UNIVERSIDAD AUTONOMA DE NUEVO LEON
FACULTAD DE CONTADURIA PUBLICA
Y ADMINISTRACION

GUIA PARA LA CONSTRUCCION DE UN DATA WAREHOUSE

POR

BERNARDO LOPEZ BERNAL

COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE MAESTRIA EN INFORMATICA ADMINISTRATIVA CON ESPECIALIDAD EN PROCESOS PRODUCTIVOS DE NEGOCIOS

JUNIO DE 2002

UNIVERSIDAD AUTONOMA DE NUEVO LEON
FACULTAD DE CONTADURIA PUBLICA Y ADMINISTRACION

GUIA PARA LA CONSTRUCCION DE UN DATA WAREHOUSE

POR

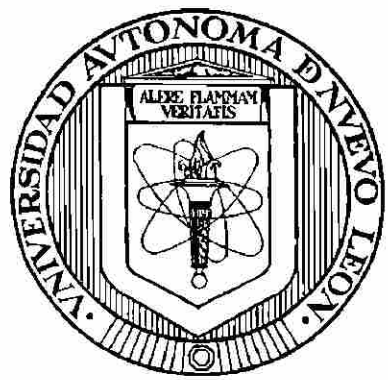
BERNARDO LOPEZ BERNAL

COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE MAESTRIA EN INFORMATICA ADMINISTRATIVA CON ESPECIALIDAD EN PROCESOS PRODUCTIVOS DE NEGOCIOS

JUNIO DE 2002

UNIVERSIDAD AUTONOMA DE NUEVO LEON

FACULTAD DE CONTADURIA PUBLICA Y ADMINISTRACION



GUIA PARA LA CONSTRUCCION DE UN DATA WAREHOUSE

Por

BERNARDO LOPEZ BERNAL

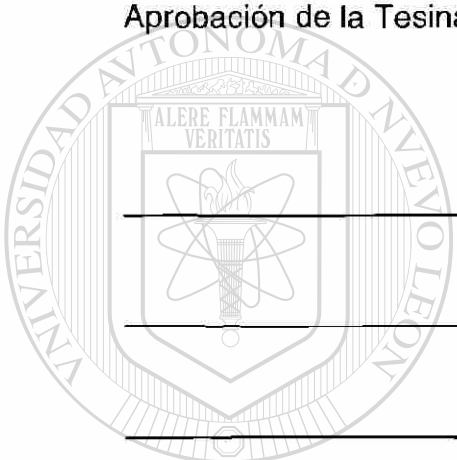
Como requisito parcial para obtener el Grado de
MAESTRIA EN INFORMATICA ADMINISTRATIVA
con Especialidad en
Procesos productivos de negocios

Junio, 2002

GUIA PARA LA CONSTRUCCION

DE UN DATA WAREHOUSE

Aprobación de la Tesina:



_____

Asesor de la Tesina

_____

_____

_____

Jefe de la División de Estudios de Postgrado o
Secretario de Postgrado o
Subdirector de Estudios de Postgrado

AGRADECIMIENTOS

Quiero agradecer a todas las personas que de una u otra manera han contribuido a la realización de este trabajo. A Mª José Adelantado del departamento de exportación de la Editorial española Gestión 2000, cuya generosa amabilidad me permitió contar con un invaluable material traído directamente desde España. A la directora de la Tesina, M.I.A. Maria de Jesús Araiza Vazquez, que contribuyó de manera esencial con su guía y apoyo constante. Al M.S. José Humberto Martínez Jiménez y al M.A. Francisco Antonio Cortes Cerda por el interés mostrado y el tiempo dedicado a la revisión de esta tesina.

También deseo agradecer el apoyo que me brindaron mis compañeros durante la realización de esta tesina y en general durante el estudio de mi maestría. A Miquel Angel, Eric, Ricardo, Braulio, y Domingo.

Agradezco de manera muy especial a mi esposa Claudia, que siempre ha contribuido de manera incondicional a la realización, no solo de esta tesina, sino de todas las metas que me he propuesto. Su apoyo y participación incondicional en cada ocasión que ha sido necesaria me han facilitado, de una manera invaluable, las labores que he emprendido. Ella es sin duda la mejor persona que he conocido y haberla elegido como esposa ha sido mi mas grande acierto.

Finalmente a mi hijo Axel Bernardo, mas que un agradecimiento le extiendo una disculpa por el tiempo que he tenido que dejar de dedicarle para llevar a cabo mis metas. Espero que en un futuro me entienda, ya que en este momento a sus escasos 2 años le importa un real cacahuate que yo tenga que cumplir algún trabajo, y me exige jugar con el constantemente (cosa que por cierto disfruto enormemente).

TABLA DE CONTENIDO

CAPITULO 1.

INTRODUCCIÓN

Se están produciendo profundas transformaciones en las empresas. Los desafíos de los distintos sectores económicos tienen clientes cada vez más exigentes, cambios cada vez más rápidos y una competencia cada vez más fuerte.

Para hacer frente a estos desafíos hay que ir más allá de la reactividad, es necesario anticipar. Anticipar los cambios, anticipar las nuevas necesidades de sus clientes, anticiparse respecto a la competencia. Para que esta anticipación sea eficaz, hay que disponer de informaciones adecuada. Todas las empresas disponen de datos que provienen de sus sistemas operativos o bien del exterior. El problema de las empresas es alcanzar los objetivos definidos por los desafíos de su sector sacando partido de los datos accesibles.

La empresa actual «se hunde» bajo los datos. Esta diluvio de información tiene como consecuencia directa un rechazo por saturación. Sin embargo, los datos representan una mina de informaciones. Son una ventaja de la que la empresa debe sacar partido. Para ello, resulta fundamental obtener una mejor comprensión del valor de la información disponible, definir indicadores de

negocio adecuados para facilitar la toma de decisiones operativas y conservar una memoria de la empresa.

Las organizaciones han usado los datos desde sus sistemas operacionales para *atender sus necesidades de información. Algunas* proporcionan acceso directo a la información contenida dentro de las aplicaciones operacionales. Otras, han extraído los datos desde sus bases de datos operacionales para combinarlos de varias formas no estructuradas, en su intento por atender a los usuarios en sus necesidades de información. Ambos métodos han evolucionado a través del tiempo y ahora las organizaciones manejan datos no limpios e inconsistentes, sobre las cuales, en la mayoría de las veces, se toman decisiones importantes. Una manera de elevar su eficiencia está en hacer el mejor uso de los recursos de información que ya existen dentro de la organización. Sin embargo, a pesar de que esto se viene intentando desde hace muchos años, en muchos de los casos, no se tiene todavía un uso efectivo de los mismos.

La razón principal es la manera en que han evolucionado las computadoras, basadas en las tecnologías de información y sistemas. La mayoría de las organizaciones hacen lo posible por conseguir buena información, pero el logro de ese objetivo depende fundamentalmente de su *arquitectura actual, tanto de hardware como de software*

Para responder a estas necesidades, la informática debe definir e integrar una arquitectura que sirva de base a las aplicaciones de ayuda a la decisión. Esta arquitectura global es el Data Warehouse. El Data Warehouse ha aparecido estos últimos años gracias a la posibilidad de convergencia entre las nuevas necesidades de informaciones de las empresas y las recientes capacidades para integrar e implementar tecnologías aptas para responder a ello.

Un Data Warehouse es una colección de datos en la cual se encuentra integrada la información de la Institución y que se usa como soporte para el proceso de toma de decisiones gerenciales. Aunque diversas organizaciones y personas individuales logran comprender el enfoque de un Data Warehouse, la experiencia ha demostrado que existen muchas dificultades potenciales.

Reunir los elementos de datos apropiados desde diversas fuentes de aplicación en un ambiente integral centralizado, simplifica el problema de acceso a la información y en consecuencia, acelera el proceso de análisis, consultas y el menor tiempo de uso de la información.

Las aplicaciones para soporte de decisiones basadas en un Data Warehousing, pueden hacer más práctica y fácil la explotación de datos para una mayor eficacia del negocio, que no se logra cuando se usan sólo los datos que provienen de las aplicaciones operacionales (que ayudan en la operación de la empresa en sus operaciones cotidianas), en los que la información se obtiene realizando procesos independientes y muchas veces complejos.

Un Data Warehouse se crea al extraer datos desde una o más bases de datos de aplicaciones operacionales. Los datos extraídos son transformados para eliminar inconsistencias y resumir si es necesario y luego, cargadas en el Data Warehouse. El proceso de transformar, crear el detalle de tiempo variante, resumir y combinar los extractos de datos, ayudan a crear el ambiente para el acceso a la información Institucional.

La innovación de la Tecnología de Información dentro de un ambiente Data Warehousing, puede permitir a cualquier organización hacer un uso más óptimo de los datos, como un ingrediente clave para un proceso de toma de decisiones más efectivo. Las organizaciones tienen que aprovechar sus recursos de información para crear la información de la operación del negocio,

pero deben considerarse las estrategias tecnológicas necesarias para la implementación de una arquitectura completa de Data Warehouse.

El lograr identificar e implementar la serie de estrategias necesarias para crear el Data Warehouse es una tarea individual de cada Organización, ya que cada una tiene sus particularidades y su propia historia. Pero existen características generales, procedimientos y cuidados a tener en cuenta para la construcción de un Data Warehouse.

El presente trabajo tiene como finalidad el desarrollar los puntos que le permitirán tener mayores posibilidades de éxito en la construcción de un Data Warehouse. Se encuentra dividido en 4 capítulos. En el capítulo 1 se define el marco teórico en el cual se circunscribe el concepto de Data Warehouse. El capítulo 2 desarrolla los conceptos teóricos en los que está fundamentado el Data Warehouse. En el capítulo 3 se establece la composición de un Data Warehouse. Y finalmente en el capítulo 4 se plantea una estrategia a seguir para la elaboración del Data Warehouse.

1.1 La Información

Nos encontramos actualmente sumergidos en la era de la información. En todos los sectores económicos, en todas las empresas, la información se convierte en «quien corta el pastel». Disponer de la información útil, tenerla en mayor abundancia que los competidores, tenerla preparada antes, disponer de ella en el momento en que el usuario la necesite en un formato comprensible y utilizable, éstos son los objetivos a lograr. Todas las técnicas y tácticas

utilizadas por los grandes estrategas se basan en la información de que disponen. En ciertos sectores, algunos hablan incluso ya de desinformación. El ejemplo más clamoroso se da en Internet, donde el contenido de las páginas está adaptado en ocasiones al usuario conectado.

En una empresa, la información está constituida por una fuente principal y fuentes externas. La fuente principal proviene del sistema llamado «de producción». Esta información interna de la empresa es en sí una verdadera mina de oro. Se completa cada vez más con datos externos a la empresa, que representan un porcentaje global que alcanza el 20% en ciertos casos. Este porcentaje depende generalmente del nivel de ubicación jerárquica de los actores, pero también del nivel de la competencia en el sector considerado. Cuanto más arriba se encuentran en la empresa quienes toman decisiones, más compararán y analizarán estas cifras respecto a los cifras provenientes del sistema de producción de la empresa.

Tanto respecto a la información externa como a la interna, actualmente se presentan tres problemas: la sobreabundancia de la información, el hecho que sea difícilmente accesible y que sea no selectiva. Esta sobreabundancia se ilustra por esta observación: «Se han producido más nuevas informaciones estos últimos treinta años que en el transcurso de los cinco milenios que nos han precedido». Asimismo, en cuanto a la accesibilidad y la selectividad de los datos, ciertas cifras estadísticas anuncian que el 27% del tiempo de un directivo, como promedio, lo pasa buscando la información, accediendo a ella y dándole formato. Ganar algunos puntos sobre este porcentaje tiene efectos directos sobre la productividad de una empresa. El problema mostrado aquí es doble: por una parte, seleccionar la información justa y útil y, por otra, referenciar esta información y almacenarla correctamente para ser capaz de recuperarla el día que se necesite. El beneficio de un sistema de decisión sólo será notable si la información es creíble, integrada, disponible en la forma deseada por el usuario, en el momento en que la necesite, sea cual sea el lugar

donde se encuentre. Todas las informaciones deben estar disponibles, bajo cualquier forma, para lo que sea, en cualquier momento.

1.2 La Empresa y su Mercado

La empresa construye un sistema de decisión con el fin de mejorar su rendimiento. El Data Warehouse debe permitirle ser proactiva en su mercado, es decir, decidir y anticipar en función de la información disponible y capitalizar sobre sus experiencias.

Cada empresa se sitúa en un mercado, en un sector económico. De manera general, todos los mercados están en plena evolución; en ciertos casos, puede incluso hablarse de mutación. Los factores relacionados con estas evoluciones son la competencia, la competitividad y la complejidad. Por lo que respecta a la justificación del Data Warehouse, hay que tener en cuenta dos fuerzas externas : la competición y la personalización.

1.2.1 La Competición

La competición tal como se vive hoy en las empresas necesita comparar sin cesar el producto propio con la competencia. La sola visión del producto a través de las informaciones internas disponibles ya no basta. Hemos pasado de

una orientación al producto a una orientación al mercado y esta visión de la competencia es fundamental en la actualidad. El objetivo es simplemente hacerlo mejor que los competidores. Los cuatro principales ejes de mejora de la situación de competencia son una mejor rentabilidad (que precisa a menudo inversiones más costosas), mayor rapidez en todas las etapas del ciclo de vida de un producto (diseño, realización, cadena de producción ...), más innovación en los productos y los servicios asociados y, generalmente, un acceso más fácil para los consumidores a los productos y a los servicios.

En el marco del Data Warehouse, el aspecto de la competición se trata mediante la integración en el sistema de decisión de datos externos introducidos o adquiridos que caracterizan el mercado y la competencia. El acercamiento entre los datos externos y los datos internos a menudo presenta graves problemas semánticos que normalmente son difíciles, o incluso imposibles, de resolver.

1.2.2 La Personalización

La personalización es la tendencia actual, que se añade a las cuatro tendencias sucesivas que hemos vivido en el tiempo: los precios, la calidad, el tiempo y los servicios. En los años de posguerra, la economía estaba resueltamente orientada al producto. Las empresas no tenían problemas para vender lo que producían, Asimismo, se daba prioridad a la producción en masa para aumentar las ventas. La preocupación siguiente (en los años setenta) era mejorar la calidad de los productos. El consumidor quería el producto y la calidad. En aquellos años, aparecieron las primeras ideas sobre normas y estándares. Hacia 1980 se toma conciencia del factor tiempo, que exige

profundos cambios en las organizaciones de las empresas y genera la automatización de un cierto número de procesos. Esta noción de tiempo se traduce a menudo en términos de reducción de plazos: plazo de diseño, plazo de entrega (time to market, time to deliver). Este objetivo de reducción de los plazos es aún hoy fundamental en los servicios. Todos los casos exitosos de reingeniería priorizan ganancias importantes en la gestión del tiempo. La tendencia de los años noventa es la mejora de los servicios asociados al producto. Estos servicios se incluyen en la etapa de compra del producto (servicios a los clientes, garantía ...), o bien en la etapa de uso de dicho producto por la incorporación cada vez más fuerte de inteligencia, a fin de hacer sus funciones más «accesibles».

La personalización es la tendencia actual. Su ambición es dar a cada cliente (usuario, consumidor, comprador ..) la impresión de ser único. El desafío para las empresas es usted, soy yo en tanto que persona y no ya en tanto que representante de una categoría de población. Hemos llegado a una lógica de segmentación llevada al extremo, donde todos los vendedores deben reaccionar como los entrañables tenderos «de la esquina» que, cada vez que entrábamos en su establecimiento, nos llamaban por nuestro nombre, nos preguntaban por la familia y que, al conocer perfectamente bien nuestras costumbres y comportamientos, proponía productos adaptados a nuestros perfiles. Hoy, este comportamiento se «simula» por la información (la masa de datos) que el sistema asocia a un cliente.

Para ilustrar esta simulación, tomemos el ejemplo de la empresa estadounidense Blockbuster, una de las mayores cadenas de venta de cintas de vídeo. Se propone al cliente apresurado (todos tenemos cada vez más prisa o creemos tenerla) una selección de diez cintas teniendo en cuenta los antecedentes y los hábitos de consumo. Este servicio constituye un valor añadido real para los clientes y ha generado para la empresa una mejora muy

sensible de sus resultados, así como un aumento de fidelidad de sus clientes. Este micromarketing, llamado también marketing de precisión, permite:

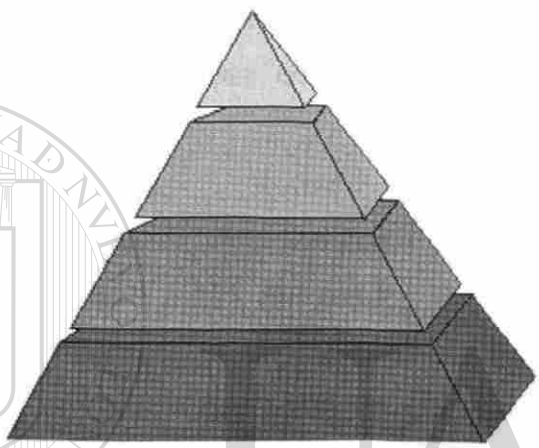
- **Aumentar el rendimiento de las acciones comerciales y mercadotecnia**. Un marketing directo estándar da un rendimiento del 2 al 4%, mientras que un marketing directo orientado da un rendimiento del 10 al 20%, A título de ejemplo, un banco francés ha conseguido más de 2 millones de francos de ahorro anual, debido a la reducción de costes de impresión y envío, trabajando «simplemente» sobre la orientación de sus mailings;

- ***Aumentar los servicios proporcionados***. *El ejemplo de Blockbuster* muestra bien que con el conocimiento de las costumbres de consumo de sus clientes, la empresa puede proponer un servicio innovador y útil basado en recomendaciones. Conocer los perfiles de consumo permite la realización de propuestas de servicios completamente adaptados y personalizados;

- **Mantener la fidelidad de la clientela.** Esta fidelidad puede conseguirse también por medio de los servicios que hacen más difícil el paso de un proveedor a otro. Este punto puede ser ilustrado por todos los nuevos servicios que se nos proponen regularmente (en cada factura) por Movistar. Se comprende así la formidable explosión de los cuestionarios y las encuestas que recibimos regularmente, solicitando detalles sobre nosotros y sobre nuestro consumo. Debido a que toda empresa debe adaptar hoy sus productos a los clientes, la palabra clave es el conocimiento del cliente. Pero no todas las empresas están en contacto directo con el cliente, como les ocurre a los proveedores de tecnologías y de productos básicos, a los fabricantes, a los distribuidores al mayor, etc. Estas empresas deben buscar, pues, en el exterior la información sobre los clientes que usan directa o indirectamente su producto.

1.3 Los Sistemas

De manera general los sistemas de información se han dividido de acuerdo al siguiente esquema:



- **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 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.** 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 ésto 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 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.

1.3.1 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.

1.3.2 Sistemas de soporte a decisiones

Por otra parte, hay otras funciones dentro de la empresa que tienen que ver con la planificación, 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 soporten. Pero estas

funciones son diferentes de las operacionales y los tipos de sistemas y la información requerida también son 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. Estos son sistemas sobre los se basa la tecnología data warehousing.

CAPITULO 2.

CONCEPTOS DE DATA WAREHOUSE

Existen una serie de conceptos que es necesario abordar para el entendimiento de lo que implica un Data Warehouse. Los conceptos serán tratados en cada uno de los siguientes temas :

2.1 Definición y Características

La definición clásica del Data Warehouse dada por Bill Inmon en su obra de referencia Using the Data Warehouse [Inmon94] es la siguiente:

«El Data Warehouse es una colección de datos orientados al tema, integrados, no volátiles e historiados, organizados para el apoyo de un proceso de ayuda a la decisión.»

De esta definición se destacan las siguientes características :

- Orientación al tema
- Datos integrados
- Datos historiados ó de tiempo *variante*
- Datos no volátiles

2.1.1 Orientación al tema

El Data Warehouse se organiza alrededor de los temas principales de la empresa. Así, los datos se estructuran por temas, contrariamente a los datos de las organizaciones tradicionales organizadas generalmente por proceso funcional. El interés de la organización es disponer de todas las informaciones útiles sobre un tema normalmente transversal a las estructuras funcionales y organizativas de la empresa. Esta orientación al tema permitirá también desarrollar el sistema de decisión (el Data Warehouso) mediante una *aproximación incrementar tema tras tema.* La integración de los diferentes temas en una estructura única es necesaria porque las informaciones comunes a varios temas no deben duplicarse. El Data Warehouse conserva así su función de punto focal. En la práctica, puede crearse una estructura suplementaria, llamada Data Mart (almacén de datos), para apoyar la orientación al tema.

El ambiente operacional se diseña alrededor de las aplicaciones y funciones tales como préstamos, ahorros, tarjeta bancaria y depósitos para una institución financiera. Por ejemplo, una aplicación de ingreso de órdenes puede acceder a los datos sobre clientes, productos y cuentas. La base de datos

combina estos elementos en una estructura que acomoda las necesidades de la aplicación.

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. Para una universidad pueden ser estudiantes, clases y profesores. Para un hospital pueden ser pacientes, personal médico, medicamentos, etc.

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.

En la **Figura No. 1** se muestra el contraste entre los dos tipos de orientaciones.

El data warehouse tiene una fuerte orientación al tema

Figura No. 1

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. Muchas de las reglas comerciales (y sus correspondientes relaciones de datos) se representan en el Data Warehouse, entre dos o más tablas.

2.1.2 Datos integrados

Un Data Warehouse es un proyecto de empresa. Por ejemplo, en la distribución, el mismo indicador de cifra de negocio interesará tanto a las fuerzas de ventas como al departamento financiero y a los compradores. Este punto de vista único y transversal constituye un fuerte valor añadido. Para llegar a ello, los datos deben estar integrados.

La consolidación de todas las informaciones respecto a un cliente dado es necesaria para dar una visión homogénea de dicho cliente a los analistas. Antes de estar integrados en el Data Warehouse, los datos deben formatearse y unificarse para llegar a un estado coherente, Un dato debe tener una descripción y una codificación únicas. Las diferencias dependen de la visión

deseada por el usuario, de la utilización que se hace, o simplemente de los programadores. Los ejemplos tradicionales para ilustrar la problemática de la integración son simplistas respecto la realidad. Integrar dos representaciones de un número real, cuatro representaciones de fecha u homogeneizar las codificaciones de una información simple y presente en forma de pocas ocurrencias es fácil. En la realidad de los proyectos, la etapa de integración es muy compleja, larga y pesada y presenta a menudo problemas de cualificación semántica de los datos a integrar. Por experiencia, representa del 60 al 90% de, la carga total de un proyecto.

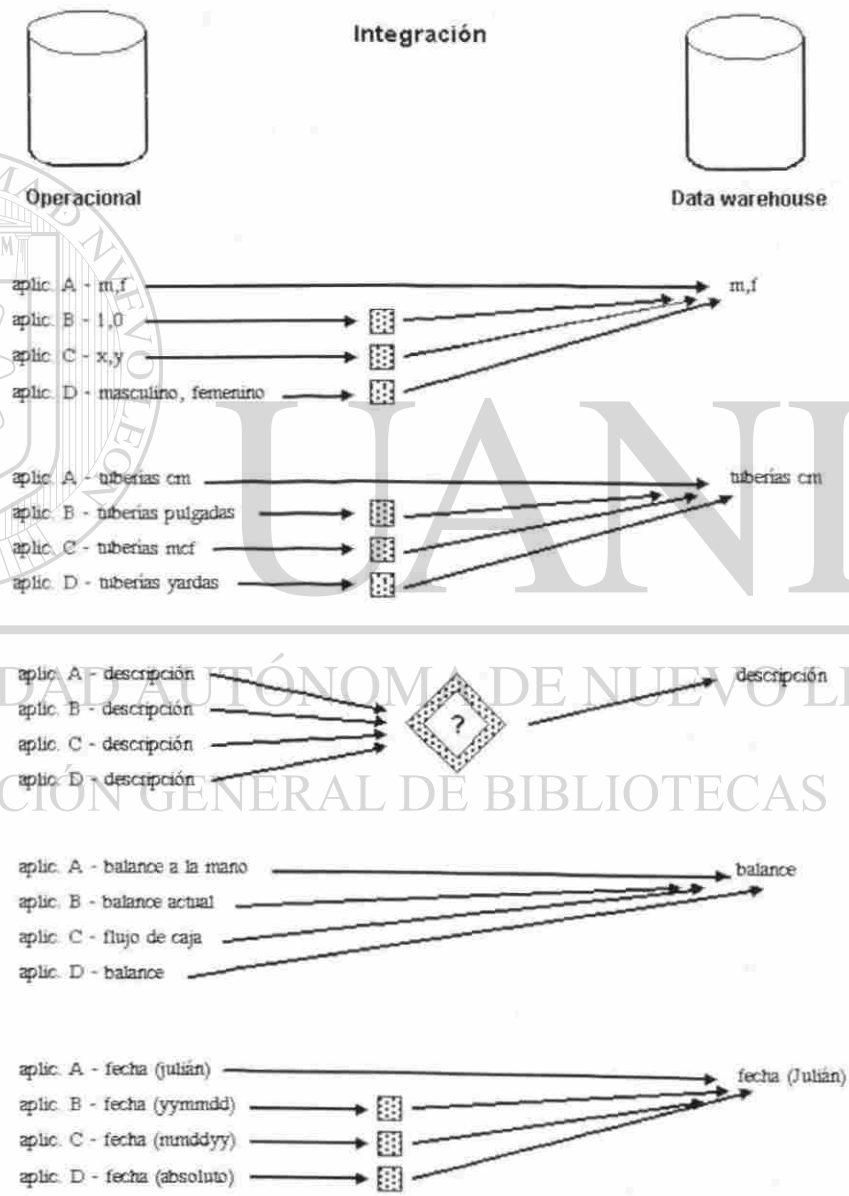
La integración de datos se muestra de muchas maneras: en convenciones de nombres consistentes, en la medida uniforme de *variables*, en la codificación de estructuras consistentes, en atributos físicos de los datos consistentes, fuentes múltiples y otros.

A través de los años, los diseñadores de las diferentes aplicaciones han tomado sus propias *decisiones sobre cómo se debería construir una aplicación*. Los estilos y diseños personalizados se muestran de muchas maneras. Se diferencian en la codificación, en las estructuras claves, en sus características físicas, en las convenciones de nombramiento y otros. La capacidad colectiva de muchos de los diseñadores de aplicaciones, para crear aplicaciones *inconsistentes, es fabulosa*.

El contraste de la integración encontrada en el Data Warehouse con la carencia de integración del ambiente de aplicaciones, se muestran en la **Figura No. 2**, con diferencias bien marcadas.

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.



Cuando los datos se mueven al data warehouse desde las aplicaciones orientadas al ambiente operacional, los datos se integran antes de entrar al depósito.

Figura No. 2

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. 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 No. 2** , 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.

2.1.3 Datos historiados o de tiempo variante

En un sistema de producción, el dato se actualiza con cada nueva transacción. El valor anterior se pierde y el dato se actualiza constantemente. Los sistemas de producción conservan bastante raramente el historial de los valores de este dato. En un Data Warehouse, el dato no debe actualizarse nunca. Representa un valor insertado en el sistema de decisión en un momento dado. El Data Warehouse almacenará así el historial, es decir, el conjunto de valores que el dato habrá tenido en su historia. Es evidente entonces que debe asociarse un referencia de tiempo al dato a fin de poder identificar un valor particular en el tiempo.

Como la información en el data warehouse es solicitada en cualquier momento (es decir, no "ahora mismo"), los datos encontrados en el depósito son historiados o 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 No. 3**).

Figura No. 3

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.

2.2.4 Datos no volátiles

La no volatilidad es en cierto modo una consecuencia de la historialización descrita anteriormente. Así, una misma consulta efectuada con tres meses de intervalo precisando naturalmente la fecha de referencia de la información buscada dará el mismo resultado. En un sistema de producción, la información es volátil, el dato se actualiza regularmente. Las consultas afectan a los datos actuales y es imposible recuperar un resultado antiguo.

Dos consecuencias se desprenden de la ausencia de actualizaciones sobre los datos de un Data Warehouse. La primera afecta a la organización interna de la base de datos que soporta el Data Warehouse. Ésta podrá adaptarse (desnormalizarse) a fin de soportar las optimizaciones para el acceso a los datos. La segunda afecta a las tecnologías necesarias para los accesos

habituales de los usuarios. En efecto, si no se realiza ninguna actualización, numerosas tecnologías costosas en tiempo de respuesta (gestión de los seguimientos, gestión de las transacciones, gestión de la concurrencia ...) integradas en los SGBD sólo sirven en las etapas de carga inicial y en las inserciones de incrementos. Por ello podrían desactivarse en las etapas de uso actual.

La perspectiva más grande, esencial para el análisis y la toma de decisiones, requiere una base de datos estable. En la **Figura No. 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.

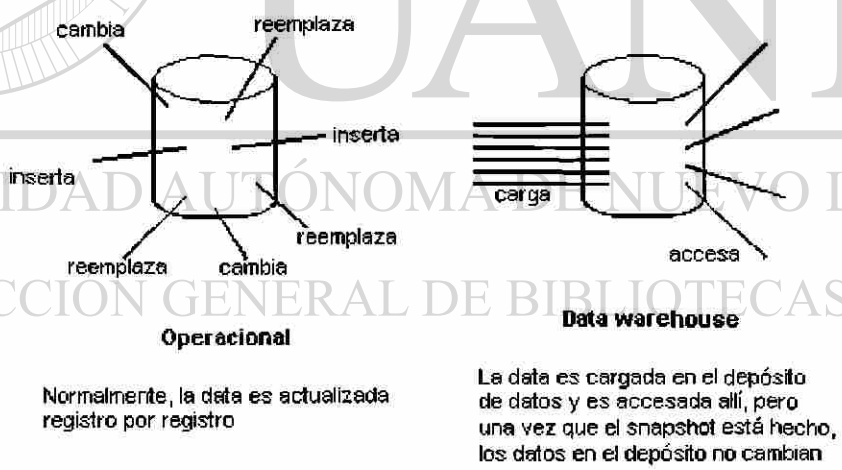


Figura No. 4

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%.

2.2 Objetivos del Data Warehouse

Como ya hemos insistido en temas anteriores, la información es vital para las empresas. Todos los datos, tanto si provienen del sistema de producción de la empresa como si se han adquirido fuera, deben organizarse, coordinarse, integrarse y almacenarse para dar al usuario una visión integrada y orientada al negocio. El Data Warehouse y sus servicios lo hacen posible.

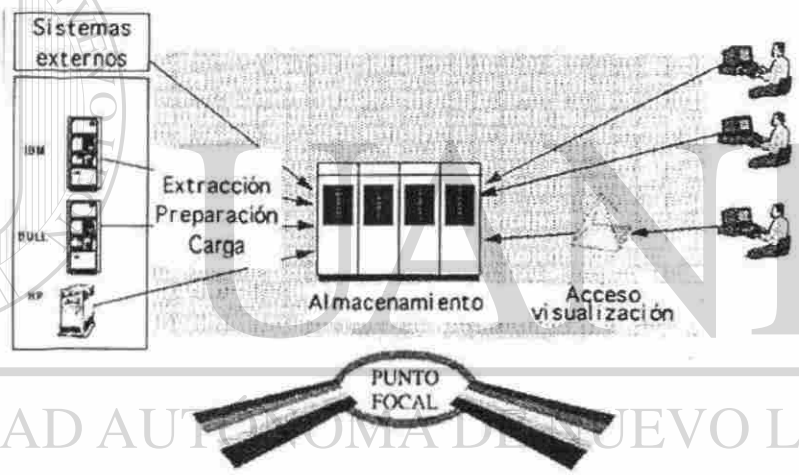


Figura No. 5

El Data Warehouse es una especie de punto focal que guarda en un único lugar toda la información útil proveniente de sistemas de producción y de fuentes externas. La **Figura No. 5** ilustra este objetivo de punto focal. Antes de cargarse en el Data Warehouse, la información debe extraerse, depurarse y prepararse. Estas etapas de alimentación son generalmente muy complejas. Una vez integrada, la información debe presentarse de manera comprensible para el usuario. Esta visión orientada al usuario se llama también visión de negocio u orientación al tema. Dos problemas derivan de esta orientación hacia el usuario final. El primero se refiere a la definición semántica de los datos

almacenados en el Data Warehouse. El segundo afecta a la implementación de la estructuración física particular de los datos. Este sistema debe ser accesible a todas las herramientas de acceso y de visualización del usuario final.

2.3 Estructura del Data Warehouse

Un Data Warehouse se estructura en cuatro clases de datos organizadas según un eje histórico y un eje sintético. La **Figura No. 6** ilustra esta estructura y sitúa las clases unas respecto a otras en un marco de arquitectura de datos.

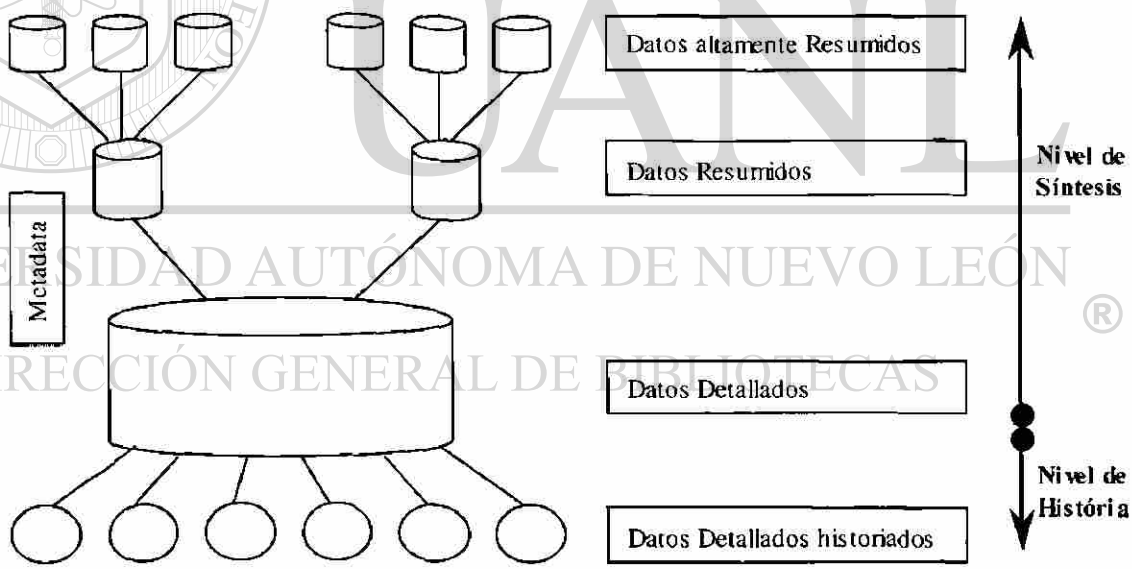


Figura No. 6

- **Los datos detallados** reflejan los eventos más recientes. Las inserciones regulares de datos surgidos de los sistemas de producción habitualmente se realizarán a este nivel. Normalmente, los datos detallados pueden ser muy voluminosos y necesitar máquinas sofisticadas para

gestionarlos y tratarlos. Es importante hacer notar que el nivel de detalle *almacenado en el* Data Warehouse no es forzosamente idéntico al nivel de detalle gestionado en los sistemas operacionales. El dato insertado en el Data Warehouse puede ser un resumen o una simplificación de informaciones sacadas del sistema de producción. Sin embargo, cuanto más fino es el nivel de detalle, más posible es segmentar estos datos dinámicamente. No porque un dato esté disponible debe integrarse a este nivel. Sólo se almacenan los datos de detalle útiles y necesarios.

- **Los datos resumidos** se utilizan a menudo ya que son los elementos de análisis que representan las necesidades de los usuarios. Constituyen ya un resultado de análisis y una síntesis de la información contenida en el sistema de decisión. Por ello deben ser fácilmente accesibles y comprensibles. La facilidad de acceso viene dada por estructuras multidimensionales que permiten a los usuarios navegar por los datos según una lógica más intuitiva. El rendimiento vinculado al acceso a estos niveles debe ser también óptimo, por lo que generalmente son soportados en discos. Para la comprensión de estos datos, es necesario poner a disposición de los usuarios la definición completa de la información que se le presenta. Por ejemplo, esta información puede estar compuesta del contenido presentado (suma de las ventas, promedio de compras, etc.) y de la unidad sobre la que se realiza el resumen (por meses, por semanas, por fábricas, por productos, etc.)

- **Los datos altamente resumidos** representan el siguiente nivel de datos encontrado en el Data Warehouse. Estos datos son compactos y fácilmente accesibles. A veces se encuentra en el ambiente de Data Warehouse y en otros, fuera del límite de la tecnología que ampara al Data Warehouse. De *cualquier forma, los datos completamente resumidos son parte del* Data Warehouse sin considerar donde se alojan físicamente.

- **La Metadata** agrupa todas las informaciones respecto al Data Warehouse y los procesos asociados. Las principales informaciones van destinadas :

1. Al usuario. Informaciones sobre la semántica de los datos utilizados y su localización en el Data Warehouse.

2. A los equipos responsables de los procesos de transformación de los datos del entorno de producción hacia el Data Warehouse. Informaciones sobre la localización del dato en los sistemas de producción y la descripción de las reglas y los procesos de transformación.

3. A los equipos responsables de los procesos de creación de los datos resumidos a partir de los datos detallados.

4. A los equipos de administración de la base de datos. Informaciones sobre la estructura de la base que implementa el Data Warehouse.

5. A los equipos de producción. Informaciones sobre los procedimientos de carga, el historial de actualizaciones, etc.

- **Los datos historiados** conservados en línea es uno de los objetivos importantes del Data Warehouse. Cada nueva inserción de datos del Sistema de Producción no destruye los anteriores valores, sino que crea una nueva ocurrencia del dato. El soporte al mecanismo de datos historiados depende del volumen de estos, de la frecuencia de acceso, del tipo de acceso y naturalmente del costo de los soportes. El disco, el disco óptico digital, las cintas, son los soportes mas habitualmente utilizados. La **Figura No. 7** es una síntesis de la Estructura de un Datawarehouse en forma de Pirámide.

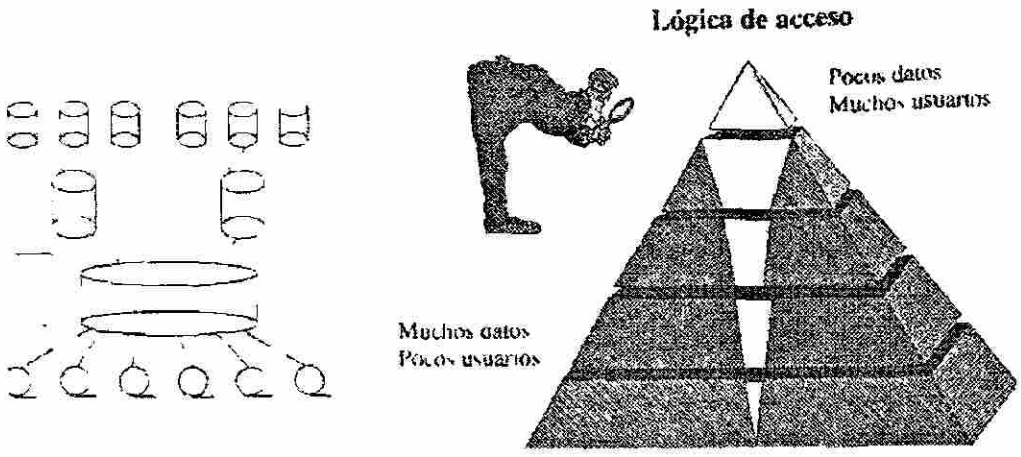


Figura No. 7

Este esquema ilustra bien la lógica de acceso utilizada más a menudo por los usuarios, el ataque a los datos por arriba (los niveles más agregados) y luego por profundizaciones sucesivas (drill down), la progresión en los datos débilmente agregados y finalmente el acceso a los datos detallados e historiados. Esta lógica de zoom corresponde a un afinado sucesivo de las necesidades de¡ usuario traducidas en criterios de selección de datos cada vez más precisos. Se accede también directamente a los datos detallados e historiados, lo que lleva generalmente a agregados de datos muy pesados que necesitan, según los volúmenes, la potencia de máquinas poderosas. «El Data Warehouse es un éxito en una empresa cuando el número de usuarios que acceden a los datos de detalle aumenta». En efecto, el valor de los datos no se da en el agregado, sino en el detalle. De hecho, está claro que, al implementar un Data Warehouse, lo más importante es el número de usuarios que acceden a los datos agregados. Pueden acceder incluso por consultas catalogadas predefinidas. La transformación del dato en conocimiento sólo se dará cuando el usuario domine sus herramientas y dirija por sí mismo sus investigaciones. Por experiencia, una vez llegado a este estadio el usuario no se queda en la agregación. Las consecuencias de esta evolución «deseada» afecta a la configuración de los componentes de Software y Hardware, que ya no deben de limitarse solo al soporte de los accesos a los datos agregados.

A fin de identificar los diferentes niveles de los datos encontrados en el Data Warehouse, considere el ejemplo mostrado en la **Figura No. 8**.

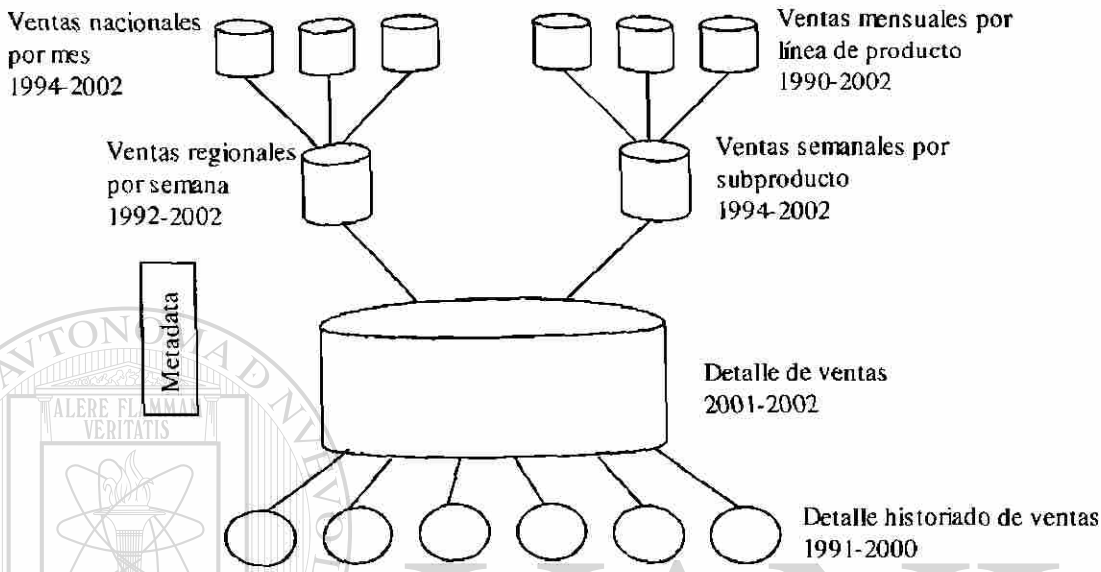


Figura No. 8

Los datos detallados historiados de las ventas, incluyen las ventas que se encuentran antes del 2001. Todos los detalles de ventas desde 1991 (o cuando el diseñador inició la colección de los archivos) son almacenados en el nivel de detalle de datos historiados.

Los datos detallados contienen información de las ventas desde 2001 al 2002 (suponiendo que 2002 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. En otras palabras, habría un retraso de tiempo de por lo menos veinticuatro horas, entre el tiempo en que en el ambiente operacional se haya hecho un nuevo ingreso de la venta y el momento cuando la información de la venta haya ingresado al Data Warehouse.

El detalle de las ventas son resumidas semanalmente por línea de subproducto y por región, para producir un almacenamiento de datos resumidos.

El detalle de ventas semanal es adicionalmente resumido en forma mensual, según una gama de líneas, para producir los datos altamente resumidos.

2.4 Arquitecturas del Data Warehouse

Para implementar un Data Warehouse, son posibles tres tipos de arquitecturas :

- Arquitectura *real*
- Arquitectura virtual, y
- *Arquitectura remota*

2.4.1 Arquitectura real

La arquitectura real es generalmente la arquitectura elegida para los sistemas de decisión. El almacenamiento de los datos del Data Warehouse se realiza en un SGBD separado del sistema de producción. Este SGBD se alimenta por extracciones periódicas. Antes de la carga, los datos sufren importantes procesos de integración, limpieza y transformación. La ventaja de esta solución es que se dispone de datos preparados para las necesidades de la decisión que responden bien a los objetivos del Data Warehouse. La principal razón para justificar la arquitectura real es la inadaptación de los datos de producción a las necesidades de los sistemas de decisión. Las estructuras de datos en un sistema de producción son complejas en cuanto a almacenamiento e implican herramientas de programación para acceder a ellas. Los datos también están codificados. En un contexto de ayuda a la decisión, el dato debe ser comprensible por el usuario. Es necesario transformar todos los códigos en datos legibles y comprensibles. Las cargas relacionadas con los accesos son también incompatibles con los objetivos de rendimiento del sistema de producción. Los datos están dispersos. El Data Warehouse debe reintegrar los datos unos con otros a fin de asegurar una coherencia semántica global. Los datos son evolutivos. No hay consolidación posible sobre un período de tiempo debido a que el dato evoluciona con las transacciones. Finalmente, el significado puede ser ambiguo. Puede depender de la aplicación que utiliza el dato. Este problema de coherencia, habitual en los sistemas de producción, debe tratarse sistemáticamente para minimizar las redundancias de información en el Data Warehouse y unificar la semántica en toda la empresa.

Los inconvenientes son el coste de almacenamiento suplementario y la falta de acceso en «tiempo real». El desplazamiento de tiempo entre los dos sistemas depende de numerosos factores como el coste de la extracción, las

necesidades funcionales, etc. Puede variar de un día a un mes según las aplicaciones.

2.4.2 Arquitectura virtual

En una arquitectura virtual, los datos del Data Warehouse residen en el sistema de producción. Se hacen visibles por productos middleware o por gateways. En esta arquitectura no hay coste de almacenamiento suplementario y el acceso se hace en tiempo real. Sin embargo, las numerosas desventajas de este tipo de arquitectura impiden frecuentemente su elección. Los datos no están preparados. El apartado anterior muestra la dificultad real que presenta la utilización de los datos de producción. Los accesos de decisión pueden perturbar el rendimiento del sistema de producción, tanto más cuanto que los procesos de transformación y de integración están aquí relacionados forzosamente con los procesos de acceso. En el caso en que la gestión de historial no esté prevista en el sistema de producción, es impensable su integración.

Según nuestros datos, no existe ninguna implementación de Data Warehouse virtual ambiciosa. Las únicas realizaciones mostradas parecen más bien prototipos.

2.4.3 Arquitectura remota

La arquitectura remota es una combinación de los dos tipos de arquitecturas descritas anteriormente. El objetivo es implementar físicamente los niveles agregados (los niveles de datos utilizados más a menudo) a fin de facilitar el acceso y conservar el nivel de detalle en los sistemas de producción dando acceso por medio de middleware o de gateways. Esta arquitectura se utiliza también muy raramente.

2.5 Elementos de una Arquitectura de Data Warehouse

El Data Warehouse es realmente es una tecnología muy entendible. De hecho, Data Warehousing puede representar mejor la estructura amplia de una empresa para administrar los datos informativos dentro de la organización. A fin de comprender cómo se relacionan todos los componentes involucrados en una estrategia data warehousing, es esencial tener una Arquitectura Data Warehouse, la cual se muestra en la **Figura No. 9**.

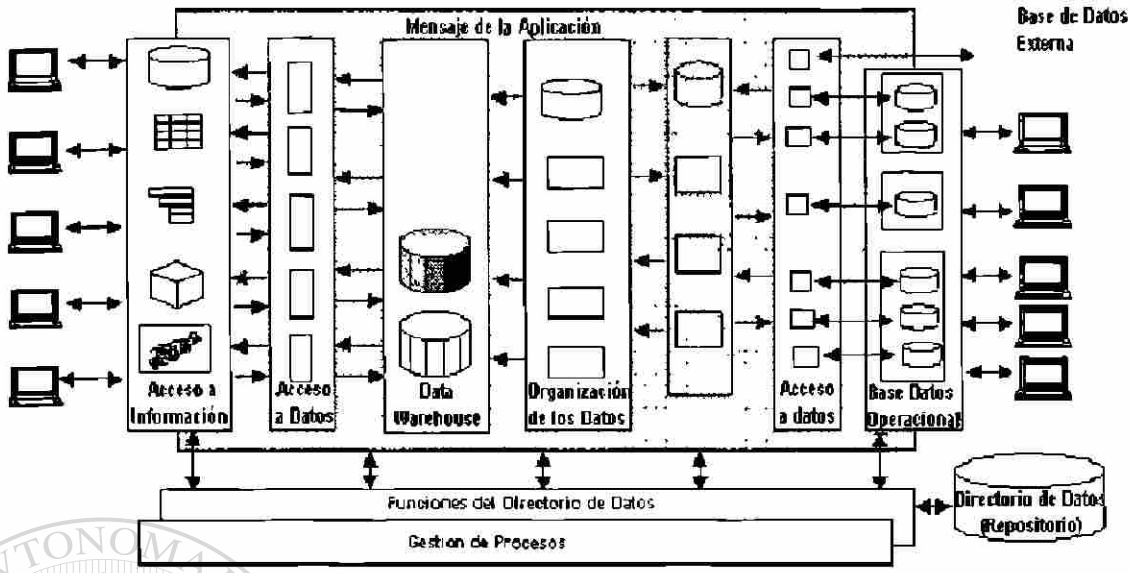


Figura No. 9

Una Arquitectura Data Warehouse (Data Warehouse Architecture - DWA) es una forma de representar la estructura total de datos, comunicación, procesamiento y presentación, que existe para los usuarios finales que disponen de una computadora dentro de la empresa. La arquitectura se constituye de un número de partes interconectadas :

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

2.5.1 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 acceder a los datos para otra gestión o propósitos informáticos. Esta dificultad en acceder 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 Warehouse es liberar la información que es almacenada en bases de datos operacionales y combinarla con la información desde otra fuente de datos, generalmente externa. Cada vez más, las organizaciones grandes adquieren datos adicionales desde bases de datos externas. Esta información incluye tendencias demográficas, econométricas, adquisitivas y competitivas (que pueden ser proporcionadas por Instituciones Oficiales – INEGI). Internet provee el acceso a más recursos de datos todos los días.

2.5.2 Nivel de acceso a la información

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

2.5.3 Nivel de acceso a los datos

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

de consulta, pero en los últimos veinte años ha llegado a ser el estándar para el intercambio de datos.

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

2.5.4 Nivel de Directorio de Datos (Metadata)

A fin de proporcionar 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 MS-SQL son metadata. También lo son las sentencias DIMENSION en un programa VB o las sentencias a crear en SQL.

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

2.5.5 Nivel de Gestión de Procesos

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

2.5.6 Nivel de Mensaje de la Aplicación

El nivel de mensaje de la aplicación tiene que ver con el transporte de información alrededor de la red de la empresa. El mensaje de aplicación se nombra 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.

2.5.7 Nivel Data Warehouse (Físico)

En el Data Warehouse (núcleo) es donde ocurre la data actual, usada principalmente para usos estratégicos. En algunos casos, uno puede pensar del Data Warehouse simplemente como una vista lógica o virtual de datos. En 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 acceder y es altamente flexible.

2.5.8 Nivel de Organización de Datos

El componente final de la arquitectura Data Warehouse es la organización de los datos. Se llama también gestión de copia o réplica, pero de hecho, incluye todos los procesos necesarios como seleccionar, editar, resumir, combinar y cargar datos en el depósito y acceder a la información desde bases de datos operacionales y/o externas. La organización de datos involucra con frecuencia una programación compleja, pero cada vez más, están creándose las herramientas Data 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.

2.6 Consideraciones para la construcción del Data Warehouse

Construir un Data Warehouse es un proceso largo y delicado. Por ello es importante resumir los factores que favorecen el éxito de un proyecto y los principales peligros a evitar, y posicionar el proyecto en el contexto de las experiencias ya vividas en el mercado.

2.6.1 Factores de éxito

Las características citadas más habitualmente respecto a un Data Warehouse con éxito son las siguientes:

- Integra datos de producción con datos externos y gestiona historiales;
- Contiene las informaciones útiles, las hace legibles y manipulabas;
- Agrupa datos de calidad (coherentes, actualizados, documentados),
- *Ofrece un acceso directo a los usuarios*;
- Aumenta el número de accesos a los datos;

- Ofrece una flexibilidad que apoya el crecimiento, tanto desde el punto de vista del número de usuarios, de las herramientas utilizadas o de los volúmenes a tratar.

Por lo que respecta a la flexibilidad necesaria para apoyar el crecimiento del número de usuarios, el Data Warehousing Institute publicó [DWHInsta] en enero de 1996 los resultados de una encuesta, llevada a cabo en 6,214 empresas, que indica una progresión media de los usuarios en el tiempo. La siguiente tabla sintetiza sus principales resultados.

	Todas las sedes	Sedes pequeñas
Numero inicial promedio	16	6
Después de 3 meses	19	12
Después de 6 meses	44	20
Después de 12 meses	99	28
Después de 24 meses	255	55

La progresión notada en este estudio es importante porque representa un factor 3 sobre los 6 primeros meses y un factor 10 a 15 sobre 2 años. Esta progresión rápida tiene como consecuencia obligar a todas las empresas a prever relativamente pronto la financiación asociada y, como consecuencia, estimar lo más pronto posible el beneficio sobre inversión previsto.

Los beneficios esperados citados más a menudo se refieren principalmente a la mejora de partes de mercado, la rentabilidad, un mejor acceso a la información que genera una ventaja competitiva, la reducción de costos o el aumento de la productividad respecto a un sistema clásico, y finalmente un mejor apoyo a los cambios organizativos.

2.6.2 Errores a evitar

La lista de los diez errores a evitar presentada en este apartado no es realmente exhaustiva. De hecho, el undécimo error es creer que sólo hay diez. Pero son sin embargo representativos de los errores cometidos más a menudo. La lista es la siguiente:

1. ***Empezar el proyecto con el promotor equivocado***. Es obligatorio iniciar un proyecto de decisión con un promotor ejecutivo que cuente con los medios de invertir en el tema de la utilización eficaz de la información.
2. ***Comprometerse con posibilidades que no pueden realizarse***. Todo proyecto de decisión debe abordarse con una lógica de iteración y para cada tema tratado; una iteración debe incluir una etapa de persuasión de los usuarios y una etapa de realización.
3. ***Comprometerse a partir de declaraciones de esperanzas ingenuas***. El Data Warehouse debe presentarse como un entorno que permite a los dirigentes tomar decisiones correctas basadas en informaciones adecuadas en lugar de un entorno de ayuda a los dirigentes para tomar las mejores decisiones. En efecto, el Data Warehouse es ante todo el único sitio, el

punto focal donde los dirigentes pueden ir a buscar la información que desean.

4. ***Cargar el Data Warehouse con datos únicamente porque están disponibles***. Identificar lo que será útil es una de las grandes dificultades del diseño. Integrar datos inútiles tiene consecuencias tecnológicas importantes, especialmente en cuanto a los volúmenes de datos a tratar y a los tiempos de respuesta.
5. ***Creer que el esquema de la base de datos que soporta el Data Warehouse se organiza de la misma manera que un esquema de base de datos tradicional de tipo transaccional.*** Las diferencias entre los dos entornos afectan a los usuarios, a los tipos de peticiones, al número de tablas, a los contenidos... Es decir, a la mayor parte de los componentes.
6. ***Elegir un jefe de proyecto orientado a la tecnología.*** Un Data Warehouse no es un problema técnico, sino la integración de un conjunto de iniciativas «orientadas a temas» en un entorno de tecnologías. La palabra tema utilizada sistemáticamente es ilustrativo de la necesaria orientación al negocio.
7. ***Concentrarse en los datos tradicionales e internos***. Como ya hemos subrayado, la integración de datos externos permite a la empresa posicionar sus productos respecto al mercado y a la competencia. Generalmente es un elemento de fuerte valor añadido cuando es posible. Respecto a los tipos de datos, está claro que cada vez más datos estructurados de tipo multimedia se integrarán en los sistemas de decisión y aportarán también un fuerte valor agregado respecto a los datos tradicionales surgidos del sistema de producción.
8. ***Creer en las promesas tecnológicas***. Características como la capacidad de evolución, la escalabilidad, son particularmente prioritarias en los proyectos de decisión, al igual que las técnicas de optimización de rendimientos y de extensión de las capacidades de almacenamiento. Atención, por ejemplo, a prever desde el inicio del proyecto un modelo de

máquina que soporte la extensión en términos de procesador y de disco si ello es necesario a corto o medio plazo.

9. ***Creer que los problemas terminan una vez terminado el Data Warehouse***. El Data Warehouse es el sistema de decisión de la empresa. Su ciclo de vida está relacionado con las evoluciones de la empresa y de su mercado. Como hemos visto anteriormente, la evolución regular del número de usuarios y del tipo de accesos realizados necesita constantes regulaciones y optimizaciones.
10. ***Centrarse en las funciones predefinidas y periódicas***. Es importante construirlas al principio, pero lo fundamental es permitir a largo plazo a los diferentes tipos de usuarios acceder, mediante la herramienta adaptada, a todos los datos del Data Warehouse (datos agregados y datos detallados). Recordemos que el valor proviene por una parte de los datos de detalle y por otra de la capacidad del usuario para apropiarse de las herramientas adaptadas a sus necesidades.

CAPITULO 3

CONSTRUCCIÓN DEL DATA WAREHOUSE

Transformar los datos en conocimiento es un proceso complejo. Este proceso de transformación e integración de los datos puede sintetizarse a través de las etapas representativas de un método industrial ilustrado por los siguientes pasos:

1. Ensamblar las materias primas (los datos de diferentes fuentes),
2. Según instrucciones específicas (el metamodelo),
3. Para realizar un producto terminado (los datos analíticos),
4. Almacenado en un almacén de datos (el Data Warehouse),
5. Para su disponibilidad fácil de cara a los Clientes (los usuarios finales)

La **Figura No. 10** ilustra el marco general de un Data Warehouse.

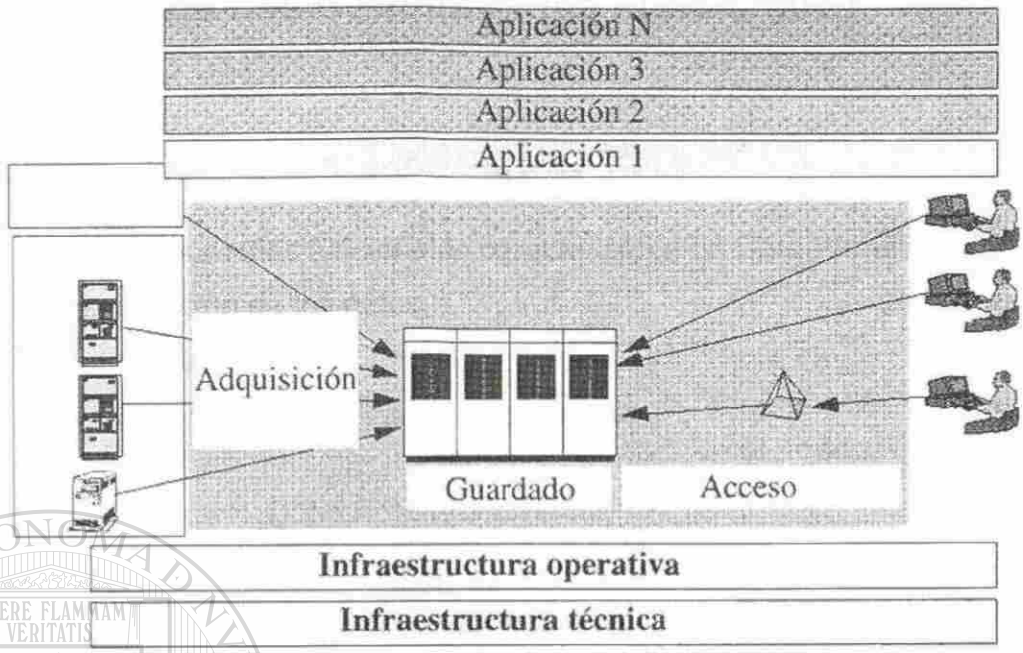


Figura No. 10

Un Data Warehouse no se construye en una sola iteración. Cada tema tratado se descompone en un conjunto de iniciativas (las aplicaciones). El perímetro de cada aplicación debe estar claramente definido (actores afectados, frecuencia y periodicidad de los análisis, objetivos y naturaleza de las vueltas sobre la actividad de la empresa ...). Las aplicaciones deben ser controlables y proporcionar resultados «tangibles» en un plazo de menos de seis meses, que corresponde al plazo medio de realización de una aplicación. Esta descomposición en aplicaciones aporta numerosas ventajas, pero genera dificultades sobre ciertos temas, como los relacionados con la infraestructura técnica y organizativa que necesitan que se imagine una visión más global. Una aplicación puede también ser un programa de decisión. En este caso, conviene asegurarse de que se integra en la infraestructura global.

3.2 Los Componentes Funcionales

Tres componentes funcionales caracterizan a un Data Warehouse:

1. La adquisición de los datos,
2. Su almacenamiento, y
3. El acceso por parte de los usuarios finales.

3.2.1 La adquisición de los datos

La adquisición de los datos se desarrolla en tres etapas:

a) La extracción,

b) La preparación, y

c) La carga.

a) La extracción de los datos consiste en recoger los datos útiles en el sistema de producción. Primero hay que identificar los datos que hayan evolucionado a fin de extraer el mínimo de datos, luego planificar estas extracciones a fin de evitar las saturaciones (red, entradas / salidas y unidad central) del sistema de producción. La integridad de los datos extraídos es obligatoria y precisa en consecuencia la sincronización de los diferentes procesos de extracción. Los problemas relacionados con esta necesaria sincronización pueden ser complejos, ya sea funcionalmente o bien técnicamente en entornos muy heterogéneos.

Para extraer los datos originales, se pueden usar varias tecnologías:

- **Pasarelas**, proporcionadas principalmente por los editores de bases de datos; estas pasarelas son generalmente insuficientes porque están orientadas esencialmente a datos y soportan con dificultad procesos (programas) de transformación complejos;
- **Utilidades de replicación**, utilizables si los sistemas de producción y el sistema de decisión son homogéneos y si la transformación a aplicar a los datos es ligera;
- **Herramientas específicas de extracción**; estas herramientas son sin duda la solución operativa al problema de la extracción, pero su precio elevado limita su uso en las primeras aplicaciones.

b) La preparación de los datos corresponde a la transformación de las características de los datos del sistema operativo en la forma definida del Data Warehouse. Esta preparación incluye la correspondencia de los formatos de datos, la limpieza, la transformación y la agregación. Este proceso de transformación accederá naturalmente a todas las informaciones almacenadas en el diccionario, especialmente la localización de las fuentes de datos y sus estructuras en el sistema de producción, la estructura objeto del Data Warehouse, las reglas de identificación, de asociación, de transformación y de agregación de los datos y las reglas de seguridad. La limpieza de los datos es una etapa sobre la que actualmente trabajan numerosos editores. Además de la calidad de los datos que aportan, las herramientas de limpieza permiten suprimir los duplicados en los archivos. La supresión de los duplicados es un proceso que se justifica rápidamente. Un ejemplo permite ilustrar este punto. Se refiere a una experiencia vivida regularmente por cada uno de nosotros cuando abrimos nuestros buzones y encontramos cartas idénticas que se refieren a la misma promoción o al mismo servicio, con la única diferencia debida al destinatario.

c) La carga es la última etapa de la alimentación del Data Warehouse. Se trata de una etapa delicada especialmente cuando los volúmenes son importantes.

Para obtener buenos rendimientos de carga, es imperativo controlar las estructuras del SGBD (tablas e índices) asociadas a los datos cargados para optimizar al máximo estos procesos. Por ejemplo, en grandes cargas, es preferible indexar las tablas una vez cargados los datos en lugar de indexarlos al mismo tiempo que se cargan. Además, las técnicas de paralelismo optimizan las cargas pesadas.

3.2.2 El almacenamiento de los datos

El componente básico de soporte del almacenamiento es el SGBD. Además del almacenamiento, el SGBD debe proponer extensiones para responder a las características del acceso a la decisión. Estas tecnologías se relacionan principalmente con el paralelismo de las consultas y con diversas optimizaciones propuestas para acelerar las selecciones y las agrupaciones de conjuntos.

Debido a la importancia del historial en un Data Warehouse, la estructuración física de los datos es también muy importante. Una partición física de las tablas en unidades menores según el criterio tiempo aporta rendimientos estables en el tiempo, facilidades para la recuperación, las indexaciones, las reestructuraciones y el archivo. También resulta más fácil almacenar estos datos en soportes menos costosos que los discos.

El último elemento relacionado con el almacenamiento concierne a los tipos de datos. Hoy, sólo el 15% de la información se almacena en formato electrónico y la principal razón de este bajo porcentaje proviene del hecho de que una información debe descomponerse primero en una serie de enteros, de

reales y de textos antes de almacenarse en un SGBD de producción. Los editores realizan muy fuertes evoluciones en lo que algunos denominan la «gesfión de contenido». Aportar una mejor respuesta a ésta pasa por la integración en el SGBD de nuevos tipos de datos que permitan almacenar y manipular estructuras multimedia compuestas de documentos, de imágenes, de sonidos, de vídeos, etc. El avance de Internet acelerará también, sin duda, esta evolución. Una vez que los SGBD sepan gestionar eficientemente estas nuevas estructuras, permitirán el almacenamiento y la manipulación de informaciones nuevas, complementarias a las informaciones tradicionales aportándoles valor en los procesos de toma de decisión.

El SGBD aporta, finalmente, la transparencia a la evolución del hardware (scalability en inglés), la independencia, ya sea en los tipos y el número de procesadores, los discos o la memoria, así como la transparencia ante la evolución de los sistemas operativos. Sin embargo hay que tener cuidado con las afirmaciones de los editores y constructores, porque esto tiene un costo y unos límites.

3.2.3 El acceso a los datos

Definir una arquitectura global que dé soporte a los accesos de decisión impone opciones tecnológicas no estructuradoras. Este método es el mismo que el que ha llevado a las empresas a orientarse hacia la arquitectura cliente/servidor, y mas recientemente a la Web, para sus aplicaciones transaccionales. Una de las bases de esta arquitectura es que permite implementar una infraestructura común a todas las aplicaciones de decisión,

dejando a los usuarios la oportunidad de utilizar las soluciones mejor adaptadas a su problemática: un grupo de usuarios deseará efectuar consultas simples o complejas sobre los datos que les interesen; otro grupo querrá efectuar análisis sobre informaciones muy estructuradas y agregadas; otro servicio necesitará hacer extrapolaciones o simulaciones a partir de la información existente.

La **Figura No. 11** ilustra la lógica de esta arquitectura, que da a los usuarios puntos de vista diferentes según las herramientas utilizadas.

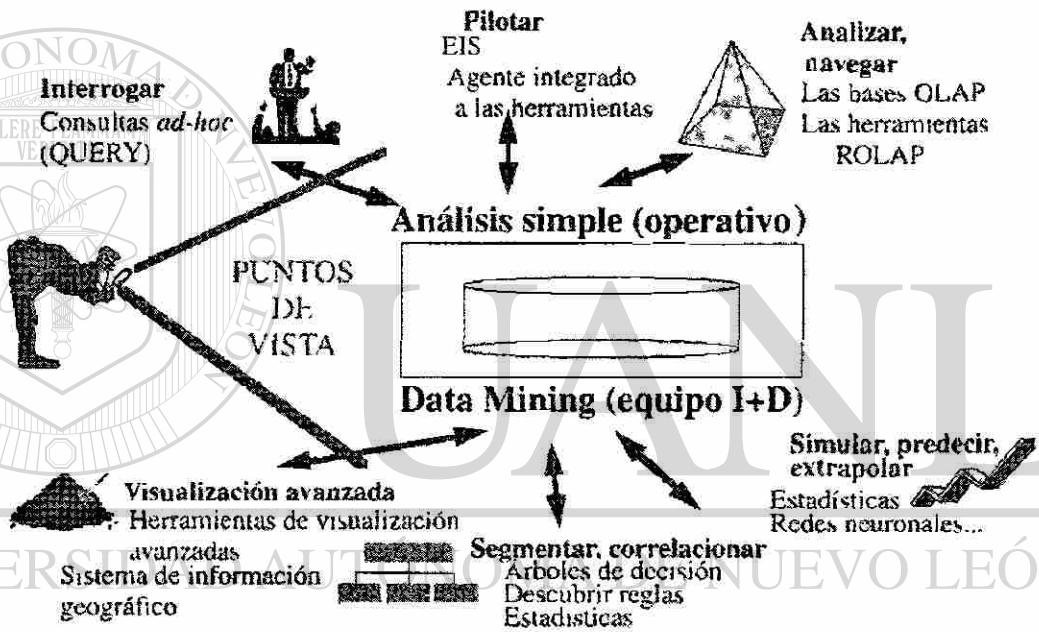


Figura No. 11

Sea cual sea el tipo de la herramienta, tendrá que adaptarse a las exigencias del usuario y a su manera de trabajar. En el mundo de la decisión, el análisis es iterativo y los resultados de la consulta actual influyen a menudo en la consulta siguiente. «Dame lo que te pido, que luego podré decirte lo que quiero realmente» es una fórmula que refleja bien este comportamiento.

3.3 Las Infraestructuras

Para responder a estas necesidades, el nuevo papel de la informática es definir e integrar una arquitectura sobre la que se basarán las aplicaciones de decisión.

Hay que considerar dos niveles de infraestructura en un Data Warehouse:

- La infraestructura técnica, es decir, el conjunto de componentes materiales y programas, y
- La infraestructura operativa, es decir, el conjunto de procedimientos y servicios para administrar los datos, gestionar los usuarios y utilizar el sistema.

3.3.1 La infraestructura técnica

Se compone de productos que implementan las tecnologías elegidas integrados en un conjunto coherente y homogéneo. Estas opciones técnicas afectan a los componentes materiales y programas junto con los componentes funcionales que son la alimentación, el almacenamiento y el acceso.

3.3.2 La infraestructura operativa

Se compone de todos los procesos que permiten, a partir de los datos de producción, crear y gestionar el sistema de decisión. Las grandes funciones de esta infraestructura conciernen a la administración de los datos, la gestión de los usuarios en el sentido de apoyo y administración y el uso del sistema de decisión. Esta última función es muy importante porque afecta a la ordenación y a la gestión de los flujos de datos de los sistemas originales al sistema destinatario (la gestión del rendimiento, de la seguridad, etc.)

CAPITULO 4

ELABORACIÓN DE UN DATA WAREHOUSE

En lugar de pretender proponer un método «llave en mano» para la elaboración de un Data Warehouse, el objetivo de esta guía es proponer un marco, que podrá servir de base para una apropiación en relación con el contexto de una empresa.

Este marco parte de los principios siguientes, puntos comunes a la mayoría de las Organizaciones actuales:

- El objetivo no es construir un sistema de decisión aislado o al margen de otro sistema, sino más bien implementar un sistema de información coherente e integrado.
- Este sistema no se construye en un solo bloque. Se descompone en aplicaciones, integrándose cada una de ellas al marco general del Data Warehouse. Su implementación es pues incremental, constituyendo cada aplicación la unidad de incremento.

Estos dos objetivos son simples de enunciar, pero más complejos de realizar, ya que implican restricciones. Renunciar a hacer frente a estas restricciones podrá llevar el proyecto hacia una deriva clásica en la informática:

construir tantos sistemas como necesidades de decisión haya en la empresa, renunciando a su integración, juzgada demasiado compleja, costosa o difícil de vender, especialmente por razones políticas. Esta deriva entraña, entre otros, dos riesgos principales.

El primero es limitar el valor de la información contenida en el Data Warehouse. Si, por ejemplo, las informaciones respecto a las ventas no son coherentes con las que afectan a las compras, porque se han modelado en dos Data Marts diferentes, el sistema perderá una parte considerable de su valor para la empresa.

El segundo es implicar un costo informático considerable a largo plazo. Para convencerse de este punto, basta con observar las realizaciones de Data Warehouse actuales. Si algunas de ellas son complejas, es precisamente porque la información necesaria para su constitución se encuentra desperdigada en una multitud de sistemas, herencia de la empresa, no integrados entre sí. Cuando los sistemas de información tienen que gestionar informaciones cada vez más diversas y voluminosas, se hace imperativo evitar que se «desintegren» y el Data Warehouse constituye una oportunidad para alcanzar este objetivo.

4.1 Estrategia de elaboración del Data Warehouse

Un proyecto de Data Warehouse puede ser descompuesto en 4 etapas.

La primera etapa, titulada ***«descubrir y definir las iniciativas»***, es a nivel de empresa – esta etapa permitirá definir el porqué del Data Warehouse,

informar sobre el beneficio esperado, sus características, sus aplicaciones para los actores afectados. Finalmente, permitirá determinar las aplicaciones a realizar (o proyectos) en este marco, estimará los beneficios esperados y dará un plan de acción para su implementación y ordenación.

La segunda etapa llamada ***«determinación de la infraestructura»*** permitirá definir la infraestructura técnica y organizativa del Data Warehouse.

La tercera etapa, conocida como ***«implementación de las aplicaciones»*** implementará las aplicaciones, una por una, y las desplegará.

La cuarta y última etapa nombrada ***«evaluación de resultados»*** la cual se refiere a evaluar los resultados obtenidos, en términos de costos y beneficios, por la implementación del Data Warehouse.

4.1.1 El descubrimiento y definición de las iniciativas

El método utilizado por esta primera etapa esta muy bien representada por la **Figura No. 12**. Primero definiremos los objetivos de cada una de las etapas presentadas en ella y posteriormente presentaremos los puntos clave necesarios para su desarrollo.

Figura No. 12

Los tres estadios superiores de la pirámide son fundamentales y afectan más a la empresa y a su estrategia que a la informática propiamente dicha; de ahí el término de «estudio estratégico» utilizado, que se encuentra bajo diferentes terminologías anglosajonas como Strategic Information Analysis ó Business Information Planning.

En nuestro contexto, ***el estudio estratégico***, se trata de :

- Informar y motivar a los actores de la empresa afectados por esta etapa inicial a fin de que se comprometan en el proceso de implementación de los proyectos de decisión: es la función de la etapa de sensibilización, de patrocinio (sponsorship), y de preparación para el cambio;

- Implicar a la dirección, a los equipos operativos y a los equipos informáticos, para que participen en la etapa de identificación y de comprensión de los objetivos de empresa que un Data Warehouse contribuiría a alcanzar; es la etapa de identificación y de comprensión de los desafíos;

- Identificar de manera más precisa los proyectos a realizar para alcanzar los objetivos identificados anteriormente.

A esta etapa le sigue la, **elaboración del plan de acción**, en la que se trata de:

- Asegurar la viabilidad de cada uno de los proyectos. En efecto, un Data Warehouse se basa en datos existentes, por lo que hay que asegurarse de su existencia y de su calidad. También deben tenerse en cuenta las restricciones técnicas (ejemplo: capacidad y voluntad de soportar grandes volúmenes de datos) u organizativas (ejemplo: impactos demasiado importantes o juzgados no realistas, inadecuación de las soluciones propuestas a los modos de trabajo de los equipos implicados por el proyecto).

- Estimar los recursos necesarios para la implementación de cada proyecto (costo de desarrollo, costo de programas y materiales, costo de administración, duración...), así como los recursos necesarios para la implementación de la infraestructura técnica y organizativa. Esta etapa permite también estimar las ganancias aportadas por cada realización, y por tanto evaluar el retorno sobre inversión de cada uno de los proyectos.
- Secuenciar y planificar la realización de los proyectos.

Estas etapas pueden parecer largas de implementar; así mismo, a fin de mejorar el tiempo, es tentador eliminarlas o reducirlas a su mínima expresión. El riesgo es llegar a la construcción de un Data Warehouse sin objetivo claramente definido, como un fin en sí mismo y no como una palanca tecnológica al servicio de los desafíos de la empresa.

Sin embargo, es posible limitar el objetivo de esta etapa a un sector o a un ámbito preciso para la empresa, lo que le quita su carácter transversal. El objetivo buscado puede ser el retorno rápido sobre inversión, o la prueba del concepto para partir hacia metas más ambiciosas. Esto equivale a constituir un prototipo operativo, con retornos sobre inversión medibles.

Los prototipos pueden destinarse a probar los beneficios funcionales del concepto o a demostrar que son tecnológicamente factibles las iniciativas futuras.

4.1.2 La determinación de la infraestructura

Se trata de determinar la infraestructura necesaria para la puesta en marcha del Data Warehouse. Esta etapa necesita evidentemente efectuar un cierto número de opciones tecnológicas, pero va más allá. Se trata también de determinar la infraestructura organizativa para su desarrollo, pero también de gestionar la conducción del cambio necesario para su constitución.

La determinación de la política tecnológica es fundamental. Muchos conceptos o características de las nuevas tecnologías, como la escalabilidad o los sistemas abiertos, son interesantes porque permiten elegir los componentes informáticos mejor adaptados a su contexto de uso. Identificar claramente esto es pues fundamental, más que elegir tal o cual producto.

La Infraestructura técnica. En este punto deben tomarse opciones tecnológicas, evidentemente en sincronía con la política más global de la empresa. Entre éstas, podemos citar :

- La elección del o de los proveedores de tecnologías: ¿hay que orientarse hacia um producto integrado, propuesto por un editor de programas como SAS, un constructor, o hacia ciertos integradores? ¿O mejor inclinarse por ensamblar las mejores ofertas para cada módulo de la arquitectura? Elegir la primera aproximación tiene la ventaja de facilitar considerablemente la definición de la política tecnológica y reducir los costos de integración en la

implementación, pero en este caso es conveniente asegurarse de que la solución elegida responderá a las necesidades actuales y futuras del proyecto. En el segundo caso, la flexibilidad aportada podrá aprovecharse para una adaptación ajustada a las necesidades de cada iniciativa y de cada grupo de usuarios. Habrá que considerar entonces el trabajo de integración necesario.

- La elección de las herramientas: ¿deben construirse, comprarse o aprovechar lo que se tiene? Esta problemática se presenta particularmente en los componentes considerados como opcionales, como son las herramientas de extracción o los referenciales. En efecto, la justificación económica, el cálculo de los riesgos inducidos por su ausencia sobre la perennidad del proyecto y la aportación en términos de mantenimiento de estas herramientas tendrán un eco importante en su elección.
- ¿Cómo se utilizará el Data Warehouse, por qué y cómo se estructurará la organización que lo usará? Las respuestas a estas preguntas permitirán determinar qué arquitectura se utilizará: centralizada (un Data Warehouse), distribuida (varios Data Marts, sin Data Warehouse físico), replicada (un Data Warehouse físico y varios Data Marts)? Asimismo, permitirán determinar si se justifican tecnologías como Internet o la intranet para el sistema, para qué tipos de necesidad y para qué usuarios.
- La elección de la estructura de almacenamiento: ¿debe ser relacional, multidimensional o híbrida (Data Warehouse relacional, Data Marts multidimensionales por ejemplo)? ¿Se dará primacía a la portabilidad o se acepta una fuerte adhesión a una herramienta, a un sistema o a un hardware para centrarse en el rendimiento o las funcionalidades?
- La elección del hardware: según los volúmenes contemplados, la población afectada, la arquitectura pensada y la flexibilidad esperada, pueden estudiarse diferentes gamas de hardware.
- La elección de las infraestructuras destinadas a la administración de sistemas, a la gestión de la seguridad, etc.

Es evidente que será necesario asegurarse de que todas las soluciones elegidas funcionan entre sí. En el mejor de los casos. descrito en los folletos comerciales, todas son compatibles. Pero la realidad aporta su paquete de decepciones: es preferible asegurarse también de que existen referencias comunes entre los editores seleccionados y de que el soporte técnico propuesto sabrá resolver las incompatibilidades eventuales entre herramientas heterogéneas.

La Infraestructura organizativa. Paralelamente a estas elecciones, será necesario determinar la logística y la organización necesarias para la concreción de las iniciativas. El equipo responsable del Data Warehouse deberá diseñar cada iniciativa, desarrollarla y utilizarla. En este estadio, destaquemos que a nivel técnico un Data Warehouse es ante todo una mecánica de gestión de flujos de datos. Los equipos de desarrollo y uso se organizan a menudo respecto a lo siguiente:

- Un primer centro de competencias tiene la responsabilidad del proceso de alimentación de datos de los sistemas de producción hacia el Data Warehouse. Los individuos que constituyen el equipo deberán tener un buen conocimiento de los sistemas afectados, tanto en el plano técnico como funcional.
- El segundo centro de competencias está encargado de la gestión y el soporte del Data Warehouse propiamente dicho. Los administradores de datos y de bases de datos especialmente constituirán este equipo.
- El último centro de competencia es responsable de los flujos de informaciones que transitan entre los usuarios y su puesto de trabajo por un lado, y el Data Warehouse por el otro.

Los implicados podrán formar parte alternativamente de una u otra de las tres celdas. Esto es particularmente válido para quienes se encargan de la administración de sistemas y de la seguridad, pero también para quienes disponen de un buen conocimiento de los ámbitos funcionales.

Conducir el cambio. Según la experiencia de la empresa en el ámbito de la ayuda a la decisión, por una parte, y de las tecnologías y herramientas utilizadas, por otra, será necesario un plan de formación. También será muy importante que los miembros del equipo sean favorables al cambio, debido a las especificidades de los proyectos de decisión. Un desarrollador, por ejemplo, no podrá trabajar ante un informe de programación, sino que deberá tomar iniciativas, saber comunicarse con los usuarios y comprender sus necesidades en términos funcionales. La formación y, más generalmente, la conducción de cambio no deben negligirse, bajo ningún concepto, para todos los actores afectados por el proyecto.

Para un equipo informática, por ejemplo, una de las características de un proyecto de decisión es que es relativamente poco complejo de implementar técnicamente, pero está sembrado de trampas. Por ejemplo, supongamos que un programa de carga mal diseñado necesita horas para su ejecución. Si el equipo controla el entorno, en unos segundos el especialista del ámbito podrá disminuir este tiempo en un factor muy significativo. Si no lo controla, el problema bloquea el proceso y el proyecto encaja un retraso.

La gestión del cambio no afecta únicamente a los informáticos; también es necesario, en esta etapa, preparar el plan de gestión del cambio e identificar el o los promotores (sponsors) cuya función será hacer que lo acepten los actores afectados. Según la dimensión del proyecto global, podrán necesitarse varios promotores. En particular, a menudo es deseable identificar un promotor por iniciativa, y cada uno de ellos se asocia en general a una entidad operativo (mercadotecnia, comercial, logística, finanzas, recursos humanos...etc.)

4.1.3 La implementación de las aplicaciones

Esta etapa se realiza para cada iniciativa que la etapa de identificación de las necesidades ha permitido delimitar. El método propuesto aquí es un método en cinco etapas:

- Una etapa de especificaciones, que define y planifica las etapas siguientes de manera más precisa y detallada que en las etapas precedentes;
- Una etapa de diseño;
- Una etapa de implementación e integración;
- Una etapa de despliegue, integrando la conducción del cambio;
- Una etapa de medidas.

Figura No. 13

Esta división en etapas representada por la **Figura No. 13** resultará familiar a quienes tengan experiencia en la implementación de proyectos informáticos. Pero hay que insistir particularmente en las dos últimas etapas.

Se ha destacado ya la importancia de la conducción del cambio, pero aún no se ha tratado la importancia del control del despliegue. Un Data Warehouse utiliza implícitamente tecnologías cliente/servidor. Si bien hay acuerdo unánime sobre los menores costes del clientelservidor comparados con

los de entornos más tradicionales, todos los estudios demuestran que esta tecnología puede inducir costos ocultos. Si estos costos son ocultos es porque no se controlan, y esta falta de control afecta especialmente a la etapa de despliegue, en general ignorada o insuficientemente industrializada. Cuidado, pues, con caer en esta trampa.

Dos artículos aparecidos en el Journal of Data Warehousing ([SKE96], [LOV96]) presentan esencialmente los éxitos en este ámbito en términos de retorno sobre inversiones, ganancias de partes del mercado, reducción de stocks, etc. Si el balance de la implementación de una aplicación Data Warehouse puede anunciarse en estos términos, es una palanca para los desarrollos siguientes. Esta etapa de medición es la que debe aportar este tipo de información. El Data Warehouse es en sí mismo la herramienta ideal para efectuar estas medidas, siempre que se haya tenido en cuenta este objetivo en su diseño, porque los datos que reúne son fechados y no volátiles.

La etapa de medición permite hacer también balance de la realización y capitalizar los éxitos y fracasos encontrados durante el desarrollo de la aplicación. Idealmente, esta etapa se repite regularmente, para hacer un seguimiento de las mediciones y determinar las necesidades de mejora de la aplicación.

Hasta ahora, la etapa de implementación se ha presentado como una sucesión secuencias de etapas. Evidentemente, también puede desarrollarse de manera iterativa, según una lógica de tipo RAD ([MAR91]).

Puede ser también muy provechoso descomponer esa etapa en 2 subetapas. La primera se destina a desarrollar un prototipo y la segunda al despliegue a una escala mas importante.

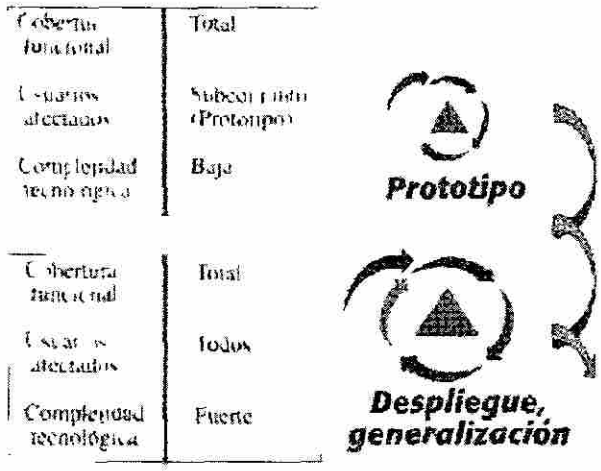


Figura No. 14

Como se muestra en la **Figura No. 14**, el objetivo de este método es empezar por un proyecto de tamaño modesto reduciendo al mínimo su complejidad tecnológica. En la distribución a gran escala, por ejemplo, el prototipo podrá afectar sólo a algunos almacenes cuyos sistemas sea relativamente homogéneos. La complejidad tecnológica se reduce y será más fácil centrarse en el aspecto funcional. Una vez realizado el prototipo, podrá ser probado y utilizado por usuarios piloto, para posteriormente, tras haber satisfecho y probado su valor, el proyecto podrá desplegarse a mayor escala y, en esta etapa, la problemática se expresará esencialmente en términos de implementación técnica y de conducción del cambio para los usuarios afectados por el despliegue.

4.1.4 La Evaluación de los Resultados

En esta etapa se deben evaluar los resultados en términos de costos y beneficios obtenidos por el Data Warehouse. Estos costos y beneficios no se deben de considerar únicamente en relación a los gastos e inversiones realizadas para la construcción del Data Waarehouse, ni tampoco en relación a los beneficios inmediatos en la toma de decisiones que permitirá hacer el Data Warehouse.

Este análisis debe ser amplio y basarse mas bien en escenarios, uno de los cuales será el escenario de tener el Data Warehouse y otro el de no contar con el Data Warehouse. Para ambos casos se debe tomar en cuenta tanto la parte económica como la parte involucrada en la satisfacción y eficiencia obtenida por parte de los usuarios en cada uno de los escenarios, y lo que esto involucra en términos económicos (rotación de personal, retrabajos, horas extras, etc).

La evaluación debe proyectarse a corto, mediano y largo plazo, para de esa manera reforzar el apoyo que se pudiera otorgar a futuros proyectos de Data Warehouse; ya que dichos proyectos necesitarán patrocinadores bien convencidos de los beneficios reales que se pueden obtener al realizarlos.

4.2 Etapa 1: Descubrimiento y definición de las iniciativas

Esta etapa esta mas bien relacionada con la parte de viabilidad del proyecto de Data Warehouse y afecta mas a la empresa que a la informática propiamente dicha. Sin embargo existen en ella aspectos informáticos de gran importancia. Dichos aspectos representarán los temas abordados en esta guía, ya que obviamente serán a los que se les prestará interés en este documento.

4.2.1 Diseño del Data Warehouse

Cuatro características del Data Warehouse tienen efectos determinantes sobre el método de diseño de un proyecto de este tipo.

La primera está vinculada a las evoluciones tecnológicas recientes. El cliente/servidor y los sistemas abiertos, por ejemplo, tecnologías implícitamente utilizadas en los sistemas Data Warehouse, han aportado evoluciones fundamentales: un sistema de información puede construirse por integración de un cierto número de componentes, pudiendo ser elegido cada uno en relación con su contexto de uso. Como las soluciones de software ya no son monolíticas, cada empresa tiene la oportunidad de definir su arquitectura, y no adaptar el problema a tal o cual tecnología del mercado. En este marco, los métodos de implementación también deben adaptarse al contexto del proyecto. En el ámbito de las metodologías y las técnicas de implementación, por

ejemplo, es posible elegir uno o más métodos, por ejemplo Merise, Information Engineering, los métodos orientados a objetos, etc. Esta diversificación puede verse como una oportunidad o como una constatación de fracaso, pero, en cualquier caso, es difícil pretender conocer métodos universales independientes del contexto de uso.

La segunda es que un Data Warehouse está mucho más cerca de la estrategia de una empresa. El Data Warehouse esta muy apegado a la estrategia de la empresa a diferencia de lo que pueden estarlo generalmente las aplicaciones de carácter transaccional. Mientras que éstas permiten a menudo la automatización de procesos existentes, o descritos formalmente por adelantado, el objetivo del Data Warehouse se expresa a menudo en términos puramente de negocio como: mantener la fidelidad de la clientelas. En su desarrollo, habrá que tener en cuenta estos aspectos, implicando al máximo a los usuarios más experimentados en el conocimiento de su empresa y/o de su negocio, pero también integrando esta dimensión en todas las técnicas utilizadas para el diseño y el seguimiento del proyecto.

La tercera se desprende del hecho que un Data Warehouse, una vez construido, debe evolucionar. El Data Warehouse debe evolucionar en función de las peticiones de los usuarios o de los nuevos objetivos de la empresa, y se sitúa, pues, en una lógica de mejora imprevisible y frecuente.

Finalmente, el nivel de madurez de cada empresa ante los sistemas de decisión puede diferir considerablemente. Para algunas empresas, el Data Warehouse está en continuidad con su adquisición de ayuda a la decisión, que les permite disponer ya de una organización y de métodos probados, mientras que para otras se trata de un ámbito aún desconocido.

A pesar que el diseño del Data Warehouse es diferente al usado en los diseños tradicionales, no es menos importante. El hecho que los usuarios finales tengan

dificultad en definir lo que ellos necesitan, no lo hace menos necesario. En la práctica, los diseñadores de Data Warehouse tienen que usar muchos "trucos" para ayudar a sus usuarios a "visualizar" sus requerimientos. Por ello, son esenciales los prototipos de trabajo.

4.2.2 Planificación del Data Warehouse

La planificación es el proceso más importante que determina el tipo de estrategias de Data Warehousing que una organización iniciará.

No existe una fórmula que garantice el éxito de la construcción de un Data Warehouse, pero hay muchos puntos que contribuyen a ese objetivo. Algunos puntos claves que deben considerarse en la planificación de un Data Warehouse son los siguientes :

1. Establecer una asociación de usuarios, gestión y grupos. Es esencial involucrar tanto a los usuarios como a la gestión para asegurar que el Data Warehouse contenga información que satisfaga los requerimientos de la empresa. La gestión puede ayudar a priorizar la etapa de la implementación del Data Warehouse, así como también la selección de herramientas del usuario. Los usuarios y la gestión justifican los costos del Data Warehouse sobre cómo será "su ambiente" y está basado primero en lo esperado y segundo, en el valor comercial real.

2. Seleccionar una aplicación piloto con una alta probabilidad de éxito. Una aplicación piloto de alcance limitado, con un reembolso medible para los

usuarios y la gestión, establecerá el Data Warehouse como una tecnología clave para la empresa. Estos mismos criterios (alcance limitado, reembolso medible y beneficios claros para la empresa) se aplican a cada etapa de la implementación de un Data Warehouse.

3. Trabajar mediante prototipos con cambios rápidos y frecuentes. La mejor manera para asegurar que el Data Warehouse reúna las necesidades de los usuarios, es hacer el prototipo a lo largo del proceso de implementación y aún más allá, así como agregar los nuevos datos y/o los modelos en forma permanente. El trabajo continuo con los usuarios y la gestión es, nuevamente, la clave.

4. Realizar una Implementación de manera incremental. La implementación incremental ayuda a la reducción de riesgos y asegura que el tamaño del proyecto permanezca manejable en cada etapa.

5. Retroalimentar constantemente y publicar los casos exitosos. La retroalimentación de los usuarios ofrece una excelente oportunidad para publicar los hechos exitosos dentro de una organización. La publicidad interna sobre cómo el Data Warehouse ha ayudado a los usuarios a operar más eficientemente puede apoyar la construcción del Data Warehouse a lo largo de una empresa. La retroalimentación del usuario también ayuda a comprender cómo evoluciona la implementación del Data Warehouse a través del tiempo para reunir requerimientos de usuario nuevamente identificados.

4.2.3 Selección del Data Warehouse a construir

Antes de intentar construir un Data Warehouse, es crítico el establecer de una estrategia que sea apropiada para las necesidades y los usuarios.

Las preguntas básicas que deben hacerse son:

- ¿ A quien va dirigido ?
- ¿ Cuál será el alcance ?
- ¿ Qué tipo de Data Warehouse debería construirse ?

Existe varias estrategias mediante las cuales las organizaciones pueden construir sus Data Warehouses, de las cuales se explicarán a continuación las mas comunes.

Estrategia del Data Warehouse virtual. Esta estrategia esta basada en el uso actual. Se crea un Data Warehouse físico para soportar los pedidos de alta frecuencia. Puede ser construido mediante:

1. La Instalación de un conjunto de facilidades para acceso a datos, directorio de datos y gestión de proceso.
2. El entrenamiento de usuarios finales.
3. El control de cómo se usan realmente las configuraciones del Data Warehouse.

Estrategia de la copia de datos operacionales. Esta estrategia se basa en obtener una copia de los datos desde un sistema operacional único y dotar al Data Warehouse de una serie de herramientas de acceso a la información. Esta estrategia tiene la ventaja de ser simple y rápida. Desafortunadamente, si los

datos existentes son de mala calidad y/o el acceso a los datos no ha sido previamente evaluado, el Data Warehouse puede presentar serios problemas.

Estrategia en base a análisis de necesidades y consensos. Esta estrategia se basa en seleccionar un número de usuarios basados en el valor de la empresa y hacer un análisis de sus puntos, preguntas y necesidades de acceso a datos. De acuerdo a estas necesidades, se construyen los prototipos de Data Warehousing y se prueban para que los usuarios finales puedan experimentar y modificar sus requerimientos. Una vez se tenga un consenso general sobre las necesidades, entonces se consiguen los datos provenientes de los sistemas operacionales existentes a través de la empresa y/o desde fuentes externas de datos y se cargan al Data Warehouse. Si se requieren herramientas de acceso a la información, se puede también permitir a los usuarios finales tener acceso a los datos requeridos usando sus herramientas favoritas propias, o facilitar la creación de sistemas de acceso a la información multidimensional de alto desempeño, usando el núcleo del Data Warehouse como base.

En conclusión, no se tiene un enfoque único para elaborar un Data Warehouse que se adapte a las necesidades de las empresas, debido a que las necesidades de cada una de ellas son diferentes, al igual que su contexto. Además, como la tecnología Data Warehousing esta en constante evolución, el enfoque mas práctico para la construcción del Data Warehouse es mantener una evolución constante en la utilización de las técnicas y herramientas que vayan surgiendo, en aras de un mejor desempeño en el almacenamiento y consulta de la información proporcionada por el Data Warehouse.

4.2.4 Administración y gestión del Data Warehouse

Un proyecto de Data Warehouse acarrea inevitablemente un problema de administración de los datos. La administración de los datos (DWA, Data Warehouse Administration) del Data Warehouse es una función que debe declinarse a varios niveles: por ejemplo, empresa, ámbito, proyecto. No hay reglas universales sobre el nombre y la naturaleza de estos niveles para obtener un sistema consolidado y coherente. Detallartemos los diferentes niveles de administración:

DWA centralizada: Las normas deben definirse a nivel de la empresa; el modelo de datos es global a la empresa.

DWA descentralizada: Data Marts Administration (DMA): en este caso, se podrá organizar la administración por ámbitos funcionales (Contabilidad, Mercadotecnia...), por ámbitos de aplicación (Facturación, Surtido ...), por entornos técnicos (Sede central, micros en agencias, IBM, DEC ...) o por sedes geográficas (Data Marts para Alemania, para Estados Unidos ...).

Administración local de proyecto: Se encuentra aquí un caso clásico de gestión de los datos de una aplicación. Los modelos de datos son locales y describen generalmente datos operativos.

Para que un Data Warehouse siga vivo, debe instituirse la función de administración de los datos para todo proyecto encargado de desarrollar una aplicación que deba alimentar el Data Warehouse.

El Data Warehouse requiere una comercialización y gestión muy cuidadosa. Debe considerarse lo siguiente:

- Un Data Warehouse es una inversión buena, sólo si los usuarios finales realmente pueden conseguir información vital más rápida y más barata de lo que obtienen con la tecnología actual. Como consecuencia, la gestión tiene que pensarse seriamente sobre cómo quieren sus depósitos para su eficaz desempeño y cómo conseguirán llegar a los usuarios finales.
- La Administración debe reconocer que el mantenimiento de la estructura de datos del Data Warehouse es tan crítico como el mantenimiento de cualquier otra aplicación de misión crítica. De hecho, la experiencia ha demostrado que el Data Warehouse llegará a ser rápidamente uno de los sistemas más usados en cualquier organización.
- La gestión debe comprender también que si se deciden por un proyecto de Data Warehouse, se crearán nuevas demandas sobre sus sistemas operacionales, como lo son:
    1. Demandas para mejorar datos
    2. Demandas para datos consistentes
    3. Demandas para diferentes tipos de datos, etc.

4.3 Etapa 2: Determinación de la Infraestructura

En esta etapa se trata de seleccionar la infraestructura desde 2 ámbitos principales : el técnico y el organizativo. Los temas que se desarrollarán en este punto están enfocados a ayudar a tomar decisiones respecto a la formación de cada una de las infraestructuras en los ámbitos mencionados.

4.3.1 Alcance del Data Warehouse

El alcance de un Data Warehouse puede ser tan amplio como toda la información estratégica de la empresa desde su inicio, o puede ser tan limitado como un Data Warehouse personal para un solo gerente durante un año.

En la práctica, en la amplitud del alcance, el mayor valor del Data Warehouse es para la empresa y lo más caro y requirente de tiempo es crearlo y mantenerlo. Como consecuencia de ello, la mayoría de las organizaciones comienzan con Data Warehouses funcionales, departamentales o divisionales y luego los expanden como usuarios que proveen retroalimentación.

La **Figura No. 15** muestra un esquema bidimensional para analizar las opciones básicas. La dimensión horizontal indica el alcance del depósito y la vertical muestra la cantidad de datos redundantes que deben almacenarse y mantenerse.

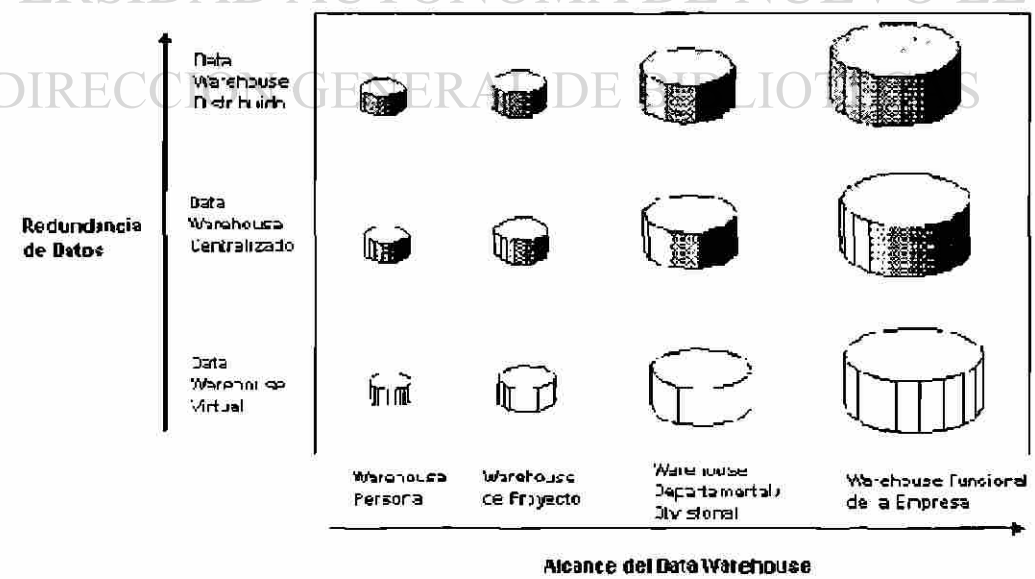


Figura No. 15

4.3.2 Arquitectura del Data Warehouse

Un Data Warehouse está integrado por un servidor de hardware y los DBMS que conforman el depósito. Del lado del hardware, se debe combinar la configuración de plataformas de los servidores, mientras se decide cómo aprovechar los saltos casi constantes de la potencia del procesador. Del lado del software, la complejidad y el alto costo de los DBMSes fuerzan a tomar decisiones drásticas y balances comparativos inevitables, con respecto a la integración, requerimientos de soporte, desempeño, eficiencia y confiabilidad.

Si se escoge incorrectamente, el Data Warehouse se convierte en una gran empresa con problemas difíciles de trabajar en su entorno, costoso para arreglar y difícil de justificar. Para conseguir que la implementación del depósito tenga un inicio exitoso, se necesita enfocar hacia tres bloques claves de construcción:

- Configuración del depósito
- Configuración del servidor
- Sistemas de Gestión de Base de Datos

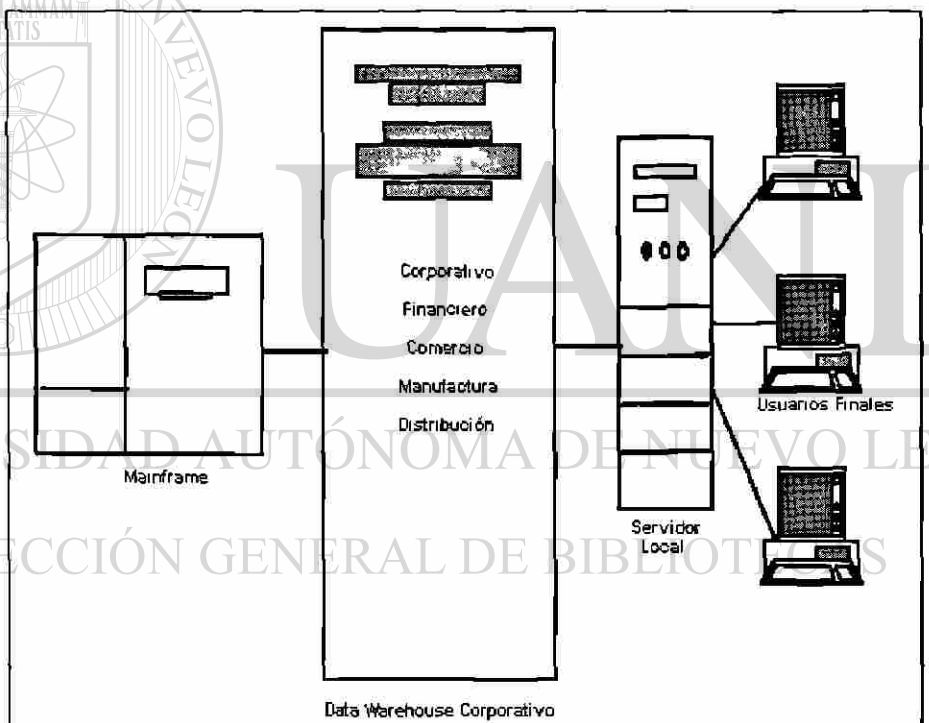
4.3.3 Configuración del depósito

La elaboración del Data Warehouse comienza con la estructura lógica y física de la base de datos del depósito más los servicios requeridos para operar

y mantenerlo. Esta elección conduce a la selección de otros dos puntos fundamentales: el servidor de hardware y el DBMS.

La plataforma física puede centralizarse en una sola ubicación o distribuirse regional, nacional o internacionalmente. A continuación se dan las siguientes alternativas de arquitectura:

Data Warehouse integrado : Consiste en generar un plan para almacenar los datos de su compañía, que podrían obtenerse desde fuentes múltiples internas y externas, y consolidarlos en una sola base de datos (el Data Warehouse). El enfoque consolidado proporciona eficiencia tanto en la potencia de procesamiento como en los costos de soporte. **(Ver Figura No. 16).**



En una arquitectura centralizada, una sola, el data warehouse integrado refleja todos los aspectos del negocio. Las bases de datos separadas son todas interrelacionadas y físicamente almacenadas en la misma plataforma.

Figura No. 16

Data Warehouse distribuido: Consiste en distribuir la información por función, con datos financieros sobre un servidor en un sitio, los datos de comercialización en otro y los datos de fabricación en un tercer lugar. (Ver la **Figura No. 17**)

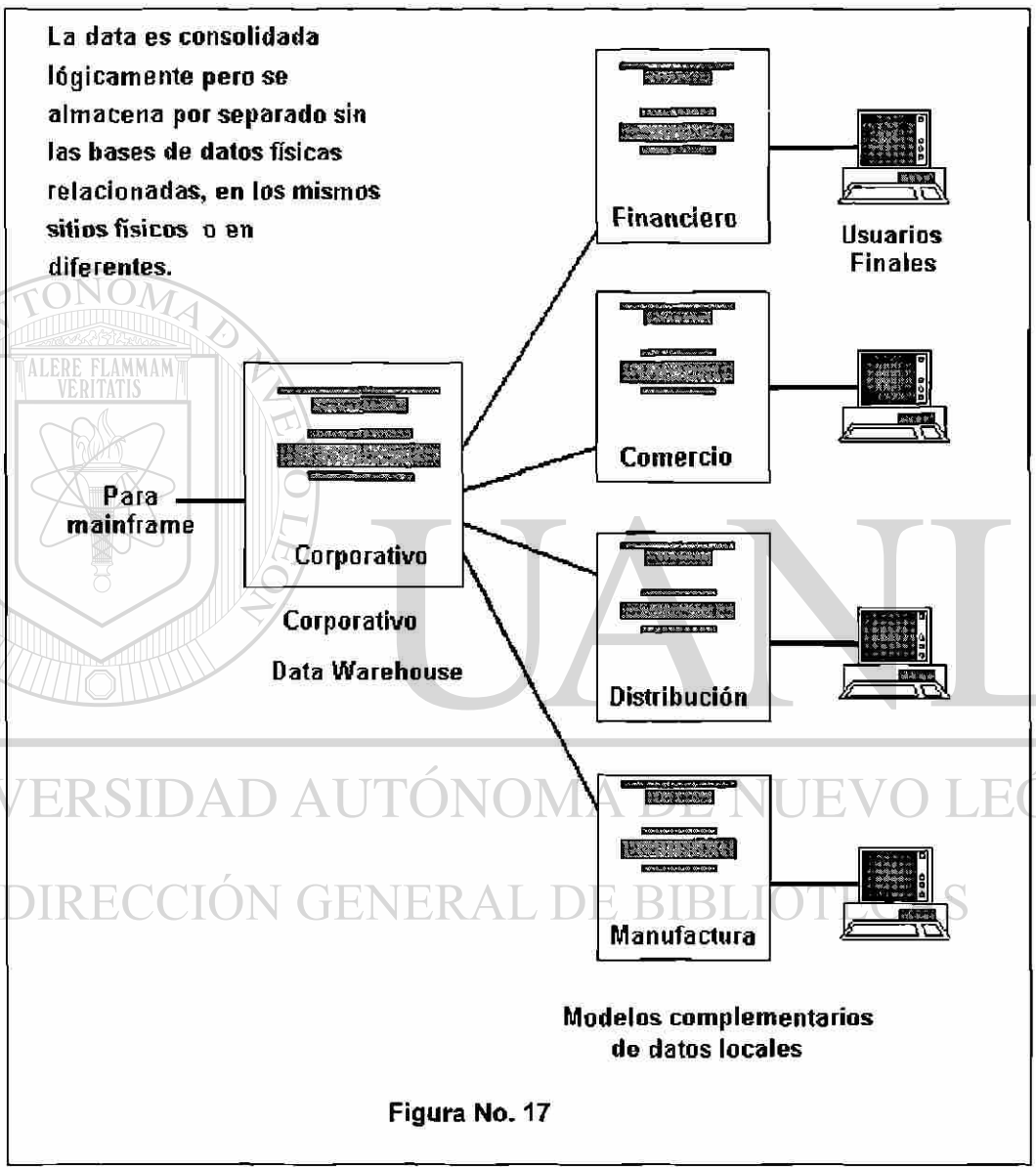
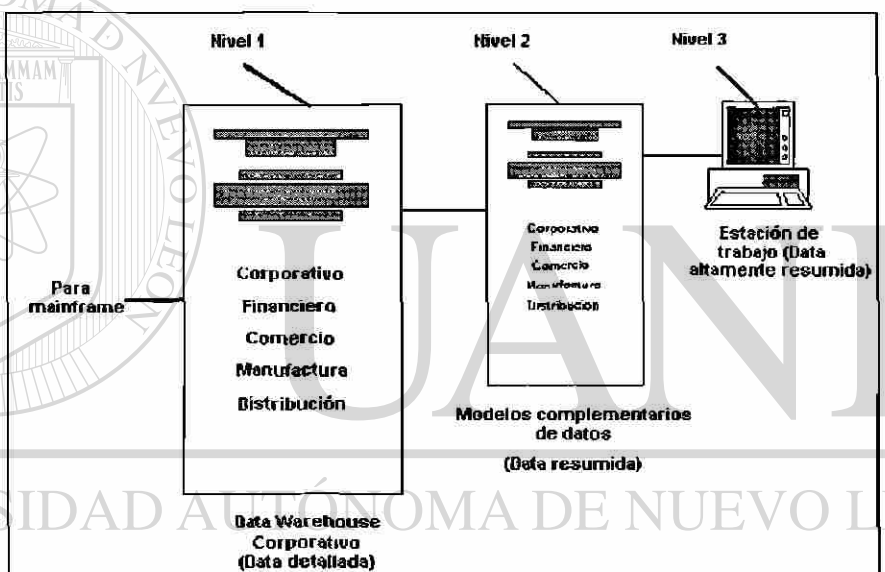


Figura No. 17

Data Warehouse por niveles: Almacena datos altamente resumidos sobre una estación de trabajo del usuario, con resúmenes más detallados en un segundo servidor y la información más detallada en un tercero. La estación de trabajo del primer nivel maneja la mayoría de los pedidos para los datos, con pocos pedidos que pasan sucesivamente a los niveles 2 y 3 para la resolución. Las computadoras en el primer nivel pueden optimizarse para usuarios de carga pesada y volumen bajo de datos, mientras que los servidores de los otros niveles son más adecuados para procesar los volúmenes pesados de datos, pero cargas más livianas de usuario. (Ver la **Figura No. 18**).



La data es dividida por niveles de detalle. El nivel 1 de servidores satisfacen la mayoría de los pedidos de los usuarios.

Figura No. 18

4.3.4 Configuración del servidor

Al decidir sobre una configuración de depósito distribuida o centralizada, también se necesita considerar los servidores que retendrán y entregarán los datos. El tamaño de su implementación (y las necesidades de su empresa para escalabilidad, disponibilidad y gestión de sistemas) influirá en la elección de la arquitectura del servidor.

Servidores de un solo procesador : Los servidores de un sólo procesador son los más fáciles de administrar, pero ofrecen limitada potencia de procesamiento y escalabilidad. Además, un servidor sólo presenta un único punto de falla, limitando la disponibilidad garantizada del depósito. Se puede ampliar un solo servidor de redes mediante arquitecturas distribuidas que hacen uso de subproductos, tales como Ambientes de Computación Distribuida (Distributed Computing Environment - DCE) o Arquitectura Broker de Objeto Común (Common Objects Request Broker Architecture - CORBA), para distribuir el tráfico a través de servidores múltiples. Estas arquitecturas aumentan también la disponibilidad, debido a que las operaciones pueden cambiarse al servidor de backup si un servidor falla, pero la gestión de sistemas es más compleja.

Multiprocesamiento simétrico : Las máquinas de multiprocesamiento simétrico (Symmetric MultiProcessing - SMP) aumentan mediante la adición de procesadores que comparten la memoria interna de los servidores y los dispositivos de almacenamiento de disco. Se puede adquirir la mayoría de SMP en configuraciones mínimas (es decir, con dos procesadores) y levantar cuando es necesario, justificando el crecimiento con las necesidades de procesamiento. La escalabilidad de una máquina SMP alcanza su límite en el número máximo de procesadores soportados por los mecanismos de conexión (es decir, el backplane y bus compartido).

Procesamiento en paralelo masivo : Una máquina de procesamiento en paralelo masivo (Massively Parallel Processing - MPP), conecta un conjunto de procesadores por medio de un enlace de banda ancha y de alta velocidad. Cada nodo es un servidor, completo con su propio procesador y memoria interna (posiblemente SMP) para optimizar una arquitectura MPP, las aplicaciones deben ser "paralelizadas" es decir, diseñadas para operar por separado, en partes paralelas. Esta arquitectura es ideal para la búsqueda de grandes bases de datos. Sin embargo, el DBMS que se selecciona debe ser uno que ofrezca una versión paralela. Y aún entonces, se requiere un diseño y afinamiento esenciales para obtener una óptima distribución de los datos y prevenir problemas.

Acceso de memoria no uniforme : La dificultad de mover aplicaciones y los DBMS a agrupaciones o ambientes realmente paralelos ha conducido a nuevas y recientes arquitecturas, tales como el acceso de memoria no uniforme (Non Uniform Memory Access - NUMA). NUMA crea una sola gran máquina, al conectar múltiples nodos en un solo (aunque físicamente distribuida) banco de memoria y un ejemplo único de OS. NUMA crea una sola gran máquina SMP al conectar múltiples nodos SMP en un solo (aunque físicamente distribuida) banco de memoria y un ejemplo único de OS. NUMA facilita el enfoque SMP para obtener los beneficios de performance de las grandes máquinas MPP (con 32 o más procesadores), mientras se mantiene las ventajas de gestión y simplicidad de un ambiente SMP estándar.

Lo más importante de todo, es que existen DBMS y aplicaciones que pueden moverse desde un solo procesador o plataforma SMP a NUMA, sin modificaciones.

4.3.5 Sistemas de Gestión de Base de Datos (SGBD)

Los Data Warehouses (conjuntamente con los sistemas de soporte de decisión [Decision Support Systems - DSS] y las aplicaciones cliente/servidor), fueron los primeros éxitos para el DBMS relacional (Relational Data Base Management Systems - RDBMS).

Mientras la gran parte de los sistemas operacionales fueron resultados de aplicaciones basadas en antiguas estructuras de datos, los depósitos y sistemas de soporte de decisiones aprovecharon el RDBMS por su flexibilidad y capacidad para efectuar consultas con un único objetivo concreto.

Los RDBMS son muy flexibles cuando se usan con una estructura de datos normalizada. En una base de datos normalizada, las estructuras de datos son no redundantes y representan las entidades básicas y las relaciones descritas por los datos (por ejemplo productos, comercio y transacción de ventas). Pero un procesamiento analítico en línea (OLAP) típico de consultas que involucra varias estructuras, requiere varias operaciones de unión para colocar los datos juntos.

El desempeño de los RDBMS tradicionales es mejor para consultas basadas en claves ("Encuentre cuenta de cliente #2014") que para consultas basadas en el contenido ("Encuentre a todos los clientes con un ingreso sobre $ 10,000 que hayan comprado un automóvil en los últimos seis meses").

Para el soporte de depósitos a gran escala y para mejorar el interés hacia las aplicaciones OLAP, los proveedores han añadido nuevas características al RDBMS tradicional. Estas, también llamadas características super relacionales, incluyen el soporte para hardware de base de datos especializada, tales como la máquina de base de datos Teradata.

Los modelos super relacionales también soportan extensiones para almacenar formatos y operaciones relacionales (ofrecidas por proveedores como RedBrick) y diagramas de indexación especializados, tales como aquellos usados por Sybase IQ. Estas técnicas pueden mejorar el rendimiento para las recuperaciones basadas en el contenido, al prejuntar tablas usando índices o mediante el uso de listas de índice totalmente invertidos.

Muchas de las herramientas de acceso a los Data Warehouses explotan la naturaleza multidimensional del Data Warehouse. Por ejemplo, los analistas de marketing necesitan buscar en los volúmenes de ventas por producto, por mercado, por período de tiempo, por promociones y niveles anunciados y por combinaciones de estos diferentes aspectos.

La estructura de los datos en una base de datos relacional tradicional, facilita consultas y análisis a lo largo de dimensiones diferentes que han llegado a ser comunes. Estos esquemas podrían usar tablas múltiples e indicadores para simular una estructura multidimensional. Algunos productos DBMS, tales como Essbase y Gentium, implementan técnicas de almacenamiento y operadores que soportan estructuras de datos multidimensionales.

Mientras las bases de datos multidimensionales (MultiDimensional Databases - MDDBs) ayudan directamente a manipular los objetos de datos multidimensionales (por ejemplo, la rotación fácil de los datos para verlos entre dimensiones diferentes, o las operaciones de drill down que sucesivamente exponen los niveles de datos más detallados), se debe identificar estas dimensiones cuando se construya la estructura de la base de datos. Así, agregar una nueva dimensión o cambiar las vistas deseadas, puede ser engorroso y costoso. Algunos MDDBs requieren un recargue completo de la base de datos cuando ocurre una reestructuración.

4.3.6 Ambiente OLTP vs OLAP

La mayor parte de los informáticos dominan aproximaciones para la implementación de sistemas de informaciones, normalmente centrados en metodologías como Merise o Information Engineering. En sus componentes relacionados con el modelado de datos, estas metodologías son a la vez precisas, potentes y bastante poco cuestionadas. A nivel de los modelos de datos, el modelo entidad-relación es el más utilizado, llevando normalmente a la creación de un modelo lógico de tipo relacional. Tanto si se conocen de manera formal los conceptos como si no, todas las teorías asociadas a estos modelos son utilizadas ampliamente en las empresas. Estas técnicas están actualmente tan ancladas en nuestras mentes que a menudo olvidamos su origen. Aparecieron cuando la informática estaba destinada a la automatización de la producción y se utilizan aún con éxito en contextos de este tipo, por ejemplo para las aplicaciones de carácter transaccional, comúnmente llamadas OLTP. Sin embargo, la informática de decisión, que algunos denominan también OLAP, justifica una revisión de los métodos de diseño de un modelo de datos.

Caracteristicas de un contexto OLTP

En la mayor parte de los sistemas transaccionales, el papel de un modelo es garantizar la persistencia de los datos. De hecho, la base de datos se diseña para conservar el rastro de eventos surgidos en la empresa.

Tomemos el ejemplo de una aplicación de gestión de pedidos. En su interior, para evitar la redundancia de información, será necesario mantener

información sobre los pedidos, los clientes y los productos en entidades distintas, relacionadas entre sí por asociaciones. Conceptos más sofisticados, como las nociones de forma normal, de clave única, de clave foránea o de restricción de integridad referencial, permiten garantizar constantemente la integridad de la base de datos. El origen de este esfuerzo de minimización de redundancias deriva principalmente del hecho que los sistemas transaccionales efectúan sus actualizaciones en tiempo real, eventualmente a través de un conjunto de aplicaciones que comparten el mismo modelo de datos.

En un sistema transaccional, el diseño se orienta a procesos y el modelo de datos interviene como apoyo de éste. Desde el punto de vista del usuario, el modelo de datos es totalmente transparente, sólo accede a él indirectamente a través de aplicaciones «empaquetadas» puestas a su disposición. Por ello, la legibilidad del modelo desde punto de vista del usuario no tiene gran importancia, excepto en ocasiones durante la fase de validación del modelo con el usuario.

Las consultas son siempre previsibles, porque se efectúan a través de una aplicación desarrollada normalmente por el mismo equipo encargado del modelo de datos. Por ello, es posible definir un modelo de datos adaptado a las consultas que el sistema será susceptible de lanzar por anticipación. Así, la mayoría de benchmarks (conjuntos de pruebas) que permiten cualificar el rendimiento de una plataforma transaccional se definen a través de un conjunto de transacciones, en general poco numerosas, que serán los únicos puntos de entrada al sistema.

En este contexto, se accede a los datos generalmente por claves, especialmente claves únicas (acceso de un cliente por su número de cuenta). Una buena indexación permite garantizar tiempos de respuesta dependiendo más del volumen de datos a tratar para realizar la transacción que del volumen global de la base de datos. Los volúmenes de datos como resultado de las

transacciones son limitados. Normalmente, el número de entradas/salidas asociadas a cada transacción es un número finito previsible. En ciertos casos, como para las consultas guiadas, este número es más difícilmente previsible; así, la búsqueda de un cliente por su nombre puede llevar a un volumen más o menos importante de resultados. En todo momento, una aplicación transaccional tiende a la productividad, y presentar a un usuario varios miles de registros en este contexto no tiene ningún sentido.

Para efectuar los tratamientos necesarios para su correcta ejecución, una transacción accederá a un número de estructuras limitado y finito. Es muy raro que una consulta transaccional necesite reunir o agregar informaciones surgidas de un gran número de tablas.

Características de un contexto OLAP

Un Data Warehouse es una base dedicada a la decisión. La información se pone a disposición de los usuarios, pero las actualizaciones no se hacen nunca en tiempo real. Por ello, las únicas actualizaciones efectuadas sobre el Data Warehouse provendrán de los sistemas de producción, en las fases de carga. Una vez efectuado este proceso de adquisición de datos, la integridad de datos del Data Warehouse no podrá volver a cuestionarse. Es posible, pues, introducir redundancias, siempre que se controlen en el proceso de alimentación.

En un contexto de decisión, las consultas manipulan regularmente conjuntos. Por ejemplo, el usuario se interesará por las ventas realizadas en la región Norte. Las consultas efectúan frecuentemente selecciones o restricciones de población, agrupaciones, cálculos, agregados, etc. Para responder a las necesidades de los usuarios, aunque el resultado de las consultas puede estar constituido únicamente por algunas líneas, a menudo será necesario manipular volúmenes importantes. En este caso, obtener

tiempos de respuesta proporcionales al volumen de datos resultado de una consulta es mucho más difícil que en transaccional. Los optimistas relativizan esta constatación, partiendo del principio que la decisión da autonomía al usuario y ello le incita a una mayor indulgencia ante los tiempos de respuesta: el Data Warehouse le permite hacer en unos minutos lo que antes hacía en varios días.

La realidad es más compleja: un usuario difícilmente comprenderá una espera de varias decenas de minutos para obtener las ventas del año pasado por regiones, porque se trata de una consulta básica que se ejecutará frecuentemente. Por el contrario, esperar cierto tiempo para obtener información sobre las «ventas de tal producto en tal almacén el último viernes antes de las vacaciones de agosto» se tolerará mejor.

En este contexto, es necesario intentar optimizar las consultas efectuadas frecuentemente. Ello es posible siempre que se predefinan físicamente los subconjuntos de datos, menos importantes en tamaño que los datos más detallados, pero suficientes para resolver las consultas más habituales.

Otra característica de la decisión es que los usuarios intentan relacionar con frecuencia elementos que a priori no se correlacionan al principio. Desearán por ejemplo relacionar las ventas con los gastos, comparar las ventas de un período respecto a otro... Para conseguirlo, son necesarias consultas complejas, que interrogan un número importante de tablas. Esta característica es tanto más actual cuanto que existen herramientas de ayuda a la decisión cada vez más sofisticadas, permitiendo al usuario formular simplemente lo que habría sido incapaz de formalizar con las herramientas de ayer. Ante esta complejidad, el Data Warehouse debe poder reaccionar en plazos razonables.

Cada vez son más raros los sistemas de decisión empaquetados, es decir, encapsulados en aplicaciones fijas y predefinidas por la informática. Cada

vez más, el usuario desea obtener los medios para su autonomía. A fin de evitar la clásica espera de reactividad relacionada con la puesta en producción, no debe depender de un informático que desarrolle la aplicación adaptada a sus necesidades. Un Data Warehouse intenta responder a las necesidades de los usuarios en términos de informaciones y no en términos de aplicaciones.

La consecuencia de esta constatación es que cuanto más legible sea el modelo, es decir, intuitivo para los usuarios, menos largo y costoso será definir una capa por encima de este sistema destinada a hacerlo comprensible y adaptado a las necesidades de los usuarios. Esta capa es de cualquier forma necesaria para dar al usuario una visión de negocio del modelo de datos y hay numerosas herramientas de ayuda a la decisión en el mercado que proponen su implementación. Pero debe ser ligera, porque cuanto más gruesa sea esta capa complementaria, más larga de definir resultará a los informáticos. Definir un modelo legible se convierte, pues, en una preocupación fundamental.

En un contexto de decisión, desde el punto de vista del administrador de base de datos, una de las mayores dificultades es gestionar lo imprevisible. En efecto, las consultas son normalmente ad hoc, generadas por el usuario a través de una herramienta, y es pues imposible optimizar cada una de ellas caso por caso. Sie embargo existen, técnicas de optimización basadas no ya en las consultas, sino más bien en los caminos de acceso para garantizar tiempos de respuesta convenientes y previsibles en la decisión.

La última característica del mundo Data Warehouse es que permite implementar normalmente un modelo de datos integrado, cuyo objetivo es ser transversal a la empresa. Este modelo se constituye habitualmente de manera incrementar, a medida de las realizaciones sucesivas de los proyectos de decisión de la empresa. Por ejemplo, el primer proyecto permitirá construir el modelo que integra las ventas para tal tipo de producto. Este modelo se enriquece poco a poco ampliando la gama de productos incluidos, integrando la

información relativa a la logística, ete. En este marco, el modelo de datos evolucionará de manera constante y regular.

4.3.7 Elección de los componentes

Una limitación de un RDBMS y un MDDB, es la carencia de soporte para tipos de datos no tradicionales como imágenes, documentos y clips de video/ audio. Si usted necesita estos tipos de objetos en su Data Warehouse, busque un DBMS relacional-objeto (Ejemplo: Ilustra de Informix).

Por su enfoque en los valores de datos codificados, la mayor parte de los sistemas de base de datos pueden acomodar estos tipos de datos, sólo con extensiones basadas en cierta referencias, tales como indicadores de archivos que contienen los objetos. Muchos RDBMS almacenan los datos complejos como objetos grandes binarios (Binary Large Objects - BLOBs). En este formato, los objetos no pueden ser indexados, clasificados, o buscados por el servidor.

Los DBMS relacional-objeto, de otro lado, almacenan los datos complejos como objetos nativos y pueden soportar las grandes estructuras de datos encontradas en un ambiente orientado a objetos. Estos sistemas de base de datos naturalmente acomodan no sólo tipos de datos especiales sino también los métodos de procesamiento que son únicos para cada uno de ellos.

Pero una desventaja del enfoque relacional-objeto, es que la encapsulación de los datos dentro de los tipos especiales de datos (una serie

de precios de stock a través del tiempo en cada registro de una tabla de stock, por ejemplo), requiere de operadores especializados para que hagan búsquedas históricas simples (por ejemplo, "Encontrar todas las existencias que han mostrado una disminución en el precio de Abril a Mayo 1996").

La selección del DBMS está también sujeta al servidor de hardware que se usa. Algunos RDBMS, como el DB2 Paralelo, Informix XPS y el Oracle Paralelo, ofrecen versiones que soportan operaciones paralelas. El software paralelo divide consultas, uniones a través de procesadores múltiples y corre estas operaciones simultáneamente para mejorar el desempeño.

Se requiere el paralelismo para el mejor desempeño en los servidores MPP grandes y SMP agrupados. No es aún una opción con MDDBS o DBMS relacional-objeto. En la tabla "Matriz para la selección de la DBMS" se resumen los pro y los contra de los diferentes tipos de DBMS para operaciones de Data Warehouse.

La tabla "Matriz de Decisión del Data Warehouse" contiene algunos ejemplos de cómo afectan estos criterios de decisión en la elección de una arquitectura de servidor/ Data Warehouse.

MATRIZ PARA LA SELECCIÓN DE LA DBMS					
Características/Función	**Relacional**	**Super Relacional**	**Multi Dimensional (Lógico)**	**Multi dimensional (Físico)**	**Objeto-Relacional**
Estructuras Normalizadas					
Tipos de datos abstractos					
Paralelismo					
Estructuras Multidimensionales					
Drill-Down					
Rotación					
Operaciones dependientes de datos					

MATRIZ DE DECISIÓN PARA EL DATA WAREHOUSE					
Para estos ambientes...			Seleccione...		
Requerimientos comerciales	**Usuarios**	**Soporte de Sistemas**	**Arquitectura**	**Servidor**	**DBMS**
Alcance: departamental **Usos:** análisis de datos	Pequeño – Ubicación única	Local mínimo – Central promedio	Consolidado – Paquete	Procesador único o SMP	MDDB
Alcance: departamental **Usos:** investigación	Pequeño – Pocas ubicaciones	Central fuerte	Centralizado	MPP	RDBMS con soporte paralelo
Alcance: departamental **Usos:** análisis más Informática	Grande – Analistas en una sola ubicación Usuarios informáticos dispersos	Local mínimo Central promedio	Seccionado – detalle en central Resumen en local	Grupos de SMP para central SP o SMP para local	RDBMS para central MDDB para local
Alcance: empresa **Usos:** análisis más informática	Grande – geográficamente disperso	Central fuerte	Centralizado	Grupos de SMP	Objeto relacional soporte Web

4.3.8 Combinación de la Arquitectura y la Gestión de la BD

Para seleccionar la combinación correcta de la arquitectura del servidor y el DBMS, primero es necesario comprender los requerimientos comerciales de su compañía, su población de usuarios y las habilidades del personal de soporte.

Las implementación del Data Warehouse varía apreciablemente de acuerdo al área. Algunos son diseñados para soportar las necesidades de análisis específico para un solo departamento o área funcional de una organización, tales como Finanzas, Ventas o Mercadotecnia. Las otras implementaciones reúnen datos a través de toda la empresa para soportar una variedad de grupos de usuarios y funciones. Por regla general, a mayor área del depósito, se requiere mayor potencia y funcionalidad del servidor y el DBMS.

Los modelos de uso del Data Warehouse es también un factor. Las consultas y vistas de reportes pre-estructuradas frecuentemente satisfacen a los usuarios informáticos, mientras que hay menos demandas sobre el DBMS y la potencia de procesamiento del servidor. El análisis complejo, que es típico de los ambientes de decisión-soporte, requiere más poder y flexibilidad de todos los componentes del servidor. Las búsquedas masivas de grandes Data warehouses favorecen el paralelismo en el DBMS y el servidor.

Los ambientes dinámicos, con sus requerimientos siempre cambiantes, se adaptan mejor a una arquitectura de datos simple, fácilmente cambiable (por ejemplo, una estructura relacional altamente normalizada), antes que una estructura intrincada que requiere una reconstrucción después de cada cambio (por ejemplo, una estructura multidimensional).

El valor de la data fresca requerida indica cuán importante es para el Data Warehouse renovar y cambiar los datos. Los grandes volúmenes de datos que se refrescan a intervalos frecuentes, favorecen una arquitectura físicamente centralizada para soportar una captura de datos eficiente y minimizar el tiempo de transporte de los datos.

Un perfil de usuario debería identificar quiénes son los usuarios de su Data Warehouse, dónde se ubican y cuántos necesita soportar. La información sobre cómo cada grupo espera usar el Data Warehouse, ayudará a analizar los diversos estilos de uso.

Conocer la ubicación física de sus usuarios ayudará a determinar cómo y a qué área necesita distribuir el Data Warehouse. Una arquitectura por niveles podría usar servidores en el lugar de las redes de área local. O puede necesitar un enfoque centralizado para soportar a los trabajadores que se movilizan y que trabajan en el depósito desde sus Lap-tops.

El número total de usuarios y sus modelos de conexión determinan el tamaño de sus servidores de depósito. Los tamaños de memoria y los canales de I/O deben soportar el número previsto de usuarios concurrentes bajo condiciones normales, así como también en las horas pico de su organización.

Finalmente, se debe factorizar la sofisticación del personal de soporte. Los recursos de los sistemas de información (Information System - IS) que están disponibles dentro de su organización, pueden limitar la complejidad o sofisticación de la arquitectura del servidor. Sin el personal especializado interno o consultores externos, es difícil de crear y mantener satisfactoriamente una arquitectura.

4.3.9 Administración de los datos

Los datos "sucios" son peligrosos. Las herramientas de limpieza especializadas y las formas de programar de los clientes proporcionan redes de seguridad.

No importa cómo esté diseñado un programa o cuán hábilmente se use. Si se alimenta mala información, se obtendrá resultados incorrectos o falsos. Desafortunadamente, los datos que se usan satisfactoriamente en las aplicaciones de línea comercial operacionales pueden ser basura en lo que concierne a la aplicación Data Warehouse.

Por lo tanto es de suma importancia tomar en cuanta una serie de conceptos que permitirán aportar la dosis de confiabilidad, oportunidad y eficiencia necesaria al Data Warehouse respecto a los datos que manipulará.

Estos conceptos se explican a continuación, y como ya lo hemos repetido insistentemente durante todo el presente documento, en este tema tampoco hay una formula predefinida a utilizar, sino mas bien algunas reglas que seguir y algunos cuidados a tener en la Administración de los datos.

Metadatos

Los metadatos es la información sobre los datos esencial para una utilización eficaz de un Data Warehouse. Forman un conjunto de información de administración y de seguimiento para el proyecto de decisión. Mas precisamente lo metadatos representan todas las informaciones necesarias para el acceso, la comprensión y la utilización de los datos del Data

Warehouse: semántica, origen, reglas de agregación, almacenamiento, formato, utilización ....

Por ejemplo, a continuación se muestra una manera de calificar este tipo de informaciones designadas con el término *metadatos* :

Tipo de Información	Significado
Semántica	¿Qué significa este dato? ¿Qué significa «cifra de negocio»? ¿Qué significa «cliente»? ¿Para qué sirve, quién la necesita? Esta pregunta se hace únicamente desde un punto de vista funcional.
Origen	¿De dónde proviene? ¿Dónde, por quién y cuándo ha sido creada, actualizada?
Regla de cálculo	¿Cómo se calcula, cómo se calcula la cifra de negocio a partir de¡ montante elemental de cada una de las facturas o de los contratos o eventualmente de los servicios anexos o de subcontrataciones, etc.? Muchos conceptos corresponden a un negocio y una regla de cálculo corresponde a una regla de gestión.
Regla de agregación	¿Cuál es el perímetro de consolidación? Por ejemplo, para el dato «cifra de negocio europeo», ¿qué significa «el conjunto de países europeos»? Esta noción puede cambiar según las empresas o según la organización.
Guardado, Formato	¿Dónde y cómo se guarda, cuál es su formato: «cifra de negocio» en pesetas, en dólares, sin tasas, precio neto, etc. ¿Cuál es su ámbito de valores?
Utilización	¿Cuáles son los programas informáticos que la utilizan? ¿Cómo y en qué máquinas está disponible? ¿Durante cuánto tiempo se conserva?

Hay una regla primordial: todos los datos deben ir documentados por un conjunto de metadatos. Por ejemplo, si interesa la «cifra de negocio consolidada en Europa para tal línea de productos y tal organización», hay que definir exactamente lo que cubre esta expresión, su modo de cálculo, etc. El dato está necesariamente vinculado a otros objetos del sistema de información; por tanto, también es necesario representar, describir y almacenar estas interacciones con otros datos.

Tipo de enlace	Definición
Ámbito, temas	El Data Warehouse se estructura más bien por tema. Por tanto, cada dato se indexará por tema o por ámbito..
Estructura organizativa, Estructura geográfica	El dato se indexará igualmente según la estructura organizativa o geográfica, porque un dato puede tener sentidos ligeramente distintos según la persona que la manipula.
Conceptos genéricos	La noción de producto se declinará en línea de productos, servicios, servicios postventa, etc.
Aplicaciones, programas	El dato será manipulado por una o más aplicaciones o programas.
Tablas, columnas	El dato se sitúa en una o más columnas, tablas y bases de datos
Sedes, máquinas	Esta rúbrica representa la localización física (sede informática y máquina) del dato.

Los metadatos servirán pues para enriquecer el dato almacenado en un Data Warehouse. Existen varias maneras de representar y almacenar estos metadatos. Para facilitar su acceso y manipulación., se podría utilizar una ficha que contenga una descripción del dato y de sus enlaces. Pero, generalmente, se utiliza un simbolismo habitual en informática: los modelos de datos en el sentido Merise del término. Los metadatos se expresan en forma de modelos

conceptuales y de modelos lógicos de datos (MCD y MLD). Así, se obtiene una visión sintética que permite navegar entre los diferentes datos.

Coherencia y fiabilidad

La administración de los datos es la función que permitirá ante todo garantizar la coherencia del modelo global. Esta coherencia no es inmediata y a menudo es necesario implementar una función dedicada a esta tarea. Esta función se encarga por ejemplo de todos los problemas de arbitraje entre los diferentes puntos de vista de la empresa (por ejemplo, la noción de margen puede diferir según los servicios).

Existe ya una función de este tipo en ciertas empresas. Antes de la llegada del Data Warehouse, algunos se dieron cuenta de la diversidad de los sistemas de la empresa: la producción, la investigación, el servicio comercial, la contabilidad, etc., tienen normalmente sus propios sistemas operativos que tienen generalmente muy mala disposición para intercambiar datos. El Data Warehousing, por su aproximación, obliga naturalmente a enfrentarse al problema. Es importante entonces implementar a nivel de grupo o a nivel de empresa funciones de federación, de pilotaje de los diferentes proyectos, para asegurarse de la coherencia de las bases. En el marco de un Data Warehouse, se encuentran los problemas habituales de administración de los datos, ya de por sí complejos de resolver, y son normalmente aún más agudos.

Recordemos ahora cuáles son estos problemas, relacionados tradicionalmente con la administración de los datos:

Redundancias, sinonimias, duplicados. La visión del cliente desde un punto de vista comercial no es la misma que desde el punto de vista de la contabilidad, la gestión administrativa o el servicio postventa. En ocasiones se puede incluso duplicar bases de datos; esto presenta un problema de administración.

Incoherencias según el origen o en el tiempo. La noción de producto puede no ser la misma en todas partes a pesar de una denominación idéntica.

No reutilización de los modelos, las estructuras o los ámbitos de valores.

No capitalización del conocimiento. Conceptos habituales, como producto, contrato o cliente se redefinen regularmente, en ocasiones de manera contradictoria.

No fiabilidad, según el origen del dato. La calidad del dato puede ser aleatoria (por ejemplo, presencia o no de un dato).

En el marco del Data Warehouse encontramos los mismos problemas con el agravante de la necesidad de consolidar y agregar los datos, por lo que será necesario un esfuerzo suplementario.

Un Ejemplo

Los datos "sucios" pueden presentarse al ingresar información en una entrada de datos (por ejemplo, "Sitsemas S. A." en lugar de "Sistemas S. A.") o de otras causas.

Cualquiera que sea, la data sucia daña la credibilidad de la implementación del depósito completo. En la **Figura No. 19** se muestra un ejemplo de formato de ventas en el que se pueden presentar errores. Afortunadamente, las herramientas de limpieza de datos pueden ser de gran ayuda. En algunos casos, puede crearse un programa de limpieza efectivo. En el caso de bases de datos grandes, imprecisas e inconsistentes, el uso de las herramientas comerciales puede ser casi obligatorio.

Decidir qué herramienta usar es importante y no solamente para la integridad de los datos. Si se equivoca, se podría malgastar semanas en

recursos de programación o cientos de miles de dólares en costos de herramientas.

FORMATO DE VENTAS

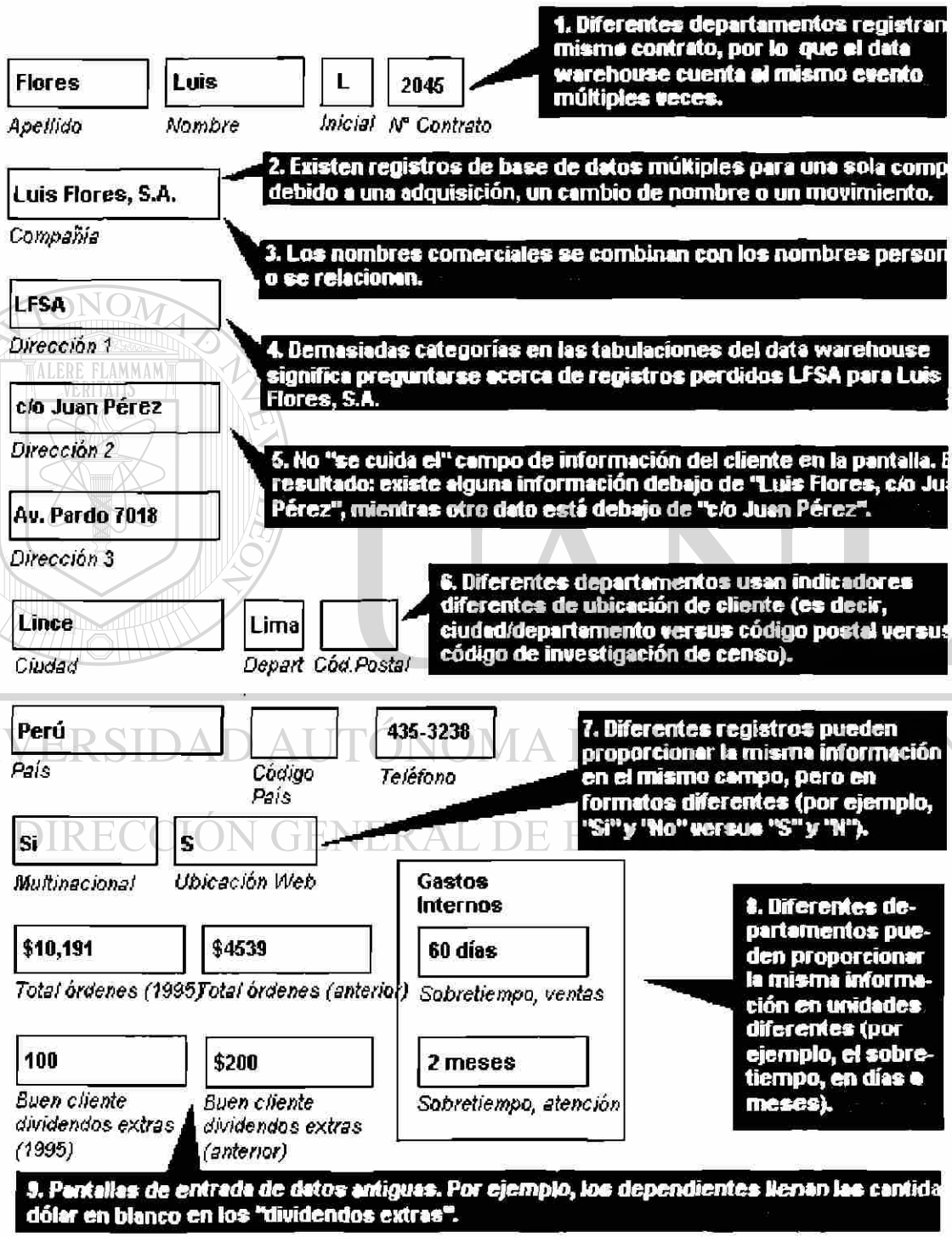


Figura No. 19

4.4 Etapa 3: Implementación de las Aplicaciones

En esta fase, el proyecto de Data Warehouse debe tener asignado el liderazgo adecuado, así como, los recursos humanos, tecnológicos y el presupuesto apropiado.

Sin embargo, deben evaluarse otros aspectos, como desarrollar un proyecto en su totalidad o por fases y además, diferenciar el tipo de proyecto a realizar.

Al llegar a esta etapa deben de coexistir dos visiones del Data Warehouse, tal y como lo muestra la **Figura No. 20** .

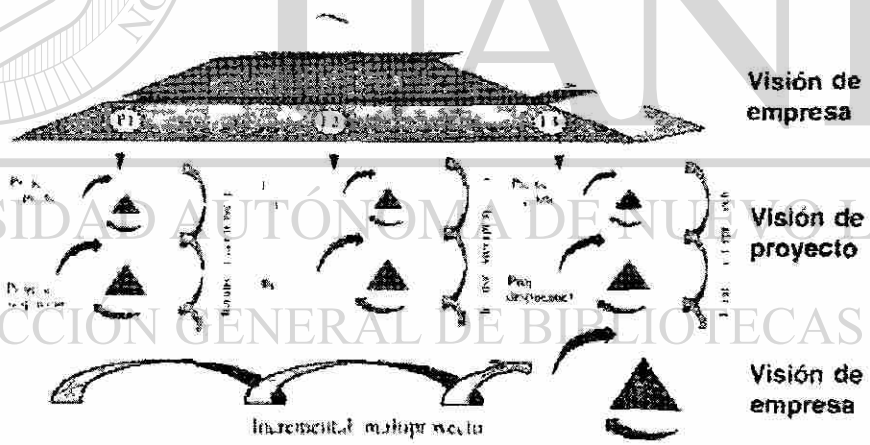


Figura No. 20

Una visión de empresa: los proyectos se identifican en la primera fase. A medida que se implementan, la infraestructura global del Data Warehouse se enriquece: desde este punto de vista, el método propuesto es incrementar. En efecto, cada proyecto es visto de manera independiente y responde a un objetivo de negocio delimitado. Pero participa también en la construcción del Data Warehouse integrándose en él.

Una visión de proyecto: los proyectos identificados en la fase inicial se convierten en aplicaciones, por ejemplo por una descomposición en dos etapas. En primer lugar, la realización del prototipo operativo permite concretar un subconjunto del proyecto identificado y desplegarlo en lugares piloto. Esta fase va seguida por el despliegue propiamente dicho del proyecto para todos los sitios afectados de la empresa.

Es necesario adaptar esta visión a cada contexto de empresa, porque es ilusorio presentar un método completo y universal sobre la implementación del Data Warehouse. Si bien la literatura se enriquece cada día con nuevos puntos de vista sobre este tema, éstos no comparten el mismo punto de vista.

4.4.1 Decisiones importantes al inicio de la implementación

Forma de implementación del Data Warehouse

Elegir la forma en que se va a implementar un proyecto de Data Warehouse no es una tarea que se pueda determinar con exactitud, depende de muchos factores, entre los que se encuentran el grado de madurez informática de la empresa, factores económicos, de organización, etc.

Sin embargo como una recomendación general para mejorar las posibilidades de éxito al embarcarse en la implementación de un Data Warehouse se puede decir que: Es más viable el desarrollo de un proyecto en fases que produzcan resultados a corto plazo que el desarrollo de un proyecto que entregue resultados al término de varios años. Por ello, el proyecto debe estar centrado en un área o un proceso.

Tipo de Proyecto a implementar

Decidir sobre el tipo de proyecto, es algo complicado, ya que tiene un alto grado de relación con las decisiones estratégicas de la empresa y con cuestiones de carácter económico.

Un proyecto especializado se encarga de soportar directamente un proceso específico, puede ser Retención de Clientes, cobranza, etc. Este tipo de proyectos, por un lado tienen un resultado mas directo y visible respecto a los objetivos buscados, pero por otro lado tienden a ser mas costosos y tardar mas tiempo su implementación.

Un proyecto base entrega capacidades genéricas de análisis a todos los usuarios que tengan acceso al Data Warehouse, pero no tiene, entre sus funcionalidades, la solución de un problema específico o el soporte especializado de un proceso específico.

Un proyecto base es más económico y fácil de acabar que uno especializado.

En base a estas consideraciones el encargado de la implementación debe decidir con cual tipo de proyecto tiene mas posibilidades de obtener logros importantes en la Organización.

4.4.2 Estrategia en la Implementación

Un objetivo importante de la fase de implementación del Data Warehouse es el minimizar los riesgos inherentes a todo proyecto que implica la introducción de nueva tecnología en una organización. Una estrategia recomendable en el proceso de implementación se plantea mediante 10 pasos, los cuales se describen a continuación :

Paso 1. Identificar uno o varios problemas ó áreas de oportunidad actual en la Organización que cumpla con las siguientes características :

- Sea de uso estratégico para el negocio.
- Genere ventaja competitiva a la organización o reduzca costos.
- Pueda ser apoyado o resuelto mediante el Data Warehouse.

De las posibles opciones se deberá seleccionar el proyecto dependiendo de las prioridades del negocio; se deberá optar por el que sea más rentable (justificación económica) y más factible técnicamente. En caso de que los factores de selección no coincidan en un mismo proyecto, se deberá optar por el que mejor siga la línea del negocio.

Paso 2. Definir el modelo lógico de datos a implementar para resolver el problema planteado. Ejemplo: Se puede dar un modelo lógico cuando se presenta al usuario la información en términos de dimensiones (clientes, productos, canales de ventas, promociones, adquirientes, etc) básicas del modelo de datos y hechos que se registrarán para estas dimensiones (medidas de ventas, de costos, de producción, de facturación, de cartera, de calidad, de servicio, etc.).

Paso 3. Reunir los datos para poblar ese modelo lógico de datos. En este punto es importante identificar las formas en que serán administrados los datos, su

alimentación, así como su adecuación a la estructura y semántica del Data Warehouse.

Paso 4. Definir el mejor diseño físico para el modelo de datos. El diseño físico debe estar orientado a generar buen rendimiento en el procesamiento de consultas, a diferencia del modelo lógico que está orientado al usuario y a la facilidad de consulta.

Paso 5. Definir los procesos de extracción, filtro, transformación de información y carga de datos que se deben implementar para poblar ese modelo de datos.

Paso 6. Definir los procesos de administración de la información que permanece en el Data Warehouse.

Paso 7. Definir las formas de consultas a la información del Data Warehouse que se le proporcionará al usuario. Para ello, debe considerarse la necesidad de resolver la necesidad y la potencia de consulta.

Paso 8. Completar el modelo de consulta base, relativo al área seleccionada.

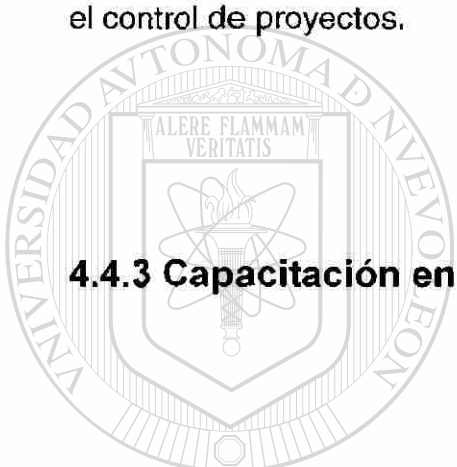
Paso 9. Implementar los procesos estratégicos del área de trabajo, es decir, implementar herramientas especializadas de scoring, herramientas especializadas para inducción de conocimiento (Data Mining), etc.

Paso 10. Completar las áreas de interés, en forma similar a lo descrito anteriormente.

Es importante destacar que estos pasos representan una guía para la implementación y pueden ser modificados y/o agregar pasos dependiendo de las necesidades de la Organización para la que se está construyendo el Data Warehouse.

Además esta guía establece un orden para la ejecución de cada una de las tareas a realizar, pero como ya habrá notado, no establece mecanismos para el seguimiento del proyecto. Estos mecanismos son importantes y en general se refiere al establecimiento de tareas, secuencia de ejecución, responsables, costos asignados, etc.

Se sugiere controlar el proyecto mediante el uso de gráficas de Gant, o cualquier otra técnica ó metodología conocida para el control de proyectos. Para una mayor agilidad y eficiencia puede utilizar software especializado para el control de proyectos.



4.4.3 Capacitación en la Implementación

La capacitación es un proceso que debe estar presente durante todo el proyecto, aun antes de la fase de implementación. Esta función no debe negligirse, ya que el proyecto de Data Warehouse lleva consigo toda una serie de cambios y nuevas tecnologías cuyo éxito depende directamente de su uso y explotación en beneficio de la Organización.

Sin embargo la capacitación toma una importancia aún mayor en la fase de implementación, ya que está íntimamente ligada con los tiempos del proyecto y en términos generales con el éxito del proyecto.

Toda persona involucrada en el proyecto generalmente requiere algún tipo de capacitación, por lo que esta se debe plantear en los diferentes ámbitos de la empresa.

En el ámbito técnico . El área de informática es una de las principales afectadas, ya que debe tiene que adiestrarse en la utilización de metodologías, técnicas, aplicaciones, etc. Pero no solo el informático debe de recibir capacitación técnica, los usuarios finales también deben ser instruidos en el uso de las diferentes herramientas que utilizarán para explotar la información derivada del Data Warehouse.

En el ámbito Organizativo . Aquí los principales involucrados son los usuarios y los mandos ejecutivos. A este nivel la capacitación se debe enfocar a instruirlos respecto a la mejor explotación de la información proporcionada. También es muy importante aclararles todos los conceptos de negocio homologados durante la creación del Data Warehouse, con los cuales trabajaran en adelante.

4.4.4 Uso de herramientas en la Implementación

Existe una gran diversidad de herramientas a utilizar en un Data Warehouse, por lo durante la implementación se dificulta el determinar cual es el uso que se le debe dar a cada una de estas herramientas.

De manera general existen algunas clasificaciones que definen la utilización de las diferentes herramientas de Data Warehouse. Cada herramienta cae en una de estas clasificaciones.

El presente tema tiene como objetivo el definir los usos adecuados para cada una de estas clasificaciones, de forma tal que el problema solo se reduzca a seleccionar la clasificación adecuada para cada caso que se presente.

En los anexos del presente documento se muestran una serie de opciones de Software agrupadas por clasificación. De esta manera una vez seleccionada la clasificación podemos consultar algunas de las herramientas de Software que apoyarían nuestro caso. La tabla "Matriz para selección de herramientas" puede serle de gran utilidad para este propósito.

MATRIZ PARA SELECCIÓN DE HERRAMIENTAS

Tipo de Herramienta	Pregunta básica	Modelo de Salida	Usuario típico
Consulta y Reporte	¿Qué sucedió?	Reportes de ventas mensuales; histórico de inventario	Necesita datos históricos. Puede tener aptitud técnica limitada
Procesamiento analítico en línea (OLAP)	¿Qué sucedió y por qué?	Ventas mensuales vs. Cambios de precio de los competidores	Necesita ir a una visión no estática de los datos. Técnicamente astuto
Sistema de Información Ejecutiva (SIE)	¿Qué necesito conocer ahora?	Libros electrónicos; Centros de comandos	Necesita información resumida o de alto nivel. Puede no ser técnicamente astuto

Herramientas de Consulta y Reporteo

Aunque las capacidades varían de un producto a otro, las herramientas de consulta y reporte son más apropiadas cuando se necesita responder a la pregunta :

- ¿"Qué sucedió"?

Ejemplo:

¿"Cómo comparar las ventas de los productos X,Y y Z del mes pasado con las ventas del presente mes y las ventas del mismo mes del año pasado?"

Existe una gran cantidad de herramientas de consulta y reporte en el mercado. Algunos proveedores ofrecen productos que permiten tener más control sobre qué procesamiento de consulta es hecho en el cliente y qué procesamiento en el servidor.

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

La herramienta de consulta genera entonces un llamado a una base de datos, extrae los datos pertinentes, efectúa cálculos adicionales, manipula los datos si es necesario y presenta los resultados en un formato claro. Se pueden almacenar las consultas y los pedidos de reporte para trabajos subsiguientes, como está o con modificaciones.

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

Lo más avanzado de estos productos lo orientará hasta las consultas que tienen sintaxis mala o que devuelven resultados imprevistos. El acceso a los datos han mejorado también con las nuevas versiones de estos productos y los vendedores ya instalan manejadores (drivers) estándares tales como ODBC y 32-bit nativo, hasta fuentes de datos comerciales.

En general, los administradores de data warehouses que usen estos tipos de productos, deben estar dispuestos a ocupar su tiempo para resolver las tareas de estructuración, como administrar bibliotecas y directorios, instalar software de conectividad, establecer nombres similares en Inglés y precalcular "campos de datos virtuales".

Herramientas de Bases de Datos Multidimensionales (OLAP)

Las primeras soluciones OLAP estuvieron basadas en bases de datos multidimensionales (MDDBS). Un cubo estructural (o un arreglo multidimensional) almacenaba los datos para que se pudieran manipular intuitivamente y claramente ver las asociaciones a través de dimensiones múltiples.

Para los usuarios que requieren algo mas que una simple vista estática de los datos, las herramientas del procesamiento analítico en línea (OLAP - On Line Analytical Processing), proveen capacidades que contestaría :

- "¿Qué sucedió? " al analizar el porqué de los resultados en diferentes dimensiones.

Sin embargo las herramientas OLAP aun presentan algunas limitaciones que afectan su utilización, configuración y desempeño. Explicaremos dos de ellas que presentan gran importancia :

- Las nuevas estructuras de almacenamiento de datos requieren bases de datos propietarias. No hay realmente estándares disponibles para accesar a los datos multidimensionales.
- Las compañías generalmente almacenan los datos de la empresa en bases de datos relacionales, lo que significa que alguien tiene que extraer, transformar y cargar estos datos en el hipercubo. Este proceso puede ser complejo y tardado pero, nuevamente, los proveedores están investigando la forma de solucionarlos. Las herramientas de extracción de datos y otras automatizan el proceso, trazando campos relacionales en la estructura multidimensional y desarrollando el MDDB sobre la marcha. Algunos proveedores ofrecen ahora la técnica ROLAP (Relational On Line Analytical Processing), que explora y opera en el Data Warehouse directamente usando llamadas SQL estándares. Las herramientas de pantallas permiten retener los pedidos multidimensionales, pero el motor ROLAP transforma las consultas en rutinas SQL. Entonces se recibe los resultados tabulados como una hoja de cálculo multidimensional.

Los retos administrativos y de desarrollo de OLAP, a diferencia de las encontradas con las herramientas de consulta y reporte, son generalmente más complejos. Definiendo el OLAP y el software de acceso a los datos, se requiere un claro entendimiento de los modelos de datos de la corporación y las funciones analíticas requeridas por ejecutivos, gerentes y otros analistas de datos.

El desarrollo de productos comerciales pueden aminorar los problemas, pero OLAP es raramente una solución clave. La arquitectura debe permitir el soporte a su fuente de datos y requerimientos. Pero una vez que se ha establecido un sistema OLAP, el soporte al usuario final será mínimo.

Sistemas de Información Ejecutivos (SIE)

Un uso típico de un SIE es :

- facilitar al usuario la recuperación y análisis de la métricas, de desempeño de la organización.

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

El precio de esta facilidad de uso es que por lo general existen algunas limitaciones sobre las capacidades analíticas disponibles con el sistema de información ejecutivo. Además, muchas de las herramientas de consulta/reporte y OLAP/multidimensional, pueden usarse para desarrollar sistemas de información ejecutivos.

El concepto de sistema de información ejecutivo es simple: los ejecutivos no tienen mucho tiempo, ni la habilidad en muchos casos, para efectuar el análisis de grandes volúmenes de datos. El SIE presenta vistas de los datos simplificados, altamente consolidados y mayormente estáticas.

Los Libros electrónicos: Son una versión en línea, contraparte del papel que muchos ejecutivos usan en reuniones con el personal. Las diapositivas presentan una visión concreta de una iniciativa organizacional o quizás los datos para dar a conocer la situación actual de un proyecto importante.

El centro de comando: Es básicamente una conjunto reducido de puntos en un amplia cantidad de reportes. El newsgroup recupera desde Internet y otros

materiales que proveen conocimientos en la organización. Los reportes del centro de comando pueden ser accedidos diariamente o con más frecuencia, si la información cambia constantemente o sólo cuando se garantiza las excepciones. Algunos productos generan alarmas cuando ocurren las excepciones especificadas.

Cuando sea apropiado, cada diapositiva del Libro electrónico o pantalla del centro de comando, debería permitir al ejecutivo recibir información adicional si lo desea (y si está disponible). A diferencia del modelo OLAP, donde el incremento de niveles de información se dan a conocer tal como el analista manipula los datos, un ejecutivo espera una descripción global. No deberían escudriñar para obtener respuestas.

Por ello, cuando los ejecutivos piden más información desde las diapositivas del libro electrónico o de las pantallas del centro de comandos, la presentación debería ser cuidadosamente elaborada para presentar principalmente información adicional amplificada. El ejecutivo debe ser capaz de pasar cada punto para "más información", sin perder alguna información crítica.

Los sistemas de información ejecutivos, generalmente tienen una programación que variará en complejidad de un producto a otro.

4.5 Etapa 4: Evaluación de los resultados

Para la evaluación económica se deberán cuantificar, mediante cualquier método, el costo de cada uno de los trabajos involucrados en la elaboración del Data Warehouse

- Cuánto cuestan las demoras
- Cuánto cuestan los errores
- Cuánto se gasta en asesorías especializadas

4.5.1 Evaluación de rendimiento de la Inversión (ROI)

Cuando se evalúan los costos, el usuario del Data Warehouse puede no tener el contenido de los costos en mente, pero las preguntas mínimas que puede comenzar a hacerse son las siguientes:

- ¿Qué clases de costos excedieron el presupuesto en más del 10% en cada uno de los 12 meses pasados?
- ¿Se aumentaron los presupuestos en más de 5% para cualquier área dentro de los últimos 18 meses?
- ¿Cómo especificar las clases de gasto entre diferentes departamentos? ¿Entre divisiones? ¿A través de las regiones geográficas?

- ¿Cómo tener márgenes de operación sobre los dos últimos años en cada área de negocio? Donde han disminuido los márgenes, ¿se han incrementado los costos?

Con frecuencia, los aspectos realmente importantes identificados por una gestión mayor, tienen un valor agregado, en el que ellos saben si tuvieron la información que estaban buscando, lo que significaría una mejora de (por ejemplo) las ventas en 0.5% a 1% - que, si su operación estuvo por los billones de dólares en un año, puede resultar en cientos de millones de dólares. En algunos casos, el costo del depósito inicial se ha recobrado en un período de 6 a 8 meses.

Al hacerse preguntas de este tipo, los usuarios comienzan a identificar las áreas en la que los costos han aumentado o disminuido significativamente y pueden evaluar cada una de estas áreas con más detalle.

4.5.2 El ROI en proyectos de Data Warehouse

A un nivel general los grandes gabinetes de análisis se han centrado en las realizaciones más significativas en el ámbito del Data Warehousing, en particular desde el ángulo económico. Así, el gabinete estadounidense Gartner Group presenta algunos ejemplos, que expresa como beneficio sobre inversión en el tiempo. Este beneficio sobre inversión se expresa en multitud de partidas:

	Retorno sobre Inversión	Periodo (en años)
Industria de gran consumo	*58	4
Compañía aérea	*5	2
Banca	*33	2
Gran distribución	*7	5
Telecomunicaciones	*9	4
Banca regional	*70	5

Hay que indicar que todos los ejemplos presentados aquí conciernen a empresas que se sitúan en el marco de un mercado de volumen, disponiendo de hecho de un número muy importante de clientes, pudiendo ser éstos empresas, pero también individuos. No cabe duda de que en este caso concreto es cuando el valor agregado de un Data Warehouse será más destacable. Esto explica que los éxitos más mediáticos afectan a este tipo de empresa, aunque el Data Warehouse es un concepto que afecta potencialmente a todo actor del mundo económico.

Un estudio sobre el mismo tema ha sido realizado por IDC. Su objetivo no era valorar tal o cual experiencia, sino reunir la información para cualificar de manera genérica la aportación de un Data Warehouse a las empresas. Se interrogó a 62 organizaciones americanas y europeas. Éstos son algunos de sus resultados:

- Beneficio sobre inversión en 3 años: la media es del 401 %, la mediana del 167%. El 90% de las empresas consultadas destacaron un beneficio sobre inversión superior al 40%. Para el 13% de las empresas, el beneficio sobre inversión sobrepasó el 1000%.

- El equilibrio sobre inversión (en inglés payback) se alcanza como promedio en 2.31 años, siendo la mediana de 1.67 años. La inversión promedio es de 2.2 millones de dólares.

A pesar de la elocuencia de estas cifras, es difícil describir de manera genérica y cualitativa los beneficios de un Data Warehouse: muy relacionados con la estrategia de la empresa, dependen necesariamente del sector de actividad. Por ejemplo, un estudio llevado a cabo por AT&T Teradata indica que, en la distribución a gran escala, las principales zonas de oportunidad son:

- Un aumento de las ventas a través de una Mercadotecnia mejor orientada;

- Una mejora de las tasas de rotación de los stocks;

- La reducción de los stocks de productos que han quedado obsoletos;

- La reducción de las pérdidas relacionadas con las rebajas,

- Devoluciones y comisiones,

- La disminución de costos de los productos de proveedores relacionados con una mejor negociación de precios.

Globalmente, el estudio indica que, para los grandes distribuidores, el beneficio sobre inversión medio es de 7 veces la inversión y puede llegar hasta 125 veces la inversión inicial.

Un último elemento contabilizado que muestra el valor agregado de un Data Warehouse aparece en un artículo sobre el Data Warehouse en la gran distribución publicado por el Journal of Data Warehousing [SKE96]. Se analiza especialmente el Data Warehouse del distribuidor W.H. Smith, que en cuanto a técnica es una base de 160 gigabytes, utilizada por 300 usuarios y que recoge 150.000 productos, 10.000 proveedores y 300 almacenes. Tras unos meses de producción, las ganancias calculadas son las siguientes:

Mejor decisión sobre la política de tarificación	$ 750,000
Reestructuración de las líneas	$ 450,000
Mejor seguimiento de la gestión de pérdidas imprevistas (robo, pérdida, etc.)	$ 100,000
Optimización del stock para la campaña de Navidad	$ 900,000
Gestión de devoluciones de mercancías	$ 800,000

ANEXOS

Lista de Software

A-1. Herramientas de consulta y Reporte

PRODUCTO	EMPRESA DISTRIBUIDORA
Access	Microsoft
Business Objects	Business Objects, Inc.
Crystal Reports, Crystal Info	Seagate Software
DB Publisher	Xense Technology Inc.
DbPower	Db-Tech Inc.
Esperant	Speedware
FOCUS Six	Information Builders, Inc.
4S-Report	Four Seasons Software, Inc
Freequery	Dimension Software Systems
Front & Center for Reporting, Nomad	Thomson Software Products
HP Information Access	Hewlett-Packard
Impromptu	Cognos Corporation
InfoAssistant	Asymetrix
InfoMaker	Powersoft Corporation
InfoQuery	Platinum Technology, Inc.
Internet DataSpot	DTL Data Technologies Ltd.
Insight	Williams & Partner
Interactive Query	New Generation software
IQ/Objects, IQ/SmartServer	IQ Software Corporation
Oracle Reports, Browser	Oracle Corporation
Paradox	Borland
Platinum Report Facility	Platinum Technology, Inc
R&R Report Writer	Concentric Data Systems
Report Writer	Raima
SAS System	SAS Institute
SQLPRO Agent	Beacon Ware, Inc.
SQR Workbench	MITI
ViewPoint	Soliton Associates
VisPro/Reports	Hock Ware
Visual Cyberquery	Cyberscience Corporation
Visual Dbase	Borland
Visual Express	Computer Associates International
Visual FoxPro	Microsoft Corporation
Visualizer Query, Charts	IBM
Voyant	Brossco Systems
WebSeQueL	InfoSpace Inc.
Xentis	GrayMatter Software Corporation

A-2. Herramientas de Bases de Datos Multidimensionales (OLAP)

PRODUCTO	EMPRESA DISTRIBUIDORA	TIPO
Acuity ES	Acuity Management Systems Ltd.	MDDB
Acumate ES	Kenan Systems Corporation	MDDB
Advance For Windows	Lighten, Inc.	MDDB
AMIS OLAP Server	Hoskyns Group plc	MDDB
BrioQuery	Brio Technology	MDDB
Business Objects	Business Objects, Inc.	Relacional
Commander OLAP, Decision, Prism	Comshare Inc.	MDDB
Control	KCI Computing	Relacional
CrossTarget	Dimensional Insight	MDDB
Cube-It	FICS Group	MDDB
Dataman	SLP Infoware	MDDB
DataTracker	Silvon Software, Inc.	Relacional
DecisionSuite	Information Advantage, Inc.	Relacional
Delta Solutions	MIS AG	MDDB
Demon for Windows	Data Command Limited	MDDB
DSS Agent	MicroStrategy	Relacional
DynamicCube.OCX	Data Dynamics, Ltd.	Relacional
EKS/Empower	Metapraxis, Inc.	MDDB
Essbase Analysis Server	Arbor Software Corporation	MDDB
Essbase/400	ShowCase Corporation	MDDB
Express Server, Objects	Oracle	MDDB
Fiscal	Lingo Computer Design, Inc.	Relacional
Fusion	Information Builders, Inc.	MDDB
FYI Planner	Think Systems	MDDB
Gentia	Planning Sciences	MDDB
Helm	Codeworks	MDDB
Holos	Holistic Systems	MDDB
Hyperion OLAP	Hyperion Software	MDDB
Informer	Reportech	MDDB/Relacional
Intelligent Decision Server	IBM	Relacional
IQ/Vision	IQ Software Corporation	Relacional
Lightship	Pilot Software, Inc.	MDDB
Matryx	Stone, Timber, River	MDDB
MDDB Server	SAS	Relacional
Media	Speedware Corporation	MDDB
Metacube	Informix	Relacional
MIKSolution	MIK	MDDB
MIT/400	SAMAC, Inc	MDDB
OLAP Office	Graphitti Software GmbH	MDDB
OpenOLAP	Inphase Software Limited	Relacional

A-3. Sistemas de Información Ejecutivos (SIE)

PRODUCTO	EMPRESA DISTRIBUIDORA
Acuity/ES	Acuity Management Systems Limited
Applixware	Applix
BusinessMetrics	Valstar Systems Ltd.
BOARD	Pragma Inform
COINS	Russell Consulting Limited
ColumbusEIS	Jitcons YO
Commander EIS	Comshare Inc.
Corporate Management/ Financial Executive Information System	Strategic Information Associates, Inc.
CorVu	CorVu Pty Ltd.
Decision Suite	Softkit
Discovery EIS	Atlantic Information Systems Ltd.
EIS	Inphase Software Limited
Electronic Balanced Scorecard	ASI Financial Services
Enterprise Periscope	Everyware Development Corp.
Eureka	European Management Systems
ExecuSense	TLG Corporation
FOCUS EIS	Information Builders, Inc.
Forest & Trees	Platinum Technologies, Inc.
iMonitor	BayStone Software
InfoManager	Ferguson Information Systems
Iridon Almanac	The Great Elk Company Limited
inSight	Arcplan Information Services
LEADER	Sterling Strategic Solutions
MagnaFORUM	Forum Systems, Inc.
Merit	GIST, s.r.o.
Open EIS Pak	Microsoft
Panorama Business Views	Panorama Business Views Inc.
Perspectives	Syntell
Qbit	Zenia Software, Inc.
Reveal	CSD Software Inc.
SAS System	SAS Institute
Show Business EIS	Show Business Software
Tiler EIS++	Avoca Systems Limited
Track	Track Business Solutions
Traffic Control EIS	Research & Planning, Inc.
VentoMap, VentoSales	Vento Software Inc.
Virtual Headquarters Management System	vHQ LLC
Visual EIS	Synergistic Software
Visual Publisher	KMA Associates International, Inc
VITAL	Braintec Corporation
Wingz	Investment Intelligence Systems Group
Wired for OLAP	AppSource Corporation
Xecutive Pulse EIS	Megatrend Systems, Ltd.

A-4. Bases de datos de Data Warehouse

PRODUCTO	EMPRESA DISTRIBUIDORA
Adabas D	Software AG
Advanced Pick	Pick Systems
DB2	IBM
Fast-Count DBMS	MegaPlex Software
HOPS	HOPS International
Microsoft SQL Server	Microsoft
Model 204	Computer Corporation of America
NonStop SQL	Tandem
Nucleus Server	Sand Technology Systems
OnLine Dynamic Server, Extended Parallel Server	Informix
OpenIngres	Computer Associates
Oracle Server	Oracle
Rdb	Oracle
Red Brick Warehouse	Red Brick Systems
SAS System	SAS
Sybase IQ	Sybase
Sybase SQL Server, SQL Server MPP	Sybase
SymfoWARE	Fujitsu
Teradata DBS	NCR
THOR	Hitachi
Time Machine	Data Management Technologies, Inc.
Titanium	Micro Data Base Systems, Inc.
Unidata	Unidata, Inc.
UniVerse	VMARK
Vision	Innovative Systems Techniques, Inc.
WX9000	White Cross Systems Inc.
XDB Server	XDB Systems, Inc.

Lista de figuras

Glosario de Términos

ADMINISTRADOR DE BASE DE DATOS O DBA

Experto técnico competente para uno (o varios) gestores de datos, capaz de escribir o de optimizar programas de extracción, de carga o de acceso en el lenguaje del motor de datos.

ADMINISTRADOR DE DATOS

Experto en el negocio que conoce la semántica de los datos de la empresa y se encarga del referencias de datos. Por ello, es capaz de arbitrar los conflictos inherentes a la constitución de definiciones únicas de los objetos de negocio de la empresa.

ANÁLISIS

Operación consistente en estudiar y descomponer los datos a fin de extraer de ellos los elementos esenciales y obtener un esquema de conjunto.

BASE DE DATOS (DATA BASE)

Conjunto de datos no redundantes, almacenados en un soporte informático, organizados de forma independiente de su utilización y accesibles simultáneamente por distintos usuarios y aplicaciones. La diferencia de una BD respecto a otro sistema de almacenamiento de datos es que éstos se almacenan en la BD de forma que cumplen tres requisitos básicos: no redundancia, independencia y concurrencia.

CATÁLOGOS

En ciertas herramientas clientes del Data Warehouse, estructura que permite al usuario trabajar sobre una vista lógica y orientada al negocio de los datos que desea ver.

CLIENTE/SERVIDOR

Arquitectura de sistemas de información en la que los procesos de una aplicación se dividen en componentes que se pueden ejecutar en máquinas diferentes. Modo de funcionamiento de una aplicación en la que se diferencian dos tipos de procesos y su soporte se asigna a plataformas diferentes.

CODIFICACION

a) Transformación de un mensaje en forma codificada, es decir, especificación para la asignación unívoca de los caracteres de un repertorio (alfabeto, juego de caracteres) a los de otro repertorio.

b) Conversión de un valor analógico en una señal digital según un código prefijado.

DATA WAREHOUSE

«Almacén de datos». Base de datos específica del mundo de la decisión destinada principalmente a analizar las palancas de negocio potenciales. Un Data Warehouse es (fuente.- Bill Inmon):

- integrado;
- orientado al tema;
- y contiene datos no volátiles.

DETECCION DE DESVIACION

Normalmente, para la detección de desviación en bases de datos grandes se usa la información explícita externa a los datos, así como las limitaciones de integridad o modelos predefinidos. En un método lineal por contraste, se enfoca el problema desde el interior de los datos, usando la redundancia implícita de los datos. Aquí se simula un mecanismo familiar a los seres humanos: después de ver una serie de datos similares, un elemento que perturba la serie se considera una excepción.

DICCIONARIO DE DATOS

Descripción lógica de los datos para el usuario. Reúne la información sobre los datos almacenados en la BD (descripciones, significado, estructuras, consideraciones de seguridad, edición y uso de las aplicaciones, etc.).

DIMENSIÓN

Eje de análisis asociado a los indicadores; corresponde normalmente a los temas de interés del Data Warehouse. ejemplo: dimensión temporal, dimensión cliente...

DIRECTORIO DE DATOS

Es un subsistema del sistema de gestión de base de datos que describe dónde y cómo se almacenan los datos en la BD (modo de acceso y características físicas de los mismos).

INCONSISTENCIA

El contenido de una base de datos es inconsistente si dos datos que deberían ser iguales no lo son. Por ejemplo, un empleado aparece en una tabla como activo y en otra como jubilado.

INTEGRIDAD

Condición de seguridad que garantiza que la información es modificada, incluyendo su creación y borrado, sólo por el personal autorizado.

INTERNET

Término usado para referirse a la red más grande del mundo, que conecta miles de redes con alcance mundial. Está creando una cultura que basándose en la simplicidad, investigación y estandarización fundamentado en usos de la vida real, está cambiando la forma de ver y hacer muchas de las tareas actuales. Mucha de la tecnología de punta en redes está proviniendo de la comunidad Internet.

INTRANET

Constituye un servicio de comunicación de los sistemas de información corporativos orientados a su personal, sobre el formato de los sistemas Web, operando sobre la red Internet. Ejemplo: El sistema contable de una empresa de ventas de productos de ferretería, tipo Home Center.

LIBRO ELECTRONICO

Guía electrónica. Documento realizado en un sistema informático, normalmente con características hipertexto y multimedia.

METADATO

Información que describe un dato. En un contexto de Data Warehouse, cualifica un dato precisando por ejemplo su semántica, las reglas de gestión asociadas, su fuente, etc.

MOLAP (MULTIDIMENSIONAL ON LINE ANALYTICAL PROCESSING)

Cf. OLAP

MPP (MASSIVE PARALLEL PROCESSING)

Caracteriza una Base de Datos dedicada a la decisión almacenando los datos en forma de tablamultidimensional. Estos SGBD son una alternativa a los SGBD relacionales.

MULTIDIMENSIONAL (SGBD)

Arquitectura de Hardware que hace colaborar a varios procesadores (hasta centenares) contando cada uno con su propia memoria.

OLAP (ON LINE ANALYTICAL PROCESSING)

Caracteriza la arquitectura necesaria para implementar un Sistema de Información de Decisión. Se opone a OLTP. El término OLAP designa a menudo las herramientas de análisis basadas en BD multidimensionales. Se habla también de herramientas MOLAP por oposición a las herramientas ROLAP.

OLTP (ON LINE TRANSACTIONNEL PROCESSING)

Tipo de entorno de tratamiento de la información en el que debe de darse una respuesta en un tiempo aceptable y consistente.

REDUNDANCIA

Repetición de los mismos datos en varios lugares.

REPOSITORIO

Base de datos central en herramientas de ayuda al desarrollo. El repositorio amplía el concepto de diccionario de datos para incluir toda la información que se va generando a lo largo del ciclo de vida del sistema, como por ejemplo: componentes de análisis y diseño (diagramas de flujo de datos, diagramas entidad-relación, esquemas de bases de datos, diseños de pantallas, etc.), estructuras de programas, algoritmos, etc. En algunas referencias se le denomina Diccionario de recursos de información.

ROLAP (RELATIONAL ON LINE ANALYTICAL PROCESSING)

Caracteriza la arquitectura necesaria para implementar un Sistema multidimensional basado en tecnologías relacionales.

SISTEMA DE GESTION DE BASE DE DATOS

Software que controla la organización, almacenamiento, recuperación, seguridad e integridad de los datos en una base de datos. Acepta pedidos de datos desde un programa de aplicación y le ordena al sistema operativo transferir los datos apropiados.

Cuando se usa un sistema de gestión de base de datos, SGDB, (en inglés DBMS), los sistemas de información pueden ser cambiados más fácilmente a medida que cambien los requerimientos de la organización. Nuevas categorías de datos pueden agregarse a la base de datos sin dañar el sistema existente.

SISTEMA DE INFORMACION (SI)

Conjunto de elementos físicos, lógicos, de comunicación, datos y personal que, interrelacionados, permiten el almacenamiento, transmisión y proceso de la información.

SMP (SYMETRIC MULTIPROCESSING)

Arquitectura de Hardware que hace colaborar varios procesadores (unas docenas) en una memoria compartida.

SQL (Structured Query Language)

Lenguaje de interrogación normalizado para bases de datos relacionales. El SQL es un lenguaje de alto nivel, no procedural, normalizado, que permite la consulta y actualización de los datos de BD relacionales. Se ha convertido en el estándar para acceder a BD relacionales. La primera versión se aprobó como norma ISO en 1987 y la segunda, conocida como SQL2 y vigente actualmente, en 1992.

Actualmente se trabaja en la norma SQL3 que soportará bases de datos orientadas a objeto y bases de datos activas. El SQL facilita un lenguaje de definición de datos y un lenguaje de manipulación de datos. Además, incluye una interfase que permite el acceso y manipulación de la BD a usuarios finales.

SPONSOR O PATROCINADOR

Individuo o grupo de individuos cuya función es obtener la adhesión y la implicación de todos los actores afectados por la implementación del Data Warehouse. Debe promover el proyecto y garantizar la sinergia entre los usuarios y los equipos informáticos. También debe gestionar los eventuales problemas políticos que la implementación de estos sistemas puede generar.

UNIX

Sistema operativo multiproceso, multiprograma y multiusuario. Software diseñado por AT&T para ingeniería de telecomunicación. Ha sido el primer sistema operativo concebido con independencia de los fabricantes. Posee una

gran facilidad para adaptarse a ordenadores con diferentes arquitecturas, siendo ampliamente autónomo respecto del hardware. Está escrito en lenguaje de alto nivel C.

Bibliografía

Libros

IDC - IDC Study on the Financial Impact of a Data Warehouse. Estudio respaldado por Andersen Consulting, Digital, Dun & Bradstreet Software & Pilot Software, EMC, Hewlett Packard, IBM, InfoiTnix, KPMG Peat Marwick, Mierosoft, NCR, Oracie, Prism, Pyramid, SAP, SAS, SGI, Software AG, Sun, Sybase, Unisys. Difundido por sus sponsors y disponible en Internet en la dirección:
http://direct.boulder.ibm.com/dss/solution/idc.html

EDS - El Data Warehouse El Data Mining, Septiembre 1997

EDS - Data Warehouse Primer, Enero 1995.

ETI - Anticipating the Cost of Maintenance, Julio 1995.

Red Brick - A Primer Information Technology and Business Management, Junio 1995.

Referencias

[DWHInsta] The Data Warehousing Institute Select Survey Results 1995-1996. Estudio basado en 6.214 respuestas.

[DWHInstb] Alan Paller.- Ten Mistakes to Avoidfor Data Warehousing Managers, Data Warehouse Institute.

[LOV96] Bruce Love.- S-trategie DSS Data Warehouse: A Case Study in Failure, Journal of Data Warehousing, julio de 1996.

[SKE96] Michael Haisten.- A History of Access and Analysis Tools, Journal ofdata Warehousing, julio de 1996.

[Inmon94] W.H. Inmon & Richard D. Hackathorn.- Using the Data Warehouse, Wiley-QED Publication, 1994.

[Mar91] James Martin - Rapid Application Development - Prentice Hall, 1991.

Internet

http://www.dw-institute.com/ - El Data Warehouse Institute es una asociación temática especializada en el Data Warehouse. Recoge más de 2.000 adheridos en 40 países.

http://www.inei.gob.pe/cpi-mapa/bancopub/libfree/lib619/INDEX.HTM
Documento HTML con información de Data Warehouse

http://www.intelligententerprise.com/ - Intelligent enterprise es la un sitio especializado en temas informáticos, el Data Warehouse, ERP, etc.

http://www.prometheus.eds.fr/ - El servicio en línea del Instituto Prométhéus de EDS, Este servicio propone la base de direcciones INFOmédiaire, que contiene cerca de mil referencias a documentos o sedes WEB. Contiene también extractos de estudios, artículos, etc.

http://www.dwinfocenter.org/ - El Data Warehousing Information Center es un centro de información dedicado al Data Warehouse