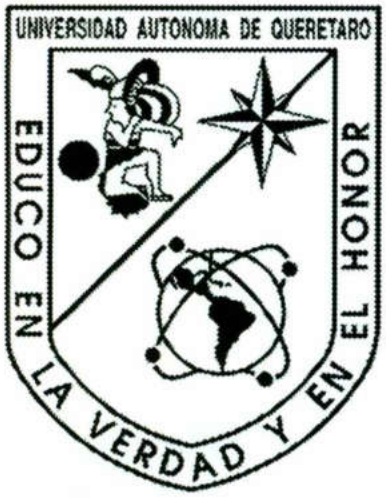




# Universidad Autónoma de Querétaro

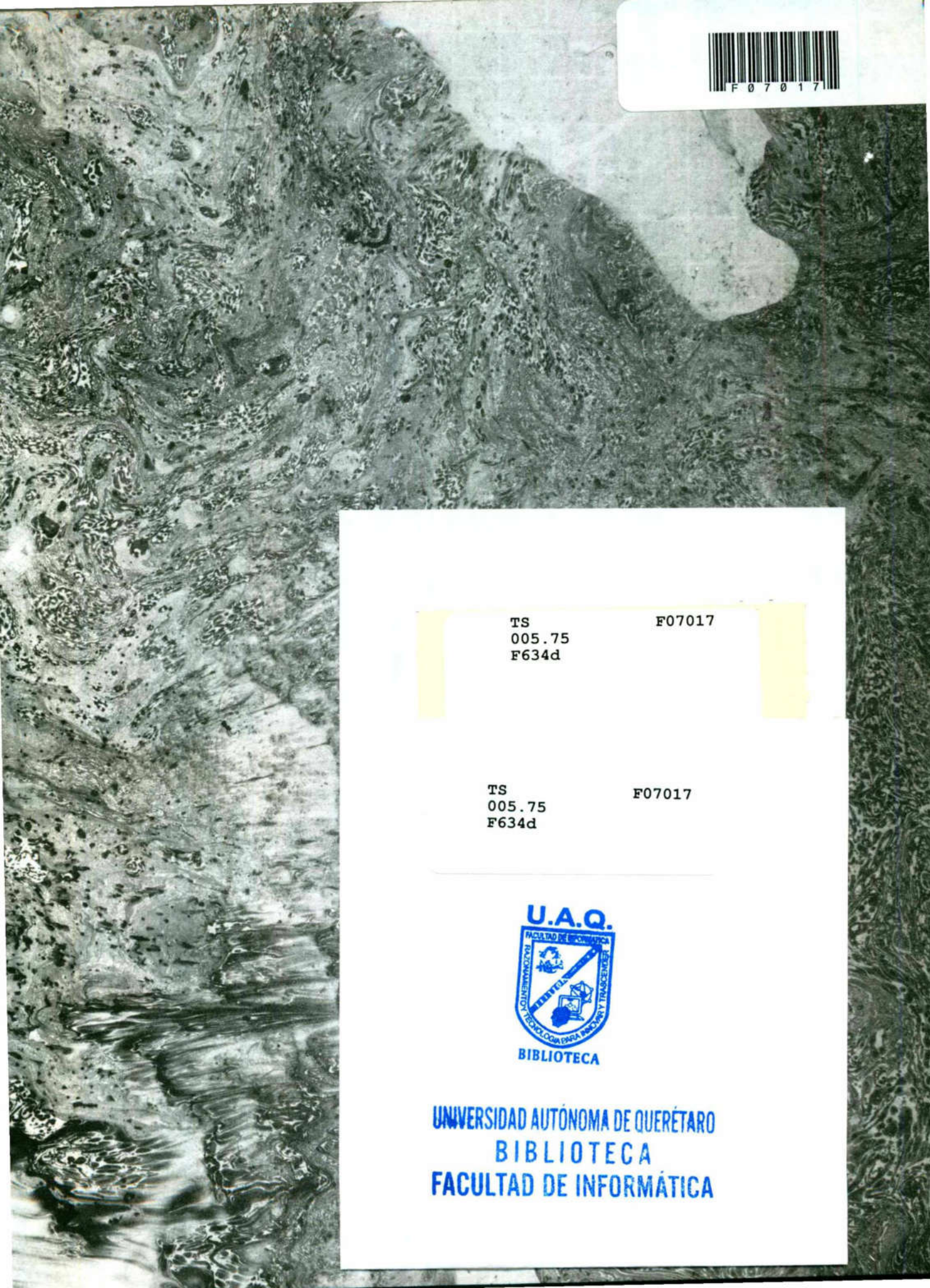
## Facultad de Informática

## Tesina: Data Warehouse

Presenta: Luis Guillermo Flores Ceja

Expediente: 80324

Asesor: I.S.C. Jabel Reséndiz González



# CARTA DE ACEPTACIÓN

Por este medio, se otorga constancia de aceptación de tesina para obtener el título de Licenciado en Informática, que presenta el pasante *LUIS GUILLERMO FLORES CEJA*, con el tema denominado *"DATA WAREHOUSE"*.

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

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

ATENTAMENTE

*I.S.C. JABEL RESÉNDIZ GONZÁLEZ*
PROF. CURSO DE TITULACIÓN

# DATAWAREHOUSE

## AGRADECIMIENTOS

A mi madre, Martha Lucila Ceja Barrera, amigos y familiares por haber creído en mi, apoyándonos en las decisiones que tomamos y estar presentes en todos nuestros logros.

Al I.S.C. Jabel Reséndiz González, catedrático de esta facultad, por haberme dado la oportunidad de ser mi asesor y apoyado en la realización de este proyecto.

A nuestra escuela, Universidad Autónoma de Querétaro, por la educación que nos ha proporcionado haciendo de nosotros verdaderos agentes de cambio.

## RESUMEN

Para desarrollar una estrategia de acercamiento al mercado se necesitan nuevos sistemas de información. Tendrán como principal objetivo ofrecer los datos informacionales generados en la propia actividad de la compañía, desde una dimensión que permita una mayor capacidad de análisis e incremente la velocidad en la toma de decisiones. Una de las tecnologías que mejor se integran y soportan el nuevo modelo de negocio es el Datawarehousing.

Un sistema DataWareHouse define un nuevo concepto para el almacenamiento de datos, integra la información generada en todos los ámbitos de una actividad de negocio (Ventas, Producción, Finanzas, Marketing, etc.) y permite un acceso y explotación de la información contenida en las bases de datos, facilitando un amplio abanico de posibilidad de análisis multivariables que permitirán la toma de decisiones estratégicas.

El proceso integra toda la información de una compañía en un único depósito. La información que se genera en una compañía proviene de diferentes fuentes, formatos y tipos, que se consolidan, se transforman y se cargan en diferentes sistemas de gestión de datos, normalmente en RDBMS (Relational Database Management Systems).

Desde un sistema DataWareHouse, la información se puede mostrar y representar de muchas maneras. La forma más común de analizar la información, es utilizando un sistema de proceso de análisis en línea (OLAP, on-line analytical processing). Los productos OLAP ofrecen un rango muy variado de capacidades de análisis avanzado, como el multidimensional y el estadístico.

Un sistema DataWareHouse soporta también sofisticadas operaciones de análisis que se conocen con el término de Data Mining (Minería de datos).

Una de las novedades que aporta el Datawarehousing como sistema de análisis de información, es la creación de la Meta Información (metadata). Se trata de un archivo al que se le considera como diccionario de estructuras de datos que el administrador del sistema define con el objetivo de asistir en los procesos de consulta a la base de datos. La metadata se adaptará a las definiciones que el usuario utilizará posteriormente en sus interrogaciones al sistema.

De esta manera se conseguirá minimizar los complejos procedimientos de definición de nombres de campos, jerarquías y relaciones entre archivos.

La implantación consiste, en una primera fase, en el análisis de las necesidades de información a las que desea acceder cada compañía. Para ello se integrarán en el sistema todos aquellos datos operacionales necesarios, además de otras fuentes de información que sea menester incorporar. Definida la estructura de la base de datos se procederá a la carga de información y se crearán las agregaciones de datos para mejorar el rendimiento del sistema en los procesos de consulta más habituales. Finalmente, se incluirán en el sistema los procedimientos que permitan la actualización de información, cuya periodicidad dependerá de las necesidades de cada usuario.

El proceso de implantación de un sistema DataWareHouse, puede adaptarse de forma gradual o departamental creando soluciones específicas para cada área con el objetivo de conseguir resultados operativos a corto plazo. Esta solución departamental se denomina Datamart.

## INDICE

# PRESENTACION

## DataWareHouse

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

Un DataWareHouse 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 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 DataWareHouse se crea al extraer datos desde una o más bases de datos de aplicaciones operacionales. La data extraída es transformada para eliminar inconsistencias y resumir si es necesario y luego, cargadas en el DataWareHouse.

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. Este nuevo enfoque ayuda a las personas individuales, en todos los niveles de la empresa, a efectuar su toma de decisiones con más responsabilidad.

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 DataWareHouse.

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

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

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

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

El DW (de ahora en adelante los términos DataWareHouse, Datawarehousing, Warehouse y DW serán utilizados en forma indistinta) convierte entonces los datos operacionales de una organización en una herramienta competitiva, por hacerlos disponibles a los empleados que lo necesiten para el análisis y toma de decisiones.

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

El concepto DataMart es una extensión natural del DataWareHouse, y está enfocado a un departamento o área especifica, como por ejemplo los departamentos de Finanzas o Marketing.

Permitiendo así un mejor control de la información que se está abarcando. Toda empresa puede ser vista en base al proceso productivo que la sustenta. El resultado de los costos y beneficios de este proceso productivo forman una cadena de valor, donde cada eslabón (proceso de negocios) adiciona valor a la empresa. De esta forma es claro, que las empresas deben buscar optimizar cada uno de sus eslabones sin perder de vista la cadena total.

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

En esta nueva tecnología cada eslabón de la cadena de valor será representado por una base de datos multidimensional, la cual permite potencialmente administrar la etapa productiva que representa.

La cadena de valor total será representada entonces por el conjunto de bases de datos.

Así, esta tecnología permite que la organización disponga, en forma integrada y estandarizada, de la información correspondiente a la operación de la empresa, así como, proporciona a los usuarios, que tienen a su cargo la toma de decisiones, las herramientas adecuadas, para que a través de consultas rápidas, ellos mismos accedan la información requerida.

Desde que se inició la era de la computadora, 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 una data no limpia e inconsistente, sobre las cuales, en la mayoría de las veces, se toman decisiones importantes.

La gestión administrativa reconoce que 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, 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.

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

Un DataWareHouse 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 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 DataWareHouse se crea al extraer datos desde una o más bases de datos de aplicaciones operacionales. La data extraída es transformada para eliminar inconsistencias y resumir si es necesario y luego, cargadas en el DataWareHouse.

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. Este nuevo enfoque ayuda a las personas individuales, en todos los niveles de la empresa, a efectuar su toma de decisiones con más responsabilidad.

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 DataWareHouse.

El trabajo consta de tres capítulos:

- En el primero, **"Aspectos Teóricos"**, se dan los conceptos y el fundamento de la tecnología data warehousing.

- En el segundo, **"Proyecto de Elaboración de un DataWareHouse"**, se definen las estrategias para su planificación, desarrollo, diseño y gestión, además de los puntos que deben considerarse en la evaluación de la inversión.

- El tercer capítulo, **"Software en un DataWareHouse"**, permite comparar las herramientas de análisis adecuadas para los usuarios del DataWareHouse. Asimismo, en los Anexos, se proporciona una relación de los diversos Software que se usan en el manejo de un DataWareHouse.

# 1. ASPECTOS TEORICOS

## 1.1 INTRODUCCION AL CONCEPTO DATA WAREHOUSING

Data warehousing es el centro de la arquitectura para los sistemas de información en la década de los '90. Soporta el procesamiento informático al proveer una plataforma sólida, a partir de los datos históricos para hacer el análisis. Facilita la integración de sistemas de aplicación no integrados. Organiza y almacena los datos que se necesitan para el procesamiento analítico, informático sobre una amplia perspectiva de tiempo.
Un DataWareHouse o Depósito de Datos es una colección de datos orientado a temas, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones gerenciales.

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

| Base de Datos Operacional | DataWareHouse |
|---|---|
| Datos Operacionales | Datos del negocio para Información |
| Orientado a la aplicación | Orientado al sujeto |
| Actual | Actual + histórico |
| Detallada | Detallada + más resumida |
| Cambia continuamente | Estable |

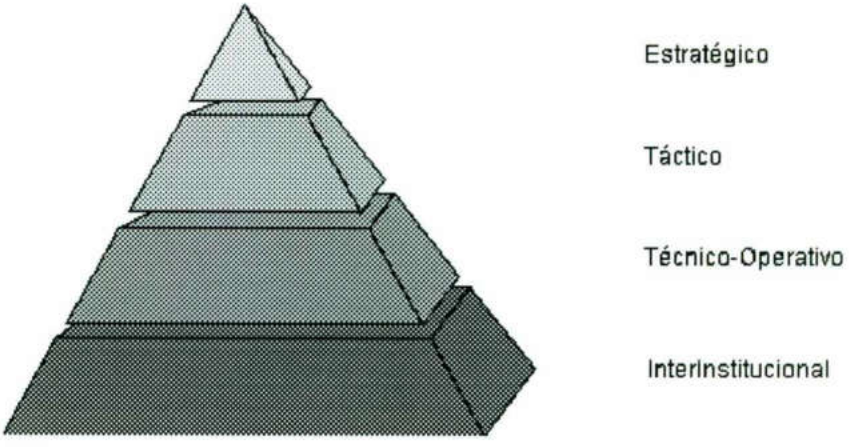
**Diferentes tipos de información**

El ingreso de datos en el DataWareHouse viene desde el ambiente operacional en casi todos los casos.

El DataWareHouse es siempre un almacén de datos transformados y separados fisicamente de la aplicación donde se encontraron los datos en el ambiente operacional.

## 1.2 SISTEMAS DE INFORMACION

En las metodologías anteriores, publicadas por el Instituto Nacional de Estadística e Informática - INEI y con el fin de proporcionar una visión más clara, 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 Información Georeferencial (GIS), Sistemas de

Simulación de Negocios (BIS y que en la práctica son sistemas expertos o de Inteligencia Artificial - AI).

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

Destacan entre ellos: los Sistemas Ofimáticos (OA), Sistemas de Transmisión de Mensajería (E-mail y Fax Server), coordinación y control de tareas (Work Flow) y tratamiento de documentos (Imagen, Trámite y Bases de Datos Documentarios).
Sistemas Técnico-Operativos, que cubren el núcleo de operaciones tradicionales de captura masiva de datos (Data Entry) y servicios básicos de tratamiento de datos, con tareas predefinidas (contabilidad, facturación, almacén, presupuesto, personal y otros sistemas administrativos). Estos sistemas están evolucionando con la irrupción de censores, autómatas, sistemas multimedia, bases de datos relacionales más avanzadas y data warehousing.

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

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

### 1.2.1 Sistemas técnico-operacionales

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

Debido a su volumen e importancia en la organización, los sistemas operacionales siempre han sido las primeras partes de la empresa a ser computarizados. A través de los años, estos sistemas operacionales se han extendido, revisados, mejorados y mantenidos al punto que hoy, ellos son completamente integrados en la organización.
Desde luego, la mayoría de las organizaciones grandes de todo el mundo, actualmente no podrían operar sin sus sistemas operacionales y los datos que estos sistemas mantienen.

### 1.2.2 Sistemas de Soporte de Decisiones

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

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

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

Mientras las necesidades de los datos operacionales se enfocan normalmente hacia una sola área, los datos para el soporte de decisiones, con frecuencia, toma un número de áreas diferentes y necesita cantidades grandes de datos operacionales relacionadas.
Son estos sistemas sobre los se basa la tecnología data warehousing.

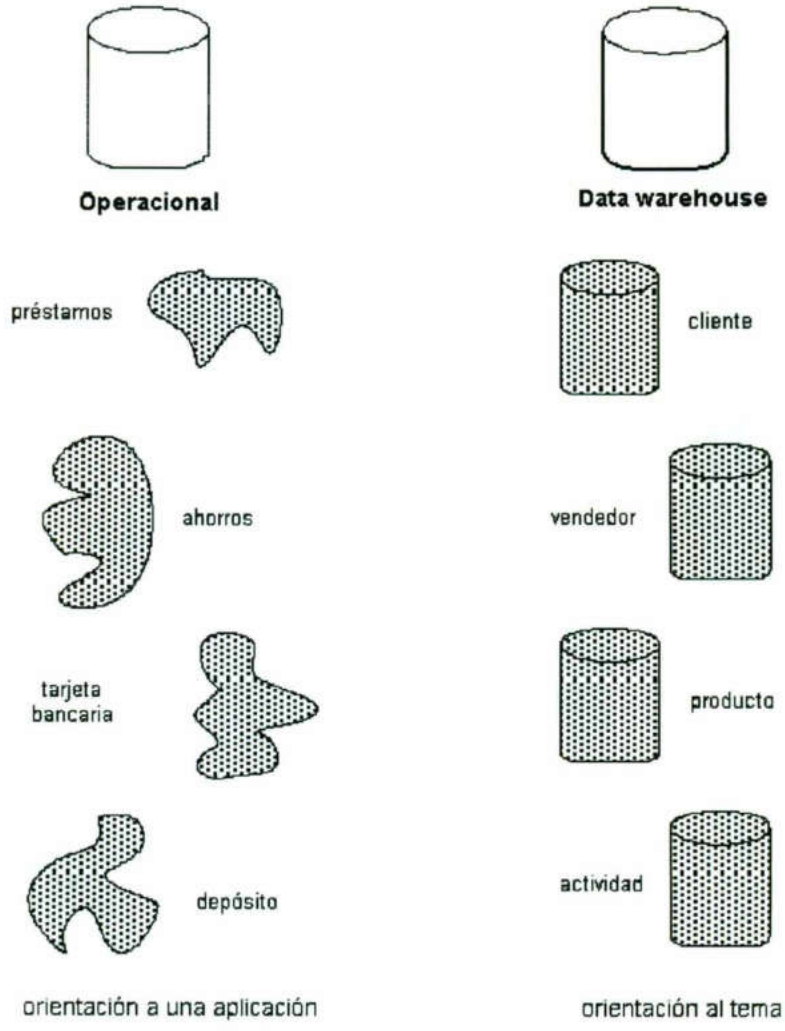
## 1.3 CARACTERISTICAS DE UN DATAWAREHOUSE

Entre las principales se tiene:

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

### 1.3.1 Orientado a Temas

Una primera característica del DataWareHouse es que la información se clasifica en base a los aspectos que son de interés para la empresa. Siendo así, los datos tomados están en contraste con los clásicos procesos orientados a las aplicaciones. En la Figura N° 1 se muestra el contraste entre los dos tipos de orientaciones.




**El data warehouse tiene una fuerte orientación al tema**

**Figura N° 1**

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 accesar 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 DataWareHouse. Las principales áreas de los temas influyen en la parte más importante de la estructura clave.
Las aplicaciones están relacionadas con el diseño de la base de datos y del proceso. En data warehousing se enfoca el modelamiento de datos y el diseño de la base de datos. El diseño del proceso (en su forma clásica) no es separado de este ambiente.

Las diferencias entre la orientación de procesos y funciones de las aplicaciones y la orientación a temas, radican en el contenido de la data a nivel detallado. En el DataWareHouse 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 DataWareHouse miden un espectro de tiempo y las relaciones encontradas en el DataWareHouse son muchas. Muchas de las reglas comerciales (y sus correspondientes relaciones de datos) se representan en el DataWareHouse, entre dos o más tablas.

### 1.3.2 Integración

El aspecto más importante del ambiente data warehousing es que la información encontrada al interior está siempre integrada.
La integración de datos se muestra de muchas maneras: en convenciones de nombres consistentes, en la medida uniforme de variables, en la codificación de estructuras consistentes, en atributos físicos de los datos consistentes, fuentes múltiples y otros.

El contraste de la integración encontrada en el DataWareHouse con la carencia de integración del ambiente de aplicaciones, se muestran en la Figura N° 2, con diferencias bien marcadas.

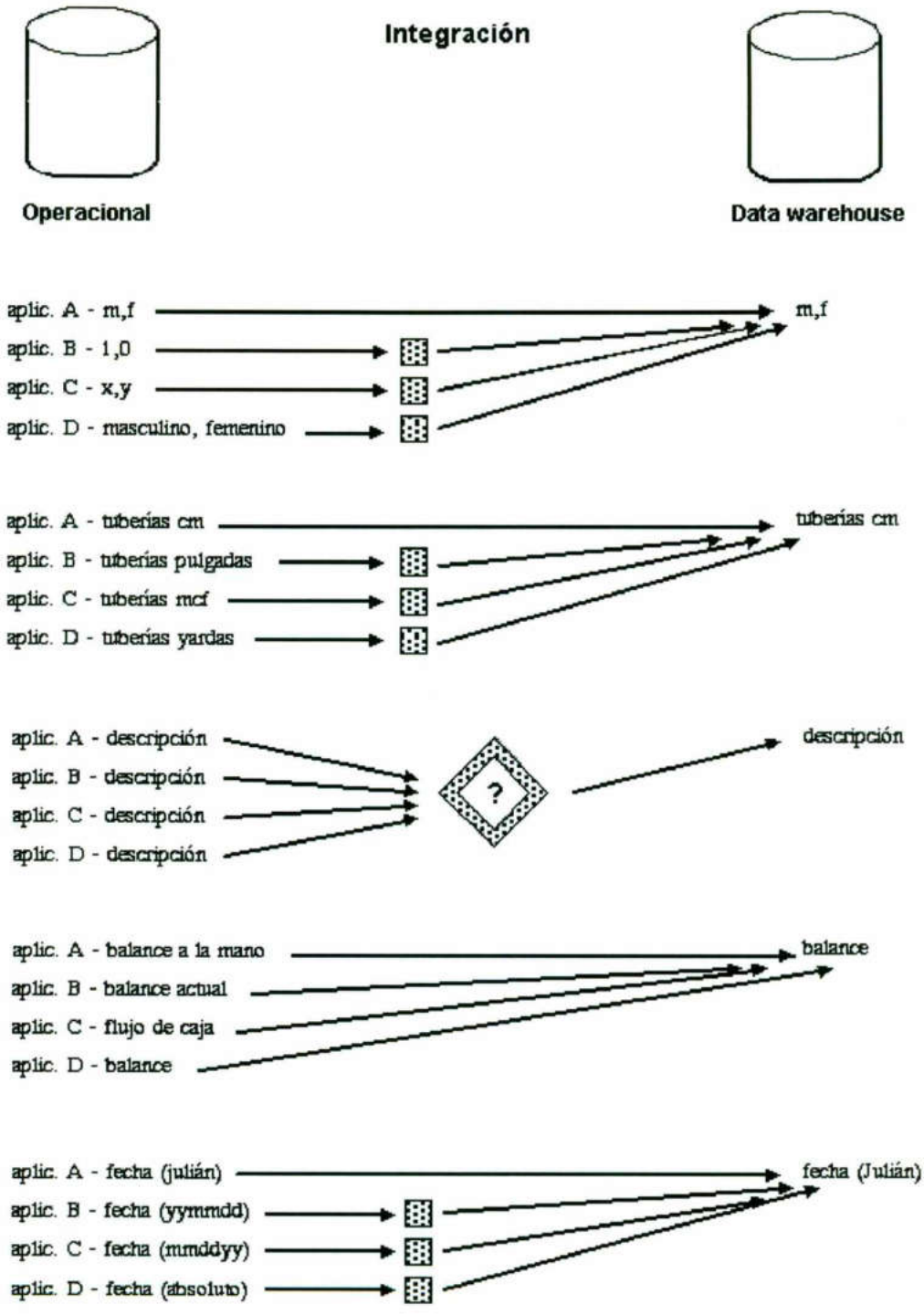
A través de los años, los diseñadores de las diferentes aplicaciones han tomado sus propias decisiones sobre cómo se debería construir una aplicación. Los estilos y diseños personalizados se muestran de muchas maneras.

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. La Figura N° 2 mencionada, muestra algunas de las diferencias más importantes en las formas en que se diseñan las aplicaciones.

Codificación. Los diseñadores de aplicaciones codifican el campo GÉNERO en varias formas. Un diseñador representa GÉNERO 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 GÉNERO llega al DataWareHouse. Probablemente "M" y "F" sean tan buenas como cualquier otra representación. Lo importante es que sea de cualquier fuente de donde venga, el GÉNERO debe llegar al DataWareHouse en un estado integrado uniforme.

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

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.




**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 N° 2**

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 DataWareHouse necesitará ser medida de la misma manera.

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

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

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

Cualquiera que sea la forma del diseño, el resultado es el mismo - la información necesita ser almacenada en el DataWareHouse 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 DataWareHouse, su enfoque deberá estar en el uso de los datos que se encuentre en el depósito, antes que preguntarse sobre la confiabilidad o consistencia de los datos.

### 1.3.3 De Tiempo Variante

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

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

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


Figura N° 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 DataWareHouse está en la estructura clave. Cada estructura clave en el DataWareHouse 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 DataWareHouse. 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 DataWareHouse, una vez registrada correctamente, no puede ser actualizada. La información del DataWareHouse 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 DataWareHouse. Los datos operacionales, siendo requeridos a partir del momento de acceso, pueden actualizarse de acuerdo a la necesidad.

### 1.3.4 No Volátil

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

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

Hay algunas consecuencias muy importantes de esta diferencia básica, entre el procesamiento operacional y del DataWareHouse. En el nivel de diseño, la necesidad de ser precavido para actualizar las anomalías no es un factor en el DataWareHouse, ya que no se hace la actualización de datos.
Esto significa que en el nivel físico de diseño, se pueden tomar libertades para optimizar el acceso a los datos, particularmente al usar la normalización y de normalización física.

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

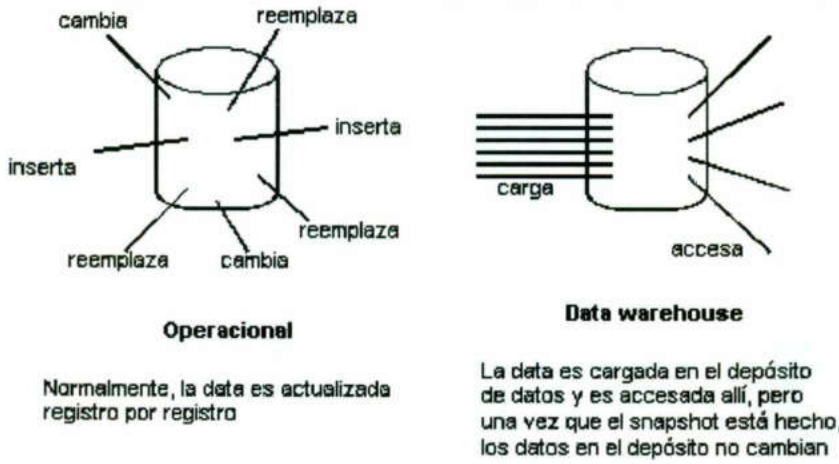



**Figura N° 4**

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 DataWareHouse no es necesario el procesamiento.

La fuente de casi toda la información del DataWareHouse es el ambiente operacional.

La primera impresión de muchas personas se centra en la gran redundancia de datos, entre el ambiente operacional y el ambiente de DataWareHouse. Dicho razonamiento es superficial y demuestra una carencia de entendimiento con respecto a qué ocurre en el DataWareHouse. 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 DataWareHouse.

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

DataWareHouse. Desde la perspectiva de los horizontes de tiempo únicos, hay poca superposición entre los ambientes operacionales y de DataWareHouse. El DataWareHouse contiene un resumen de la información que no se encuentra en el ambiente operacional. Los datos experimentan una transformación fundamental cuando pasa al DataWareHouse.

La mayor parte de los datos se alteran significativamente al ser seleccionados y movidos al DataWareHouse. Dicho de otra manera, la mayoría de los datos se alteran física y radicalmente cuando se mueven al depósito. No es la misma data que reside en el ambiente operacional desde el punto de vista de integración.

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

## 1.4 ESTRUCTURA DEL DATAWAREHOUSE

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

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

- Detalle de datos actuales
- Detalle de datos antiguos
- Datos ligeramente resumidos
- Datos completamente resumidos
- Meta data

**Detalle de datos actuales.-** En gran parte, el interés más importante radica en el detalle de los datos actuales, debido a que:

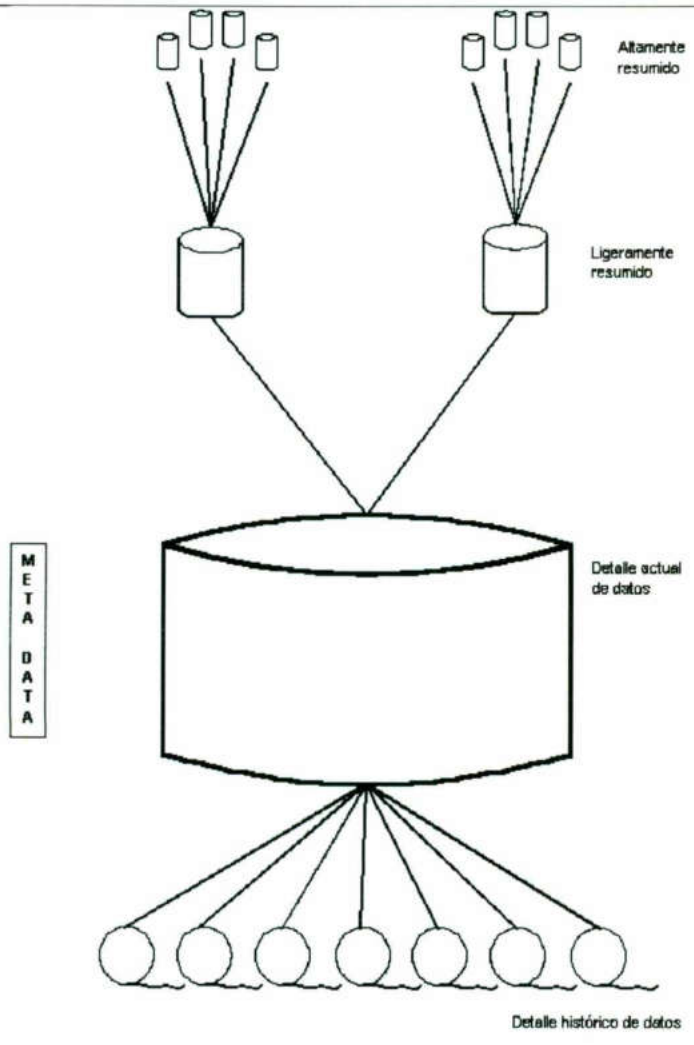
- Refleja las ocurrencias más recientes, las cuales son de gran interés.
- Es voluminoso, ya que se almacena al más bajo nivel de granularidad.
- Casi siempre se almacena en disco, el cual es de fácil acceso, aunque su administración sea costosa y compleja.

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

**Datos ligeramente resumidos.-** La data ligeramente resumida es aquella que proviene desde un bajo nivel de detalle encontrado al nivel de detalle actual. Este nivel del DataWareHouse casi siempre se almacena en disco. Los puntos en los que se basa el diseñador para construirlo son:

- Que la unidad de tiempo se encuentre sobre la esquematización hecha.
- Qué contenidos (atributos) tendrá la data ligeramente resumida.

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




**Estructura de los datos en un Data Warehouse**

**Figura Nº 5**

A veces se encuentra en el ambiente de DataWareHouse y en otros, fuera del límite de la tecnología que ampara al DataWareHouse. (De todos modos, los datos completamente resumidos son parte del DataWareHouse sin considerar donde se alojan los datos físicamente.)

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

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

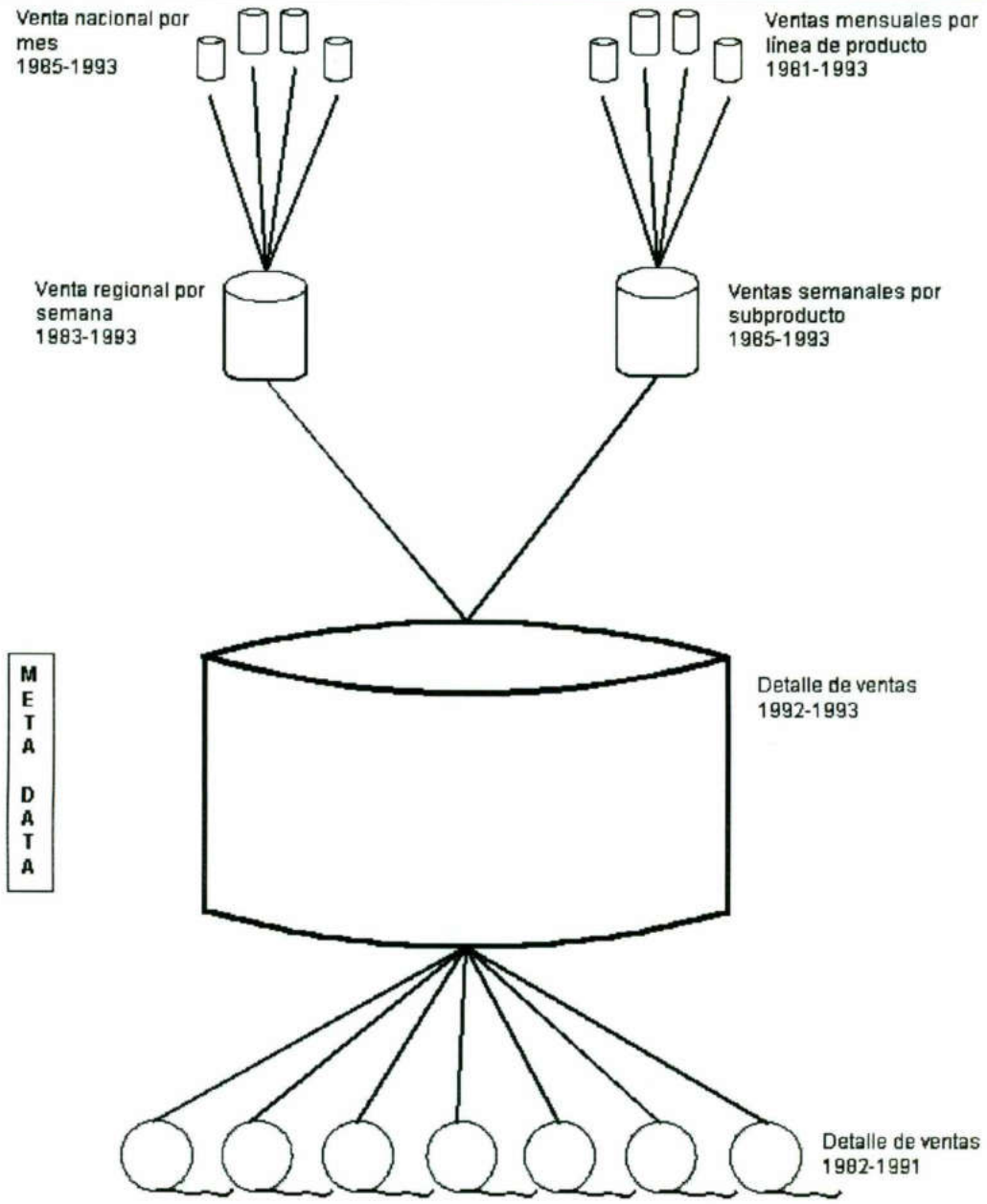
- Un directorio para ayudar al analista a ubicar los contenidos del DataWareHouse.
- Una guía para el mapping de datos de cómo se transforma, del ambiente operacional al de DataWareHouse.
- Una guía de los algoritmos usados para la esquematización entre el detalle de datos actual, con los datos ligeramente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

La metadata juega un papel mucho más importante en un ambiente data warehousing que en un operacional clásico.
A fin de recordar los diferentes niveles de los datos encontrados en el DataWareHouse, considere el ejemplo mostrado en la Figura N° 6.

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

El detalle actual contiene información desde 1992 a 1993 (suponiendo que 1993 es el año actual).

En general, el detalle de ventas no se ubica en el nivel de detalle actual hasta que haya pasado, por lo menos, veinticuatro horas desde que la información de ventas llegue a estar disponible en el ambiente operacional.




**Ejemplo de niveles de esquematización que podría encontrarse en un data warehouse**

**Figura N° 6**

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 DataWareHouse.

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

El detalle de ventas semanal es adicionalmente resumido en forma mensual, según una gama de líneas, para producir los datos completamente resumidos.
La metadata contiene (al menos):

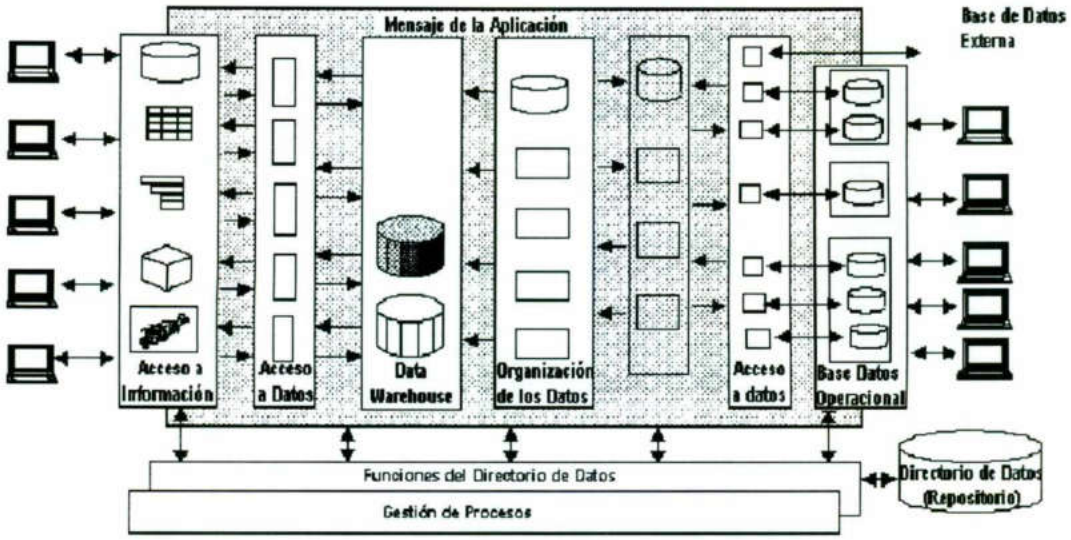
- La estructura de los datos
- Los algoritmos usados para la esquematización
- El mapping desde el ambiente operacional al DataWareHouse
- La información adicional que no se esquematiza es almacenada en el DataWareHouse.

En muchas ocasiones, allí se hará el análisis y se producirá un tipo u otro de resumen. El único tipo de esquematización que se almacena permanentemente en el DataWareHouse, es el de los datos que son usados frecuentemente. En otras palabras, si un analista produce un resumen que tiene una probabilidad muy baja de ser usado nuevamente, entonces la esquematización no es almacenada en el DataWareHouse.

## 1.5 ARQUITECTURA DE UN DATAWAREHOUSE

Una de las razones por las que el desarrollo de un DataWareHouse crece rápidamente, es que realmente es una tecnología muy entendible. De hecho, data warehousing puede representar mejor la estructura amplia de una empresa para administrar los datos informacionales 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 DataWareHouse. La Arquitectura de un DataWareHouse se muestra en la Figura N° 7.




**ARQUITECTURA DE UN DATAWAREHOUSE**

**Figura N° 7**

### 1.5.1 Elementos constituyentes de una Arquitectura DataWareHouse

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

La arquitectura se constituye de un número de partes interconectadas:

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

- Nivel de organización de datos

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

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

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

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

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

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

**Nivel de acceso a la información**

El nivel de acceso a la información de la arquitectura DataWareHouse, es el nivel del que el usuario final se encarga directamente.

En particular, representa las herramientas que el usuario final normalmente usa día a día. Por ejemplo: Excel, Lotus 1-2-3, Focus, Access, SAS, etc.

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

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

**Nivel de acceso a los datos**

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

Uno de los adelantos claves de los últimos años ha sido el desarrollo de una serie de "filtros" de acceso a datos, tales como EDA/SQL para accesar a casi todo los Sistemas de Gestión de Base de Datos (Data Base Management Systems - DBMSs) y sistemas de archivos de datos, relacionales o no.

Estos filtros permiten a las herramientas de acceso a la información, accesar también a la data almacenada en sistemas de gestión de base de datos que tienen veinte años de antigüedad.

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

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

El nivel de acceso a los datos entonces es responsable de la interfase entre las herramientas de acceso a la información y las bases de datos operacionales. En algunos casos, esto es todo lo que un usuario final necesita.

Sin embargo, en general, las organizaciones desarrollan un plan mucho más sofisticado para el soporte del data warehousing.

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

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

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

A fin de tener un depósito totalmente funcional, es necesario tener una variedad de metadata disponibles, información sobre las vistas de datos de los usuarios finales e información sobre las bases de datos operacionales. Idealmente, los usuarios finales deberían de accesar a los datos desde el DataWareHouse (o desde las bases de datos

operacionales), sin tener que conocer dónde residen los datos o la forma en que se han almacenados.

**Nivel de Gestión de Procesos**

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

**Nivel de Mensaje de la Aplicación**

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

**Nivel DataWareHouse (Físico)**

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

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

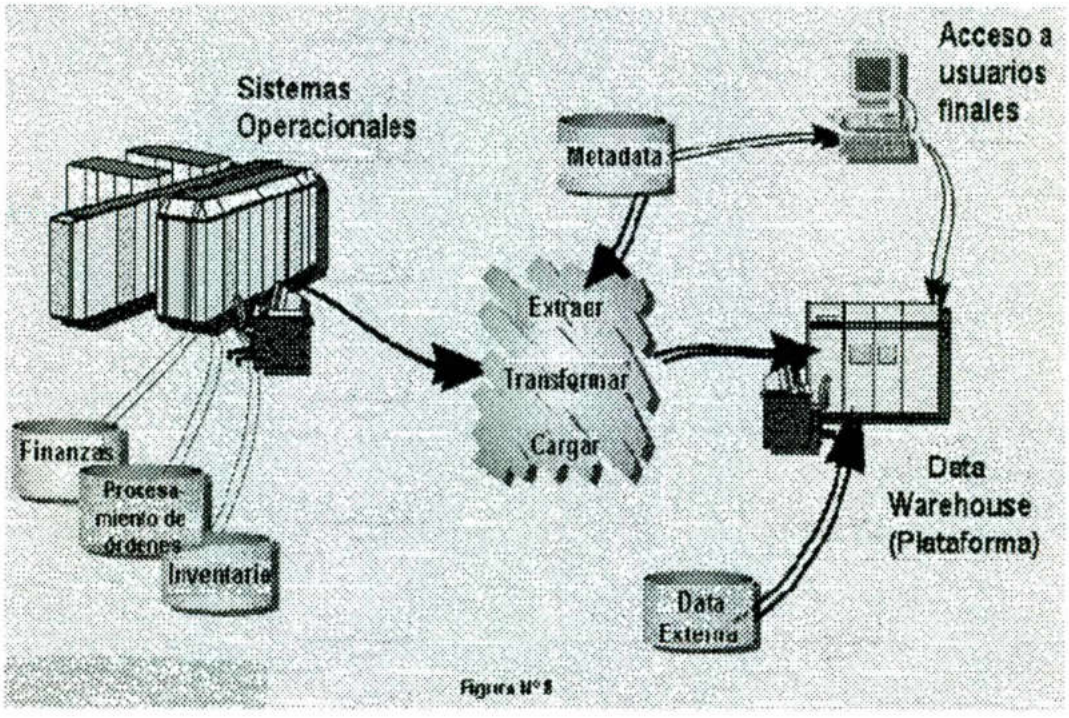
Nivel de Organización de Datos

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

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

### 1.5.2 Operaciones en un DataWareHouse

En la Figura N° 8 se muestra algunos de los tipos de operaciones que se efectúan dentro de un ambiente data warehousing.




**a) Sistemas Operacionales**

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

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

**b) Extracción, Transformación y Carga de los Datos**

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

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

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

**c) Metadata**

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

La metadata consiste de definiciones de los elementos de datos en el depósito, sistema(s) del (os) elemento(s) fuente. Como la data, se integra y transforma antes de ser almacenada en información similar.

**d) Acceso de usuario final**

Los usuarios accesan al DataWareHouse por medio de herramientas de productividad basadas en GUI (Graphical User Interface - Interfase gráfica de usuario). Pueden proveerse a los usuarios del DataWareHouse muchos de estos tipos de herramientas.
Estos pueden incluir software de consultas, generadores de reportes, procesamiento analítico en línea, herramientas data/visual mining, etc., dependiendo de los tipos de usuarios y sus requerimientos particulares. Sin embargo, una sola herramienta no satisface todos los requerimientos, por lo que es necesaria la integración de una serie de herramientas.

**e) Plataforma del DataWareHouse**

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

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

Muchas de las organizaciones quieran o no escogen una plataforma por diversas razones:

El Sistema X es nuestro sistema elegido o el Sistema Y está ya disponible sobre un sistema UNIX que nosotros ya tenemos.

Uno de los errores más grandes que las organizaciones cometen al seleccionar la plataforma, es que ellos presumen que el sistema (hardware y/o DBMS) escalará con los datos.

El sistema de depósito ejecuta las consultas que se pasa a los datos por el software de acceso a los datos del usuario.

Aunque un usuario visualiza las consultas desde el punto de vista de un GUI, las consultas típicamente se formulan como pedidos SQL, porque SQL es un lenguaje universal y el estándar de hecho para el acceso a datos.

**f) Datos Externos**

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

**Evolución del Depósito**

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

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

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

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

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

## 1.6 TRANSFORMACION DE DATOS Y METADATA

### 1.6.1 Transformación de Datos

Uno de los desafíos de cualquier implementación de DataWareHouse, es el problema de transformar los datos. La transformación se encarga de las inconsistencias en los formatos de datos y la codificación, que pueden existir dentro de una base de datos única y que casi siempre existen cuando múltiples bases de datos contribuyen al DataWareHouse.

En la Figura N° 9 se ilustra una forma de inconsistencia, en la cual el género se codifica de manera diferente en tres bases de datos diferentes. Los procesos de transformación de datos se desarrollan para direccionar estas inconsistencias.

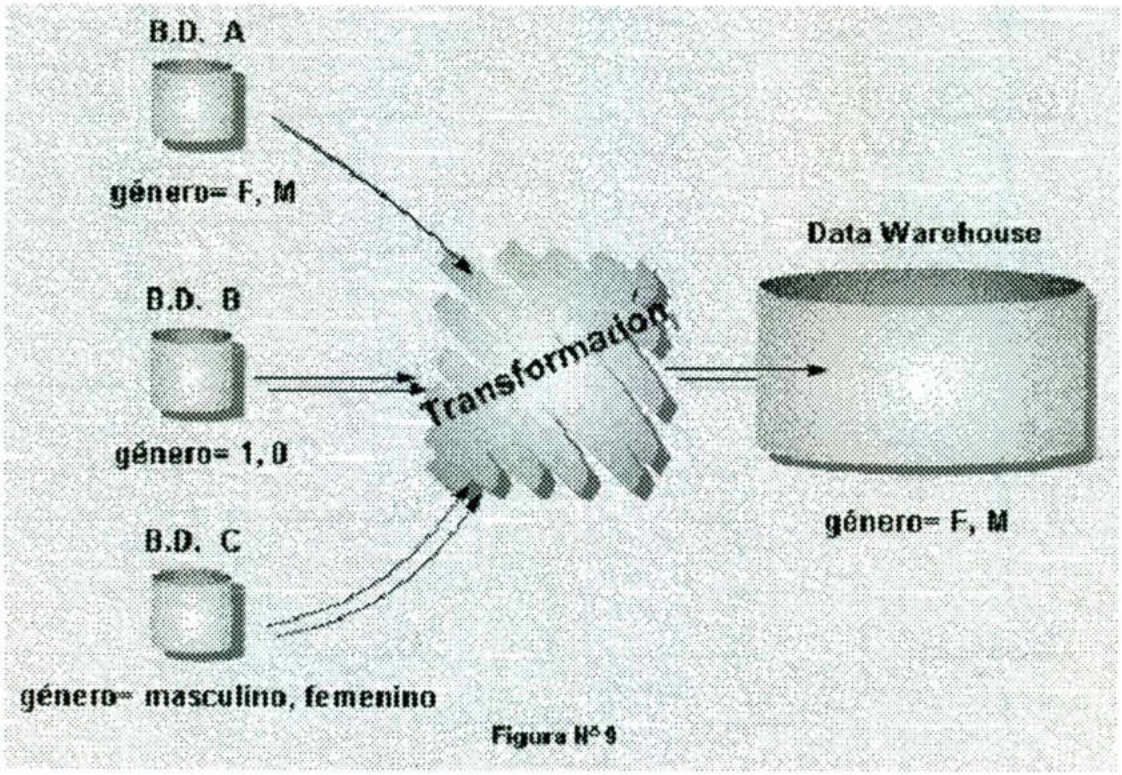


Figura N° 9

La transformación de datos también se encarga de las inconsistencias en el contenido de datos. Una vez que se toma la decisión sobre que reglas de transformación serán establecidas, deben crearse e incluirse las definiciones en las rutinas de transformación.
Se requiere una planificación cuidadosa y detallada para transformar datos inconsistentes en conjuntos de datos conciliables y consistentes para cargarlos en el DataWareHouse.

### 1.6.2 Metadata

Otro aspecto de la arquitectura de DataWareHouse es crear soporte a la metadata. Metadata es la información sobre los datos que se alimenta, se transforma y existe en el DataWareHouse. Metadata es un concepto genérico, pero cada implementación de la metadata usa técnicas y métodos específicos.
Estos métodos y técnicas son dependientes de los requerimientos de cada organización, de las capacidades existentes y de los requerimientos de interfase de usuario.

Hasta ahora, no hay normas para la metadata, por lo que la metadata debe definirse desde el punto de vista del software data warehousing, seleccionado para una implementación específica.

Típicamente, la metadata incluye los siguientes ítems:

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

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

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

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

integrado, son los esfuerzos que darán como resultado el mantenimiento del DataWareHouse.

## 1.7 FLUJO DE DATOS

Existe un flujo de datos normal y predecible dentro del DataWareHouse. La Figura N° 10 muestra ese flujo.

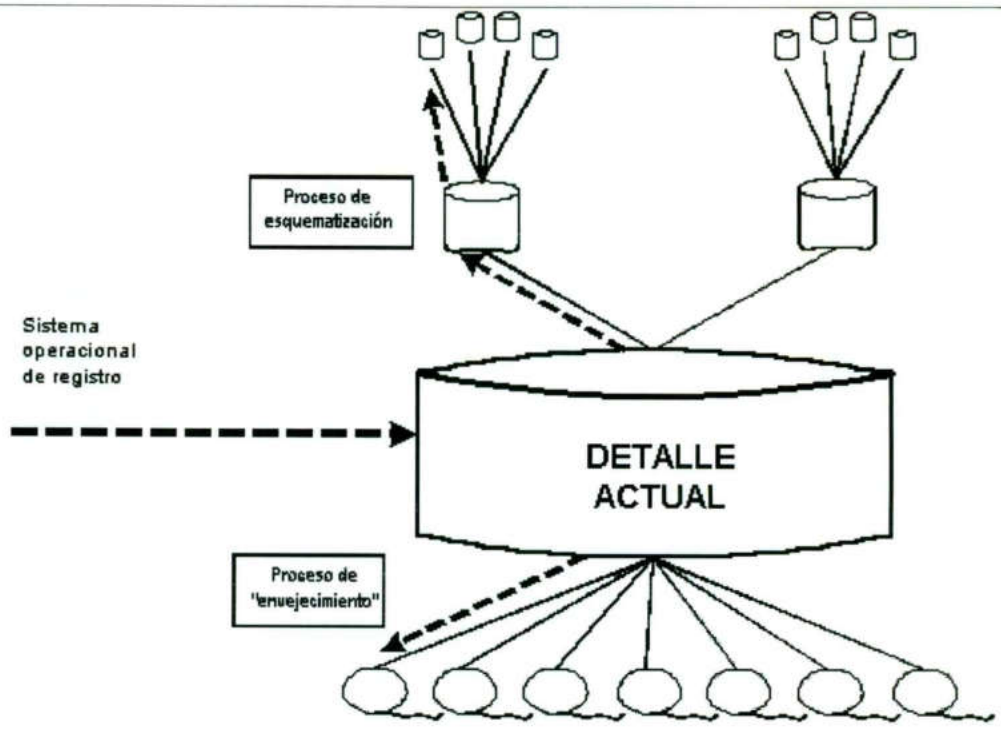
Los datos ingresan al DataWareHouse desde el ambiente operacional. (Hay pocas excepciones a esta regla).

Al ingresar al DataWareHouse, la información va al nivel de detalle actual, tal como se muestra. Se queda allí y se usa hasta que ocurra uno de los tres eventos siguientes:

- Sea eliminado
- Sea resumido
- Sea archivado

Con el proceso de desactualización en un DataWareHouse se mueve el detalle de la data actual a data antigua, basado en el tiempo de los datos. El proceso de esquematización usa el detalle de los datos para calcular los datos en forma ligera y completamente resumidos.

Hay pocas excepciones al flujo mostrado. Sin embargo, en general, para la mayoría de datos encontrados en un DataWareHouse, el flujo de la información es como se ha explicado.




**Flujo de datos en el data warehouse**

**Figura N° 10**

## 1.8 MEDIOS DE ALMACENAMIENTO PARA INFORMACION ANTIGUA

El símbolo mostrado en la Figura N° 11 para medios de almacenamiento de información antigua es la cinta magnética, que puede usarse para almacenar este tipo de información. De hecho hay una amplia variedad de medios de almacenamiento que deben considerarse para almacenar datos más antiguos. En la figura se muestra algunos de esos medios.

Dependiendo del volumen de información, la frecuencia de acceso, el costo de los medios y el tipo de acceso, es probable que otros medios de almacenamiento sirvan a las necesidades del nivel de detalle más antiguo en el DataWareHouse.


**Los medios de almacenamiento para la porción voluminosa del data warehouse puede ser de una amplia variedad de tipos de almacenamiento**

**Figura N° 11**

## 1.9 USOS DEL DATAWAREHOUSE

Los datos operacionales y los datos del DataWareHouse son accesados por usuarios que usan los datos de maneras diferentes.

| Uso de Base de Datos Operacionales | Uso de DataWareHouse |
|---|---|
| Muchos usuarios concurrentes | Pocos usuarios concurrentes |
| Consultas predefinidas y actualizables | Consultas complejas, frecuentemente no anticipadas. |
| Cantidades pequeñas de datos detallados | Cantidades grandes de datos detallados |
| Requerimientos de respuesta inmediata | Requerimientos de respuesta no críticos |

**Maneras diferentes de uso de datos**

Los usuarios de un DataWareHouse necesitan accesar a los datos complejos, frecuentemente desde fuentes múltiples y de formas no predecibles.

Los usuarios que accesan a los datos operacionales, comúnmente efectúan tareas predefinidas que, generalmente requieren acceso a una sola base de datos de una aplicación. Por el contrario, los usuarios que accesan al DataWareHouse, efectúan tareas que requieren acceso a un conjunto de datos desde fuentes múltiples y frecuentemente no son predecibles. Lo único que se conoce (si es modelada correctamente) es el conjunto inicial de datos que se han establecido en el depósito.

Por ejemplo, un especialista en el cuidado de la salud podría necesitar accesar a los datos actuales e históricos para analizar las tendencias de costos, usando un conjunto de consultas predefinidas. Por el contrario, un representante de ventas podría necesitar accesar a los datos de cliente y producto para evaluar la eficacia de una campaña de marketing, creando consultas base o ad-hoc para encontrar nuevamente necesidades definidas.

**Sólo pocos usuarios accesan a los datos concurrentemente**

En contraste a la producción de sistemas que pueden manejar cientos o miles de usuarios concurrentes, al DataWareHouse accesa un limitado conjunto de usuarios en cualquier tiempo determinado.

**Los usuarios generan un procesamiento no predecible complejo**

Los usuarios del DataWareHouse generan consultas complejas. A veces la respuesta a una consulta conduce a la formulación de otras preguntas más detalladas, en un proceso llamado drill down. El DataWareHouse puede incluir niveles de resúmenes múltiples, derivado de un conjunto principal, único, de datos detallados, para soportar este tipo de uso.

En efecto, los usuarios frecuentemente comienzan buscando en los datos resumidos y como identifican áreas de interés, comienzan a accesar al conjunto de datos detallado.

Los conjuntos de datos resumidos representan el "Qué" de una situación y los conjuntos de datos detallados permiten a los usuarios construir un cuadro sobre "Cómo" se ha derivado esa situación.

**Las consultas de los usuarios accesan a cantidades grandes de datos**

Debido a la necesidad de investigar tendencias y evaluar las relaciones entre muchas clases de datos, las consultas al DataWareHouse permiten accesar a volúmenes muy grandes tanto de data detallada como resumida. Debido a los requerimientos de datos históricos, los DataWareHouse evolucionan para llegar a un tamaño más grande que sus orígenes operacionales (de 10 a 100 veces más grande).

**Las consultas de los usuarios no tienen tiempos de respuesta críticos**

Las transacciones operacionales necesitan una respuesta inmediata porque un cliente puede estar esperando una respuesta. En el DataWareHouse, por el contrario, tiene un requerimiento de respuesta no-crítico porque el resultado frecuentemente se usa en un proceso de análisis y toma de decisiones.

Aunque los tiempos de respuesta no son críticos, los usuarios esperan una respuesta dentro del mismo día en que es hecha la consulta.

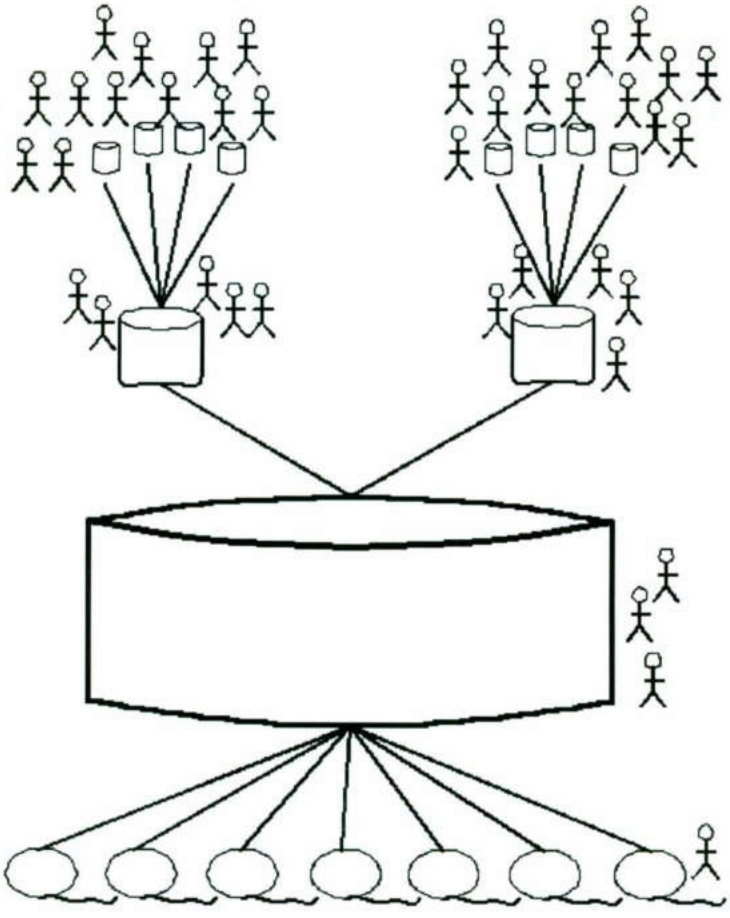
Por lo general, los diferentes niveles de datos dentro del DataWareHouse reciben diferentes usos. A más alto nivel de esquematización, se tiene mayor uso de los datos.

En la Figura N° 12 se muestra que hay mayor uso de los datos completamente resumidos, a diferencia de la información antigua que apenas es usada.

Hay una buena razón para mover una organización al paradigma sugerido en la figura, la utilización del recurso. La data más resumida, permite capturar los datos en forma más rápida y eficiente.

Si en una tarea se encuentra que se hace mucho procesamiento a niveles de detalle del DataWareHouse, entonces se consumirá muchos recursos de máquina. Es mejor hacer el procesamiento a niveles más altos de esquematización como sea posible.

Para muchas tareas, el analista de sistemas de soporte de decisiones usa la información a nivel de detalle en un pre-DataWareHouse. La seguridad de la información de detalle se consigue de muchas maneras, aun cuando estén disponibles otros niveles de esquematización. Una de las actividades del diseñador de datos es el de desconectar al usuario del sistema de soporte de decisiones del uso constante de datos a nivel de detalle más bajo.



**A más altos niveles de esquematización, más uso de los datos**

**Figura N° 12**

**El diseñador de datos tiene dos predisposiciones:**
Instalar un sistema chargeback, donde el usuario final pague por los recursos consumidos.

Señalar el mejor tiempo de respuesta que puede obtenerse cuando se trabaja con la data a un nivel alto de esquematización, a diferencia de un pobre tiempo de respuesta que resulta de trabajar con los datos a un nivel bajo de detalle.

Para ilustrar cómo un DataWareHouse puede ayudar a una organización a mejorar sus operaciones, se muestra un ejemplo de lo que es el desarrollo de actividades sin tener un DataWareHouse.

**Ejemplo: Preparación de un reporte complejo**
Considere un problema bastante típico en una compañía de fabricación grande en el que se pide una información (un reporte) que no está disponible.

El informe incluye las finanzas actuales, el inventario y la condición de personal, acompañado de comparaciones del mes actual con el anterior y el mismo mes del año anterior, con una comparación adicional de los 3 años precedentes. Se debe explicar cada desviación de la tendencia que cae fuera de un rango predefinido.

Sin un DataWareHouse, el informe es preparado de la manera siguiente:

La información financiera actual se obtiene desde una base de datos mediante un programa de extracción de datos, el inventario actual de otro programa de extracción de otra base de datos, la condición actual de personal de un tercer programa de extracción y la información histórica desde un backup de cinta magnética o CD-ROM.

Lo más interesante es que se ha pedido otro informe que continúe al primer informe (debido a que las preguntas se originaron a partir del anterior). El hecho es, que ninguno de los trabajos realizados hasta aquí (por ejemplo, diversos programas de extracción) se

pueden usar para los próximos o para cualquier reporte subsiguiente. Imagine el tiempo y el esfuerzo que se ha desperdiciado por un enfoque anticuado. (Ver Figura N° 13).

Las inconsistencias deben identificarse en cada conjunto de datos extraídos y resolverse, por lo general, manualmente. Cuando se completa todo este procesamiento, el reporte puede ser formateado, impreso, revisado y transmitido.

Nuevamente, el punto importante aquí es que todo el trabajo desempeñado para hacer este informe no afecta a otros reportes que pueden solicitarse es decir, todos ellos son independientes y caros, desde el punto de vista de recursos y productividad.

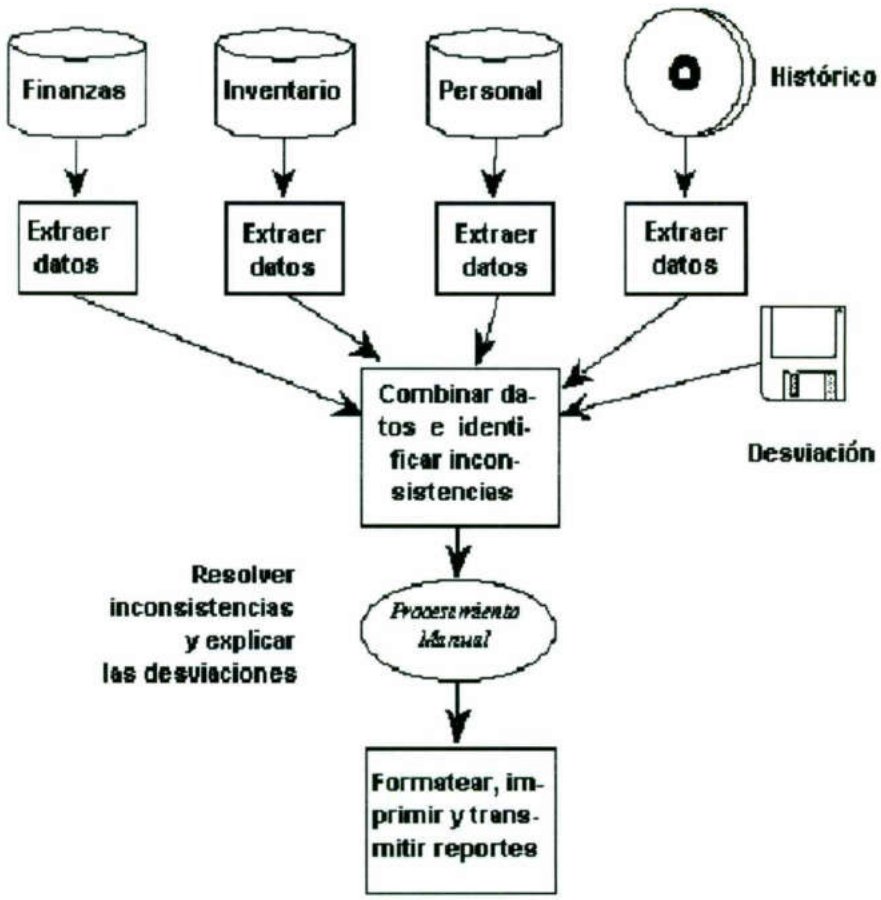



**Figura N° 13**

Al crear un DataWareHouse y combinar todos los datos requeridos, se obtienen los siguientes beneficios:

- Las inconsistencias de los datos se resuelven automáticamente cuando los elementos de datos se cargan en el DataWareHouse, no manualmente, cada vez que se prepara un reporte.
- Los errores que ocurrieron durante el proceso complejo de la preparación del informe, se minimizan porque el proceso es ahora mucho más simple.
- Los elementos de datos son fácilmente accesibles para otros usos, no sólo para un reporte particular.
- Se crea una sola fuente.

## 1.10 CONSIDERACIONES ADICIONALES

Hay algunas consideraciones adicionales que deben tenerse en cuenta al construir y administrar el DataWareHouse.

La primera consideración es respecto al índice. La información de los niveles de esquematización más altos pueden ser libremente indexados, mientras que las de los niveles más bajos de detalle, por ser tan voluminosa, pueden ser indexados moderadamente.

Por lo mismo, los datos en los niveles más altos de detalle pueden ser reestructurados fácilmente, mientras que el volumen de datos en los niveles más inferiores es tan grande, que los datos no pueden ser fácilmente reestructurados.

Por consiguiente, el modelo de datos y el diseño clásico fundamentan que el DataWareHouse se aplique casi exclusivamente al nivel actual de detalle.

En otras palabras, las actividades de modelamiento de datos no se aplican a los niveles de esquematización, en casi todos los casos.

Otra consideración estructural es la partición de la información en el DataWareHouse. El nivel de detalle actual es casi siempre particionado.

La partición puede hacerse de dos maneras: al nivel de DBMS y al nivel de la aplicación. En la partición DBMS, se conoce las particiones y se administra por consiguiente. En el caso de la partición de las aplicaciones, sólo los programadores de las mismas conocen las particiones y la responsabilidad de su administración es asignada a ellos.

Al interior de las particiones DBMS, mucho de los trabajos de infraestructura se hacen automáticamente. Pero existe un elevado grado de rigidez asociada con la gestión automática de las particiones. En el caso de las particiones de las aplicaciones del DataWareHouse, la mayor parte del trabajo recae sobre el programador, pero el resultado final es que la gestión de datos es más flexible.

## 1.11 EJEMPLO DE UN DATAWAREHOUSE

En la Figura N° 14 se muestra un ejemplo hipotético de un DataWareHouse estructurado para un centro de producción industrial.

Se muestra sólo el detalle actual, no así los niveles de esquematización ni los archivos de detalle más antiguos.

Además, se observa que hay tablas del mismo tipo divididas a través del tiempo.

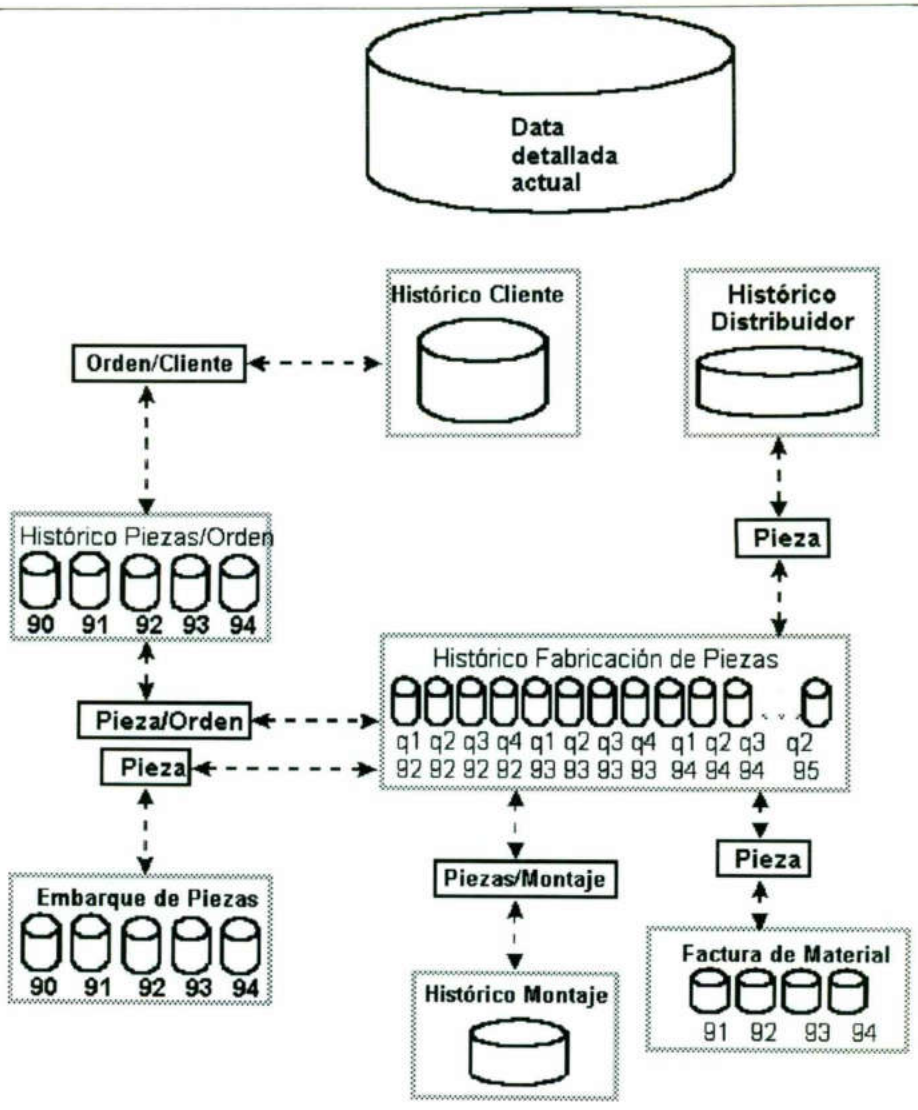



**Figura N° 14**

Por ejemplo, para el histórico de la fabricación de las piezas, hay muchas tablas separadas físicamente, representando cada una un trimestre diferente. La estructura de los datos es consistente con la tabla de la elaboración de las piezas, aunque físicamente hay muchas tablas que lógicamente incluyen el histórico.

Para los diferentes tipos de tablas hay diferentes unidades de tiempo que físicamente dividen las unidades de información. El histórico de fabricación está dividido por

trimestres, el histórico de la orden de piezas está dividido por años y el histórico de cliente es un archivo único, no dividido por el tiempo.

Así también, las diferentes tablas son vinculadas por medio de un identificador común, piezas u órdenes de piezas (la representación de la interrelación en el ambiente de depósito toma una forma muy diferente al de otros ambientes, tal como el ambiente operacional).

## 1.12 EXCEPCIONES EN UN DATAWAREHOUSE

Mientras que los componentes del DataWareHouse trabajan de acuerdo al modelo descrito para casi todos los datos, hay pocas excepciones útiles que necesitan ser discutidas.

Una de ellas es la data resumida pública, que es la data que ha sido calculada fuera del DataWareHouse pero es usada a través de la corporación. La data resumida pública se almacena y administra en el DataWareHouse, aunque su cálculo se haya hecho fuera de él.

Un ejemplo clásico de data resumida pública es el archivamiento trimestral hecho por cada compañía pública. Los contadores trabajan para producir cantidades como rentas trimestrales, gastos trimestrales, ganancias trimestrales y otros. El trabajo hecho por los contadores está fuera del DataWareHouse.

Sin embargo, esas cantidades referenciales producidas por ellos se usan ampliamente dentro de la corporación para marketing, ventas, etc. Una vez que se haya hecho el archivo, los datos se almacenan en el DataWareHouse.

Otro excepcional tipo de datos a veces encontrados en un DataWareHouse es el detalle de los datos permanentes, que resulta de la necesidad de una corporación para almacenar la data a un nivel detallado permanentemente por razones éticas o legales.

Si una corporación expone a sus trabajadores a sustancias peligrosas hay una necesidad de detalle de datos permanente. Si una corporación produce un producto que involucra la

seguridad pública, tal como la construcción de las partes de aviones, hay una necesidad de datos permanentes. Si una corporación se compromete con contratos peligrosos, hay una necesidad de detalle de datos permanentes.

La organización simplemente no puede dejar los detalles porque en futuros años, en el caso de una demanda, una notificación, un edificio en disputa, etc., se incrementaría la exposición de la compañía. Por lo tanto hay un único tipo de datos en el DataWareHouse conocido como detalle de datos permanentes.

El detalle de datos permanentes comparte muchas de las mismas consideraciones como otro DataWareHouse, excepto que:

- El medio donde se almacena la data debe ser tan seguro como sea posible.
- Los datos deben permitir ser restaurados.
- Los datos necesitan un tratamiento especial en su indexación, ya que de otra manera los datos pueden no ser accesibles aunque se haya almacenado con mucha seguridad.

# 2. PROYECTO DE ELABORACION DE UN DATAWAREHOUSE

## 2.1 FASE: ORGANIZACION

La planificación es el proceso más importante que determina la clase de tipo de estrategias datawarehousing que una organización iniciará.

### 2.1.1 FACTORES EN LA PLANIFICACION DE UN DATAWAREHOUSE

No existe una fórmula de garantía real para el éxito de la construcción de un DataWareHouse, pero hay muchos puntos que contribuyen a ese objetivo.
A continuación, se indican algunos puntos claves que deben considerarse en la planificación de un DataWareHouse:

**1. Establecer una asociación de usuarios, gestión y grupos**
Es esencial involucrar tanto a los usuarios como a la gestión para asegurar que el DataWareHouse contenga información que satisfaga los requerimientos de la empresa.

La gestión puede ayudar a priorizar la fase de la implementación del DataWareHouse, así como también la selección de herramientas del usuario. Los usuarios y la gestión justifican los costos del DataWareHouse 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 DataWareHouse como una tecnología clave para la empresa. Estos mismos criterios (alcance limitado, reembolso medible y beneficios claros para la empresa) se aplican a cada fase de la implementación de un DataWareHouse.

**3. Construir prototipos rápida y frecuentemente**

La única manera para asegurar que el DataWareHouse 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. Implementación incremental**

La implementación incremental reduce riesgos y asegura que el tamaño del proyecto permanezca manejable en cada fase.

**5. Reportar activamente 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 DataWareHouse ha ayudado a los usuarios a operar más efectivamente puede apoyar la construcción del DataWareHouse a lo largo de una empresa.

La retroalimentación del usuario también ayuda a comprender cómo evoluciona la implementación del DataWareHouse a través del tiempo para reunir requerimientos de usuario nuevamente identificados.

## 2.1.2 ESTRATEGIAS PARA EL DESARROLLO DE UN DATAWAREHOUSE

Antes de desarrollar un DataWareHouse, es crítico el desarrollo de una estrategia equilibrada que sea apropiada para sus necesidades y sus usuarios.

Las preguntas que deben tenerse en cuenta son:

- ¿Quién es el auditorio?
- ¿Cuál es el alcance?
- ¿Qué tipo de DataWareHouse debería construirse?

Existe un número de estrategias mediante las cuales las organizaciones pueden conseguir sus DataWareHouse.

**1ra.:** Establecer un ambiente "DataWareHouse virtual", el cual puede ser creado por:

- Instalación de un conjunto de facilidades para acceso a datos, directorio de datos y gestión de proceso.
- Entrenamiento de usuarios finales.
- Control de cómo se usan realmente las instalaciones del DataWareHouse.
- Basados en el uso actual, crear un DataWareHouse físico para soportar los pedidos de alta frecuencia.

**2da.:** Construir una copia de los datos operacionales desde un sistema operacional único y posibilitar al DataWareHouse 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, entonces se puede crear una serie de problemas.

**3ra.:** Finalmente, la estrategia data warehousing óptima es seleccionar el 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 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 DataWareHouse.

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 alta performance, usando el núcleo del DataWareHouse como base.

En conclusión, no se tiene un enfoque único para construir un DataWareHouse 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 va evolucionando, se aprende cada vez más y más sobre el desarrollo de DataWareHouse, que resulta en que el único enfoque práctico para al almacenamiento de datos es la evolución de uno mismo.

### 2.1.3 ESTRATEGIAS PARA EL DISEÑO DE UN DATAWAREHOUSE

El diseño de los DataWareHouse es muy diferente al diseño de los sistemas operacionales tradicionales. Se pueden considerar los siguientes puntos:

**1ra.:** Los usuarios de los DataWareHouse usualmente no conocen mucho sobre sus requerimientos y necesidades como los usuarios operacionales.

**2da.:** El diseño de un DataWareHouse, con frecuencia involucra lo que se piensa en términos más amplios y con conceptos del negocio más difíciles de definir que en el diseño de un sistema operacional. Al respecto, un DataWareHouse está bastante cerca a Reingeniería de los Procesos del Negocio (Business Process Reengineering).

**3ra.:** Finalmente, la estrategia de diseño ideal para un data warehousing es generalmente de afuera hacia adentro (outside-in) a diferencia de arriba hacia abajo (top-down).

A pesar que el diseño del DataWareHouse 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 DataWareHouse tienen que usar muchos "trucos" para ayudar a sus usuarios a "visualizar" sus requerimientos. Por ello, son esenciales los prototipos de trabajo.

### 2.1.4 ESTRATEGIAS PARA LA GESTION DE UN DATAWAREHOUSE

Los DataWareHouse requieren una comercialización y gestión muy cuidadosa. Debe considerarse lo siguiente:

1ra.: Un DataWareHouse 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.

**2da.:** La administración debe reconocer que el mantenimiento de la estructura del DataWareHouse es tan crítico como el mantenimiento de cualquier otra aplicación de misión-crítica. De hecho, la experiencia ha demostrado que los DataWareHouse llegarán a ser rápidamente uno de los sistemas más usados en cualquier organización.

**3ra.:** La gestión debe comprender también que si se embarcan sobre un programa data warehousing, se crearán nuevas demandas sobre sus sistemas operacionales, que son:

- Demandas para mejorar datos
- Demandas para una data consistente
- Demandas para diferentes tipos de datos, etc.

## 2.2. FASE: DESARROLLO

### 2.2.1 ¿PORQUE CONSTRUIR BLOQUES DE DATAWAREHOUSE?

Para ampliar un negocio, se necesita que la información sea comprensible. Para muchas compañías, esto significa un gran DataWareHouse que muestre, junto a los datos no filtrados y dispersos, nuevas formas creativas de presentación.

Las herramientas para capturar y explorar los datos al detalle evolucionan, así como nuestra capacidad para encontrar las formas de explotar los datos recolectados.
En los últimos 10 años se han combinado dos factores para ayudar a la difusión de los DataWareHouse. Ellos son:

1° Se ha reconocido los beneficios del procesamiento analítico en línea (On Line Analytical Processing - OLAP), más allá de las áreas tradicionales de marketing y finanzas.

Las organizaciones saben que los conocimientos inmersos en las masas de datos que rutinariamente recogen sobre sus clientes, productos, operaciones y actividades comerciales, contribuyen a reducir los costos de operación y aumentar las rentas, por no mencionar que es más fácil la toma de decisiones estratégicas.

2° El crecimiento de la computación cliente/servidor, ha creado servidores de hardware y software más poderosos y sofisticados que nunca. Los servidores de hoy compiten con las mainframes de ayer y ofrecen arquitecturas de memoria tecnológicamente superiores, procesadores de alta velocidad y capacidades de almacenamiento masivas.

Al mismo tiempo, los Sistemas de Gestión de Base de Datos (Data Base Management Systems - DBMS(s)) modernos, proporcionan mayor soporte para las estructuras de datos complejas.

De esta renovación de hardware y software surgen los DataWareHouse multiterabyte que ahora se ve en ambientes de cliente/servidor.

### 2.2.2 CONSIDERACIONES PREVIAS AL DESARROLLO DE UN DATAWAREHOUSE

Hay muchas maneras para desarrollar DataWareHouse como tantas organizaciones existen. Sin embargo, hay un número de dimensiones diferentes que necesitan ser consideradas:

- Alcance de un DataWareHouse
- Redundancia de datos
- Tipo de usuario final

La Figura N° 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.

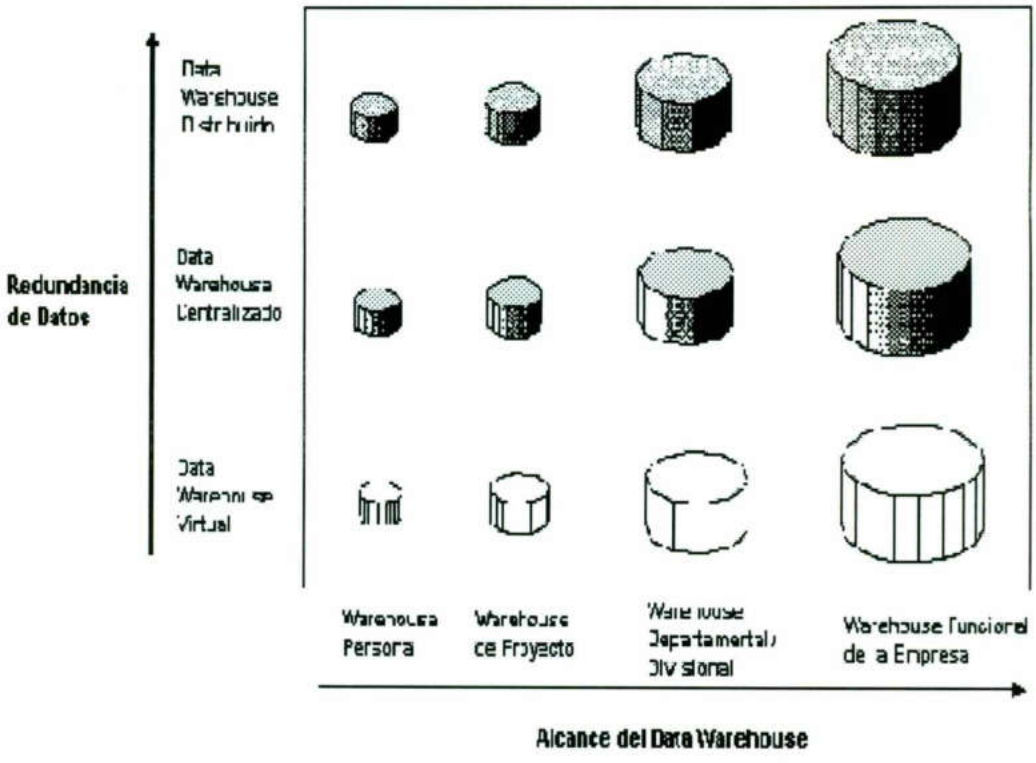



Figura N° 15

### 2.2.2.1 ALCANCE DEL DATAWAREHOUSE

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

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

### 2.2.2.2 REDUNDANCIA DE DATOS

Hay tres niveles esenciales de redundancia de datos que las empresas deberían considerar en sus opciones de DataWareHouse:

- DataWareHouse "virtual" o "Point to Point"
- DataWareHouse "centrales"
- DataWareHouse "distribuidos"

No se puede pensar en un único enfoque. Cada opción adapta un conjunto específico de requerimientos y una buena estrategia de almacenamiento de datos, lo constituye la inclusión de las tres opciones.

**1. DataWareHouse "Virtual" o "Point to Point"**

Una estrategia de DataWareHouse virtual, significa que los usuarios finales pueden accesar a bases de datos operacionales directamente, usando cualquier herramienta que posibilite "la red de acceso de datos".

Este enfoque provee flexibilidad así como también la cantidad mínima de datos redundantes que deben cargarse y mantenerse. Además, se pueden colocar las cargas de consulta no planificadas más grandes, sobre sistemas operacionales.

Como se verá, el almacenamiento virtual es, frecuentemente, una estrategia inicial, en organizaciones donde hay una amplia (pero en su mayor parte indefinida) necesidad de

conseguir la data operacional, desde una clase relativamente grande de usuarios finales y donde la frecuencia probable de pedidos es baja.

Los depósitos virtuales de datos proveen un punto de partida para que las organizaciones determinen qué usuarios finales están buscando realmente.

**2. DataWareHouse "Centrales"**

El concepto de DataWareHouse centrales es el concepto inicial que se tiene del DataWareHouse. Es una única base de datos física, que contiene todos los datos para un área funcional específica, departamento, división o empresa.

Los DataWareHouse centrales se seleccionan por lo general donde hay una necesidad común de los datos informáticos y un número grande de usuarios finales ya conectados a una red o computadora central. Pueden contener datos para cualquier período específico de tiempo. Comúnmente, contienen datos de sistemas operacionales múltiples.

Los DataWareHouse centrales son reales. Los datos almacenados en el DataWareHouse son accesibles desde un lugar y deben cargarse y mantenerse sobre una base regular. Normalmente se construyen alrededor de RDBMs avanzados o, en alguna forma, de servidor de base de datos informático multidimensional.

**3. DataWareHouse Distribuidos**

Los DataWareHouse distribuidos son aquellos en los cuales ciertos componentes del depósito se distribuyen a través de un número de bases de datos físicas diferentes.

Cada vez más, las organizaciones grandes están tomando decisiones a niveles más inferiores de la organización y a la vez, llevando los datos que se necesitan para la toma de decisiones a la red de área local (Local Area Network - LAN) o computadora local que sirve al que toma decisiones.

Los DataWareHouse distribuidos comúnmente involucran la mayoría de los datos redundantes y como consecuencia de ello, se tienen procesos de actualización y carga más complejos.

#### 2.2.2.3 TIPO DE USUARIO FINAL

De la misma forma que hay una gran cantidad de maneras para organizar un DataWareHouse, es importante notar que también hay una gama cada vez más amplia de usuarios finales.

En general, se puede considerar tres grandes categorías:

- Ejecutivos y gerentes. "Power users" o "Buzo de Información" (analistas financieros y de negocios, ingenieros, etc.).
- Usuarios de soporte (de oficina, administrativos, etc.).
- Cada una de estas categorías diferentes de usuario tienen su propio conjunto de requerimientos para los datos, acceso, flexibilidad y facilidad de uso.

### 2.2.3 ELEMENTOS CLAVES PARA EL DESARROLLO DE UN DATAWAREHOUSE

Los DataWareHouse exitosos comienzan cuando se escogen e integran satisfactoriamente tres elementos claves.

Un DataWareHouse 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 DataWareHouse 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:

- Arquitectura total del depósito
- Arquitecturas del servidor
- Sistemas de Gestión de Base de Datos

### 2.2.3.1 DISEÑO DE LA ARQUITECTURA

**a) Arquitectura del Depósito**

El desarrollo del DataWareHouse 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 ítems 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:

**1°** Un plan para almacenar los datos de su compañía, que podría obtenerse desde fuentes múltiples internas y externas, es consolidar la base de datos en un DataWareHouse integrado. El enfoque consolidado proporciona eficiencia tanto en la potencia de procesamiento como en los costos de soporte. (Ver Figura N° 16).

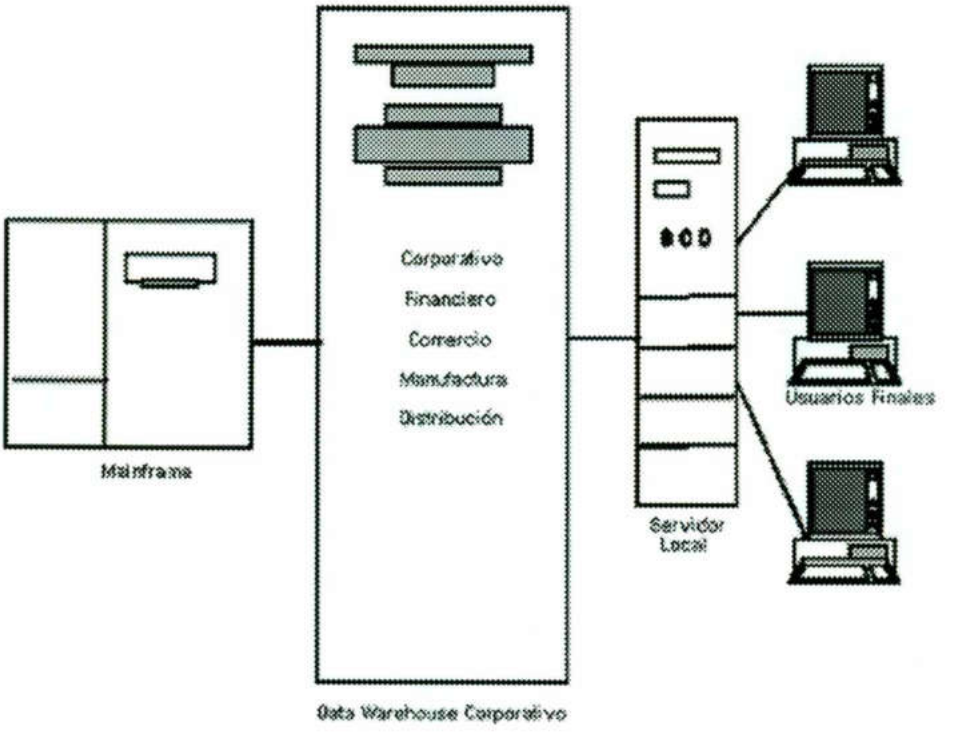



**Figura N° 16**

**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 fisicamente almacenadas en la misma plataforma.**

2° La arquitectura global distribuye 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. Esto se puede observar claramente en la figura 17.

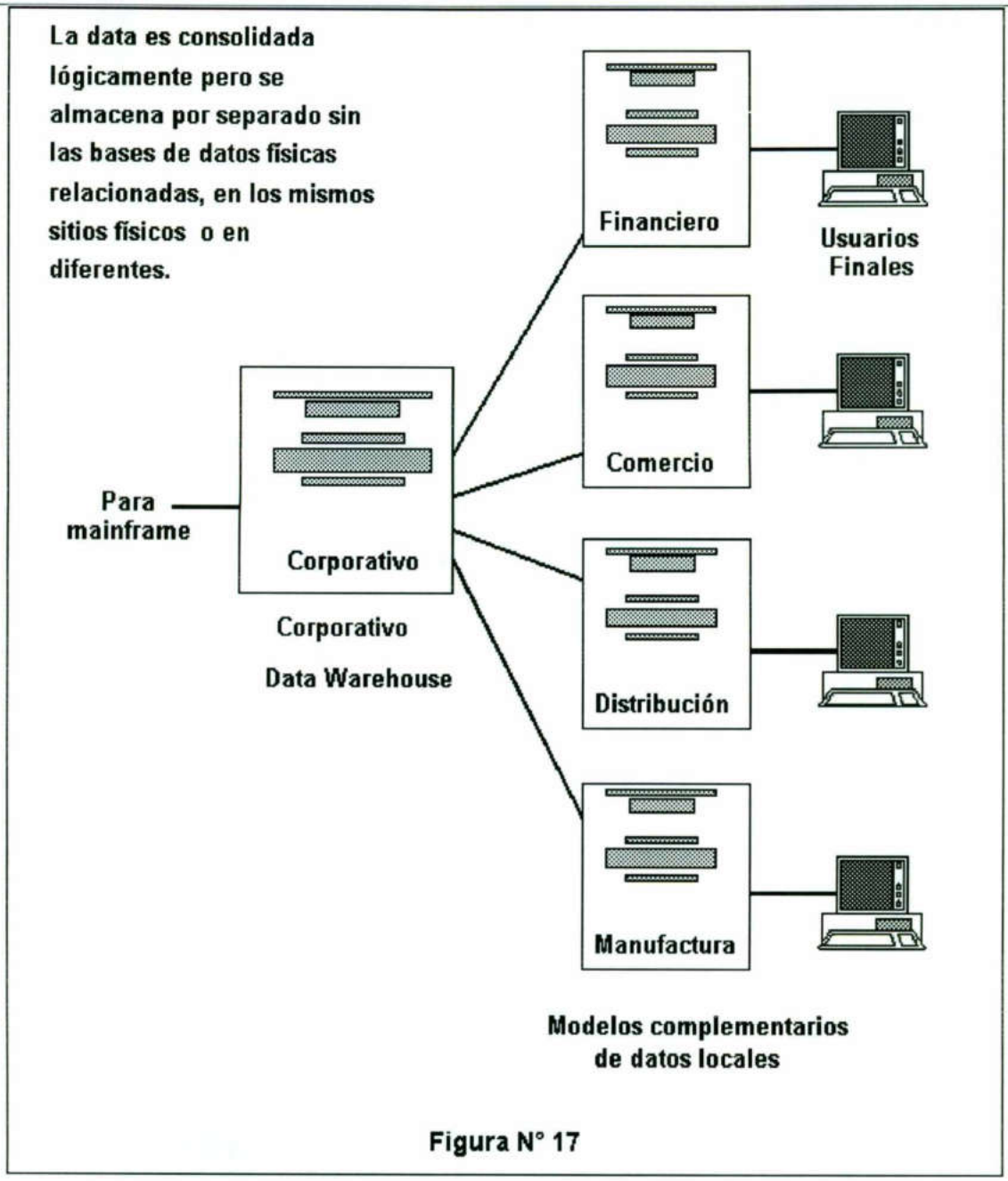
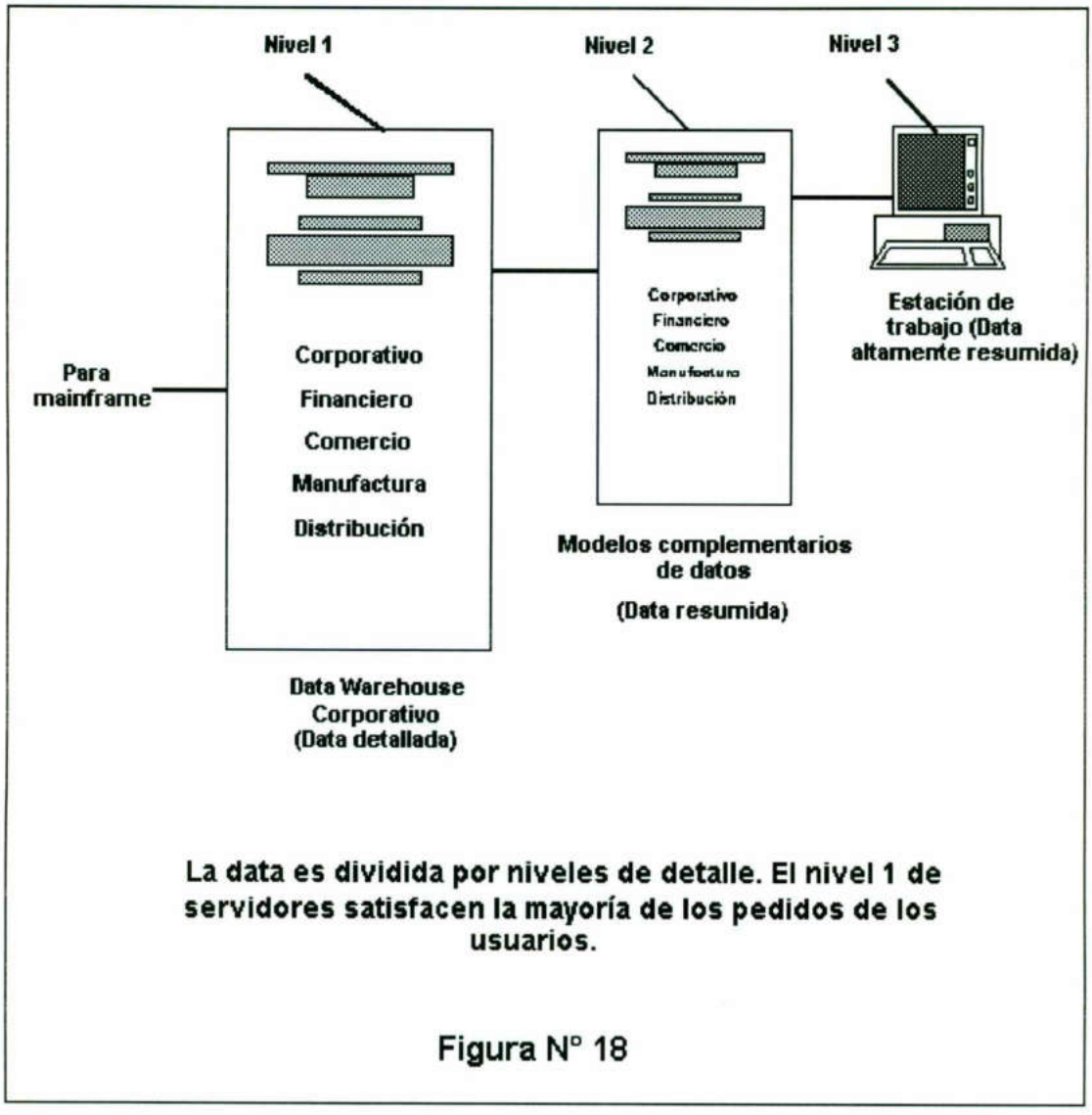



**Figura N° 17**

3° Una arquitectura 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 figura N° 18).




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

Figura N° 18

**b) Arquitectura del servidor**

Al decidir sobre una estructura 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.

**1° 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.

**2° 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).

**3° 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 (posiblemente SMP) y memoria interna. 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 "hot spots" o "data skew" (donde una cantidad desproporcionada del procesamiento es cambiada a un nodo de procesamiento, debido a la partición de los datos bajo su control).

**4° 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 SMP al conectar múltiples nodos SMP en un solo (aunque fisicamente 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.

### 2.2.3.2 SISTEMAS DE GESTION DE BASES DE DATOS

Los DataWareHouse (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.

La performance 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 súper 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 pre juntar tablas usando índices o mediante el uso de listas de índice totalmente invertidos.

Muchas de las herramientas de acceso a los DataWareHouse explotan la naturaleza multidimensional del DataWareHouse. 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.

### 2.2.3.3 NUEVAS DIMENSIONES

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.

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 simples previamente (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 la performance.

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.

### 2.2.3.4 COMBINACION DE LA ARQUITECTURA CON EL SISTEMA DE GESTION DE BASE DE DATOS

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 implementaciones de los DataWareHouse varían 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 marketing. 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 de los DataWareHouse son también un factor. Las consultas y vistas de reportes preestructuradas 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 DataWareHouse 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 DataWareHouse renovar y cambiar los datos. Los grandes volúmenes de datos que se refrescan a intervalos frecuentes, favorecen una arquitectura fisicamente 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 DataWareHouse, dónde se ubican y cuántos necesita soportar. La información sobre cómo cada grupo espera usar los DataWareHouse, ayudará a analizar los diversos estilos de uso.

Conocer la ubicación fisica de sus usuarios ayudará a determinar cómo y a qué área necesita distribuir el DataWareHouse. 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 laptops.

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 punta 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 que requiere paralelismo en la plataforma del servidor (MPP o SMP agrupado, por ejemplo).

### 2.2.3.5 PLANES DE EXPANSION

Como su depósito evoluciona y los datos que contiene llegan a ser más accesible, los empleados externos al depósito podrían descubrir también el valor de sus datos. Al enlazar

su DataWareHouse a otros sistemas (tanto internos como externos a la organización), se puede compartir información con otras entidades comerciales con poco o sin desarrollo. Los mensajes E-mail, servidores Web y conexiones Intranet/Internet, pueden entregar listas por niveles a sus proveedores o según su condición, a sus socios de negocio.

Como los DataWareHouse continúan creciendo en sofisticación y uso, los datos acumulados dentro de una empresa llegarán a ser más organizados, más interconectados, más accesibles y, en general, más disponibles a más empleados.
El resultado será la obtención de mejores decisiones en el negocio, más oportunidades y más claridad de trabajo.

### 2.2.4 CONFIABILIDAD DE LOS DATOS

La data "sucia" es peligrosa. 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 warehousing.

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.

### 2.2.4.1 LIMPIEZA DE LOS DATOS

La limpieza de una data "sucia" es un proceso multifacético y complejo. Los pasos a seguir son los siguientes:

**1°** Analizar sus datos corporativos para descubrir inexactitudes, anomalías y otros problemas.

**2°** Transformar los datos para asegurar que sean precisos y coherentes.

**3°** Asegurar la integridad referencial, que es la capacidad del DataWareHouse, para identificar correctamente al instante cada objeto del negocio, tales como un producto, un cliente o un empleado.

**4°** Validar los datos que usa la aplicación del DataWareHouse para realizar las consultas de prueba.

**5°** Producir la metadata, una descripción del tipo de datos, formato y el significado relacionado al negocio de cada campo.

**6°** Finalmente, viene el paso crucial de la documentación del proceso completo para que se pueda ampliar, modificar y arreglar los datos en el futuro con más facilidad.

En la práctica, se tendría que realizar múltiples pasos como parte de una operación única o cuando use una sola herramienta. En particular, limpiar la data y asegurar la integridad referencial son procesos interdependientes.

Las herramientas comerciales pueden ayudar en cada uno de estos pasos. Sin embargo, es posible escribir sus propios programas para hacer el mismo trabajo.

Los programas de limpieza de datos no proporcionan mucho razonamiento, por lo que las compañías necesitan tomar sus decisiones en forma manual, basados en información importante y reportes de auditoria de datos.

Cada vez que se carga un nuevo conjunto de datos, la limpieza de datos comúnmente constituye cerca del 25 por ciento de lo que puede ser un proceso de cuatro semanas.

A continuación, se darán algunos ejemplos de las experiencias de las empresas que han realizado limpieza de datos para un ambiente data warehousing.

**Ejemplo 1:**

**CompuCom Systems**, un gran integrador de sistemas basados en Dallas, implementó un registro de 12 millones, en un depósito de 10 Gb para el soporte de decisiones internas y de los clientes, según el orden y la condición y producir información por medio del Web.

**CompuCom** implementó algunas rutinas de mejoramiento de datos en lenguajes de cuarta generación (4GL), asociado con su base de datos Progress, la cual corre sobre un HP 9000. El incremento incluye desciframiento de valores de columnas en descripciones inglesas cortas o menemotecnia. El código de limpieza de datos, tales como las conversiones de fecha y datos, están escritas en lenguaje C.

La ventaja de esto es que CompuCom ahora posee estas rutinas y puede usarlas en otras aplicaciones.

Los usuarios ayudaron a definir los requerimientos de limpieza de datos, ya que son ellos los que mejor conocen los datos y pueden informar sobre qué tipo de datos sucios deben salir y cómo limpiarlos.

La compañía no usa una herramienta de limpieza comercial porque gran parte de sus datos está en la misma forma básica. Así, la compañía puede fácilmente usar de nuevo las rutinas escritas.

La desventaja principal ha sido la cantidad de tiempo de desarrollo (alrededor de una semana) que se necesitó para crear las rutinas. Aunque tienen cierta dificultad de tiempo para mantenerse al día con la demanda y han buscado paquetes de software [comercial], no han encontrado aún, en el mercado, algo que se ajuste mejor a sus requerimientos.

**Ejemplo 2:**

**Ohio Casualty Insurance** (Hamilton, OH) experimentó por dos años con la limpieza in-house, usando programas COBOL, antes de usar la herramienta comercial, Integrity Data Reengineering Tool de Vality Technology.

El DataWareHouse de **Ohio Casualty** combina registros asociados con alrededor de 1 millón de pólizas de seguro personales, incluyendo auto y pólizas de casa propia. Como una prueba, la compañía comenzó con 3,500 pólizas de sus empleados.

Sin embargo, es difícil tratar de programar para todas las situaciones en que se puede caer. Después de tomar un año en desarrollar programas genéricos de extraer/ transformar/cargar, se necesitó otro año, para programar en Cobol y editar el manual, para conseguir los datos de las pólizas correctos para el depósito.

La herramienta Vality Integrity Data Reengineering ayuda a atacar el primer conjunto de datos de los clientes - alrededor de 15, 000 pólizas en el centro comercial Denver de la compañía. Aunque el personal de Ohio Casualty todavía necesita investigar las anomalías que ha descubierto el producto Vality, no se ha requerido ninguna programación o redacción del manual de los datos. Los datos estuvieron listos para el depósito en alrededor de seis semanas.

**Ejemplo 3:**

**Intel (Hillsboro)** es un ejemplo de compañía que ha realizado exitosamente una limpieza de datos in-house, aunque con ciertos problemas. Inicialmente pretendió encargar su limpieza de datos a una agencia de servicios, para un depósito de aproximadamente 1 millón de registros tomados desde cinco sistemas operacionales.

La agencia de servicios prometió identificar las relaciones entre los diversos grupos dentro de las compañías clientes. Además, la agencia proveería información industrial para las organizaciones de clientes, tales como el número de empleados, las rentas y el crecimiento,

las cuales serían valiosas para las ventas de Intel. Desafortunadamente, la agencia de servicio no hizo un buen trabajo de identificar las relaciones entre los clientes, lo que dio como resultado el hecho de que algunas personas estuvieron asociadas con compañías equivocadas.

Intel tomó la cinta de la agencia de servicio y luego corrió los datos con el paquete de análisis estadístico SAS, del Instituto SAS, para identificar y corregir los problemas con las relaciones con un tope de 10 agrupaciones (es decir, las primeras compañías en una relación jerárquica única).

La compañía luego usó las herramientas de base de datos Oracle para propiciar el análisis y la limpieza. Ya que la nueva data llegaba todo el tiempo, algunas de las rutinas de limpieza de Oracle fueron implementadas como procedimientos almacenados para que puedan correr automáticamente contra la nueva data.

Intel aún persiste en encargar las tareas de la limpieza de los datos. Sin embargo, la compañía planea mantener la limpieza in-house hasta que encuentre una agencia de servicio aceptable.

**Ejemplo 4:**

**CrediCard** (São Paulo, Brasil), un gran emisor de tarjetas de crédito en Sudamérica, consiguió herramientas de limpieza y mejora de datos como parte de la implementación de un DataWareHouse por Market Knowledge, una filial de Equifax.

El personal de comercialización de CrediCard usa aproximadamente 200 rutinas para efectuar operaciones de limpieza, tales como la eliminación de datos malos o sin uso, corrección de valores equivocados y estandarización de formatos diversos.

Además, ellos pueden mejorar los datos al realizar operaciones como corrección de cantidades monetarias por la inflación y la devaluación, creando un campo de edad virtual basado en la fecha de nacimiento de una persona y añadiendo datos de censos a los

registros entrantes. Estas rutinas (por ejemplo, corrección de inflación) favorecen particularmente a los requerimientos brasileños.

Ellos además están diseñados para el uso del personal de comercialización no-técnico. Las rutinas de limpieza de los datos, las cuales son programadas como comandos SQL, empleó sólo alrededor de tres personas por semana para crearlas, una porción mínima de un proyecto de dos años y medio.

Las herramientas para mejorar los datos, más automatizadas y más inteligentes, representan alrededor de $ 120,000 del total del proyecto de $ 840,000.

### 2.2.4.2 Tipos de Limpieza de Datos

**a) Limpieza de datos moderada**

Si decide no programar funciones de limpieza de datos o contratar un consultor para hacer el trabajo, puede inhibirse también de la compra de una herramienta específica para esa tarea. El software de gestión del DataWareHouse puede ser suficiente para limpiar y validar según sus propósitos.

Muchos proyectos de DataWareHouse usan productos como Warehouse Manager de Prism Solutions o Passport de Carleton, para una gama de tareas de gestión de DataWareHouse, que incluyen:

- Extracción de los datos desde las bases de datos operacionales
- Preparación de los datos para cargarlos en una base de datos del depósito,
- Administración de la metadata.

Estos productos cuestan desde $ 75,000 a más de $ 200,000, dependiendo del tamaño y la complejidad del proyecto y pueden también limpiar, transformar y validar.

**Ejemplo:**

**La Universidad Emory** (Atlanta) hace la limpieza de toda la data para su depósito de 6 Gb con programas en Cobol generados por Prism Warehouse Manager. Además de tener problemas típicos, tales como formatos múltiples de fecha, la data con frecuencia contiene campos no inicializados que retienen valores arbitrarios. Dos miembros del personal utilizan como cuatro horas de un día de trabajo en las tareas de limpieza de datos.

Emory ha considerado usar herramientas de limpieza de datos especializados, pero la escuela está eliminando la data sucia hasta ahora, lo suficientemente bien, que no ve el valor adicional en otros productos comerciales para justificar la compra.

Sin embargo, tienen una buena oportunidad de que las herramientas mencionadas anteriormente de Prism y Carleton no limpien todo lo que se necesite. Ellos pueden encontrar anomalías comunes que pueden manejarse mediante simples tablas de búsqueda de información (por ejemplo, reconocer que Avenida y Av. representan la misma información), pero podrían no salir exitosos con irregularidades más importantes e impredecibles, porque estas herramientas no están diseñadas para hacer tipos de limpieza de gran intensidad.

Si los datos que requieren limpieza consisten predominantemente de nombres (incluyendo nombres de compañía) y direcciones, las compañías tales como Harte-Hanks Communications e Innovative Systems proveen no solamente herramientas de software, sino que actualizan periódicamente los archivos de datos para ayudar a combinar las variantes de los nombres de las compañías, detectar códigos postales que no corresponden a las direcciones proporcionadas y encontrar anomalías similares.

Estas herramientas pueden ser apropiadas en otros campos (aparte de nombres y direcciones) que sean conocidos para ser corregidos (por ejemplo, cantidades de dólar devaluados que han sido validados por las cuentas) o contengan información independiente

que no será usada como una llave o índice (por ejemplo, las anotaciones de contacto de los vendedores).

Las soluciones orientadas al nombre y la dirección pueden costar en cualquier parte desde $ 30,000 a más de $ 200,000, dependiendo del tamaño del DataWareHouse en cuestión. Además se necesita, una herramienta de extraer/ transformar/cargar (Extract, Transform, Load - ETL), tales como el Warehouse Manager o Passport.

Lamentablemente, en el país no existen empresas que se especialicen en estas actividades. Sólo corporaciones internacionales como las de Arthur Andersen han efectuado limpieza de datos en nuestro medio en bancos privados y muy pocos organismos públicos.

**b) Limpieza de datos intensa**

Para trabajos de limpieza intensos, se deben considerar herramientas que se han desarrollado para esas tareas. Existen dos grandes competidores: Enterprise/Integrator de Apertus Technologies y la herramienta Integrity Data Reengineering de Vality.

**Enfoque Top-Down**

La empresa Enterprise/Integrator toma un enfoque top-down, en la que usted propone las reglas para limpiar los datos. Esta es una estrategia directa, donde usted impone sus conocimientos sobre su negocio en los datos.

La empresa Enterprise/Integrator ofrece no solamente limpieza de datos, sino también extracción, transformación, carga de datos, repetición, sincronización y administración de la metadata. Es bastante caro (de $130,000 a $250,000), pero se puede ahorrar dinero si elimina la necesidad de otras herramientas de gestión de DataWareHouse.

La desventaja principal del enfoque top-down de Enterprise/Integrator es que usted tiene que conocer, o ser capaz de deducir las reglas del negocio y de la limpieza de datos.

Apertus provee ejemplos para trabajar con muchas estructuras comerciales y excepciones comunes. Aún así, crear reglas es consumo de tiempo y esté seguro de encontrar algunas excepciones no esperadas. Estos pueden manejarse manualmente mediante un sistema de excepto - manipulación, pero es un proceso que consume tiempo.

**Enfoque Bottom-Up:**

La herramienta Integrity Data Reengineering de Vality tiene un enfoque bottom-up. Analiza los datos caracter por caracter y automáticamente emergen los modelos y las reglas del negocio. Integrity proporciona un diseño de la data para ayudar a normalizar, condicionar y consolidar los datos. Este enfoque tiende a dejar pocas excepciones para manejarse manualmente y el proceso tiende a consumir menos tiempo.

Al igual que Enterprise/Integrator, Integrity puede tomar en cuenta las relaciones comerciales que no son obvias a partir de los datos, tales como fusiones y adquisiciones que han tenido lugar desde que fueron creados los datos. Pero con cualquier herramienta, estas reglas deben imponerse con un modelo top-down.

Integrity incide exclusivamente sobre la limpieza de los datos, comenzando desde los archivos básicos. No extrae los datos desde bases de datos operacionales, carga los datos en la base de datos del depósito, duplica y sincroniza los datos o administra la metadata.

Por ello, además de costar $ 250,000, Integrity podría requerir también una herramienta como Warehouse Manager o Passport. Sin embargo, pueden ser suficientes los utilitarios disponibles con la base de datos para una simple extracción/carga.

### 2.2.5 FACTORES DECISIVOS PARA DECIDIR EL DESARROLLO DE UN DATAWAREHOUSE

La data sucia es un serio peligro para el éxito de un proyecto de DataWareHouse. Dependiendo del alcance del problema, simplemente podría no ser posible dirigirlo rápidamente y abaratarlo. Los principales factores son:

- El tiempo que toma la programación interna
- El costo de las herramientas

Los gerentes de proyectos de DataWareHouse necesitan evaluar el problema con realismo, los recursos internos disponibles para distribuirlos y seleccionar la solución que se adapte a la planilla y presupuesto del proyecto, o modificar la planilla y el presupuesto para solucionar el problema.

## 2.3 FASE: IMPLEMENTACION

En esta fase, el proyecto de DataWareHouse debe tener asignado el liderazgo adecuado, así como, los recursos humanos, recursos 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.

### 2.3.1 ELEMENTOS A CONSIDERAR EN LA IMPLEMENTACION

**a) Proyecto Total o Proyecto en Fases**

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.

**b) Modelo lógico de datos**

El modelo lógico de datos debe tener un alcance más alto y cubrir todas las áreas de interés, así como los procesos más estratégicos de cada una de ellas.

Ejemplo: Puede cubrir las áreas de mercadeo, crédito y comercialización y los procesos de segmentación, scoring para retención, scoring para crédito y gestión de clientes, productos y canales de ventas.

**c) Proyecto Especializado o Proyecto Base**

Decidir sobre qué tipo de proyecto, es algo complicado. Un proyecto especializado soporta directamente un proceso específico.

Un proyecto base entrega capacidad genérica de análisis a todos los usuarios que tengan acceso al DataWareHouse, 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, más costoso y difícil de terminar.

## 2.3.2 ESTRATEGIAS PARA EL PROCESO DE IMPLEMENTACION

Deben definirse las siguientes:

1° Identificar el problema en el cual el uso estratégico de la información detallada, permita conseguir una solución para generar una ventaja competitiva o un ahorro de costos.

Ejemplo: Un problema puede ser la ausencia de un modelo para estudios de retención de clientes.

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.).

3° Reunir los datos para poblar ese modelo lógico de datos.

4° Tomar iniciativas de complementación de información para asegurar la calidad de los datos requeridos para poblar el modelo de datos.

Estas definiciones deben estar acompañadas de un servidor apropiado para el DataWareHouse, así como elementos de comunicaciones, nodos cliente, el manejador de la base de datos del DataWareHouse y otros hardware y software requeridos para la implementación del proyecto.

### 2.3.3 ESTRATEGIAS EN LA IMPLEMENTACION

Deben plantearse las siguientes:

1° 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.

2° Definir los procesos de extracción, filtro, transformación de información y carga de datos que se deben implementar para poblar ese modelo de datos.

3° Definir los procesos de administración de la información que permanece en el DataWareHouse

4° Definir las formas de consultas a la información del DataWareHouse que se le proporcionará al usuario. Para esto, debe considerarse la necesidad de resolver un problema y la potencia de consulta.

5° Completar el modelo de consulta base, relativo al área seleccionada.

6° Implementar los procesos estratégicos del área de trabajo, es decir, implementar herramientas especializadas de scoring, herramientas especializadas para inducción de conocimiento (Data Mining), etc.

7° Completar las áreas de interés, en forma similar a lo descrito anteriormente.

## 2.4 FASE: EVALUACIÓN

### 2.4.1 EVALUACION DE RENDIMIENTO DE LA INVERSION

Cuando se evalúan los costos, el usuario del DataWareHouse 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.

### 2.4.1.1 COSTOS Y BENEFICIOS

Se han identificado diversos costos y beneficios en la elaboración de un proyecto de construcción de un DataWareHouse, tales como:

**a) Costos**

- Costos preliminares
- Planificación
- Diseño
- Modelamiento/Ingeniería de Información
- Costos iniciales
- Plataforma de hardware
- Software de base de datos
- Herramientas de transferencia y limpieza de datos
- Costos en procesamiento
- Mantenimiento de datos
- Desarrollo de aplicaciones
- Capacitación y soporte

**b) Beneficios**

- Beneficios Tácticos
- Impresión y emisión de reporte reducido
- Demanda reducida para consultas de clientes

- Entrega más rápida de información a los usuarios
- Beneficios Estratégicos (Potencialidad)
- Aplicaciones y herramientas de acceso para los usuarios finales
- Decisiones con mayor información
- Toma de decisiones más rápida
- Capacidad de soporte a la información organizacional

### 2.4.2 BENEFICIOS A OBTENER

**a) Para la empresa**

El DataWareHouse hace lo posible por aprovechar el valor potencial enorme de los recursos de información de la empresa y volver ese valor potencial en valor verdadero.

**b) Para los usuarios**

El DataWareHouse extiende el alcance de la información para que puedan accesar directamente en línea, lo que a la vez contribuye en su capacidad para operar con mayor efectividad las tareas rutinarias o no.

Los usuarios del DataWareHouse pueden accesar a una riqueza de información multidimensional, presentado coherentemente como una fuente única confiable y disponible a ellos por medio de sus estaciones de trabajo.

Los usuarios pueden usar sus herramientas familiares, hojas de cálculo, procesadores de textos y software de análisis de datos y análisis estadístico para manipular y evaluar la información obtenida desde el DataWareHouse.

**c) Para la Organización en Tecnologías de Información**

El DataWareHouse enriquece las capacidades del usuario autosuficiente y hace lo factible para ofrecer nuevos servicios a los usuarios, sin interferir con las aplicaciones cotidianas de producción.

La pugna constante por resolver las necesidades de usuarios que piden acceso a los datos operacionales, finaliza con la implementación de un DataWareHouse. La mayoría de los usuarios no necesita accesar más a los datos actuales, porque ellos tienen información más útil disponible desde el DataWareHouse.

Un DataWareHouse aumenta el valor de las inversiones en tecnologías de información, en aplicaciones y bases de datos operacionales.

Como estas bases de datos alimentan información, al evolucionar el DataWareHouse, llegan a ser imprescindibles no solamente para las operaciones diarias, sino además como la fuente de información del negocio de amplio rango.

# 3. SOFTWARE EN UN DATAWAREHOUSE

## 3.1 HERRAMIENTAS DE CONSULTA Y REPORTE

Existe una gran cantidad de poderosas 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 (o más preciso, generadores de acceso-llamada a base de datos).

Más que aprender SQL o escribir un programa para accesar a la información de una base de datos, las herramientas de consulta al igual que la mayoría de herramientas visuales, le permiten apuntar y dar un click a los menús y botones para especificar los elementos de datos, condiciones, criterios de agrupación y otros atributos de una solicitud de información.

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

Se puede 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. 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?").

Para hacer consultas más accesibles a usuarios no-técnicos, los productos tales como Crystal Reports de Seagate, Impromptu de Cognos, Reportsmith de Borland, Intelligent Query de IQ Software, Esperant de Software AG y GQL de Andyne, ofrecen interfases gráficas para seleccionar, arrastrar y pegar.

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 drivers estándares tales como ODBC y 32-bit nativo, hasta fuentes de datos comerciales.

En general, los administradores de DataWareHouse que usen este tipo 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".

Una vez que se han creado las pantallas SQL, puede necesitar desarrollar un conjunto de consultas y reportes estándares, aunque algunos productos ofrecen librerías de plantillas prediseñadas y reportes predefinidos que se pueden modificar rápidamente.

## 3.2 HERRAMIENTAS DE BASE DE DATOS MULTIFIMENCIONALES / OLAP

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

Las primeras soluciones OLAP estuvieron basadas en bases de datos multidimensionales (MDDBS). Un cubo estructural (dos veces un hipercubo o un arreglo multidimensional)

almacenaba los datos para que se puedan manipular intuitivamente y claramente ver las asociaciones a través de dimensiones múltiples.

Los productos pioneros tal como Essbase de Arbor Software soportan directamente las diferentes vistas y las manipulaciones dimensionales requeridas por OLAP.

Limitaciones del enfoque de bases de datos multidimensionales:
**1.** Las nuevas estructuras de almacenamiento de datos requieren bases de datos propietarias. No hay realmente estándares disponibles para accesar a los datos multidimensionales.

Los proveedores como Arbor, vieron esto como una oportunidad para crear de facto normas para editar MDDB APIs, propiciando herramientas terceristas y estableciendo asociaciones estratégicas.

Muchas de estas herramientas de consulta y de soluciones data-mining soportan directamente Essbase, Oracle Express y otros formatos MDDB comunes. El Commander OLAP, herramienta cliente/servidor de Comshare, se sitúa sobre la parte superior de un DataWareHouse multidimensional Essbase y soporta el acceso dinámico y la manipulación de los datos.

**2.** La segunda limitación de un MDDB concierne al desarrollo de una estructura de datos. 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 consumidor de tiempo 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 OLAP relacional (Relational On Line Analytical Processing - ROLAP), que explora y opera en el DataWareHouse 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 cálculos multidimensional o en alguna otra forma que soporte rotación, drilling down y reducción.

Así como la extracción de los datos, el desarrollo y evolución de la estructura MDDB puede cambiarse. Los administradores ROLAP deben afrontar algunas veces las tareas (agobiantes) de desarrollar las rutinas SQL para agregar e indexar los datos ROLAP, así como, asegurar la traducción correcta de los pedidos multidimensionales en la ventana de comandos SQL.

Los defensores de ROLAP argumentan que se usan estándares abiertos (SQL) y que se esquematiza (nivel de detalle) los datos para hacerlos más fácilmente accesibles. Por otra parte, argumentan que una estructura multidimensional nativa logra mejor performance y flexibilidad, una vez que se desarrolla el almacén de los datos.
Lo bueno es que estas tecnologías evolucionan rápidamente y/o pueden proveer una pronta solución OLAP. Algunos productos ejemplos son PowerPlay de Cognos, Business Objects con el software del mismo nombre, Brio Query de Brio Technology y una serie de DSS Agent/DSS Server de MicroStrategy.

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.

Los usuarios de estos productos deben decidir sobre si los datos del procesamiento analítico en línea, deberían almacenarse en bases de datos multidimensionales especialmente diseñadas o en bases de datos relacionales. Esto depende de las necesidades de la organización. En el Anexo 1-B, se indica si un producto almacena datos en bases de datos relacionales o en una base de datos multidimensional (MDDB).

## 3.3 SISTEMAS DE INFORMACION EJECUTIVOS

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

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

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 EIS presenta vistas de los datos simplificados, altamente consolidados y mayormente estáticas.

**Categorías de Ambientes EIS:**

**1.** El libro electrónico es una versión en línea, electrónica, contraparte del papel que muchos ejecutivos usan en reuniones con el personal. Las diapositivas electrónicas presentan una visión concreta de una iniciativa organizacional o quizás los datos para dar a conocer la situación actual de un proyecto importante.

**2.** El centro de comando es básicamente una colección de puertos en un amplio conjunto de reportes, el newsgroup recupera desde Internet y otros materiales que proveen conocimientos en la organización.

Los reportes del centro de comando pueden ser accesados 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 ejecutivos pueden administrar su propio libro electrónico y centro de comandos o los administradores pueden mantener y modificar el EIS de acuerdo a las especificaciones del ejecutivo. Los sistemas de información ejecutivos, generalmente tienen una programación que variará en complejidad de un producto a otro.

Los pioneros en el mercado de EIS incluyen Comshare, creadores del Commander EIS y Pilot Software, desarrolladores del Pilot Command Center.

## 3.4 HERRAMIENTAS DATA MINING

Data mining es una categoría de herramientas de análisis open-end. En lugar de hacer preguntas, se toma estas herramientas y se pregunta algo "interesante", una tendencia o una agrupación peculiar, por ejemplo. El proceso de data mining extrae los conocimientos guardados o información predictiva desde el DataWareHouse sin requerir pedidos o preguntas específicas.

Las herramientas Mining usan algunas de las técnicas de computación más avanzadas como:

- Redes neurales.
- Detección de desviación.
- Modelamiento predictivo y programación genética para generar modelos y asociaciones.
- Mining es un dato-conducido, no una aplicación-conducida.

El Intelligent Miner de IBM para AIX soporta sofisticadas técnicas mining, así como las funciones de preparación de los datos para extraer información desde bases de datos Oracle o Sybase y cargarlos en DB2 para mining. Con su opción Data Mine para el motor Red

Brick Warehouse 5.0, Red Brick integra la funcionalidad de un data mining y la arquitectura de almacenamiento.

Otros ejemplos de herramientas data mining comerciales incluyen Darwin de Thinking Machines, herramientas de visualización de datos en MDDB de SAS Institute, SGI MineSet y Focus 6 Serie de Visualización y Análisis de Information Builders.

## 3.5 SISTEMAS DE GESTION DE BASES DE DATOS

Estos softwares proporcionan procesamiento en paralelo y/o algo fuera de los aspectos ordinarios, que puedan ser especialmente interesantes para la gente de desarrollo de DataWareHouse y de sistemas de soporte de decisiones.

## 3.6 ELECCION DE HERRAMIENTAS

Hay algunas reglas obvias a seguir cuando se eligen herramientas de análisis. Las herramientas se combinan según las necesidades de los usuarios finales, capacidad técnica empresarial y la fuente de datos existente.

**1°** Si se elige un proveedor de depósito que además ofrece herramientas integradas, probablemente se ahorrará un tiempo de desarrollo significativo al elegir un conjunto de herramientas compatibles.

De otro modo, seleccione un conjunto de herramientas que soporte su fuente de datos original. Sin ese soporte, se debería optar por una solución OLAP relacional debido a que provee una arquitectura abierta.

**2°** Después que se ha seleccionado un conjunto de herramientas compatible con su fuente de datos, determine cuánto análisis necesita realmente.
Si usted simplemente necesita saber "cuánto" o "cuántos", será suficiente una herramienta básica de consultas y reportes.

Si usted requiere un análisis más avanzado que explique la causa y los efectos de las ocurrencias y las tendencias, busque una solución OLAP.

Las herramientas data mining sofisticadas requieren expertos en técnicas de análisis de datos y se necesitan para pronósticos avanzados, clasificación y creación del modelo.

**3°** Como con cualquier tecnología, para el mejor desempeño de su compañía, se puede optar por una solución única o un conjunto de soluciones. Su personal debe comprender los requerimientos de tecnología, desarrollar soluciones que reúnan esos requerimientos y mantener y mejorar efectivamente los sistemas. Los software de negocio inteligentes son sólo herramientas. Todavía se necesita gerentes y ejecutivos que capten los conocimientos derivados y tomen decisiones intuitivamente. En otras palabras, estos software requieren todavía inteligencia comercial propia.

# CONCLUSIONES

Un sistema DataWareHouse define un nuevo concepto para el almacenamiento de datos, integra la información generada en todos los ámbitos de una actividad de negocio (ventas, producción, finanzas, Marketing, etc.) y permite un acceso y explotación de la información contenida en las bases de datos, facilitando un amplio abanico de posibilidad de análisis multivariables que permitirán la toma de decisiones estratégicas.

El proceso integra toda la información de una compañía en un único depósito. La información que se genera en una compañía proviene de diferentes fuentes, formatos y tipos, que se consolidan, se transforman y se cargan en diferentes sistemas de gestión de datos, normalmente en RDBMS (Relational Database Management Systems).

Desde un sistema DataWareHouse, la información se puede mostrar y representar de muchas maneras. La forma más común de analizar la información, es utilizando un sistema de proceso de análisis en línea (OLAP, on-line analytical processing). Los productos OLAP ofrecen un rango muy variado de capacidades de análisis avanzado, como el multidimensional y el estadístico.

Un sistema DataWareHouse soporta también sofisticadas operaciones de análisis tales como los sistemas scoring y aplicaciones de detección de fraude. Todas estas funciones de análisis se conocen con el término de Data Mining.

Una de las novedades que aporta el Datawarehousing como sistema de análisis de información, es la creación de la Meta Información (metadata). Se trata de un fichero al que se le considera como diccionario de estructuras de datos que el administrador del sistema define con el objetivo de asistir en los procesos de consulta a la base de datos.

La metadata se adaptará a las definiciones que el usuario utilizará posteriormente en sus interrogaciones al sistema.

De esta manera se conseguirá minimizar los complejos procedimientos de definición de nombres de campos, jerarquías y relaciones entre ficheros.

La implantación consiste, en una primera fase, en el análisis de las necesidades de información a las que desea acceder cada compañía. Para ello se integrarán en el sistema todos aquellos datos operacionales necesarios, además de otras fuentes de información que sea menester incorporar.

Definida la estructura de la base de datos se procederá a la carga de información y se crearán las agregaciones de datos para mejorar el rendimiento del sistema en los procesos de consulta más habituales. Finalmente, se incluirán en el sistema los procedimientos que permitan la actualización de información, cuya periodicidad dependerá de las necesidades de cada usuario.

El proceso de implantación de un sistema DataWareHouse, puede adaptarse de forma gradual o departamental creando soluciones específicas para cada área con el objetivo de conseguir resultados operativos a corto plazo.

Un sistema DataWareHouse es una eficaz herramienta de organización y análisis de los complejos volúmenes de información que las compañías generan, que posteriormente permite el desarrollo de estrategias más efectivas y rentables.

Pero la definición del nuevo modelo de datos y el método de carga y mantenimiento de la información, requiere un personal especializado que atienda las necesidades de cada empresa.

Por último se puede decir que un proyecto datawarehousing se considera exitoso, cuando su objetivo final comienza a concretarse, es decir que la gente de la empresa use el DW para satisfacer sus necesidades empresariales.

Como ya hemos visto, son variados los cambios que comenzarán a producirse al implementar un DW.

Es importante entonces anticiparse a estos cambios, considerar sus implicancias y planificarlos en la empresa. Las siguientes situaciones, gatillan el comienzo de estos cambios:

- La gente de la empresa depende del DW como un recurso primario de información.
- La gente de empresa se vuelve menos dependiente de los sistemas operacionales y de sus bases de datos para sus necesidades de información.
- Se ve reducida o eliminada la demanda por programación especializada para encontrar la información necesaria.
- Los usuarios y uso del DW crecen, con un correspondiente incremento en la demanda de soporte.
- La complejidad de cambios en los sistemas operacionales se incrementa, y su efecto sobre el DW debe ser considerado.

## BIBLIOGRAFIA


En libros:

- **A Multidimensional DataWareHouse Development Methodology**
  AUTORES: J. M. Cavero, C. Costilla, E. Marcos, M. G. Piattini y A. Sánchez
  LIBRO: Managing Data Mining Technologies in Organizations: Techniques and Applications
  EDITORES: Parag C. Pendharkar
  EDITORIAL: Idea Group Publishing
  FECHA EDICIÓN: 2002

- **A Methodology for DataWareHouse Design: Conceptual Modelling** – (pp.185-197)
  AUTORES: J. M. Cavero, E. Marcos, M. G. Piattini y A. Sánchez
  LIBRO: Data Warehousing and Web Engineering
  EDITORES: Shirley Becker
  EDITORIAL: IRM Press
  FECHA EDICIÓN: 2002

En Internet:

- http://www.ideasa.net/DataWareHouse.htm
- http://www.dw-institute.com/
- http://www.dwinfocenter.org/