas actividades de soporte, la fuente de continuos costos operacionales para un Datawarehouse. Se pueden distinguir tres tipos de costos de operación:

- *Evolutivos:* ajustes continuos del Datawarehouse a través del tiempo, como cambios de expectativas y, cambios producto del aprendizaje del Recurso Humano del proyecto mediante su experiencia usando el Datawarehouse.

- *Crecimiento:* Incrementos en el tiempo en volúmenes de datos, del número de usuarios del Datawarehouse, lo cual conllevará a un incremento de los recursos necesarios como a la demanda de monitoreo, administración y sintonización del Datawarehouse (evitando así, un incremento en los tiempos de respuesta y de recuperación de datos, principalmente).

- *Cambios*: El Datawarehouse requiere soportar cambios que ocurren tanto en el origen de datos que éste usa, como en las necesidades de la información que éste soporta.

Los dos primeros tipos de costos de operación, son básicos en la mantención de cualquier sistema de información, por lo cual no nos resultan ajenos; sin embargo, se debe tener especial cuidado con los costos de operación por cambios, ya que ellos consideran el impacto producto de la relación del OLTP y del Ambiente Empresarial, con el Datawarehouse.

Valor Del Datawarehouse

El valor de un Datawarehouse queda descrito en tres dimensiones:

1. Mejorar la Entrega de Información: información completa, correcta, consistente, oportuna y accesible. Información que la gente necesita, en el tiempo que la necesita y en el formato que la necesita.

2. Mejorar el Proceso de Toma de Decisiones: con una mayor soporte de información se obtienen decisiones más rápidas; así también, la gente de negocios adquire mayor confianza en sus propias decisiones y las del resto, y logra un mayor entendimiento de los impactos de sus decisiones.

3. Impacto Positivo sobre los Procesos Empresariales: cuando a la gente se le da acceso a una mejor calidad de información, la empresa puede lograr por sí sola:

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

Balance de Costos v/s Valor.

Lograr una cuantificación económica de los factores de valor no es fácil ni natural a diferencia de los factores de costos, agregar valor económico a los factores de valor resulta ser en extremo complejo y subjetivo. Una alternativa a ello, es hacer una valoración desde la perspectiva de costos evitables, relacionados con los "costos de no disponer en la organización de información apropiada", tanto a un nivel técnico como de procesos empresariales (en especial, para el proceso de Toma de Decisiones).

Datawarehouse es una estrategia de largo plazo. Al querer implementar un Datawarehouse, se debe evaluar el costo y el valor considerando un período de tiempo razonable para obtener beneficios. El retorno sobre la inversión de un Datawarehouse, se comienza a percibir bastante más tarde del tiempo en el cual se realizó la inversión inicial. Si se calcula costo/valor desde una perspectiva de corto plazo, los costos serán significativamente más altos en proporción al valor.

Cambios y el Datawarehouse.

Cuando se implementa un Datawarehouse, el impacto de cambios es compuesto. Dos orígenes primarios de cambios existen:

- Cambios en el ambiente empresarial: Un cambio en el ambiente empresarial puede cambiar las necesidades de información de los usuarios. Así, el contenido del Datawarehouse se puede ver afectado y las aplicaciones DSS y EIS pueden requerir cambios.

- Cambios en la tecnología: Un cambio en la tecnología puede afectar la manera que los datos operacionales son almacenados, lo cual implicaría un ajuste en los procesos de Extracción, Transporte y Carga para adaptar las variaciones presentadas.

Un cambio de cualquiera de ellos impacta los sistemas operacionales. Un cambio en el ambiente operacional puede cambiar el formato, estructura o significado de los datos operacionales usados como origen para el Datawarehouse. De esta forma serían impactados los procesos de Extracción, Transformación y Carga de datos.

1.9 Impactos Datawarehouse

El éxito de Datawarehouse no está en su construcción, sino en usarlo para mejorar procesos empresariales, operaciones y decisiones. Posicionar un Datawarehouse para que sea usado efectivamente, requiere entender los impactos de implementación en los siguientes ámbitos:

Impactos Humanos.

Efectos sobre la gente de la empresa:

- Construcción del Datawarehouse: Construir un Datawarehouse requiere la participación activa de quienes usarán el Datawarehouse. A diferencia del desarrollo de aplicaciones, donde los requerimientos de la empresa logran ser relativamente bien definidos producto de la estabilidad de las reglas de negocio a través del tiempo, construir un Datawarehouse depende de la realidad de la empresa como de las condiciones que en ese momento existan, las cuales determinan qué debe contener el Datawarehouse. La gente de negocios debe participar activamente durante el desarrollo del Datawarehouse, desde una perspectiva de construcción y creación.
- Accesando el Datawarehouse: El Datawarehouse intenta proveer los datos que posibilitan a los usuarios accesar su propia información cuando ellos la necesitan. Esta aproximación para entregar información tiene varias implicancias :
    a) La gente de la empresa puede necesitar aprender nuevas destrezas.
    b) Análisis extensos y demoras de programación para obtener información será

eliminada. Como la información estará lista para ser accesada, las expectativas probablemente aumentarán

c) Nuevas oportunidades pueden existir en la comunidad empresarial para los especialista de información.

d) La gran cantidad de reportes en papel serán reducidas o eliminadas.

e) La madurez del Datawarehouse dependerá del uso activo y retroalimentación de sus usuarios.

- Usando aplicaciones DSS/EIS: usuarios de aplicaciones DSS y EIS necesitarán menos experiencia para construir su propia información y desarrollar nuevas destrezas.

Impactos Empresariales.

- Procesos Empresariales Y Decisiones Empresariales. Se deben considerar los beneficios empresariales potenciales de los siguientes impactos:

  a) Los Procesos de Toma de Decisiones pueden ser mejorados mediante la disponibilidad de información. Decisiones empresariales se hacen más rápidas por gente más informada.

  b) Los procesos empresariales pueden ser optimizados. El tiempo perdido esperando por información que finalmente es incorrecta o no encontrada, es eliminada.

  c) Conexiones y dependencias entre procesos empresariales se vuelven más claros y entendibles. Secuencias de procesos empresariales pueden ser optimizados para ganar eficiencia y reducir costos.

  d) Procesos y datos de los sistemas operacionales, así como los datos en el Datawarehouse, son usados y examinados. Cuando los datos son organizados y estructurados para tener significado empresarial, la gente aprende mucho de los sistemas de información. Pueden quedar expuestos posibles defectos en aplicaciones actuales, siendo posible entonces mejorar la calidad de nuevas aplicaciones.

- Comunicación e Impactos Organizacionales.

  Apenas el Datawarehouse comienza a ser fuente primaria de información empresarial consistente, los siguientes impactos pueden comenzar a presentarse:

a) La gente tiene mayor confianza en las decisiones empresariales que se toman. Ambos, quienes toman las decisiones como los afectados conocen que está basada en buena información.

b) Las organizaciones empresariales y la gente de la cual ella se compone queda determinada por el acceso a la información. De esta manera, la gente queda mejor habilitada para entender su propio rol y responsabilidades como también los efectos de sus contribuciones; a la vez, desarrollan un mejor entendimiento y apreciación con las contribuciones de otros.

c) La información compartida conduce a un lenguaje común, conocimiento común, y mejoramiento de la comunicación en la empresa. Se mejora la confianza y cooperación entre distintos sectores de la empresa , viéndose reducida la sectorización de funciones.

d) Visibilidad, accesibilidad, y conocimiento de los datos producen mayor confianza en los sistemas operacionales.

Impactos Técnicos De Datawarehouse.

Considerando las etapas de construcción, soporte del Datawarehouse y soporte de sistemas operacionales, se tienen los siguientes impactos técnicos:

- Nuevas destrezas de desarrollo: cuando se construye el Datawarehouse, el impacto más grande sobre la gente técnica está dada por la curva de aprendizaje, muchas destrezas nuevas se deben aprender, incluyendo:
  - b) Conceptos y estructura Datawarehouse.
  - c) El Datawarehouse introduce muchas tecnologías nuevas (ETT, Carga, Acceso de Datos, Catálogo de Metadatos, Implementación de DSS/EIS), y cambia la manera que nosotros usamos la tecnología existente. Nuevas responsabilidades de soporte, nuevas demandas de recursos y nuevas expectativas, son los efectos de estos cambios.
  - d) Destrezas de diseño y análisis donde los requerimientos empresariales no son posibles de definir de una forma estable a través del tiempo.
  - e) Técnicas de desarrollo incremental y evolutivo.

f) Trabajo en equipo cooperativo con gente de negocios como participantes activos en el desarrollo del proyecto.

Nuevas responsabilidades de operación: Cambios sobre los sistemas y datos operacionales deben ser examinados más cuidadosamente para determinar el impacto que estos cambios tienen sobre ellos, y sobre el Datawarehouse.

1.10 Aplicaciones

Comercio Minorista

Utilizan grandes sistemas de Procesamiento Paralelo Masivo para acceder a meses o años de historia transaccional tomada directamente en los puntos de venta de cientos, o miles, de sucursales. Con esta información detallada pueden efectuar en forma más precisa y eficiente actividades de compra, fijación de precios, manejo de inventarios, configuración de góndolas, etc.

Las promociones y las ofertas de cupones son seguidas, analizadas y corregidas. Modas y tendencias son cuidadosamente administradas a efectos de maximizar utilidades y reducir costos de inventario. El stock es reasignado por sucursales o regiones según ventas y tendencias. Estos sistemas con capacidad de procesar gran cantidad de datos detallados permiten implementar eficientemente prácticas de mercadería "en consignación", en esta modalidad la cadena minorista paga al proveedor recién cuando los productos son vendidos y pasados por el lector de códigos de barras (scanner) del punto de venta.

Esta información detallada permite ejercer mayor poder de negociación sobre los proveedores, dado que el comercio minorista puede llegar a saber más que el fabricante sobre sus productos: quién lo compra, dónde, cuándo, con que otros productos, etc.

En su libro "Made in América: My Story" el fundador de Wal*Mart, Sam Walton, escribe: "...me dicen que es la base de datos comercial más grande del mundo. Lo que me gusta es la clase de información que puedo obtener de ella al instante ¡todos esos números!,

llevamos 65 semanas de historia de cada artículo que vendemos. Esto significa que puedo elegir cualquiera y decir exactamente cuantos vendimos... no en promedio, sino en cualquier región, distrito o sucursal. Es difícil que un proveedor sepa más acerca de su producto de lo que sabemos nosotros. Nos da el poder de la ventaja competitiva." Para poner esto en perspectiva debemos considerar que las sucursales a las que hace referencia Sam Walton son unas 2500 y que cada una de ellas tiene una variedad de entre 50.000 y 80.000 artículos, todas las noches 20 millones de actualizaciones se realizan en el Data Warehouse. Wal*Mart es un excelente ejemplo práctico del concepto planteado por A. Tofler en su libro "Powershift": el poder se desplaza del fabricante al minorista por el manejo de la información.

Otras instalaciones de Data Warehousing de magnitud en la industria minorista son las de Kmart, Sears, Meijer, Kohl's Department Stores, American Stores (Jewel/OSCO/Lucky/Savon/ACME/SuperSaver), Mervyn's, Buttrey Food & Drug, QVC Home Shopping, Canadian Tyre, WH Smith Books (Gran Bretaña), Great Universal (GB), Supermercados Casino (Francia), Migrosgenossenschaftsbund (Suiza), Otto Versand (Alemania).

Manufactura de Bienes de Consumo Masivo

Las empresas de este sector necesitan hacer un manejo cada vez más ágil de la información para mantenerse competitivas en la industria. Los Data Warehouses se utilizan para predecir la cantidad de producto que se venderá a un determinado precio y, por consiguiente, producir la cantidad adecuada para una entrega "justo a tiempo". A su vez se coordina el suministro a las grandes cadenas minoristas con inmensas cantidades de productos "en consignación", que no son pagados hasta que estos productos son vendidos al consumidor final.

Las cadenas minoristas y sus proveedores utilizan sus Data Warehouses para compartir información, permitiéndole a las empresas de manufactura conocer el nivel de stock en las góndolas y eventualmente hacerse responsables de la reposición de inventario de la cadena

minorista. Como es de esperar esto reduce fuertemente la intermediación. También se utilizan para campañas de marketing, planificación de publicidad y promociones y se coordinan las ofertas de cupones y promociones con las cadenas minoristas.

Un ejemplo interesante es el de Whirlpool. Este fabricante global de electrodomésticos con base en Benton Harbor, Michigan, utiliza su Data Warehouse para hacer un seguimiento directo de sus casi 15 millones de clientes y de sus más de 20 millones de aparatos instalados. Las mayores aplicaciones del sistema son para marketing, ventas, mantenimiento, garantía y diseño de productos. Permite mantener stock de partes más ajustados y mejorar las condiciones de negociación con los proveedores de las mismas. Si, por ejemplo, un determinado motor se identifica como poseedor de una tasa de falla superior, Whirlpool puede utilizar la información para hacer renegociaciones de garantía con el proveedor.

Como anécdota interesante se puede mencionar que durante el verano de 1993 los ingenieros de Whirlpool detectaron una tasa de falla muy alta en una manguera de conexión en una serie de lavarropas que se estaba vendiendo. A partir de allí se detuvo la producción, se identificaron los clientes y se enviaron técnicos a reemplazar la parte defectuosa antes de que entrara en falla. Esto no solo tuvo un impacto muy importante en satisfacción de clientes sino que se redujeron los costos de garantía por el reemplazo planificado y, especialmente, ¡se evitaron costosos reclamos por daño a la propiedad debidos a pérdidas de agua!

Otras empresas del sector que cuentan con Data Warehouses de importancia son: Coca Cola, Nike, Procter & Gamble, Hallmark, Maybelline, Helene Curtis, 3M, Owens Corning Glass, Karsten Ping Golf Clubs, Walt Disney.

Transporte de Cargas y Pasajeros

Se utilizan Data Warehouses para almacenar y acceder a meses o años de datos de clientes y sistemas de reservas para realizar actividades de marketing, planeamiento de capacidad, monitoreo de ganancias, proyecciones y análisis de ventas y costos, programas de calidad y servicio a clientes.

Las empresas de transporte de cargas llevan datos históricos de años, de millones de cargamentos, capacidades, tiempos de entrega, costos, ventas, márgenes, equipamiento, etc..

Las aerolíneas utilizan sus Data Warehouses para sus programas de viajeros frecuentes, para compartir información con los fabricantes de naves, para la administración del transporte de cargas, para compras y administración de inventarios, etc. Hacen un seguimiento de partes de repuesto, cumplimiento con las regulaciones aeronáuticas, desempeño de los proveedores, seguimiento de equipaje, historia de reservas, ventas y devoluciones de tickets, reservas telefónicas, desempeño de las agencias de viajes, estadísticas de vuelo, contratos de mantenimiento, etc.

Algunas empresas que cuentan con Data Warehouses de magnitud: Cornrail, Union Pacific, Norfolk Southern, American President Lines, Delta, Lufthansa, QANTAS, British Airways, American Airlines, Canadian Airlines, SNFC.

Telecomunicaciones

Estas empresas utilizan sus Data Warehouses para operar en un mercado crecientemente competitivo, desregulado y global que, a su vez, atraviesa profundos cambios tecnológicos. Se almacenan datos de millones de clientes: sus circuitos, facturas mensuales, volúmenes de llamados, servicios utilizados, equipamiento vendido, configuraciones de redes, etc. así como también información de facturación, utilidades, y costos son utilizadas con propósitos de marketing, contabilidad, reportes gubernamentales, inventarios, compras y administración de redes.

Muchas otras industrias y actividades utilizan actualmente, o están comenzando a instalar, Data Warehouses: entidades gubernamentales, especialmente para el control impositivo, empresas de servicios públicos, de entretenimiento, editoriales, fabricantes de automóviles, empresas de petróleo y gas, laboratorios farmacéuticos, droguerías, etc.

En la industria informática NCR dispone de los Data Warehouses de mayor magnitud y antigüedad. Sus mayores instalaciones se encuentran en distintos centros de la compañía en Estados Unidos. La de NCR El Segundo, California, es una de las más antiguas del mundo, su primera aplicación fue el seguimiento histórico y detallado de la base de clientes: llamados de servicios, productos instalados, performances, etc. Esta instalación es herencia de Teradata, compañía fundada en 1979 para la producción de sistemas de procesamiento paralelo masivo destinados a aplicaciones de soporte a la toma de decisiones y posteriormente adquirida por NCR.

En NCR San Diego, California, se encuentra el centro de desarrollo de los computadores WorldMark. Sobre los mismos se realizó la demostración del Data Warehouse más grande del mundo: 10 Terabytes de información (=10.000 Gigabytes=10.000.000 Megabytes), para poner esto en términos manejables debemos considerar que toda la información escrita de la Biblioteca del Congreso de los Estados Unidos se podría almacenar en unos 20 Terabytes.

En NCR Dayton, Ohio, la compañía dispone de un Data Warehouse de 1 Terabyte (=1000 Gigabytes) destinado fundamentalmente a tareas de marketing, producción y finanzas. A la fecha tiene almacenados 281.154 documentos, agrupados en 36 grupos de interés temático, que pueden ser accedidos 24 Hs. al día, los siete días de la semana, por 16.100 usuarios distribuidos en 46 países. A principios del año 1996 el sistema estaba respondiendo un promedio de 242.707 consultas mensuales.

Capitulo 2. Datamining

Data Mining, ***la extracción de información oculta y predecible de grandes bases de datos***, es una poderosa tecnología nueva con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse). Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información (knowledge-driven). Los ***análisis prospectivos*** automatizados ofrecidos por un producto así van más allá de los eventos pasados provistos por herramientas retrospectivas típicas de sistemas de soporte de decisión. Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas y a los cuales los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

Muchas compañías ya colectan y refinan cantidades masivas de datos. Las técnicas de Data Mining pueden ser implementadas rápidamente en plataformas ya existentes de software y hardware para acrecentar el valor de las fuentes de información existentes y pueden ser integradas con nuevos productos y sistemas pues son traídas en línea (on-line). Una vez que las herramientas de Data Mining fueron implementadas en computadoras cliente servidor de alta performance o de procesamiento paralelo, pueden analizar bases de datos masivas para brindar respuesta a preguntas tales como, "¿Cuáles clientes tienen más probabilidad de responder al próximo mailing promocional, y por qué? y presentar los resultados en formas de tablas, con gráficos, reportes, texto, hipertexto, etc.

2.1 Definición

Es un proceso de descubrimiento que permite al usuario conocer la esencia y las relaciones entre sus datos. Data mining descubre patrones y tendencias en el contenido de esta información. También se utilizan términos con igual o similar significado, tales como Knowledge Mining from Data Base (<< Minería de conocimiento de base de datos>>), Knowledge Extracción (<< Extracción del conocimiento>>), Data Archeology (<<Arqueología de Datos>>) y Data Dredging (<<Excavación de Datos>>). Algunos autores también se refieren a esta actividad bajo el nombre de KKD : Knowledge Discovery in Database (<< Descubrimiento de conocimiento en base de datos>>), no obstante, otros afirman que Data Mining es solo un paso al proceso total del KKD.

Para evitar confusiones en este trabajo se considerará que el data Mining es sinónimo de KKD.

A continuación se detallan las etapas del proceso de KKD:

1. Limpieza en los datos: Este paso intenta remover tanto ruido como inconsistencia que pueda encontrarse en los datos sobre los que se trabajará. Como se vera mas adelante no siempre es deseable que se lleve cabo este paso.

2. Integración de los Datos: Se combinan las diferentes fuentes de datos en un único repositorio. Este repositorio global se denomina Data Warehouse.

3. Selección de Datos: A partir de los datos integrados en el repositorio, se toman para trabajar solamente aquellos que pueden resultar útiles para el análisis que se va a realizar.

4. Transformación de Datos: Es la agrupación de los datos útiles de forma apropiada para utilizar los algoritmos de Data Mining.

5. Data Mining: Es la aplicación de métodos inteligentes para obtener patrones útiles de datos

6. Evaluación de Patrones: De todos los patrones obtenidos, hay que diferenciar aquellos que son verdaderamente útiles, es decir, los que representan algún tipo de conocimiento o medición interesante.

7. Presentación del Conocimiento: Es la aplicación de técnicas especiales utilizadas en la representación a los usuarios de conocimiento extraído.

Las herramientas de Data Mining predicen tendencias y comportamientos, posibilitando tomar decisiones proactivas y conducidas por un conocimiento acabado de la información (<< knowledge drive decisions>>). Las herramientas de Data mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo de procesamiento, que los usuarios de esta información no están dispuestos a aceptar. Estas herramientas explorar las bases de datos en busca de patrones ocultos.

Con el pasar de los años, junto con el avance de la tecnología, el costo de almacenar y mantener los datos ha disminuido drásticamente. Debido a esto las empresas y organismos llevan registros de todas sus operaciones de manera realmente precisa y completa. La acumulación constante de datos genera repositorios muy grandes que se administran con herramientas que no han sido diseñadas para manejar estos cambios.

El auge que ha alcanzado actualmente el Data Mining es debido a que en el presente nos encontramos ante enormes cantidades de datos y con la urgente necesidad de transformarlos en información útil y conocimiento. Se dice que sin Data Mining somos <<ricos en datos>> pero <<pobres en información>>.

Muchas compañías (de telecomunicaciones, ventas on line y supermercados) recogen y refinan cantidades masivas de datos. Las técnicas de Data Mining pueden ser implantadas rápidamente en plataformas existentes de software y hardware para aumentar el valor y la utilidad de las fuentes de información existentes.

El nombre de Data Mining deriva de la similitud entre buscar en grandes bases de datos información valiosa sobre negocios, y minar una montaña para encontrar metales preciosos. Ambos procesos requieren examinar una inmensa cantidad de material, o investigar inteligentemente hasta encontrar exactamente donde residen los valores. No obstante, la terminología utilizada no es exacta, ya que en realidad debería llamarse << Información Mining>>, ya que lo que se busca extraer es información y no datos. Pero ese nombre no ha tenido tanta aceptación.

2.2 Patrones de Información

Como se menciona anteriormente. el objetivo del Data Mining es obtener patrones ¿ pero que tipo de patrones? , para responder esta pregunta primero hay que hacer un división importante. Las tareas de Data Mining pueden ser descriptivas o predictivas. Las descriptivas caracterizan las propiedades generales de los datos en una base de datos. Por el contrario, las predictivas realizan inferencias en los datos para poder realizar predicciones. En todos los casos en los que se desee aplicar técnicas de Data Mining es muy importante tener en claro de que tipo es el mas conveniente para utilizar. Esto depende de los resultados que se quieren obtener.

Funcionalidades y lista de patrones que pueden ser descubiertos utilizando Data Mining:

a. Descripción de clases: la idea de trabajar con este tipo de patrones es poder asociar datos en clases o conceptos. Hay tres formas para enfrentar este enfoque. La primera se denomina Caracterización de los datos (Data Caracterización) y se basa en realizar una sumarización de las características generales de una clase en particular de datos. Los resultados de este tipo de análisis suele presentarse en la forma de reglas de caracterización. La segunda es la Discriminación de datos(Data Discrimination) que es una comparación entre las características generales de los objetos de una clase en particular respecto a las de otro conjunto contrastante. En este caso se habla de reglas de discriminación.
b. Análisis de Asociación: Es el descubrimiento de reglas de asociación que muestran condiciones del tipo atributo-valor que ocurren con frecuencia dentro de un conjunto de datos.
c. Clasificación y Predicción: Es el proceso por el cual se busca un conjunto de modelos o funciones que describan y distingan clases de datos o conceptos, con el objetos de utilizar dichos modelos para predecir de que clase son ciertos objetos. El modelo derivado se basa en el análisis de un conjunto de datos de entrenamiento (es decir, datos de los cuales si se conoce su clase).

d. Análisis de Clusters: En el caso anterior se hacia referencia a analizar clases conocidas, pero el clusring analiza objetos sin consultar clases conocidas. Por este motivo, esta técnica se clasifica como aprendizaje no supervisado. En general, las clases no se presentan en los datos de entrenamiento simplemente porque no se conocen. El análisis de clusters se utiliza justamente para identificar las clases. El proceso trabaja agrupando objetos según el principio de <<maximizar la similitud dentro de una clase y minimizar la similitud entre clases>>.

e. Análisis de Infrecuentes: Una base de datos puede contener objetos que no obedecen al comportamiento general o al modelo de los datos. Esos objetos son los infrecuentes (Outliers). La mayoría de los métodos de Data Mining descartan estos objetos como si se trataran de ruido o excepciones. Pero hay que tener en cuenta que, en algunos casos, este tipo de información puede ser más útil que los sucesos regulares. Este tipo de análisis suele llamarse Outliers Mining.

f. Análisis Evolutivo: Describe y modela la monotonía o la tendencia que posee determinados objetos que tienen un comportamiento variable en el tiempo. Un ejemplo clásico de aplicación de estos patrones es el de desarrollar un sistema que identifique las regularidades en un conjunto de acciones de la bolsa, para poder determinar así los ciclos de alzas y bajas. No obstante, debe tenerse en cuenta que los sistemas de Data Mining son propensos a generar muchos más patrones o reglas de los que se esperan al inicio del proceso. Esto no necesariamente quiere decir que hay tanto conocimiento oculto en el conjunto original de datos, ya que hay que tener en cuenta que no todos los patrones generados son útiles. Es necesario entonces poder discriminar que patrones son útiles y cuales no lo son.

Un patrón es interesante si cumple con las siguientes condiciones:

- Es fácilmente comprensible para las personas
- Es válido, con cierto grado de certeza, para otro conjunto de datos, ya sea nuevo o de prueba.
- Tiene una utilidad potencial

- Expresa un conocimiento novedoso y no trivial

Un patrón también será útil si sirve para validar una hipótesis que le usuario pretende confirmar. No es difícil notar que las condiciones observadas anteriormente son, en su mayoría, subjetivas ya que dependen en gran medida, de las creencias que el usuario tiene de los datos. Para establecer un criterio mas uniforme se crearon varias medias objetivas para determinar si un patrón es, o no, interesante. Principalmente se basa en la estructura de patrones descubiertos y en las estadísticas que los apoyan. En el caso de las reglas de asociación, las medidas objetivas que se utilizan son el soporte y la confianza.

2.3 Reglas de asociación:

La búsqueda de reglas de asociación se utiliza para encontrar asociaciones interesantes o relaciones de correlación entre un conjunto grandes de datos.

La forma clásica de ver la extracción de reglas de asociación es a través del análisis del análisis del carrito de compras, es decir, intentar obtener información sobre los hábitos de compras de los clientes a partir de la información de las transacciones diarias que se realizan. Analicemos el clásico ejemplo de un supermercado. A través del registro de transacciones, donde se indican que items se vendieron se podría determinar que productos son comprados simultáneamente por una cantidad importante de clientes. De esta forma podría descubrirse, por ejemplo, que los días sábados es muy posible que la gente que compra cerveza también compre pañales. Este tipo de información resulta sumamente útil y se extrajo de los datos que la empresa ya poseía pero de la cual no podía extraer <<conocimiento>>de este tipo utilizando herramientas comunes.

Si pensamos en el universo como el conjunto de todos los artículos disponibles en una tienda, entonces cada uno posee una variable lógica que representa su presencia o ausencia dentro de un grupo de elementos. Cada carrito de compras puede entonces ser representado por un vector lógico de valores asignados a esas variables. Es posible analizar estos vectores de forma conjunta para determinar los patrones de compras y reflejar aquellos items que se encuentran frecuentemente asociados o comprados juntos, sin olvidar que lo que realmente interesa es la información << no trivial >> que pueda extraerse. Estos patrones pueden representarse en la forma de reglas de asociación.

Por ejemplo, la información que indica que los clientes que compran cerveza también suelen adquirir pañales al mismo tiempo se representa por la siguiente regla de asociación:

Cerveza => pañales

[soporte = 4%, confianza = 70%]

El soporte y la confianza son dos medidas del interés de la regla. Estos parámetros indican qué tan útil puede resultar la regla extraída. Un soporte del 4% indica que el 4% de todas las transacciones analizadas mostraron que la cerveza y los pañales fueron comprados juntos. Una confianza del 70% indica que el 70% de los clientes que compraron cerveza y los pañales. Generalmente una regla de asociación se considera útil si satisface un mínimo de soporte y un mínimo de confianza preestablecidos. Ambos umbrales son fijados por el usuario o por un experto en el tema.

Definición 1 – [Soporte y Confianza] una regla A => B en el conjunto de transacciones D tiene soporte s, cuando es el porcentaje de las transacciones de D que, conteniendo a A , también contienen a B. Esto se expresa como la probabilidad condicional P(A|B):

$$\text{Confianza } (A \Rightarrow B) = P(A \mid B)$$

Definición 2 – [Conjunto de items Frecuentes] Dado un soporte S, un conjunto de items frecuentes es el conjunto de items que aparece al menos S veces dentro del conjunto de transacciones. Un conjunto de items frecuente es un conjunto de items que están en la misma transacción y que se repiten al menos el mínimo de soporte en todas las transacciones consideras. El proceso para extraer las reglas de asociación de una base de datos consta de dos partes:

a. Encontrar todos los conjuntos de items frecuentes
b. Generar reglas de asociación fuertes entre los items frecuentes. Las reglas de asociación permiten dada una confianza y un soporte predefinido determinar si

existen una relación entre dos conjuntos de items. Por ejemplo si tenemos una lista de transacciones de un supermercado, donde cada transacción es una compra realizada por un cliente cualquiera, mediante las reglas de asociación se puede descubrir que cuando un cliente lleva un x producto también lleva el producto y, con un soporte y una confianza determinados. Una vez obtenidos los conjuntos de items frecuentes para obtener las reglas de asociación se puede expresar la probabilidad condicional en términos de la cuenta soporte de un conjunto de items:

Confianza (A => B) = P (B |A) = cuenta_soporte (A U B) / cuenta_soporte (A)
Donde cuenta_soporte (A U B) es el numero de transacciones que contiene los conjuntos de items A y B o sea AUB. Y cuenta_soporte (A) es el numero de transacciones que contiene un conjunto de items A. A partir de esta ecuación, las reglas de asociación pueden ser generadas según el algoritmo 1

Algoritmo 1 – [Generación de reglas de Asociación]
Input: Conjunto de items frecuentes
Output: Conjunto de reglas de Asociación
Método:

1. Para cada conjunto de items frecuentes L, generar todos los subconjuntos no vacíos de L.
2. Para cada subconjuntos S NO vació de L, generar la regla <<s → (L-s)>> si (cuenta_soporte(L) / cuenta soporte(s))>=confianza_minima, si confianza_minima es el umbral mínimo de confianza.

2.3.1 Tipos de Reglas de Asociación

El análisis del carrito de compras es un ejemplo de solamente una de las formas que pueden tener las reglas de asociación. En realidad, hay varias clases posibles de reglas. Estas se clasifican siguiendo el criterio explicado a continuación:

* Según el tipo de valores que se manejan en la regla:

Si una regla está basada en la presencia o ausencia de items, se dice que es una regla de asociación Booleana. Ejemplo:

cerveza => pañales

Si una regla describe una asociación entre atributos o items cuantitativos, entonces se dice que es una regla de asociación Cuantitativa. Para este tipo de reglas, los valores se dan particionados en intervalos (indicados por dos puntos '..'). El siguiente es un ejemplo de este tipo de reglas, dónde X es una variable que representa a un cliente:

edad(X,»20..25") ^ ingresos(X,»20000..35000") _ compra(X,minicomponente)

* Según las dimensiones de los datos que intervienen en la regla:
Si los items o atributos de una regla solamente referencian una dimensión (donde una dimensión se representa mediente un predicado), entonces es una regla de Dimensión Simple. Ejemplo:

compra(X,cerveza) _ compra(X,pañales)

En cambio, cuando una regla referencia dos o más dimensiones, se dice que la regla es Multi-Dimensional. En el siguiente ejemplo se utilizan dos dimensiones, *compra y edad*:

compra(X,cerveza) ^ edad (X,»20..45") _ compra(X,pañales)

* Según los niveles de abstracción empleados en el conjunto de reglas:
Algunos métodos de extracción de reglas permiten encontrar conocimiento en diferentes niveles de
abstracción. Por ejemplo:

edad(X,»19..25") _ consulta(X,»www.onlineub.com»)
edad(X,»19..25") _ consulta(X,»www.onlineub.com/default.htm»)

En el primer ejemplo, la regla hace referencia al sitio consultado, mientras que la segunda hace referencia concreta a que página del sitio se accede. Ambas reglas se diferencian por el nivel de abstracción. De esta forma pueden diferenciarse entre reglas de asociación Multinivel y de Nivel Único.

* Según extensiones a la extracción de asociaciones:
La extracción de asociaciones puede ser extendida al análisis de correlaciones, dónde la ausencia o presencia de items correlativos puede ser identificada. También puede extenderse para extraer *maxpatterns*4 y elementos frecuentemente cercanos. Estas extensiones pueden ser utilizadas para disminuir la cantidad de itemsets frecuentes generados en el mining.

2.3.2 Algoritmo Apriori

El algoritmo implementado para el proyecto esta basado en el algoritmo APRIORI [AGR93]. Este es un algoritmo que permite detectar los conjuntos de items más frecuentes en distintas transacciones, a través de la generación de candidatos y de reglas de asociación booleanas.

El nombre de este algoritmo hace referencia a que, en cada paso, se utiliza el conocimiento de la propiedad llamada Apriori de los itemsets frecuentes. Esta propiedad indica que para que un itemset sea frecuente todos sus subconjuntos no vacíos también deberán ser frecuentes.

El algoritmo tiene dos pasos: el paso JOIN (junta) y el paso PRUNE (optimización). El paso JOIN consiste en calcular los conjuntos de K items frecuentes, y el paso PRUNE se encarga de depurar aquellos conjuntos de K items que incluyen algún K-I conjunto de items no frecuente dentro, reduciendo la cantidad de K items frecuentes, optimizando así el proceso.

Es importante aclarar que el algoritmo Apriori va calculando los conjuntos de K items frecuentes, a partir del conjunto K-1 de items frecuentes, por lo tanto, si asegurándose que K-1 tiene solo conjuntos de items frecuentes (gracias al paso PRUNE), se logra reducir el procesamiento en el paso JOIN.

Algoritmo 2 - [Algoritmo Apriori]

```
INPUT: Conjunto de transacciones (D), Soporte mínimo
(min_sup)
OUTPUT: Itemsets frecuentes en D (L1)
ESTRUCTURAS DE DATOS: Conjunto de n itemsets frecuentes (Ln),
Candidatos a itemsets frecuentes(Cn)
L1 = {Lista de itemsets en D}
PARA (k=2 ; Lk-1<>0 ; k++)
     Ck = APRIORI_GEN(Lk-1,min_sup)
     PARA (todas las transacciones t ( D)
          Ct = SUBCONJ(Ck ,t) // Toma todos los candidatos
     que contienen a t
          PARA (todos los candidatos c ( Ct)
               c.count++
          FIN PARA
     FIN PARA
     Lk = { c ( Ck / c.count ( min_sup }
FIN PARA
DEVOLVER L=Uk Lk

PROCEDIMIENTO APRIORI_GEN (Lk-1,min_sup)
     PARA (todos los itemsets l1 ( Lk-1)
          PARA (todos los itemsets l2 ( Lk-1)
               SI  (l1[1]=l2[1])^  (l1[2]=l2[2])^...^  (l1[k-
          1]=l2[k-1]) ENTONCES
```

```
                    c = l1 JOIN l2 //Paso JOIN
                    SI (TIENE_SUBCONJ_FRECUENTES (c,Lk-1))
                         BORRAR c //Paso PRUNE
                    SI NO
                         AGREGAR c A Ck
                    FIN SI
               FIN SI
          FIN PARA
FIN PARA
DEVOLVER Ck
FIN PROCEDIMIENTO

FUNCION TIENE_SUBCONJ_FRECUENTES (c,Lk-1)
     PARA (cada subconjunto (k-1) s ( C)
          SI ( S ( Lk) ENTONCES
               DEVOLVER VERDADERO
          SI NO
               DEVOLVER FALSO
          FIN SI
     FIN PARA
FIN FUNCION
```

A continuación se brinda un ejemplo del funcionamiento del algoritmo Apriori:

Ejemplo 1:

Sea un conjunto de Transacciones T={T1,T2,...,T9} donde cada Ti está constituído por un conjunto de

ítems Ii. El soporte exigido es del 22% (2/9).

T1 : I1, I2, I5

T2 : I2, I4

T3 : I2, I3

T4 : I1, I2, I4

T5 : I1, I3

T6 : I2, I3

T7 : I1, I3

T8 : I1, I2, I3, I5

T9 : I1, I2, I3

En la primera pasada del algoritmo cada ítem es miembro del conjunto de candidatos y se busca el número de ocurrencias de cada uno en todas las transacciones (Tabla 2.1).

Itemset	Soporte
{|1}	6
{|2}	7
{|3}	6
{|4}	2
{|5}	2

Tabla 2.1:Items candidatos del 1 elemento

Itemset
{|1}
{|2}
{|3}
{|4}
{|5}

Tabla 2.2 Itemsets frecuentes de 1 elemento

La cantidad mínima de apariciones indicada por el soporte es 2, es decir que en el primer paso no se borrarán itemsets ya que todos los candidatos son frecuentes (Tabla 2.2). A

continuación se procede a generar la lista de cantidades para dos ítems y se calcula el soporte para cada uno de ellos (Tabla 2.3). Los marcados en negrita son aquellos que no cumplen con el soporte necesario y se eliminarán (Tabla 2.4).

Itemset	Soporte
{|1,|2}	4
{|1,|3}	4
{|1,|4}	**1**
{|1,|5}	2
{|2,|3}	4
{|2,|4}	2
{|3,|4}	**0**
{|3,|5}	**1**
{|4,|5}	**0**

Tabla 2.3: Itemsets candidatos de 2 elementos

Itemset
{|1,|2}
{|1,|3}
{|1,|5}
{|1,|5}
{|2,|3}
{|2,|4}
{|2,|5}

Tabla 2.4: Itemsets frecuentes de 2 elementos

A partir de los itemsets frecuentes se generan los candidatos de tres ítems (Tabla 2.5). Pero al realizarse la junta (paso JOIN del algoritmo) debe tenerse en cuenta la propiedad Apiori (todos los subconjuntos de un conjunto frecuente también deben ser frecuentes) porque

gracias a ella se pueden eliminar combinaciones que no tienen posibilidades de ser frecuentes (paso PRUNE del algoritmo), ahorrando tiempo de proceso.

Itemset	Soporte
{|1,|2,|3}	**2**
{|1,|2,|5}	**2**
{|1,|3,|5}	No se puede ser frecuente por {|3,|5}
{|2,|3,|4}	No se puede ser frecuente por {|3,|4}
{|2,|3,|5}	No se puede ser frecuente por {|3,|5}
{|2,|4,|5}	No se puede ser frecuente por {|4,|5}

Tabla 2.5:Itemsets cantidades de 3 elementos

Posteriormente se comprueba el soporte de los itemsets candidatos (sólo de aquellos que no fueron eliminados por el prune) y se determinan cuales son los itemsets frecuentes (Tabla 2.6).

Itemset
{|1,|2,|3}
{|1,|2,|5}

Tabla 2.6: Itemsets frecuentes de 3 elementos

El proceso continúa en forma análoga hasta que el conjunto de items candidatos tenga cero elementos y ya no puedan determinarse nuevos itemsets frecuentes.
Se han propuesto las siguientes optimizaciones a efectos de mejorar la eficiencia del algoritmo. Algunas de ellas son:

- **Hashing:** Utilizar una técnica de hashing para reducir el tamaño de los itemsets candidatos.

- **Reducción de Transacciones:** Se basa en reducir el número de transacciones tenidas en cuenta en las futuras iteraciones del algoritmo.
- **Particiones:** Funciona particionando los datos en dos grupos. En un primer paso se determinan los itemsets frecuentes locales a cada partición. Posteriormente se determina cuales de esos son itemsets frecuentes globales.
- **Muestreo:** Trabaja con el concepto de determinar los itemsets frecuentes de un subconjunto (muestra) del total de las transacciones. Se pierde precisión a cambio de eficiencia.
- **Conteo Dinámico de Itemsets:** Es una modificación del Apriori que consume menos pasadas por las transacciones, ya que los nuevos itemsets candidatos pueden ser agregados en cualquier momento durante el proceso.

2.4. Fundamentos del Data Mining

Las técnicas de Data Mining son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los

datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. Data Mining toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. Data Mining está listo para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

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

Las bases de datos comerciales están creciendo a un ritmo sin precedentes. Un reciente estudio del META GROUP sobre los proyectos de Data Warehouse encontró que el 19% de los que contestaron están por encima del nivel de los 50 Gigabytes, mientras que el 59% espera alcanzarlo en el segundo trimestre de 1997. En algunas industrias, tales como ventas al por menor (retail), estos números pueden ser aún mayores. MCI Telecommunications Corp. cuenta con una base de datos de 3 terabytes + 1 terabyte de índices y overhead

corriendo en MVS sobre IBM SP2. La necesidad paralela de motores computacionales mejorados puede ahora alcanzarse de forma más costo - efectiva con tecnología de computadoras con multiprocesamiento paralelo. Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que consistentemente son más performantes que métodos estadísticos clásicos.

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

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

2.4.1 Descubrimiento:

El descubrimiento aparece sin una idea predeterminada acerca de lo que la búsqueda va a encontrar. No hay intervención en el proceso por parte del usuario final; el proceso de descubrimiento de data warehouse examina los datos de entrada buscando similaridades y ocurrencias en los datos, lo cual le permite agruparlos e identificar patrones. En el entorno electrónico del data warehouse , este proceso se debe poder realizar en un corto período de tiempo. La rapidez en la obtención de los resultados es crucial para la adopción de un producto de data mining.

En algunos sistemas operacionales se captura tanta información que el contenido de algunos elementos de información, aparentemente sin importancia, se pierde en un todo. De esta forma, la esencia de los datos aparece como importante, y los valores de los elementos de información nos pueden producir una gran rentabilidad cuando son tratados mediante un data mining.

2.4.2 Relaciones:

El descubrimiento de las relaciones es fundamental para tener éxito en marketing. En sistemas operacionales o data warehouse, el arquitecto de datos y el personal de diseño han definido meticulosamente entidades y relaciones. Una entidad es un conjunto de información que contiene hechos acerca de un grupo de datos asociado. El proceso de descubrimiento en un ejercicio de data mining arroja luz sobre las relaciones escondidas en las profundidades de muchas capas de datos corporativos.

2.4.3 Patrones:

Los beneficios del descubrimiento de patrones a un negocio agregan valor real a un ejercicio de data mining.

Nadie puede predecir con precisión que la persona X va a ejecutar la actividad YY en estrecha proximidad con la actividad que Z. De cualquier forma, usando técnicas de data mining y un análisis sistemático en datos del data warehouse, esta predicción puede ser apoyada por el descubrimiento de patrones de conducta.

Existe un componente temporal en el descubrimiento de patrones. Una conducta paralela durante un periodo de dos semanas puede descubrir un patrón que podría llegar a ser la razón para la implementación de una nueva compaña de marketing. En cambio la detección de esta conducta a lo largo de un período de seis meses daría más credibilidad a la idea de que se ha descubierto un patrón. Los patrones están estrechamente ligados al hábito; en otras palabras, la probabilidad de que una actividad se ejecute en proximidad estrecha a otra se descubre en la identificación de un patrón.

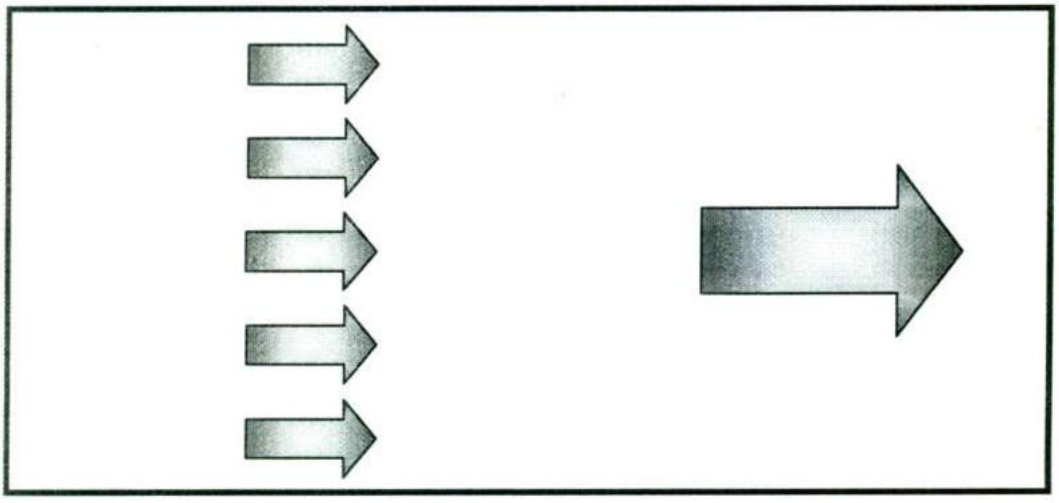
Data mining deja que las compañías vean su información de formas desconocidas previamente. Algunos sistemas operacionales, en la tarea de satisfacer a diario los requisitos del negocio, crean grandes cantidades de datos. Los datos son complejos y las relaciones entre los elementos no se hallan fácilmente mediante la simple observación.

En la figura (2.4) se muestra la verdadera naturaleza de data mining. Usando el data mining como herramienta, los datos de convierten el información.

Los datos están constituidos por una serie de caracteres que por si mismos no significan nada. Al agruparlos constituyendo elementos de datos, comienzan a tomar sentido. Cuando

se realiza un esfuerzo de data mining con estos elementos de datos es cuando pueden proporcionar información valiosa. Las empresas tienen sólo una cantidad limitada de dinero y tiempo para consumir en iniciativas de data mining. Cuando las compañías ven los beneficios a corto y largo plazo de todo el tiempo y dinero invertidos compensaran con creces a largo plazo.

Figura (2.4). Resumen del proceso de Data Mining



2.5 Elementos del Data mining

El Data mining se constituye como el proceso, máximamente automa-tizado, intermedio entre la información y la toma de decisio-nes por parte de la dirección de la organización. El Data Mining se aplica sobre las bases de datos corpo-rativas, Data Warehouse, o sobre aquellas otras específicas de propósito departamental (DataMarts), una vez que éstas tienen adecuadamente estructurada, transformada y limpia toda la información de interés.

A partir de tal información, las herramientas y técnicas del Data Mining contemplarán los siguientes elementos:

- *Agentes Inteligentes:* Se encargan de analizar la información para detectar patrones y relaciones de forma automática o interactiva con el analista. De esta forma, estos

agentes son capaces de identificar grupos, comportamientos, reglas, mode-los cuyo descubrimiento e, incluso, planteamiento habría supuesto un enorme esfuerzo de trabajo metódico. (Inteligencia Artificial).

- *Detección de Alarmas:* Que consiste en la ejecución periódica o permanente de agentes inteligentes con el objetivo de detec-tar y reconocer momentos, situaciones o acciones susceptibles de desencadenar una acción o decisión extraordinaria fuera del ciclo ordinario.

- *Análisis Multidimensional:* Consistente en la estructuración y presentación de la información bajo aquellas perspectivas, ejes o dimensiones de interés. A los anteriores elementos, medidas y dimensiones, el análisis multidimensional añade un tercer elemento como es la jerarquía de la dimensión.

- *Consultas e Informes (Query and Report):* Toda base de datos debe tener herramientas que permitan realizar consultas y obtener como resultado informes; es

- por ello, que toda plata-forma de Data Mining debe contar con este tipo de utilidades o herramientas. El estándar SQL ha evolucionado incorporando interfaces gráficas muy avanzadas, intuitivas y fáciles de usar.

- *DSS/EIS:* Estas herramientas de soporte a la toma de decisio-nes son las precursoras del actual DM. Aquellas DSS nacieron para tratar, transformar y presentar la información en tablas y gráficos de forma que la dirección pudiera comprender la situación y tomar decisiones. (Las DSS están orientadas al analista). Las EIS surgieron para permitir a la dirección el acceso de forma fácil, visual y gráfica a la información previamente identificada como crítica y relevante.

- *Visualización de datos:* Dado que el DM conlleva la explora-ción y observación de la información, el aspecto de su visua-lización se hace muy importante.

- *Tratamiento de datos:* Al ser los datos y la información la materia prima del DM el tratamiento de éstos es un elemento muy importante. Las soluciones DM deben contemplar módulos de Tratamiento de Datos a efectos de simplificar al máximo los interfaces de datos e información. Estos módulos pueden ser transparentes al usuario, o pueden constituirse en opciones inteligentes guiadas.

2.6. Beneficios de Data mining

Uno de los principales beneficios es la habilidad para convertir presentimientos en hechos. Se puede usar el data mining para apoyar a refutar presentimientos que la gente tiene acerca de cómo va el negocio. Se puede usar añadir credibilidad a estos presentimientos y garantizar la dedicación de más recursos y tiempo a las áreas más productivas de las operaciones de una compañía. Este beneficio trata de situaciones donde una compañía comienza un esfuerzo en data mining con una idea de qué buscan. Esta se llama data mining orientado. El data mining puede descubrir patrones inesperados en conducta, patrones que no se consideraban cuando el ejercicio de búsqueda comenzó. Esto se llama data mining inesperado (out-of-the-blue).

2.6.1 Escalabilidad de la solución electrónica

Los más importantes dentro del terreno de data mining proveen soluciones robustas y escalables. Una solución robusta es aquella con buen rendimiento y que puede enseñar resultados en periodos de tiempo aceptables. La duración de este periodo aceptable depende de la experiencia pasada del usuario y de sus expectativas. Las soluciones de data mining robustas proporcionan resultados en periodos de tiempo apropiados.

Los productos de los proveedores con más éxito en soluciones software de data mining pueden analizar desde pequeñas cantidades de datos hasta cantidades muy grandes. La habilidad para trabajar con un amplio rango de conjuntos de datos de entrada es la parte del fenómeno llamado escalabilidad. Otra componente de la escalabilidad es la capacidad de instalar una solución de data mining en un computador personal autónomo, en un pequeño grupo de computadoras a nivel empresarial. La transición desde uno hasta muchos usuarios ha de ser tanto transparente y uniforme para los usuarios como fácil de desplegar para los

profesionales responsables del esfuerzo de data mining a nivel empresarial o de grupo de trabajo.

2.7 Alcance de Data Mining

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

- **Predicción automatizada de tendencias y comportamientos**. Data Mining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos. Un típico ejemplo de problema predecible es el marketing apuntado a objetivos (targeted marketing). Data Mining usa datos en mailing promocionales anteriores para identificar posibles objetivos para maximizar los resultados de la inversión en futuros mailing. Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.
- **Descubrimiento automatizado de modelos previamente desconocidos.** Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar *datos anormales* que pueden representar errores de tipeado en la carga de datos.

2.8 Herramientas del Data Mining

La gestión de consultas, visualización de datos y genera-ción de informes se fundamentan en una serie de herramientas básicas usadas bien de forma independiente o de forma conjunta e integrada. Estas herramientas son:

- Agrupamiento ("Clustering"):

También llamada Segmentación, esta herramienta permite la identificación de tipologías o grupos donde los elementos guardan similitud entre sí y diferencias con aquellos de otros grupos.

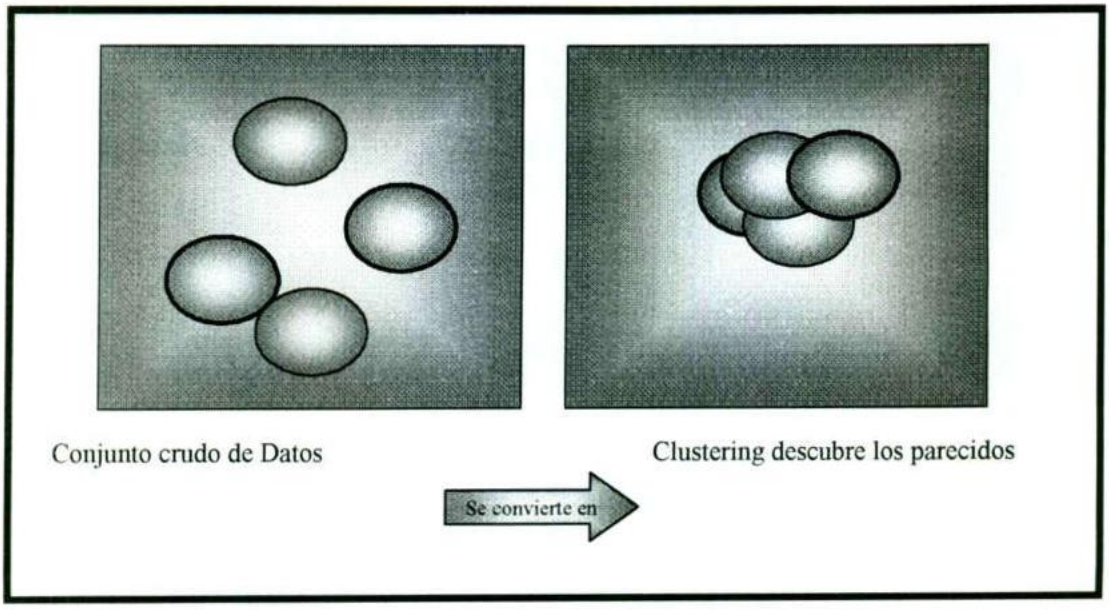
Para alcanzar las distintas tipologías o grupos existen-tes en una base de datos, estas herramientas requieren, como entrada, información sobre el colectivo a segmentar. Esta información corresponderá a los valores concretos, para cada elemento en un momento del

tiempo, de una serie de variables ("Segmentación estática") o a través del comportamiento en el tiempo de cada uno de los elementos del colectivo ("Segmentación dinámica").

Como resultado del tratamiento de la información, estas herramientas presentan los distintos grupos detectados junto con los valores característicos de las variables.

Este tipo de herramientas se basan en técnicas de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros. Para otro tipo de elementos a agrupar o segmentar, como texto y documentos, se usan técnicas de reconocimiento de conceptos.

Figura 2.8 La técnica de clustering descubre las similaridades



- Asociación ("Association Pattern Discovery"):

Este tipo de herramientas establece las posibles relacio-nes o correlaciones entre distintas acciones o sucesos aparen-temente independientes, pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros.

Normalmente este tipo de herramientas se fundamenta en técnicas estadísticas como los análisis de correlación y de variación.

- Secuenciamiento ("Sequential Pattern Discovery"):

Esta herramienta permite identificar como, en el tiempo, la ocurrencia de una acción desencadena otras posteriormente. Es muy similar a la anteriormente analizada si bien, en este caso, el tiempo es una variable crítica e imprescindible a introducir en la información a analizar.

- Reconocimiento De Patrones ("Pattern Matching"):

Estas herramientas permiten la asociación de una señal o información de entrada con aquella o aquellas con las que guarda mayor similitud y que están catalogadas en el sistema. Estas herramientas son usadas por elementos que son tan habituales como un procesador de texto o un despertador. Los patrones pueden ser cualquier elemento de información que deseemos. En el ámbito particular del DM estas herramientas pueden ayudarnos en la identificación de problemas e incidencias y de sus posibles soluciones toda vez que dispongamos de la base de información necesaria en la que buscar. Estas herramientas se sustentan en las técnicas de Redes Neuronales y Algoritmos Matemáticos.

- Previsión ("Forecasting"):

La Previsión establece el comportamiento futuro más probable dependiendo de la evolución pasada y presente. Esta herramienta tiene su uso fundamental en el trata-miento de Series Temporales y las técnicas asociadas disponen de una importante madurez. Las herramientas de Previsión utilizan bien la propia información histórica, o bien, la información histórica rela-tiva a otras variables de las cuales la primera depende.

- Simulación:

Las herramientas de Simulación forman parte también del conjunto de herramientas veteranas de la investigación cientí-fica. Como ejemplo están las herramientas de diseño y produc-ción asistidas por ordenador, "CAD"-"CAM", en las cuales se revisan los diseños sometiéndoles a una amplísima serie de condiciones reales normales y extremas.
Ello permite no sólo ajustar y adaptar el diseño sino poste-riormente establecer márgenes y límites de funcionamiento.

La simulación se puede definir como la generación de múltiples escenarios o posibilidades sujetos, normalmente, a unas reglas o esquemas con el objeto de analizar la idoneidad y comportamiento de una decisión o prototipo en un marco de posibles condiciones futuras o para analizar todas las posi-bles variaciones o alternativas a una decisión o situación y también se usa para el cálculo numérico.

- Optimización:

Al igual que la Previsión y la Simulación, las herramien-tas de Optimización tienen una amplia tradición de uso. La optimización ha sido y es extensivamente usada en la resolución de los problemas asociados a la logística de dis-tribución y a la gestión de "Stocks" en los negocios y en la determinación de parámetros teóricos a partir de los experi-mentos en la investigación científica. La optimización resuelve el problema dela minimización o maximización de una función que depende de una serie de varia-bles, encontrando los valores de éstas que satisfacen esa condición de máximo, típicamente beneficios, o mínimo, normal-mente costes.

Habitualmente estos problemas conllevan, adicionalmente, una serie de "ligaduras" o restricciones de forma que no todas las posibles soluciones son aceptables, ello se traduce en que debemos reducir nuestro universo de búsqueda a aquellas solu-ciones que satisfagan tales restricciones.

- Clasificación ("Clasification", "Prediction" O "Scoring"):

La clasificación agrupa todas aquellas herramientas que permiten asignar a un elemento la pertenencia a un grupo o clase. Ello se instrumenta a través de la dependencia de la pertenencia a las clases en los valores de una serie de atri-butos o variables. A través del análisis de un colectivo de elementos, o casos de los cuales conocemos la clase a la que pertenecen, se establece un mecanismo que establece la pertenencia a tales clases en función de los valores de las distintas variables y nos permite establecer el grado de discriminación o influencia de éstas. También se utiliza para estas herramientas la denomina-ción de Predicción o Evaluación para aquellos casos donde se aplican técnicas, normalmente numéricas, que establecen para cada elemento un valor dependiente de los valores que tengan las variables en tal elemento. Las herramientas de Clasificación hacen uso de técnicas como algoritmos matemáticos, análisis discriminante y de variaciones, sistemas expertos y sistemas de conocimiento e inducción de reglas.

*Como se ha podido apreciar, normalmente es necesaria la con-junción e integración de varios tipos de herramientas a efec-tos de brindar una solución completa a nuestros problemas.

2.9 Técnicas del Data Mining

Las técnicas de Data Mining pueden redituar los beneficios de automatización en las plataformas de hardware y software existentes y puede ser implementadas en sistemas nuevos a medida que las plataformas existentes se actualicen y nuevos productos sean desarrollados. Cuando las herramientas de Data Mining son implementadas en sistemas de procesamiento paralelo de alta performance, pueden analizar bases de datos masivas en minutos. Procesamiento más rápido significa que los usuarios pueden automáticamente experimentar con más *modelos* para entender datos complejos. Alta velocidad hace que sea práctico para los usuarios analizar inmensas cantidades de datos. Grandes bases de datos, a su vez, producen mejores predicciones.

Las bases de datos pueden ser grandes tanto en profundidad como en ancho:

- **Más columnas**. Los analistas muchas veces deben limitar el número de variables a examinar cuando realizan análisis manuales debido a limitaciones de tiempo. Sin embargo, variables que son descartadas porque parecen sin importancia pueden proveer información acerca de modelos desconocidos. Un Data Mining de alto rendimiento permite a los usuarios explorar toda la base de datos, sin preseleccionar un subconjunto de variables.

- **Más filas.** Muestras mayores producen menos errores de estimación y desvíos, y permite a los usuarios hacer inferencias acerca de pequeños pero importantes segmentos de población.

Las técnicas más comúnmente usadas en Data Mining son:

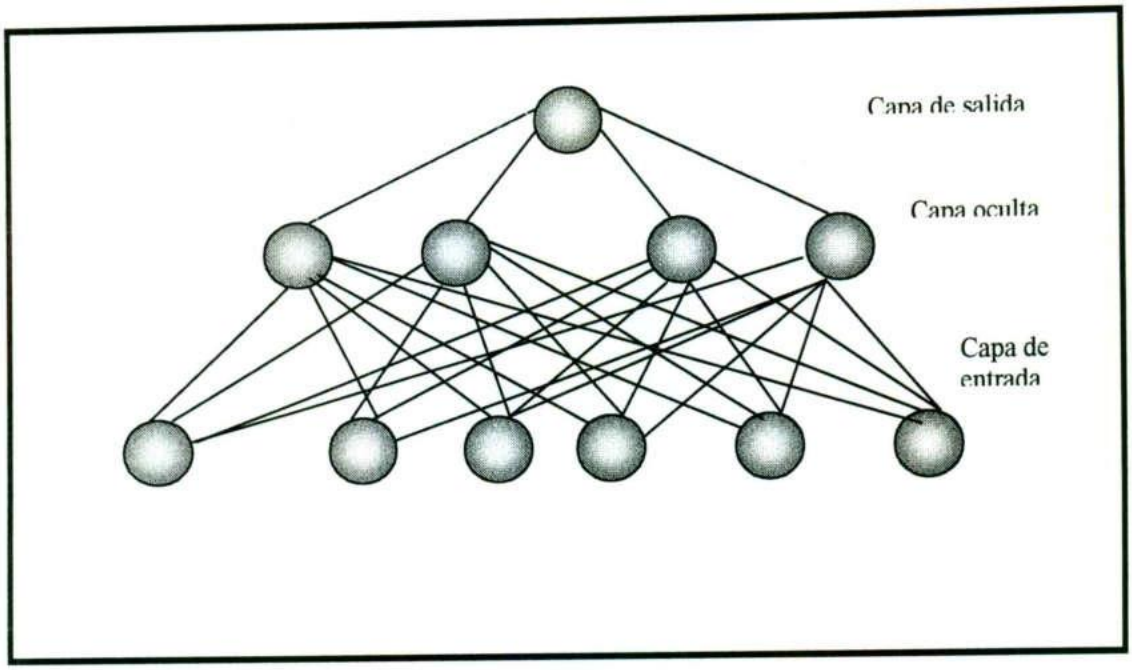
- Métodos Estadísticos:

La estadística es tradicionalmente la técnica que se ha usado para el tratamiento de grandes volúmenes de datos numé-ricos y nadie pone en duda su efectividad al poseer un amplí-simo conjunto de modelos de análisis para cubrir el tratamien-to de todo tipo de poblaciones y series de datos.

- Redes Neuronales ("Neural Networks"):

La minería basada en redes neuronales esta especialmente indicada para identificar patrones prever tendencias basadas en comportamientos previamente identificados. Una tendencia identifica un movimiento en el habito basado en el comportamiento anterior. La base de este tipo de procesamiento es lo que se aprendió trabajando en el sistema nervioso central. El conocimiento se puede aprender de una serie de datos ampliamente dispares, complejos o imprecisos. Hay 3 capas en la red: La capa inferior recibe los datos de entrada, la capa oculta (intermedia) realiza el trabajo y la exterior presenta las salidas del analista. La capa oculta o intermedia procesa los datos de entrada y entrega los resultados en la capa de patrones o tendencias a la capa exterior. La capa de entrada, la de salida y la oculta están compuestas por nodos estos nodos son otra forma de llamar a los elementos del procesamiento, que se asemejan a las neuronas del cerebro; de ahí la terminología red neuronal. Cuando esta red se entrena mediante información en la capa de entrada asume un misterioso componente de la humanidad a medida que se hace experta tomando datos de elementos de datos sin aparente relación y devolviendo resultados a la capa de salida. La figura muestra la estructura de una red neuronal y como cada nodo en cada capa esta conectado con todos los nodos de las capas adyacentes.

Figura 2.9 Estructura de una Red Neuronal



- Lógica Difusa (Fuzzy Logic"):

La Lógica Difusa surge de la necesidad de modelizar la realidad de una forma más exacta evitando precisamente el determinismo o la exactitud. La Lógica permite el tratamiento probabilístico de la catego-rización de un colectivo. La Lógica Difusa es aquella técnica que permite y trata la existencia de barreras difusas o suaves entre los distintos grupos en los que categorizamos un colectivo o entre los distintos elementos, factores o proporciones que concurren en una situación o solución.

- Algoritmos Genéticos ("Genetic Algorithms"):

Los Algoritmos Genéticos son otra técnica que debe su inspiración, de nuevo, a la Biología como las Redes Neurona-les. Estos algoritmos representan la modelización matemática de como los cromosomas en un marco evolucionista alcanzan la estructura y composición más óptima en aras de la superviven-cia. Entendiendo la evolución como un proceso de búsqueda y optimización de la adaptación de las especies que se plasma en mutaciones y cambios en los genes o cromosomas. Los Algoritmos Genéticos hacen uso de

2.10. Metodología de aplicación:

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

- Muestreo

Extracción de la población muestral sobre la que se va a aplicar el análisis. En ocasiones se trata de una muestra aleatoria, pero puede ser también un subconjunto de datos del Data Warehouse que cumplan unas condiciones determinadas. El objeto de trabajar con una muestra de la población en lugar de toda ella, es la simplificación del estudio y la disminución de la carga de proceso. La muestra más óptima será aquella que teniendo un error asumible contenga el número mínimo de observaciones.

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

- El nivel de confianza de la muestra (usualmente el 95% o el 99%).
- El tamaño máximo de la muestra (número máximo de registros), en cuyo caso el sistema deberá informar del el error cometido y la representatividad de la muestra sobre la población original.
- El error muestral que está dispuesto a cometer, en cuyo caso el sistema informará del número de observaciones que debe contener la muestra y su representatividad sobre la población original.

Para facilitar este paso s debe disponer de herramientas de extracción dinámica de información con o sin muestreo (simple o estratificado). En el caso del muestreo, dichas

herramientas deben tener la opción de, dado un nivel de confianza, fijar el tamaño de la muestra y obtener el error o bien fijar el error y obtener el tamaño mínimo de la muestra que nos proporcione este grado de error.

- Exploración

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

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

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

- Manipulación

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

- Modelización

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

- Valoración

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

2.11.Áreas De Aplicación Del Data Mining:

La utilización del Data Mining significa un enorme paso hacia delante en la automatización de tareas: el Análisis, el Con-trol y la Planificación. Las necesidades y problemáticas de carácter general a cualquier organización o empresa a las que el Data Mining puede dar solución son:

- Marketing Y Comercial:

Dentro de las áreas de Marketing y Comercial la necesidad más notoria es la adecuada concepción y gestión de la cuatri-logía Cliente-Canal-Campaña-Producto ("CCCP") cuando el clien-te está identificado y de la trilogía Canal-Campaña-Producto ("CCP") cuando es anónimo. Para ambos modelos, CCCP y CCP, el análisis y estableci-miento de previsiones, para el amplio abanico de series temporales que se pueden concebir, es también un elemento importante para las áreas de Marketing y Comercial dado que ello permite tanto el establecimiento de objetivos y precios como la detección de comportamientos estacionales o cíclicos.

- FINANZAS:

Pueden ser utilizadas para analizar aquella o aquellas composiciones de la cartera que maximicen los beneficios futuros minimizando el riesgo. Lógicamente, con anterioridad, deberá obtenerse la previ-sión de aquellas variables de las cuales dependan nuestros resultados futuros.

- Gestión De Planes Y Proyectos:

El tratamiento de la información relativa a la planifica-ción, ejecución y control de los distintos planes y proyectos puede ser de suma utilidad a efectos de analizar puntos críti-cos de decisión y el efecto de éstos.

Capitulo 3. Olap

OLAP es un tipo de tecnología que permite a los usuarios mejorar la visión que tienen de sus datos de una manera rápida, interactiva y fácil de usar.

La clave de una base de datos OLAP son sus dimensiones. Cuando se conocen las distintas dimensiones, la intersección de estas produce un lugar llamado *celda*.

Una celda es un punto unitario de datos que aparece en la intersección definida al seleccionar un valor en cada una de las dimensiones.

Dentro de OLAP existen dos enfoques:

- OLAP multidimensional o MOLAP: almacenar datos en forma multidimensional para accesarlos de manera multidimensional. Existe un límite en cuanto al número de dimensiones que se pueden manejar. Útil si se pueden descomponer los datos en grupos pequeños.
- OLAP relacional o ROLAP: almacenar los datos en un modelo relacional, para lograr un acceso rápido.

OLAP combina datos multidimensionales y herramientas de análisis.

Una de las ideas principales es poder realizar *slice* y *dice* de una manera más rápida e inteligente.

Slice (rebanada) y *dice* (cubo) hacen particiones de los datos en una base de datos multidimensional de acuerdo a valores de ciertas dimensiones.

Dentro de las tendencias actuales de bases de datos está la creación de objetos/herramientas reutilizables que sean sensibles a los datos y el desarrollo de herramientas de generación de informes y de análisis.

3.1 Antecedentes

La tecnología OLAP (On Line Analytical Processing) fue definida en 1993 por E.F. CODD. La creación ocurrió como resultado de la fuerte necesidad de analizar los datos de manera fácil y flexible y, simultaneamente, visiones múltiples del negocio en diferentes niveles de detalles. Los Bancos de Datos Multidimensionales fueron la respuesta para atender a esas necesidades analíticas. Al inicio del los años 90 empezaron a surgir los primeros prototipos de Bancos de Datos Multidimensionales. Después de algunos años de mejoramientos de la tecnología, los Bancos de Datos Multidimensionales fueron sometidos al análisis de CODD y su equipo en 1993. CODD entonces definió 12 reglas, patrones, homologó la tecnología y bautizó los Bancos de Datos Multidimensionales con el nombre de OLAP (derivado del término OLTP – On Line Transactional Processing – que fue atribuido a los Bancos de Datos Relacional al inicio de la década de 1970, cuando CODD definió los patrones para el modelo Relacional). A partir de la homologación de CODD, la tecnología empezó a ser utilizada y conocida en 1994, y los suministradores de la tecnología han creado productos capaces de almacenar más y más, además de varios otros recursos que facilitan el análisis. Entonces, entre 1995 y 96, empezó la utilización de los Bancos de Datos Multidimensionales como Data Marts y la tecnología avanzó a buen paso. Se pueden accesar y manejar los bancos OLAP a través de aplicaciones personalizadas o aún:

- vía Internet/Intranet
- vía aplicaciones predefinidas para hacer análisis diversos

La tecnología OLAP hoy día es ampliamente utilizada en la elaboración de Data Marts, con desdoblamientos para ROLAP/modelado NxN en el Relacional) y HOLAP (Hybrido OLAP que combina OLAP con ROLAP). La utilización de OLAP híbrido, el H-OPLAP, solamente es necesaria cuando se trata de bancos muy grandes, lo que ocurre mucho en la Venta al por Menor, Bancos y Compañías de Seguros

3.2 Definición y Arquitectura de OLAP

OLAP es la sigla de "On-line Analytical Processing", para diferenciar de OLTP ("On-line Transation Processing"). OLAP describe una clase de servidores de bases de datos que están diseñados para permitir acceso y análisis ad-hoc de los datos (figura 3.2).

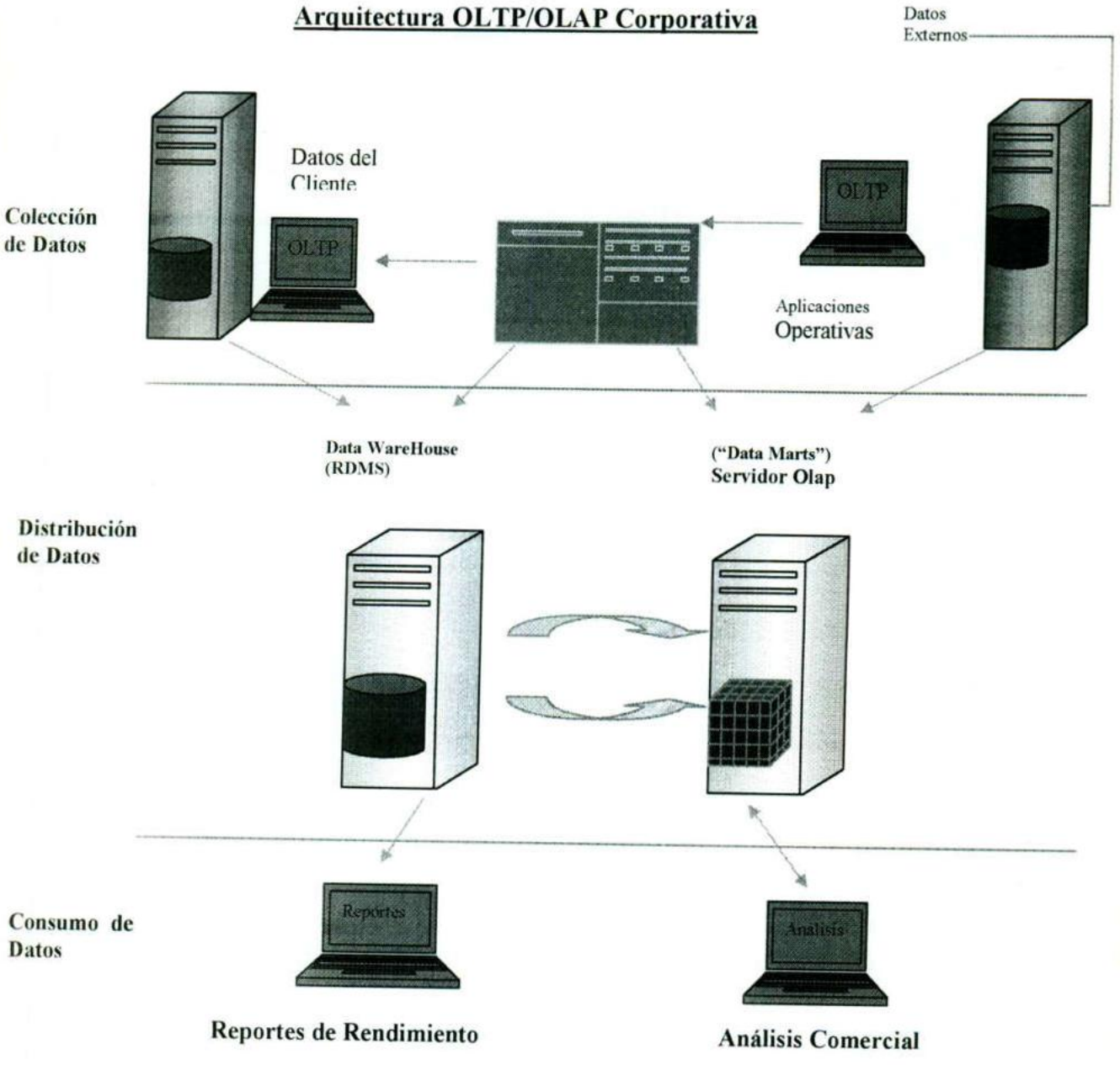


Figura 3.2. Arquitectura OLTP/OLAP Corporativa

Mientras que las transacciones residen en Bases de Datos Relacionales (BASE DE DATOS RELACIONAL) o en otro tipo de archivos, OLAP logra su máxima flexibilidad y poder utilizando la tecnología de Bases de Datos Multidimensionales (BDM). Es por esto que últimamente BDM y OLAP se los utiliza como sinónimos. Esta nueva y sofisticada tecnología provee a los usuarios con poderosas funciones para el análisis, síntesis y consolidación de datos (análisis de datos multidimensional) con un mínimo conocimiento de la estructura de los mismos.

Las aplicaciones OLTP están caracterizadas por varios usuarios creando, actualizando o recuperando registros individuales . Las aplicaciones OLAP son usadas por analistas y gerentes que frecuentemente quieren altos grados de agregación de los datos y desde distintas perspectivas y focos (totales de ventas por región, por línea de producto, ver la tendencia de lo presupuestado vs. Lo real, etc.). Las bases de datos OLAP son normalmente actualizadas en forma batch, y en general desde distintas fuentes de datos, y proveen un poderoso back-end analítico para múltiples aplicaciones de usuarios.

Mientras que las Bases de Datos Relacionales son buenas recuperando una pequeña cantidad de registros rápidamente, no son buenas para recuperar grandes volúmenes y haciendo sumarizaciones *al vuelo*. Bajos tiempos de respuesta y un abusivo uso de los recursos de los sistemas son las características comunes de las Aplicaciones de Soporte de Decisión desarrolladas con tecnología relacional.

Muchos de los problemas analíticos que se tratan de resolver con tecnología relacional, son en sí mismos de naturaleza multidimensional. Por ejemplo, los queries SQL para crear totales de productos vendidos por región, ventas de regiones por producto, etc.; involucran el barrido de todos o casi todos los registros en una base de datos de marketing y puede tomar horas de proceso. Un server OLAP puede resolver estos queries en segundos.

OLTP	OLAP
Automatizado	Sumarizado
Datos Actuales	Datos Históricos
Un registro a la vez	Muchos Registros a la vez
Orientado a lo operativo	Orientado a un tema
Datos Operativos	Datos Multidimensionales
Datos Actualizables	Datos Históricos (solo lectura)

3.3 Tipos de OLAPs

3.3.1 Sistemas MOLAP

La arquitectura MOLAP usa unas bases de datos multidimensionales para proporcionar el análisis, su principal premisa es que el OLAP está mejor implantado almacenando los datos multidimensionalmente. Por el contrario, la arquitectura ROLAP cree que las capacidades OLAP están perfectamente implantadas sobre bases de datos relacionales

Un sistema MOLAP usa una base de datos propietaria multidimensional, en la que la información se almacena multidimensionalmente para ser visializada multidimensionalmente.

El sistema MOLAP utiliza una arquitectura de dos niveles: La bases de datos multidimensionales y el motor analítico.

- La base de datos multidimensional es la encargada del manejo, acceso y obtención del dato.

El nivel de aplicación es el responsable de la ejecución de los requerimientos OLAP. El nivel de presentación se integra con el de aplicación y proporciona un interfaz a través del cual los usuarios finales visualizan los análisis OLAP. Una arquitectura cliente/servidor permite a varios usuarios acceder a la misma base de datos multidimensional.

La información procedente de los sistemas operacionales, se carga en el sistema MOLAP, mediante una serie de rutinas batch. Una vez cargado el dato elemental en la Base de Datos multidimensional (MDDB), se realizan una serie de cálculos en batch, para calcular los datos agregados, a través de las dimensiones de negocio, rellenando la estructura MDDB.
Tras rellenar esta estructura, se generan unos índices y algoritmos de tablas hash para mejorar los tiempos de accesos a las consultas.
Una vez que el proceso de compilación se ha acabado, la MDDB está lista para su uso. Los usuarios solicitan informes a través del interfase, y la lógica de aplicación de la MDDB obtiene el dato.
La arquitectura MOLAP requiere unos cálculos intensivos de compilación. Lee de datos precompilados, y tiene capacidades limitadas de crear agregaciones dinámicamente o de hallar ratios que no se hayan precalculados y almacenados previamente.

3.3.2 Sistemas ROLAP

La arquitectura ROLAP, accede a los datos almacenados en un Data Warehouse para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales.
El sistema ROLAP utiliza una arquitectura de tres niveles. La base de datos relacional maneja los requerimientos de almacenamiento de datos, y el motor ROLAP proporciona la funcionalidad analítica.

- El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención del dato.
- El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.

El motor ROLAP se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP

Después de que el modelo de datos para el Data Warehouse se ha definido, los datos se cargan desde el sistema operacional. Se ejecutan rutinas de bases de datos para agregar el dato, si así es requerido por el modelos de datos.
Se crean entonces los índices para optimizar los tiempos de acceso a las consultas.
Los usuarios finales ejecutan sus análisis multidimensionales, a través del motor ROLAP, que transforma dinámicamente sus consultas a consultas SQL. Se ejecutan estas consultas SQL en las bases de datos relacionales, y sus resultados se relacionan mediante tablas cruzadas y conjuntos multidimensionales para devolver los resultados a los usuarios.
La arquitectura ROLAP es capaz de usar datos precalculados si estos están disponibles, o de generar dinámicamente los resultados desde los datos elementales si es preciso. Esta arquitectura accede directamente a los datos del Data Warehouse, y soporta técnicas de optimización de accesos para acelerar las consultas. Estas optimizaciones son, entre otras, particionado de los datos a nivel de aplicación, soporte a la desnormalización y joins múltiples

3.3.3 ROLAP vs. MOLAP (Comparativa)

Cuando se comparan las dos arquitecturas, se pueden realizar las siguientes observaciones:

- El ROLAP delega la negociación entre tiempo de respuesta y el proceso batch al diseño del sistema. Mientras, el MOLAP, suele requerir que sus bases de datos se precompilen para conseguir un rendimiento aceptable en las consultas, incrementando, por tanto los requerimientos batch.
- Los sistemas con alta volatilidad de los datos (aquellos en los que cambian las reglas de agregación y consolidación), requieren una arquitectura que pueda realizar esta consolidación ad-hoc. Los sistemas ROLAP soportan bien esta consolidación dinámica, mientras que los MOLAP están más orientados hacia consolidaciones batch.

- Los ROLAP pueden crecer hasta un gran número de dimensiones, mientras que los MOLAP generalmente son adecuados para diez o menos dimensiones.
- Los ROLAP soportan análisis OLAP contra grandes volúmenes de datos elementales, mientras que los MOLAP se comportan razonablemente en volúmenes más reducidos (menos de 5 Gb)

Por ello, y resumiendo, el ROLAP es una arquitectura flexible y general, que crece para dar soporte a amplios requerimientos OLAP. El MOLAP es una solución particular, adecuada para soluciones departamentales con unos volúmenes de información y número de dimensiones más modestos

3.4 Datos Multidimensionales

Las Bases de Datos Relacionales están organizadas en *listas de "registros"*, cada grupo de listas constituyen una Tabla, que en la terminología relacional se denomina *matriz plana de ítem de datos*. Cada " registro" contiene información relativa al mismo la cual está organizada en "campos". Un ejemplo típico podría ser una lista de clientes con campos para el nombre, el número de cliente, su teléfono y su dirección (figura 3.4) :

NOMBRE	CODIGO	TELEFONO	DIRECCION
Julio Sánchez	1428	785-2255	Cuba 3251
Estela Gómez	1553	331-2054	Sarmiento 1325
Gustavo López	1429	701-0388	Pampa 3456
Liliana García	1523	773-5689	Quintana 123

Figura 3.4: Una tabla relacional está basada en un formato simple de filas y columnas.

En el ejemplo anterior, la tabla contiene cuatro columnas (llamadas "campos") y cuatro filas (llamadas "registros"). Si bien la tabla contiene varias columnas de información, cada pieza de información se refiere sólo a un cliente en particular. En esencia, esta tabla posee una única dimensión. Si se tratara de crear una matriz de dos dimensiones donde el nombre

del cliente se describe hacia abajo y cualquier otro campo (ej.: código) se describe a la derecha, se podrá ver rápidamente que sólo existe una correspondencia biunívoca (1 a 1):

En el siguiente caso en donde la correspondencia entre los campos de la tabla (3.5) relacional no son 1 a 1:

PRODUCTO	REGIÓN	VENTAS
Afeitadora	Este	50
Afeitadora	Oeste	60
Afeitadora	Centro	80
Plancha	Este	140
Plancha	Oeste	20
Plancha	Centro	43
Televisión	Este	56
Televisión	Oeste	70
Televisión	Centro	42
Vides	Este	110
Vides	Oeste	65
Vides	Centro	32

Esta Tabla 3.5 Relacional tiene más de un producto por región y más de una región por producto. Por lo tanto se presta a una representación multidimensional.

Una forma más clara de ver estos datos, podría ser en un formato de una matriz bidimensional (Tabla 3.6):

	ESTE	OESTE	CENTRO
Afeitadora	50	60	80
Plancha	140	20	43
Televisión	72	40	56
Video	110	65	32

Tabla 3.6 Matriz Bidimensional

Si se empieza a hablar sobre grandes bases de datos donde se deben recuperar cientos o miles de productos, el tiempo de proceso que se requiere en una Base de Datos Relacional para barrer todos los registros y acumular un total, comienza a ser inaceptable. Una Base de Datos Relacional típica puede barrer unos pocos cientos de registros por segundos. Una Base de Datos Multidimensional típica puede acumular totales por filas o por columnas en un promedio de 10.000 por segundo o más. Como podemos ver en un simple ejemplo, es fácil generar queries que puedan tomarse minutos u horas para completar su proceso usando tecnología relacional, pero sólo segundos usando tecnología OLAP multidimensional.

3.5 Consolidación: La Clave para obtener Tiempos de Respuesta Rápidos y Consistentes

El tiempo de respuesta de un query en una BDM, a pesar de su velocidad, también depende de cuantos números están involucrados en un cálculo. Lo que la mayoría de los usuarios

requieren es un rápido tiempo de respuesta con independencia del tipo de query. Por lo tanto, la única forma de obtener en forma consistente tiempos de respuesta rápidos es pre-calculando (**consolidando**) todos los totales y subtotales lógicos. Esto es de hecho lo que la mayoría de los Sistemas de Información realizan con sus tablas relacionales. La diferencia es que una Base de Datos Multidimensional puede realizar operaciones aritméticas por filas y columnas utilizando álgebra matricial y vectorial, mientras que una Base de Datos Relacional debe ser barrida (por algún criterio de selección, columna o índice). Una Base de Datos Multidemensional consolida cientos o miles de veces más rápido que una BASE DE DATOS RELACIONAL, lo que le posibilita no sólo una notable superioridad en la performance, sino que también (y como consecuencia de lo anterior) incluir situaciones en las que por el gran volumen de datos no se resolvían en el entorno relacional, o que por dicha limitación se le buscaba una solución parcial.

Supongamos que queremos obtener, elevados tiempos de respuesta independientemente del query que realicemos. Veamos que ocurre en ambos entornos :

a) En una BASE DE DATOS RELACIONAL el tiempo de respuesta es proporcional a la cantidad de registros que se lean. Si efectuamos un query que obtenga un total como "Total de Ventas para el Este" llevará cuatro veces más tiempo que obtener el total de "Planchas Vendidos en el Este" (obviamente en el primer caso tiene que leer cuatro registros y en éste último sólo uno). Y si queremos saber "Cuál es el total de Ventas para todas las Regiones" se deben acumular los doce totales parciales (de cuatro Productos por tres Regiones). Con lo cuál levará doce veces más de tiempo que la obtención de un total parcial de un producto determinado en una región dada.

Una BASE DE DATOS RELACIONAL típica puede leer alrededor de 200 registros por segundo y grabar unos 20 nuevos registros por segundo.

b) Con un servidor OLAP multidimensional, podemos realizar las mismas consolidaciones utilizando la ventaja del servidor de realizar operaciones entre filas y columnas. Mientras

que una BASE DE DATOS RELACIONAL puede acceder unos cientos de registros por segundo, una buena BDM debe tener la capacidad de ***consolidar*** de veinte mil a treinta mil celdas de datos por segundo, incluyendo la grabación de los totales en la Base. Justamente es la capacidad de realizar consolidaciones a altas velocidades donde reside principalmente el poder de esta sofisticada tecnología (las ***celdas*** también se denominan ***combinaciones***, por ser el resultado de las combinaciones de Productos y Regiones, es decir de los ***miembros*** entre distintas ***dimensiones***).

En una Base de Datos OLAP Multidimensional (BDOM) un usuario no tiene que saber nada respecto de la tecnología de la bases de datos para poder observar esta representación bidimensional de los datos con totales generales por filas y columnas ya que es la forma más natural, clara y resumida como se podría representar la información (ver matriz más abajo). Y así de resumida como se puede observar en esta hoja es igualmente menor el espacio que ocupa en disco, ya que por su naturaleza de ser una "matriz plana de ítems de datos" debe repetir las descripciones (o códigos) para representar cada una de las instancias de las combinaciones de Productos y Regiones, mientras que en una BDOM los valores resultan de la intersección de los miembros de cada dimensión. El almacenamiento físico de los datos en disco y el esquema de indexación para localizar los datos son la clave de la velocidad de las bases de datos multidimensionales (Tabla 3.5.a).

	ESTE	OESTE	CENTRO	Total
Afeitadora	50	60	80	190
Plancha	140	20	43	203
Televisión	72	40	56	168
video	110	65	32	207
Total	372	185	211	768

Tabla 3.5.a Base de datos multidimensional

Terminologías Multidimensionales

- Las ***celdas*** (***Cells***) que contienen los datos fuente originales (en cursiva) se denominan ***Inputs***.
- Los totales calculados (en cursiva negrita) son llamados ***Outputs***
- Este, Oeste y Centro son Input Members de la Dimensión Región
- El Total de la Región es el ***Output Member*** de la ***Dimensión*** Región
- Televisor, Video, Afeitadora y Plancha son ***Input Members*** de la ***Dimensión*** Producto
- El Total de Productos es el ***Output Member*** de la ***Dimensión*** Producto
- Los valores representados corresponden a una ***Variable***. Para este caso la variable es Ventas y está ***dimensionada por*** Producto y Región.
- Las ***Variables*** representan típicamente mediciones/indicadores numéricos tales como Ventas, Costos, Ingresos, Gastos, Margen, etc.
- El valor hallado en la intersección de cada región y producto ocupa una ***celda***, tal como ocurre en una planilla de cálculo.
- Las celdas son a veces denominadas indistintamente como ***combinaciones***. En nuestro ejemplo hay 20 ***combinaciones*** (celdas).

En una BDR, las columnas pueden representar tanto dimensiones como variables, por lo que, para representar las 20 combinaciones necesita información redundante.

3.6 Jerarquías simples dentro de las Dimensiones

En nuestro ejemplo existe una ***Jerarquía Simple*** en ambas dimensiones, las que podemos representar de la siguiente manera:

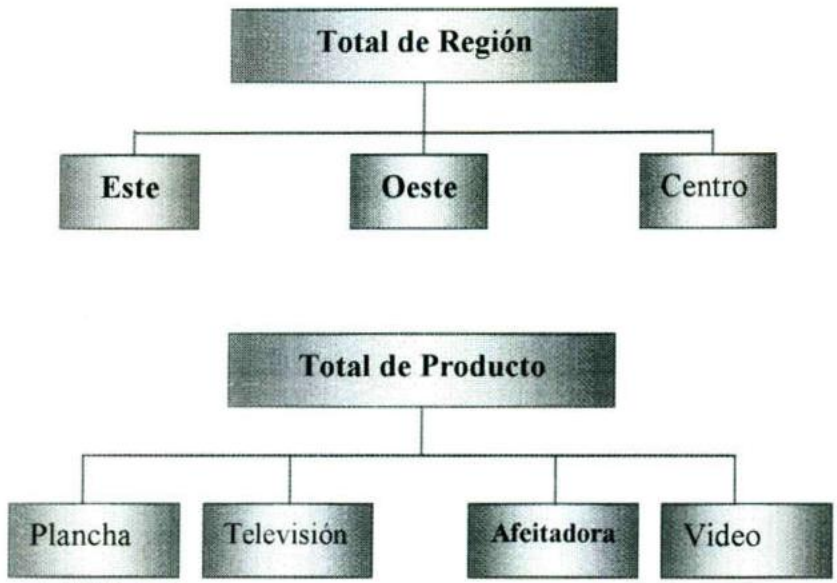


Fig. 3.6 Jerarquía Simple

Productos individuales hacen ***roll up*** en el Total de Productos y las regiones individuales hacen ***roll up*** en el Total de al Región.

La jerarquía representada es simple porque cada input hace ***roll up*** en un único total. Es posible que una dimensión como Producto pueda tener distintas formas de hacer ***roll up*** para más de un total. Las Jerarquías simples pueden contener varios niveles. Si un servidor OLAP no puede soportar múltiples niveles de jerarquía dentro de una misma dimensión, se deberán definir dimensiones separadas para las Ciudades, Provincias y Regiones.

La razón por la que se necesitan múltiples niveles de jerarquía o dimensiones adicionales es que no se pueden mezclar Ciudades, Provincias y Regiones en una misma dimensión a

menos que se las puedan tratar en una dimensión jerárquica. Si no se poseen jerarquías en una dimensión, se trataría de crear una base de datos bidimensional como la que sigue (figura 3.6.a):

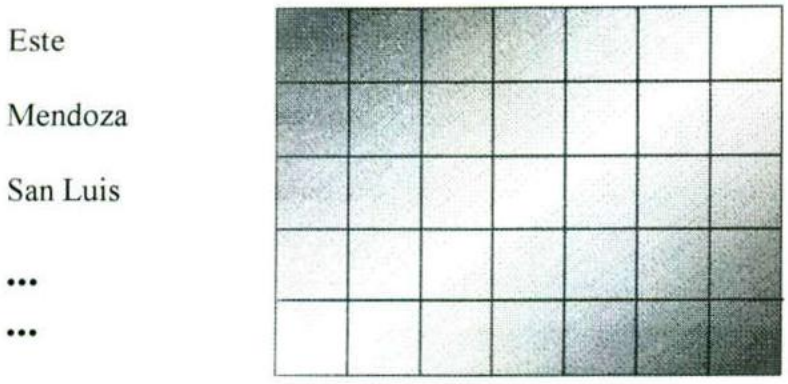
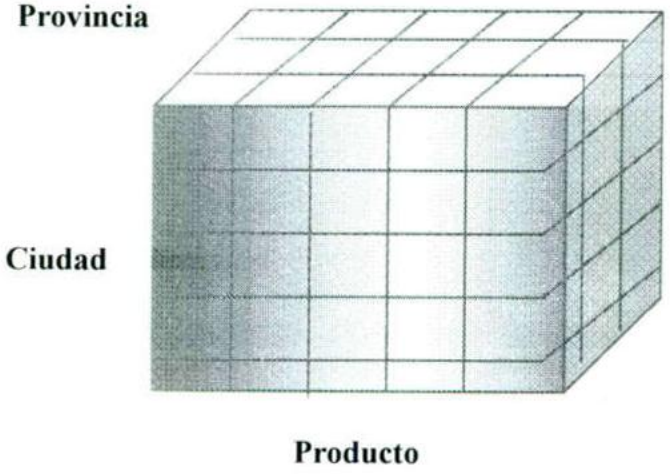


Figura 3.6.a Base de Datos Bidimensional

Obtener el total por filas funciona bien, los totales que se obtendrían serían los correctos. Pero los totales para un producto particular sería incorrecto ya que se sumarían los de cada Ciudad, con los de cada Provincia, con los de cada Región. La solución a este problema es tener dimensiones separadas para Ciudades, Provincias y Regiones (Figura 3.6.b).

Figura 3.6.b: Dimensiones separadas para Ciudades, Provincias y Regiones



Otro problema que ocurriría con la solución precedente es que muchas celdas no contendrían datos, la forma correcta de resolver éste problema es usando ***Dimensiones Jerárquicas***. Figura 3.6.c

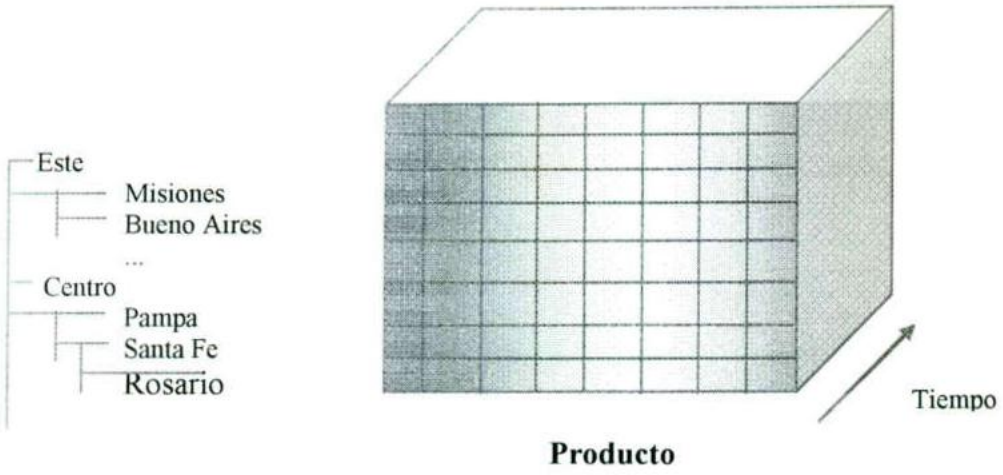


Figura 3.6.c Dimensiones Jerárquicas

El BDM tiene la inteligencia propia para acumular todas las Ciudades correspondientes por Provincia, todas las Provincias por Regiones; pero para el total general no acumulará Ciudades, Provincias y Regiones porque producirá un resultado incorrecto ya que cada nivel superior incluye los datos del inmediato inferior. El Total general surgirá de las acumulaciones de cada una de las Regiones.

Desde el punto de vista aplicativo, el uso de dimensiones jerárquicas nos provee de la funcionalidad del ***drill down*** (perforar para abajo) facilidad que nos posibilita descender a sucesivos niveles de detalle (funcionalidad necesaria para el desarrollo de cualquier EIS). Otra funcionalidad que potencia la aplicación del ***drill down***, es la posibilidad de posicionamiento dentro de una jerarquía : en la dimensión geográfica podemos posicionarnos directamente justo debajo (***just below***) de Oeste y obtener las ventas realizadas por todas las Provincias de la Región Oeste. Frecuentemente estas facilidades son implementadas de manera tal de que el usuario simplemente hace click en un ítem de la pantalla y la aplicación despliega los datos del siguiente nivel de detalle. En la sección titulada "Drilling to Relational" veremos porqué esto tiene sentido de ser usado en un

server OLAP para realizar consolidaciones, pero sin embargo se mantiene el menor nivel de datos de detalle en una BDR.

3.7 Variables

Las ***Variables*** son medidas numéricas : Ventas, Costos, Precios, Facturaciones, etc.. Algunos servers OLAP tratan a las variables como una dimensión especial, y hay varias buenas razones para esto. Pensemos ***variables*** como ***dimensionada por*** cierta dimensión en la base de datos. Por ejemplo "Ventas" puede estar ***dimensionada por*** Región, Producto y Tipo de Cliente. Por otro lado "Precio" (supongamos que así sea en nuestro modelo) debe ser idéntica para todas las Regiones y Tipo de Cliente, entonces sólo necesitamos ***dimensionarla*** por Producto. Si las ***variables*** son tan solo una dimensión normal , nos veríamos forzados a dimensionar "Precio" por todas las otras dimensiones, y por lo tanto tendríamos una cantidad innecesaria de celdas en nuestra base de datos. Tratando a la

Variables como un caso especial de ***dimensiones***, sólo se seleccionan las dimensiones relevantes para cada variable.

Este concepto se denomina ***Variables Dimensionadas Independientemente*** y es una herramienta esencial para optimizar la performance en una BDM reduciendo su tamaño al mínimo lógico, como también la complejidad de su carga. No todos los servers OLAP soportan variables dimensionadas en forma independiente.

Otra característica que no todos los servers OLAP poseen es la posibilidad de definir ***variables*** a través de operaciones matemáticas complejas entre otras variables. Este tipo de variable se denominan ***variables complejas***. En general la relación entre los miembros de una dimensión se establece sólo con adiciones. En determinados modelos (tanto en la misma estructura de una dimensión, como entre distintas variables) podemos requerir : diferencias, promedios, promedios ponderados, fórmulas financieras o cualquier otra expresión aritmética compleja. Podemos tener en nuestra BDM las Unidades Vendidas, el Precio Unitario, el Costo Unitario y podemos calcular :

Costo = (Unidades Vendidas) * (Costo Unitario)

Ventas = (Unidades Vendidas) * (Precio Unitario)

Margen = Ventas - Costo

Que un servidor OLAP posea estas capacidades en la mayoría de los casos representa una enorme ventaja en tiempo y costo, ya que de tener que hacer externamente los cálculos puede representar la elaboración y corrida de varios programas complejos más la transmisión y carga de los datos resultantes.

Las ***Variables*** son también especiales porque pueden incorporar distintas reglas de consolidación. Por ejemplo cuando las ventas de cada Región tienen ***roll up*** en el Total de Región, los valores son sumados aritméticamente. Pero el Precio no tiene un comportamiento aditivo, puede ser ponderado o computado a partir de alguna fórmula compleja. Un caso similar es cuando se convierten los datos de una ***periodicidad*** en otra (*diaria* a *semanal*), distintos tipos de variables deben ser tratadas en forma distinta. Convertir Ventas diarias en Ventas semanales se logra simplemente sumando cada uno de los días de cada semana, pero convertir Precios diarios en semanales seguramente *no* se obtendrá de la misma manera.

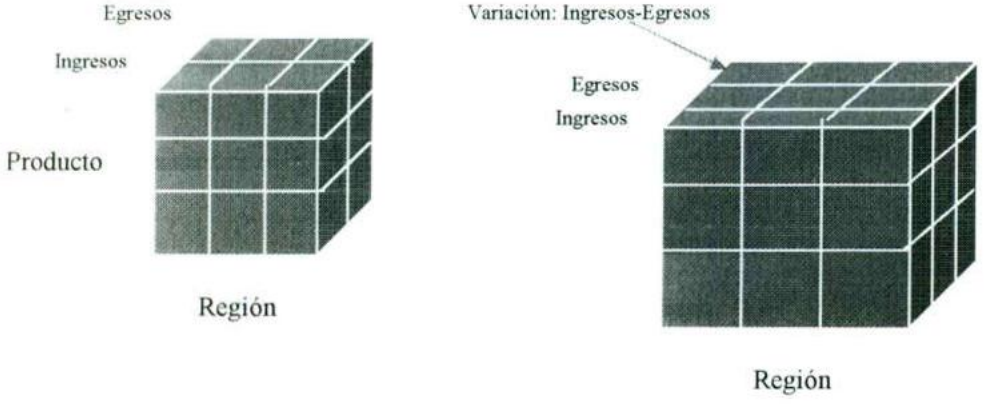
Las ***Variables*** también pueden contener información de cómo convertir una divisa en otra, si posee una leyenda descriptiva larga, unidad en que se define, etc.. Todos estos ***Atributos de las Variables*** son normalmente almacenados en un diccionario de datos.

Otro concepto que quisiéramos introducir en este punto es el de ***variable virtual***. Esta tipo de variable es requerida desde el punto de vista del usuario y es calculada *al vuelo* en tiempo de ejecución. Por ejemplo una base de datos puede contener variables para Ingresos y Egresos, y se necesita crear una variable "Margen Bruto" que sea la resta de Egresos sobre Ingresos y almacenarla en la base de datos. Alternativamente se puede definir Margen Bruto como una ***variable virtual***, lo que significa que es calculada al vuelo usando la fórmula: Margen Bruto = Ingresos - Egresos. No todos los servidores OLAP soportan este tipo de variables. Una ***variable virtual*** no ocupa espacio en la base, por lo que es extremadamente útil para reducir el tamaño de la base de datos y los tiempos de

consolidación de todas las *combinaciones* (desde ya con el precio de un pequeño incremento de overhead en el tiempo de ejecución de query que la involucre).

3.8 Vector Aritmético

Datos que son inherentemente organizados en vectores pueden ser manipulados más rápido que los mismo datos en una tabla relacional. Por ejemplo, podemos fácilmente restar el *plano* para Actual del *plano* para Presupuestado para crear el *plano* para la Variación :



En un servidor de datos OLAP Multidimensional, este vector aritmético puede ser expresado en una operación. En el caso de una representación relacional:

- ✓ Cada registro de la base debe ser accedido
- ✓ El actual debe ser restado del presupuestado
- ✓ Y cada diferencia debe ser grabada como un nuevo registro

Obviamente resulta claro que en el entorno relacional el proceso es bastante más largo. El vector aritmético permite consistentemente rápidas computaciones de ***variables virtuales***. Con esta facilidad la Variación se podría definir como una variable virtual.

3.9 Bases de Datos N-Dimensionales

Una base de datos bidimensional es fácil de entender y visualizar. Extenderemos el concepto a tres o más dimensiones. Con la representación multidimensional pasamos de una matriz de dos dimensiones a una de tres (figura 3.9) :

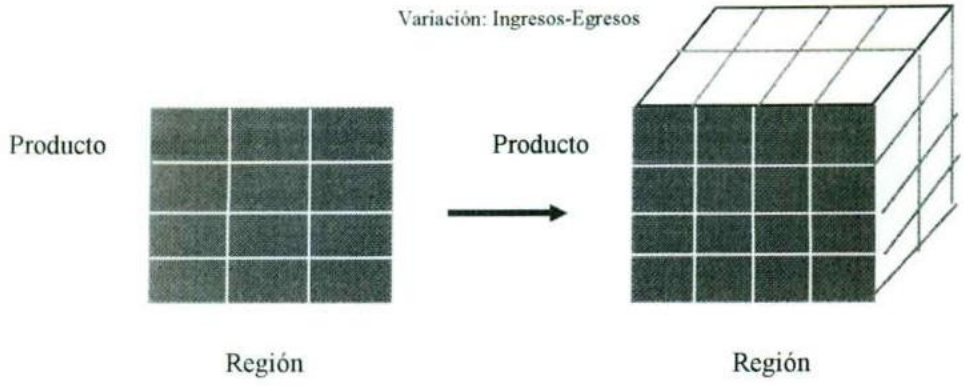


Figura 3.9: Base de datos de N Dimensiones

Esta matriz tiene 24 ***celdas*** (4 productos X 3 regiones X 2 tipos de valores) corresponde a los 24 registros en la representación relacional.

En este caso particular de tres dimensiones, se puede ver claramente que de acuerdo a la combinación de dimensiones que se quieran analizar, el "cubo" se puede "rotar" para presentar la vista de datos requerida. Justamente esta es una de las habilidades principales en la navegación y manipulación de estructuras de datos en una BDM : *el cubo se puede rotar, girar, cortar porciones o extraer parte de su estructura.*

3.10 Limitación Práctica en el Tamaño de una Base de Datos

Existe un concepto erróneo común en el mercado que el tamaño de una base de datos OLAP está limitada primariamente por el máximo número de dimensiones que soporta. La real limitación es casi siempre el número de ***celdas***, no la cantidad de dimensiones. Además, no todas las dimensiones son creadas de la misma manera. Algunos productos sólo soportan simples jerarquías dentro de las dimensiones, mientras que otros soportan múltiples jerarquías complejas. Entraremos en más detalle en la sección "Todas las

Dimensiones no son creadas iguales". Basta decir que una base de datos de ocho dimensiones en un producto OLAP, se puede reducir a tres o cuatro con otro.

En general, a medida que el número de dimensiones se incrementa, el número de ***celdas*** aumenta exponencialmente. Por ejemplo , una base de datos bidimensional con 100 Productos y 100 Regiones generará 10.000 celdas (combinaciones). Si agregamos una tercera dimensión con 52 semanas llegamos a 520.000 celdas (100 X 100 X 52). Agregándole una cuarta dimensión para Actual, Presupuestado, Variación y Proyectado; nos da 2.080.000 celdas (100 X 100 X 52 X 4). Y si agregamos una quinta dimensión para almacenar 10 "Tipos de Clientes" nos da un total de 20.800.000 celdas. Y así sucesivamente. Con base de base de datos de 16 dimensiones con sólo 5 miembros en cada dimensión alcanzamos cerca de los 152 billones (152.587.890.625) celdas !

La mayoría de los servers OLAP que se ofrecen en un determinado valor máximo de celdas sin tener en cuenta si el máximo de dimensiones que ofrecen es coherente con dicho límite. Por ejemplo, un proveedor OLAP declara que su tecnología soporta 32 dimensiones, y que tiene un límite de alrededor de dos billones de celdas. Con sólo dos miembros en cada dimensión, una base de datos de 32 dimensiones deberá soportar 2^{32} (o sea 4,3 billones) de celdas. Es obvio que una producto de estas características tiene una seria limitación en la cantidad de dimensiones y no sólo en la cantidad de celdas. En la práctica ocurre que la mayoría de las dimensiones (ej. Productos y Regiones) poseen más de dos miembros.

Un server OLAP puede manejar esta consolidación sorprendentemente rápido con poco esfuerzo. El factor clave utilizado por algunos servers OLAP, es el uso de un tipo de datos Time-Series de alta performance en lugar de utilizar una dimensión tiempo. Con este tipo de datos Time-Series, conversiones de diario a semanal, mensual y/o cuatrimestral pueden hacerse al-vuelo en tiempo de ejecución. Debido a que este tipo de conversiones sobre los periodos de tiempo pueden hacerse rápidamente, no hay necesidad de pre-consolidar estos periodos.

Tipo de Datos Times-Series. El tiempo es probablemente la dimensión más común en Bases de Datos Multidimensionales. De hecho todos quieren ver las tendencias -- de ventas, de finanzas, de mercado y demás. Los usuarios quieren ver las tendencias en todos los

aspectos de sus negocios, comparar resultados actuales con los de años anteriores, convertir el período actual en período *year-to-date* (lo que va del año). Una serie de números representando una variable en particular (como ventas) a través del tiempo se llama una Time-Series. Por ejemplo, los números de ventas de 52 semanas es una Time-Series, como lo son los números de las ganancias correspondientes a 12 meses, 5 días de balance de caja y demás. En una celda de una planilla electrónica de cálculo se puede almacenar un único número. Supongamos que se pueda almacenar 10 años de historia diaria en cada celda. Esta es la idea del tipo de datos Time-Series. La incorporación del tipo de datos Time-Series le permite almacenar una serie entera de números (representando, por ejemplo, información totalizada para periodicidad diaria, semanal o mensual) en cada celda. Si el server OLAP tiene un tipo de datos Time-Series, se puede almacenar toda su información histórica en una celda en lugar de tener que especificar un dimensión separada para el tiempo.

A diferencia de otras dimensiones, el tiempo tiene algunas cualidades y reglas especiales. Antes que nada, el tiempo siempre tiene una particular periodicidad, esto es el intervalo de tiempo entre los números de la serie. Periodicidades comunes son: diaria, semanal, mensual, cuatrimestral, etc.. Secundariamente, los datos de tipo Time-Series deben incluir reglas de conversión a otras periodicidades. Estos son los llamados Atributos de las Time-Series.

Antes de entrar en detalle sobre el tipo de datos Time-Series observemos cómo tendríamos que tratar con datos de una Time-Series si no existiera un especial tipo de datos Time-Series. Al no poseer un tipo de datos Time-Series, se debe definir explícitamente al "Tiempo" como una de sus dimensiones en la Base de Datos. Como se querrá trabajar aritméticamente su fila y columna en forma correcta, se tendrá que tomar una periodicidad para toda la Base de Datos (como "mensual") y expresar todo en esta periodicidad. Los miembros de la dimensión "meses" deberían ser nombrados explícitamente -- como Enero, Febrero, Marzo, ..., Diciembre.

Capitulo 4. Datamarts

4.1 Definición

Se define como una vista de un Dara warehouse orientada a un aspecto de negocio. Contiene mucha menos cantidad de datos que el data warehouse y es el objeto del procesamiento analítico por parte del usuario final. En un entorno de soluciones de data warehouse corporativo, habrá datamarts que se ajusten a distintos aspectos de la organización, tales como finanzas, producción y facturación. Los datamarts permiten a estas partes de la organización tomar decisiones más informadas. Los datamarts son, por lo general, más baratos y mucho más pequeños que todo un datawarehouse a nivel corporativo. Las organizaciones inteligentes implementan una serie de datamarts como soluciones a corto plazo; el ciclo de vida de muchos de estos marts es de dos a tres años.

A medida que se desarrolla el data warehouse de la organización a nivel corporativo, la comunidad de usuarios tiene siempre acceso a sus propios datamarts, que van evolucionando en el tiempo. Los sistemas operacionales clásicos se concentran en los requisitos de alto nivel que atienden a las necesidades de todos los usuarios. Cuando el sistema global está preparado, se satisfacen las necesidades de bajo nivel de los distintos segmentos de la comunidad de usuarios. Este tipo de implementación se llama *top-down* (desde arriba hasta abajo). Un ejemplo típico de desarrollo de *top-down* es una aplicación de gestión financiera. Los módulos específicos se planean para que traten las transacciones a medida que se van moviendo por el ciclo de vida, desde los presupuestos a los gastos. Los datamarts se desarrollan en el sentido contrario, llamado Botton – up(desde abajo hacia arriba). Las necesidades específicas de pequeñas partes de la organización se atienden usando el enfoque bottom – up. Como el datamart se centra en una fracción de todo el data warehouse corporativo, los analistas, desarrolladores y usarios pueden pensar que esta es definitivamente la manera de hacer las cosas basándose en el tamaño y solamente en el tamaño. Esta aproximación puede ser fácil de vender a la dirección a corto plazo, pero, según nos ha enseñado la experiencia, no es la forma correcta de proceder. Indudablemente, las adquisiciones actuales y futuras de hardware deben seguir los planes de un cambio de

datamarts a datawarehouse, pero las necesidades de negocio de la dirección de la organización pueden anteponer otros factores.

Los datamarts son bases de datos multidimensionales orientadas a una materia especifica, con un ciclo de vida esperado de tres años. Los fabricantes de herramientas que construyen, y después gestionan, datamarts deben dar una solución rápida a un coste eficiente, que puedan ser usadas en un proyecto de cualquier tamaño. El resultado de este esfuerzo constructivo debe estar disponible para la comunidad de usuarios en un tiempo adecuado. Este tiempo adecuado significa que debe estar disponible de dos a cuatro meses. Los posibles usuarios de los datamarts corporativos pueden perder fácilmente el interés cuando la espera sea cada vez mayor. El software de datamart debe apoyar a otros sistemas operacionales y otros repositorios para soporte a la toma de decisión durante la fase de construcción de estos.

El modelo de datamart vendrá dado por la forma en que el usuario necesita ver la información. Más que prestar mucha atención al aspecto fisico de los datos, el modelo de datamart refleja qué quieren ver los usuarios y cómo quieren que se les presente. Los constructores del datamart han de entrevistar a los usuarios y, con estos datos y el conocimiento que poseen de la tecnología, construir un modelo que se ajuste perfectamente a los requisitos del usuario. Este es un proceso iterativo, que a veces dura toda la vida del datamart. Un proceso iterativo es aquel que se repite, pero esto no significa que la iteración se repite porque la ejecución anterior era errónea. Un proceso iterativo simplemente se ejecuta una y otra vez.

4.2 Datamarts autónomos

Algunas veces, en una compañía muy descentralizada, algunas partes de la comunidad de negocio han fundado, desarrollado e implantado una solución datamart sin prácticamente involucrar al personal dedicado a la gestión de soluciones informáticas. A éstos se les conoce como *datamarts autónomos*. Hay demasiadas discrepancias entre estos datamars en la forma de estructurar y codificar los datos. Es virtualmente imposible mezclar los

contenidos de estos datamarts cuando se buscan formas de compartir datos entre las diferentes secciones de la organización.

4.3 Datamarts subconjunto

Muchos datamarts son un subconjunto de la información de un data warehouse mayor, diseñado, construido, mantenido y distribuido de forma centralizada a los grupos de usuarios para la ayuda a la toma de decisión a lo largo de toda la compañía. En los años noventa, con el impulso en aras de la gestión centralizada para reducir la redundancia administrativa que es inherente a un modelos descentralizado, las compañías están construyendo datamarts que se alimentan de los datos que producen una combinación de sistemas operacionales y de sistemas de ayuda a la toma de decisión; esto es sólo posible en compañías que se comprometan a implantar los sistemas desde un punto central bajo los auspicios y supervisión de un grupo a nivel corporativo de desarrolla e implantación.

4.4 Base de datos multidimensional

Los datamarts son multidimensionales. Una base de datos multidimensional permite el uso de queries para el soporte a la decisión usando un rango de combinaciones de criterios. Muchos queries que utilicen los datos del datamarts son ad hoc. Un query ad hoc es aquella cuyos criterios de selección son escogidos por el usuario cuando el query es formulado. Muchos queries que sehacen contra los sistemas operacionales están latadas o preprogramadas. Un query enlatado es aquel preparado para ejecutarse a petición del usuario y que devuelve los datos en un formato predeterminado; siempre acceden a las mismas tablas. El datamart debe permitir queries de n-formas con una red de índices construida de tal manera que el operador pueda usar una herramienta OLAP y :

- Realizar un informe sobre el contenido de una tabla de datamart usando cualquier columna como criterio de selección.
- Reunir los datos de dos o más tablas del datamart, enlazando los objetos usando relaciones de claves ajenas (foreign key); un query puede unir (join) las tablas X, Y, y Z , al minuto siguiente, el

mismo usuario puede unir las tablas A, B, y C de una manera que nunca más vuelva a usarse.

El datamart sirve para establecer el procesamiento analítico interactivo (on line analytical processing) OLAP en un sistema de ayuda a la toma de decisión. Como resultado, el arquitecto del sistema de ayuda a la toma de decisión escucha a los usuarios y recopila información en temas como:

- Qué datos quieren extraer del datamart
- Cómo quieren que se les presente la información
- Qué nivel de resumen o agregación quieren en os datos
- Sobre qué tablas se realizarán joins más frecuentes en el procesamiento de queries OLAP

Armando con las respuestas a estas preguntas, el analista empezará el proceos de diseño del datamart y llegará a un diseño fisico que se ajuste a las necesidades del producto más valioso de todo el proceso, el usuario.

4.4.1 Las dimensiones afectan al diseño

Cada dimensión en un datamart requiere el almacenamiento de una nueva columna. Pongamos que hay una tabla en un datamart dedicado a ventas que tiene las siguientes columnas:

Persona_venta cant_venta valor_venta región

Y se desea añadir una dimensión temporal a esta tabla. Se debe añadir una columna que contenga esta información temporal. El formato de esta columna estará basado en hasta que punto quiere el usuario que se agrupen los datos. Si el usuario quiere que se agrupen por meses del año, así como por el cuatrimestre fiscal, habrá que hacer lo siguiente:

1.Añadir una columna a la tabla y darle formato de forma que sea fácil la extracción del mes y del cuatrimestre fiscal de la fila. Mas que posiblemente, el formato elegido será

YYYMMDD, de tal forma que los dos primeros caracteres, esto es, el número de mes, pueden ser fácilmente convertidos a un cuatrimestre fiscal. La columna se llamará fecha_tran. La lógica necesaria para extraer la estación se puede expresar fácilmente en SQL con la sentencia:

```
select decode (substr(fecha_tran, 5,2),
'1','1','2','1','3','1','4','2','5','2','6','2','7','3','8','3','9','3','10','4','11','4','12','4')...
```

2. Añadir dos columnas a la tabla, de forma que la fecha de transacción se guarde en una y, durante el proceso de transformación, el número de mes se convierta en el identificador del cuatrimestre fiscal. Esta tabla contendrá dos columnas FECHA_TRAN y CUA_FISCAL. Si se adopta este plan, dos filas de la tabla de resumen de transacciones del datamart tendran estos dos valores de columna, tal y como aparecen en la tabla 4.1

Tabla 4.1. Valores de las columnas, fecha de transacción y cuatrimestre iscal.

Fecha_Tran(YYYMMDD)	Cua_Fiscal
19971203	4
19980103	1
19980218	1

4.4.2 Los requisitos del drill-down afectan al diseño

A medida que se introducen datos en el datamart, sufren habitualmente algún tipo de agregación. La siguiente frase <<la compañía engreso 26 millones de dólares el año pasado>>, no significa mucho para el usuario de DSS. Cuando se presenta a los usuarios este tipo de información, inevitablemente querrán obtener la información que hay detrás de este número multimillonario. Este proceso de excavar más profundamente en los datos es llamado drill-dow. El equipo de implementación del datamart tiene la responsabilidad de decidir:

1. ¿qué nivel de drill-down quiere el usuario del datamart tener disponible?

2. Basándose en la respuesta a la primera pregunta, ¿qué información se debe guardar en el datamart y que información se calculará en tiempo de ejecución?

La tabla 4.2 muestra como de complejo puede resultar un drill-down en un datamart de ventas.

Las decisiones de diseño afectan al rendimiento de los quiries contra el datamart. Usando la tabla 4.3 como ejemplo, el datamart puede guardar datos agregados hasta el nivel de ciudad y calcular el nivel de jurisdicción a medida que se procesa el query. Un punto a tener en cuanta es dónde acaba la capacidad de drill-down. Cuando se contemple acabar con este proceso, el analista del DSS deberá hablar con el usuario del datamart y determinar hasta que punto la perdida de detalle no afecta a la capacidad de tomar decisiones basadas en información. En el diseño que aparece en la tabla 4.3 una vez que el usuario ha hecho drill-down hasta nivel de ciudad, ya no hay donde ir.

Tabla 4.2 Ejemplo de drill-down

País	Area	Jurisdicción	Ciudad
USA	West	California	San Jose
			Los Gatos
			Cupertino
		Washington	Seattle
			Tacoma
	East	Ohio	Dublin
			Toledo
			Cincinnati
Canadá	East	New Brunswick	Chatham
			Bathurst

	New foundland	Conerbrook
		Porte aux Basques
West	Alberta	Calgary
	Manitoba	Shilo
		Brandon

Queries contra datamarts

Los criterios de los queries pueden ser agrupados en cuatro categorías principales:

1. Operaciones de inclusión, donde los datos se seleccionan según satisfagan una o más comparaciones. Esto incluye los tres mecanimos que aparecen en negrita en la siguiente lista:

```
--igualdad
select amt
        from dw_fin
  where acc_num = '555199';
--igualdad dentro de un conjunto
select sum(bal)
        from dw_fin
  where acc_num in ('555199','555210');
--Rango acotado
select sum (atm)
        from sw_fin
  where acc_num between '231999' and '556112'
```

2. Operaciones de exclusión, en las que se eliminan datos, basándose en que no satisfagan una o más comparaciones. Esta operación, habitualmente, usa algún tipo de construcción negativa. Son las que aparecen en negrita en el siguiente listado:

```
-- no igual
select sum (atm)
     from dw_fin
     where  acc_num <> '555199';
--no está en un conjunto
select sum(bal)
     from dw_fin
          where acc_num not in ('555199', '555210');
          ---no esta en un rango acotado
          select sum(atm)
              from sw_--fin
          where    acc_num    not    between    '231999'
and'556112';
```

3. Una combinación de operaciones de inclusión y exclusión, donde algunos datos son eliminados y cualificados y forman parte del conjunto resultado. El siguiente listado ilustra este tipo de query.

```
Select sum(bal)
     From dw_fin
   Where acc_num <> '555342'
     And acc_num between '555009' and '555821';
```

4. Funciones aritméticas, tales como MIN, MAX, o AVG, donde se aplica una función a los campos numéricos del query, junto con cualquier combinación de operaciones de inclusión o exclusión. Habitualmente, los campos de tipo carácter se dejan tal y

como están, o se usan como origen de las operaciones de agrupamiento. Las siguientes sentencias SQL son ejemplos de este tipo de selección de query:

```
Select sum(bal)
            From dw_fin
        Where substr(acc_num,1,3) <> '555'
       And ro_char(trans_date,'mon')='feb';
        Select count(*)
             From dw_fin
         Where acg(tran_amt) < 1000
             And    to_char    (trans_date,'Mon')    in
('Feb','Mar');
```

El datamart está fuertemente indexado en previsión de que los usuarios realizan sus queries sobre grandes cantidades de datos usando un amplio rango de criterios de selección.

4.5 Agregación

Las agregaciones desempeñan un papel vital en el datamart de soporte a la decisión. Agregación es un proceso por el cual muchos registros detallados se combinan dentro de un solo registro del datamart. Los datos numéricos que se almacenan en cada registro agregado representan lasuma de los correspondientes campos de todos los registros operacionales que resume. Se puede pensar en las columnas de un datamart como en rebanadas de datos operacionales, con una dimensión añadida que implica el tiempo y un nivel de resumen. Los datos ya no son atómicos, como lo eran cuando cada fila contenía la información correspondiente a una, y solo una, transacción. A media que los datos se introducen en el datamart, el analista decide con que nivel de granularidad se empieza. La función SQL SUM realiza la operación de agregación, junto con la sentencia GROUP BY, a medida que se introducen los datos en el datamart. Se identifican nuevos tipos de agregación a medida que los usuarios empiezan a trabajar con los datamarts.

4.6 Data Wherehouse frente a datamart

El datamart es el objetivo de muchas de las etapas de un data warehouse; es donde el usuario del DSS ocupará la mayor parte de su jornada. La gran cantidad de dato de un datawarehouse corporativo puede hacer fracasar algunas herramientas de ayuda a la decisión. Cuando más pequeño sea y menos datos contenga el datamart, está más preparado para optimizar el acceso del analista a la información. Hay varios fabricantes que venden software de datamart que está diseñado para ayudar a realizar el viaje que hacen los datos desde los sistemas operacionales al entorno de data warehouse (Fig.4.3.).

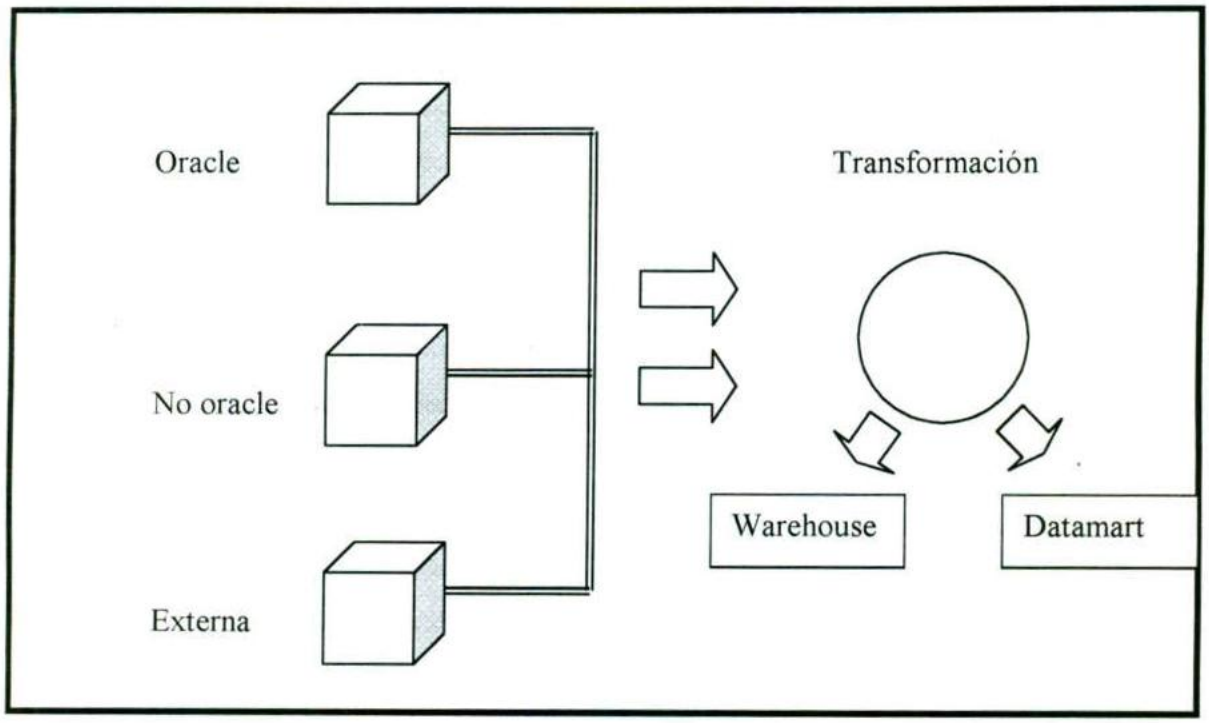


Figura 4.3 Fabricantes de software, sistemas operacionales y entorno data warehouse.

Algunos proveedores de datamart leen los diccionarios de datos de los sistemas operacionales, analizan las relaciones, trabajan a través de las complejidades inherentes y proveen al analista de una serie de elecciones acerca de cómo se pueden cargar los datos dentro del datamart Proporcionar asistencia relacionada con esta actividad costosa en

tiempo es una característica fundamental de un proveedor de soluciones maduro de datamart.

Una ventaja importante al invertir tiempo y dinero en el desarrollo de datamart es que cuestan una pequeña fracción de lo que cuesta un data warehouse a nivel corporativo. La experiencia demuestra que el desarrollo de un warehouse cuesta de tres a cinco millones de dólares y tarda hasta tres años en desarrollarse. Estos números están al alcance de grandes corporaciones, pero los datamarts proveen de soluciones a pequeñas compañías, así como a segmentos localizados de toma de decisión de la compañía. El data warehouse puede ser un objetivo a largo plazo para algunos equipos pequeños, pero, como ya se ha mencionado anteriormente este capitulo, los datamarts producen rápidamente salidas que implementan una aproximación bottom-up. Proveen al consumidor de la información necesaria para tomar decisiones de negocio inteligentes en períodos de tiempo más cortos.

Los datamarts son una nueva especie de soluciones de data ware house. Al ser bottom – up, son impulsados por consumidores en cooperación con un grupo de expertos en tecnologías de la información. Muchos data warehouse corporativos han fracasado, así como muchos de los sistemas operacionales precedentes.

La industria del datawarehuse está explotando, especialmente en el áea de gestión de datamart. La siguiente generación de herramientas es tan sensible al datawarehouse que automatiza su colección de metadatos a medida que el proyecto progresa. Los metadatos son los datos acerca de los datos; describen qué tipo de datos hay almacenados en el datawarehouse y le guía a través de la red de relaciones que tienen unos datos con otros.

4.6.1 Integridad Referencial

La integridad referencial (IR) desempeña un gran papel dentro del datamart. La IR es un mecanismo usado en las bases de datos relaciónales para reforzar las relaciones entra los datos de diferentes tablas. La IR refuerza, además, las reglas del negocio, tales como:

- No se introducirá la dirección dentro de la tabla de direcciones de cliente hasta que la calle donde está el edificio haya sido introducida

- No se introducirá ningún número de pieza en el inventario hasta que se haya introducido el fabricante de dicha pieza.
- Ningún pequeño negocio debe usar el método rápido para archivar los impuestos de bienes y servicios hasta que haya recibido notificación de la Agencia Tributaria Canadiense

Los constructores de datamart prestan especial atención a este esquema porque proporciona un acceso óptimo a los datos en muchos datamarts o repositorios completos data warehouse. Para establecer el esquema, es necesario definir previamente las claves primarias (primary) y ajenas (foreign) en el origen de datos Oracle antes de poder diseñar e implantar el esquema. Hablemos del establecimiento de las claves primarias y ajenas en el datamart.

4.6.2 Claves Primarias

Una clave primaria es un campo o combinación de campos de una tabla que identifica de forma única cada fila de esa tabla. La sentencia SQL para establecer una clave primaria puede formar parte de la sentencia de la creación de la tabla, tal como se muestra a continuación:

```
Create table dw_finsum
(dwf_id number constraint dwf_pk primary key, ...

o, despues de haber creado la tabla, usando la construcción
alter table:

alter table dw_finsum add constraint dwf_pk primary key
(dwf_id);
```

Oracle crea un índice sobre una o más columnas definidas como parte de la clave primaria: se pueden usar algunas especificaciones adicionales cuando se crea una restricción de clave primaria.

4.6.3 Claves foráneas

Cuando una columna se ha definido como clave primaria en una tabla y se incluye además en otra tabla diferente, esa columna es llamada clave foránea. Las columnas que se definen como clave foránea sólo pueden referenciar a columnas que ya se han definido en otras tablas como parte de la clave primaria. La sintaxis para definir una clave foránea es:

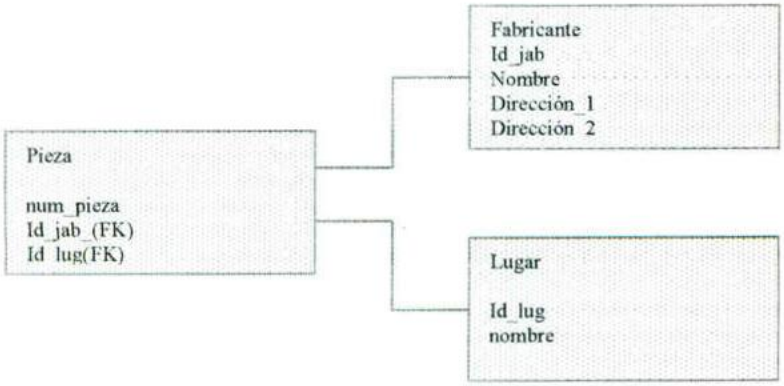
```
Alter table dw_finsum add constraint dw_acc_fk
    Foreign key (acc) references account;
```

Dicho de otra forma, una clave foránea es una clave primaria de una tabla almacenada en otra tabla. En un modelo de base de datos relacional, las claves foráneas definen las relaciones entre las columnas de tablas diferentes.

Una vez definidas las claves primarias y las claves ajenas, el arquitecto del DSS puede empezar a trazar las tablas en el datamart, apoyando la existencia de claves primarias y ajenas como la forma en que los datos están relacionados unos con otros.

4.7 Las mejores herramientas de datamart

Con el asalto de los proveedores de soluciones de datamarts, los consumidores examinan las siguientes cuestiones antes de decir que conjunto de herramientas es apropiado para ellos:



1.- Los consumidores quieren un conjunto de herramientas a bajo coste que les puedan proporcionar soluciones antes de los seis meses, con un precio de menos de 250.000 dólares

2.- Los consumidores quieren obtener un conjunto de herramientas con una única compra. Los consumidores insisten en que cualquier que sea el fabricante con el que hagan los negocios, este debe de agrupar las funciones del núcleo del datamart en un motor único.

3.- Los consumidores quieren una herramienta que les facilite la introducción rápida de los metadatos ya existentes en los sistemas operacionales.

4.- Los consumidores quieren un grupo de herramientas que puedan aprovecharse de las ventajas del procesamiento paralelo que les proporcionan los computadores con varias CP,

que sirven como huésped de varias implementaciones de grandes data warehouse y datamarts.

4.8 Funciones del núcleo de una herramienta datamart

Los consumidores están buscando una serie de módulos principales cuando evalúan a los fabricantes cuyos productos ofrecen soluciones datamart al mercado. Estos son algunos módulos del núcleo:

4.8.1 Extracción

Los proveedores de software de datamart deben desempeñar un papel para facilitar el proceso de extracción. La extracción esta muy relacionada con la transformación porque la representación de los datos puede ser muy diferente entre sistemas no relacionados que sirven como origen de los datos al datamart. La forma mas fácil de introducir los datos en un datamart es usando la sentencia SQL estándar INSERT. El origen de los datos del datamart debe ser un sistema gestor de base de datos que se pueda leer usando los comandos SQL estándar.

4.8.2 Transformación.

La transformación garantiza que, a medida que los datos se introducen en el datamart, son conformes a un sistema estándar de códigos y abreviaturas. Las decisiones que se toman son sobre como indicar los campos código y que descripciones utilizar si los códigos se extraen de tablas en diferentes sistemas que usan el mismo código, pero distintas formas de representación. La mayoría de las transformaciones pueden ser llevadas a cabo directamente por funciones y operaciones SQL, pero algunos fabricantes de datamarts han decidido complementar la funcionalidad SQL con sus propios mecanismos propietario.

4.8.3 Carga

Introducir los datos dentro del repositorio Oracle puede resultar un reto. A medida que se incrementa el tamaño del datamart, la complejidad del problema se magnifica. Los que adoptan las herramientas datamarts buscan ayuda con este proceso de carga. Quieren tener

la capacidad de planificar las cargas y especificar que datos deben ser reemplazados y donde estar dentro del datamart.

Cuando se cargan datos en un datamart Oracle, se pueden llevar directamente a las tablas objetivo o ser puestos en tablas intermedias; allí, los datos pueden ser procesados y llevados luego a las tablas objetivo reales. El uso de un conjunto de tablas intermedias puede ser la forma mas eficiente de cargar información en el datamart. La estructura de las tablas intermedias es, a menudo, radicalmente diferente a la de las tablas finales.

Una herramienta de datamart también debe ser capaz de apoyar el procesamiento paralelo de varios procesadores de altas prestaciones, dado que tantas cargas involucran una cantidad enorme de datos. Una maquina de altas prestaciones es aquella que une una gran capacidad de procesamiento junto con operaciones rápidas de entrada/salida en paralelo. Las maquinas con mas de un procesador (llamadas MPP o SMP, dependiendo de cómo estén configuradas) son candidatas ideales para albergar los datos de un data warehouse.

Aportación Personal

Hoy en día las empresas cuentan en su mayoría con la automatización de sus procesos, manejando gran cantidad de datos en forma centralizada y manteniendo sus sistemas en línea. En esta información descansa el know-how de la empresa, constituyendo un recurso corporativo primario y parte importante de su patrimonio.

El nivel competitivo alcanzado en las empresas les ha exigido desarrollar nuevas estrategias de gestión. En el pasado, las organizaciones fueron típicamente estructuradas en forma piramidal con información generada en su base fluyendo hacia lo alto; y era en el estrato de la pirámide más alto donde se tomaban decisiones a partir de la información proporcionada por la base, con un bajo aprovechamiento del potencial de esta información. Estas empresas, han reestructurado y eliminado estratos de estas pirámides y han autorizado a los usuarios de todos los niveles a tomar mayores decisiones y responsabilidades. Sin embargo, sin información sólida para influenciar y apoyar las decisiones, la autorización no tiene sentido.

Esta necesidad de obtener información para una amplia variedad de individuos fue la que me llevo a investigar las diferentes Tecnologías de Bases de Datos y tomar de ellas sus conceptos, organización, ventajas y desventajas con el fin de tener un criterio amplio y así aplicar estos conocimientos en el área laboral.

Estas tecnologías entonces se convierten en herramientas competitivas, por hacer disponible la información a los empleados que lo necesiten para el análisis y toma de decisiones.

El objetivo será el de satisfacer los requerimientos de información interna de la empresa para una mejor gestión. El contenido de los datos, la organización y estructura son dirigidos a satisfacer las necesidades de información de los analistas. Toda empresa puede ser vista en base al proceso productivo que la sustenta. El resultado de los costos y beneficios de este proceso productivo forman una cadena de valor, donde cada eslabón (proceso de negocios) adiciona valor a la empresa. De esta forma es claro, que las empresas deben buscar optimizar cada uno de sus eslabones sin perder de vista la cadena total.

Al manejar eficientemente la información de cada área de la empresa, se pueden tomar mejores decisiones y así efectuar acciones apropiadas y finalmente conseguir un mejor control sobre la producción empresarial.

Con estas tecnologías no solo podemos unir los eslabones de la una cadena de información sino también descubrir, orientar y presentar datos de manera que pueda ser útil para el tomador de decisiones.

Bibliografía

Referencias:

- En Libros

Oracle Data Warehousing. Michael J. Cory, Michael Abbey. Oracle Press Vr. 7.3

Introducción a los sistemas de bases de datos. C.J.Date. Addison Wesley Langman, Quinta edición.

Data ware Housing "La integración de la información para la mejor toma de decisiones" . Harjinder s. Gill y Prakash C. Rao. Pretice Hall

- En Internet

http://www.cs.us.es/~delia/sia/html98-99/pag-alumnos/web5/indice.html
http://www.lania.mx/spanish/actividades/newsletters/1997-otono-invierno/mineria.html

http://agamenon.uniandes.edu.co/sistemas/6702.htm

http://www.wntmag.com/atrasados/1999/35_oct99/articulos/database.htm

http://agamenon.uniandes.edu.co/sistemas/6605.htm

http://ifem.tuportal.com/datatrab.htm

http://www.sysameri.com/osc/wp4sp.htm

http://www.map.es/csi/silice/DW2251.html

http://www.wntmag.com/atrasados/1999/35_oct99/articulos/database.htm

http://www.comsoft.com.ar/olap.doc

http://www.inf.udec.cl/revista/edicion3/cwolff.htm