

**Universidad Nacional del Nordeste  
Facultad de Ciencias Exactas y Naturales y  
Agrimensura**

**Monografía de Adscripción:  
Data Warehouse**

**Rojas, Mariana Isabel      LU: 38382  
Prof. Director: Mgter. David Luis La Red Martínez  
Licenciatura en Sistemas de Información  
Corrientes-Argentina  
2009**

## Índice General

|   |    |
|---|----|
| Introducción.....                             | 4  |
| ¿Qué es un Data Warehouse?.....               | 5  |
| Sistemas de Información.....                  | 5  |
| Datos operacionales y datos informativos..... | 7  |
| Características.....                          | 8  |
| Orientado a Temas.....                        | 8  |
| Integración.....                              | 10 |
| De Tiempo Variante.....                       | 11 |
| No Volátil.....                               | 12 |
| Organización de un Proyecto.....              | 14 |
| Planeamiento.....                             | 15 |
| Requerimiento.....                            | 16 |
| Análisis.....                                 | 16 |
| Diseño.....                                   | 16 |
| Construcción.....                             | 17 |
| Despliegue.....                               | 17 |
| Expansión.....                                | 17 |
| Impactos Data Warehouse.....                  | 17 |
| Estructura de un Datawarehouse.....           | 20 |
| Inteligencia de Negocio.....                  | 23 |
| Características.....                          | 23 |
| Niveles de realización de BI.....             | 23 |
| Extracción, transformación y carga (ETL)..... | 24 |
| Extracción.....                               | 24 |
| Transformación.....                           | 25 |
| Carga.....                                    | 25 |
| Ambiente Data Warehouse.....                  | 26 |
| Data Mart.....                                | 26 |
| Metadatos.....                                | 26 |
| Modelado de Datos.....                        | 27 |
| El modelo relacional.....                     | 27 |
| El modelo dimensional.....                    | 28 |
| Ventajas del modelo dimensional.....          | 31 |
| Herramientas de acceso y uso.....             | 32 |
| OLAP (On Line Analytical Processing).....     | 32 |
| Drill Down y Roll Up.....                     | 33 |
| Slice y Dice.....                             | 34 |
| Data Mining (Minería de Datos).....           | 35 |
| Sitios de Internet consultados.....           | 37 |

## Índice de Figuras

|  |    |
|--|----|
| Figura 1. Sistemas de Información. Esquema. ....                                       | 6  |
| Figura 2. Características del Datawarehouse. Orientado a una materia. ....             | 9  |
| Figura 3. Características del Datawarehouse. Integración. ....                         | 11 |
| Figura 4. Características del Datawarehouse. De Tiempo Variante. ....                  | 12 |
| Figura 5. Características del Datawarehouse. No Volátil. ....                          | 13 |
| Figura 6. Sistema Nervioso Digital. Transformación. ....                               | 15 |
| Figura 7. Planeación necesaria para el sistema de Data Warehouse. ....                 | 15 |
| Figura 8. Requerimientos para la solución del Data Warehouse.....                      | 16 |
| Figura 9. Estructura de los Datos de un Data Warehouse.....                            | 21 |
| Figura 10. Ejemplo de Niveles de Esquematización que podría encontrarse en un DW. .... | 22 |
| Figura 11. Extracción, transformación y carga. ETL.....                                | 25 |
| Figura 12. Modelado de Datos. Esquema en Estrella. ....                                | 30 |
| Figura 13. Modelado de Datos. Esquema Copo de Nieve. ....                              | 31 |
| Figura 14. Herramientas de Acceso y Uso. Drill Down y Roll Up. ....                    | 34 |
| Figura 15. Herramientas de Acceso y Uso. Slice y Dice. ....                            | 35 |

## Introducción

En la actualidad, el dinámico mundo de los negocios plantea la necesidad de disponer de un acceso rápido y sencillo a información para la toma de decisiones. Dicha información debe estar estructurada y elaborada de acuerdo a parámetros de calidad, a fin de posibilitar una adaptación ágil y precisa a las fluctuaciones del ambiente externo.

Las empresas disponen, para la gestión de sus procesos de negocio, de sistemas transaccionales corporativos que manejan enormes cantidades de datos, organizados de forma tal que puedan ser utilizados por las aplicaciones operacionales existentes. Los niveles gerenciales necesitan a menudo tomar decisiones de alto nivel, cruciales para el funcionamiento de la empresa.

Frecuentemente se basan en su experiencia, utilizando un enfoque subjetivo del proceso decisorio. Este enfoque no es apto para las condiciones del mundo actual en el que los sistemas de gestión de calidad vigentes han demostrado la importancia de la toma de decisiones basada en cifras, datos y hechos.

El Data Warehouse permite que los gerentes tomen decisiones siguiendo un enfoque racional, basados en información confiable y oportuna. Consiste básicamente en la transformación de los datos operacionales en información útil para decidir. El uso del Data Warehouse permite también encontrar relaciones ocultas entre los datos y predecir el comportamiento futuro bajo condiciones dadas.

La filosofía de trabajo del Data Warehouse es diferente a la de los sistemas transaccionales. Se modelan los datos a partir de dimensiones, en lugar del tradicional modelado relacional, y las herramientas de acceso a los datos se basan en una tecnología de procesamiento analítico (OLAP), distinta al procesamiento transaccional (OLTP) de los sistemas operacionales.

Los datos operacionales que sirven de entrada al Data Warehouse generalmente están dispersos en distintos sistemas de la organización, desarrollados en diferentes entornos de desarrollo, por diferentes personas y en diferentes momentos. Es tarea fundamental del Data Warehouse recolectarlos, unificarlos y depurarlos según las necesidades del negocio, eliminando inconsistencias y conservando sólo la información útil para los objetivos empresariales. Esto se lleva a cabo mediante procesos que se ejecutan periódicamente y conducen a mantener la información actualizada.

Las aplicaciones de usuario final que acceden al Data Warehouse brindan a los gerentes la posibilidad de ver la información a diferentes niveles de agregación (detallados o resumidos) y filtrar las consultas por distintas variables (“rebanar” y “picar” la información).

Finalmente, el Data Warehouse permite aplicar herramientas como el Data Mining, para encontrar relaciones entre los datos a fin de comprender las causas de variabilidad presentes y realizar pronósticos con el apoyo de modelos estadísticos.

En la sociedad actual, la información constituye un activo esencial de cualquier organización proporcionando beneficios significativos, tangibles y cuantificables. Como consecuencia, la integración de un Data Warehouse a la empresa representa una ventaja competitiva en el mundo de los negocios.

## ¿Qué es un Data Warehouse?

Un **almacén de datos** (del inglés *data warehouse*) es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.

Data warehousing es el centro de la arquitectura para los sistemas de información desde 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 Data Warehouse 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 Data Warehouse haciendo un contraste de cómo los datos de un negocio almacenados en un data warehouse, difieren de los datos operacionales usados por las aplicaciones de producción.

El ingreso de datos en el Data Warehouse viene desde el ambiente operacional en casi todos los casos. El Data Warehouse es siempre un almacén de datos transformados y separados físicamente de la aplicación donde se encontraron los datos en el ambiente operacional [1].

## Sistemas de Información

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 un mundo de rápidos cambios.

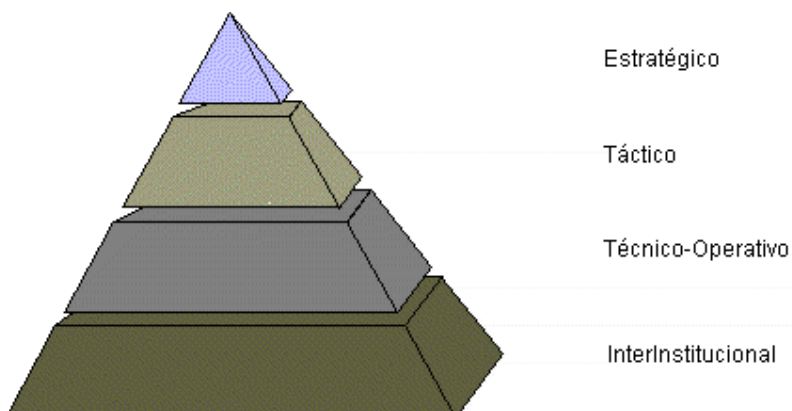
Las funciones como "planificación de marketing", "planeamiento de ingeniería" y "análisis financiero", requieren, además, de sistemas de información que las soporte. Pero estas funciones son diferentes de las operacionales y los tipos de sistemas y la información requerida son también diferentes. Las 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, toman un número de áreas diferentes y necesitan cantidades grandes de datos operacionales relacionadas.

Son estos sistemas sobre los que se basa la tecnología Data Warehousing.

Los sistemas de información se han dividido de acuerdo al siguiente esquema:



**Figura 1. Sistemas de Información. 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. 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 sensores, autómatas, sistemas multimedia, bases de datos relacionales más avanzadas y data warehousing.

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

Sin embargo, la tecnología data warehousing basa sus conceptos y diferencias entre dos tipos fundamentales de sistemas de información en todas las organizaciones: los *sistemas técnico-operacionales* y los *sistemas de soporte de decisiones*. Este último es la base de un data warehouse [1].

## Datos operacionales y datos informativos

El proceso automatizado de un negocio utiliza *datos operacionales*, los que constituyen el conjunto de registros de las transacciones del negocio.

Estos datos son creados durante la ejecución de estos procesos y son almacenados en un archivo o en una base de datos. Frecuentemente contienen valores incorrectos, son muy detallados y son de mínimo uso en los negocios debido a su gran volumen, ubicación y formatos.

En conclusión, es difícil para los usuarios del negocio tener acceso a los datos operacionales debido a las limitaciones de performance y tecnología.

Lo que el usuario del negocio necesita como entrada a sus actividades de análisis son *datos informativos*.

Estos son una combinación de datos operacionales que han sido modificados, depurados, transformados, consolidados y organizados desde diversas fuentes externas al proceso del negocio.

Este tipo de información generalmente es específico para un conjunto de usuarios del negocio que lo hacen significativo y útil para su análisis.

Ambos tipos de datos y ambos tipos de uso son muy importantes, pero es difícil cumplir con ambos propósitos en el mismo sistema.

Los datos operacionales son específicos para cada aplicación y usualmente son almacenados de manera separada por otras aplicaciones. Estos datos son útiles en la medida en que se aprovechen para satisfacer el proceso de las aplicaciones predefinidas. Mayormente se requieren sólo datos actuales y estos deben ser mantenidos al día haciendo actualizaciones frecuentes en la base de datos. En cambio, para los datos informativos, el usuario necesita datos que crucen por varias

aplicaciones, que estén reorganizados por temas de negocio, que contengan valores históricos, que se encuentren disponible para análisis durante períodos largos y que sea accesible de manera fácil y flexible.

Los datos operacionales son manejados, precisamente, por los sistemas operacionales o transaccionales (On Line Transactional Processing, OLTP), los cuales se concentran en la administración y la medición de indicadores empresariales (capital e inversión), indicadores financieros (márgenes de utilidades, rotación de inventarios), indicadores de ventas (identificación de clientes persistentes), etc.

Por su parte, los datos informativos son los que conforman un Data Warehouse, el cual tiene como fin comprender, medir y administrar parámetros empresariales estratégicos, como el crecimiento del ingreso y rentabilidad, la participación del mercado y los segmentos del cliente. En el siguiente cuadro se muestran las diferencias entre los datos operacionales y los datos informativos.

| <b>Datos Operacionales</b>  | <b>Datos Informativos</b>  |
|---|--|
| <ul style="list-style-type: none"> <li>• Orientados a una aplicación</li> <li>• Integración limitada</li> <li>• Constantemente actualizados</li> <li>• Sólo valores actuales</li> <li>• Soportan operaciones diarias</li> </ul> | <ul style="list-style-type: none"> <li>• Orientados a un tema</li> <li>• Integrados</li> <li>• No volátiles</li> <li>• Valores a lo largo del tiempo</li> <li>• Soportan decisiones de administración</li> </ul> |

## **Características**

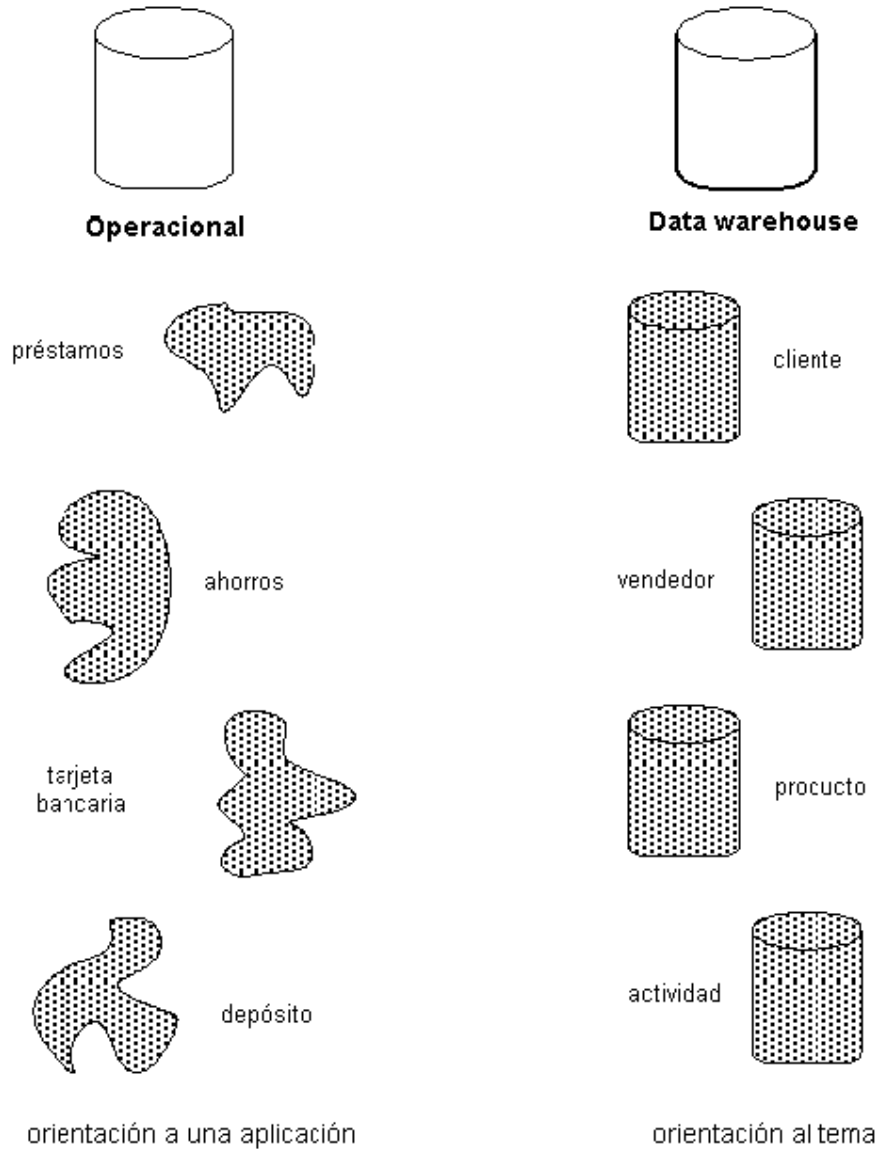
### **Orientado a Temas**

Una primera característica del data warehouse 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 2 se muestra el contraste entre los dos tipos de orientaciones.

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

En el ambiente data warehousing se organiza alrededor de sujetos tales como cliente, vendedor, producto y actividad. Por ejemplo, para un fabricante, éstos pueden ser clientes, productos, proveedores y vendedores. Para una universidad pueden ser estudiantes, clases y profesores. Para un hospital pueden ser pacientes, personal médico, medicamentos, etc.

La alineación alrededor de las áreas de los temas afecta el diseño y la implementación de los datos encontrados en el data warehouse. Las principales áreas de los temas influyen en la parte más importante de la estructura clave [4].



**Figura 2. Características del Datawarehouse. Orientado a una materia.**

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

Las diferencias entre la orientación de procesos y funciones de las aplicaciones y la orientación a temas, radican en el contenido de la data a nivel detallado. En el data warehouse se excluye la información que no será usada por el proceso de sistemas

de soporte de decisiones, mientras que la información de las orientadas a las aplicaciones, contiene datos para satisfacer de inmediato los requerimientos funcionales y de proceso, que pueden ser usados o no por el analista de soporte de decisiones.

Otra diferencia importante está en la interrelación de la información. Los datos operacionales mantienen una relación continua entre dos o más tablas basadas en una regla comercial que está vigente. Las del data warehouse miden un espectro de tiempo y las relaciones encontradas en el data warehouse son muchas [4].

## **Integración**

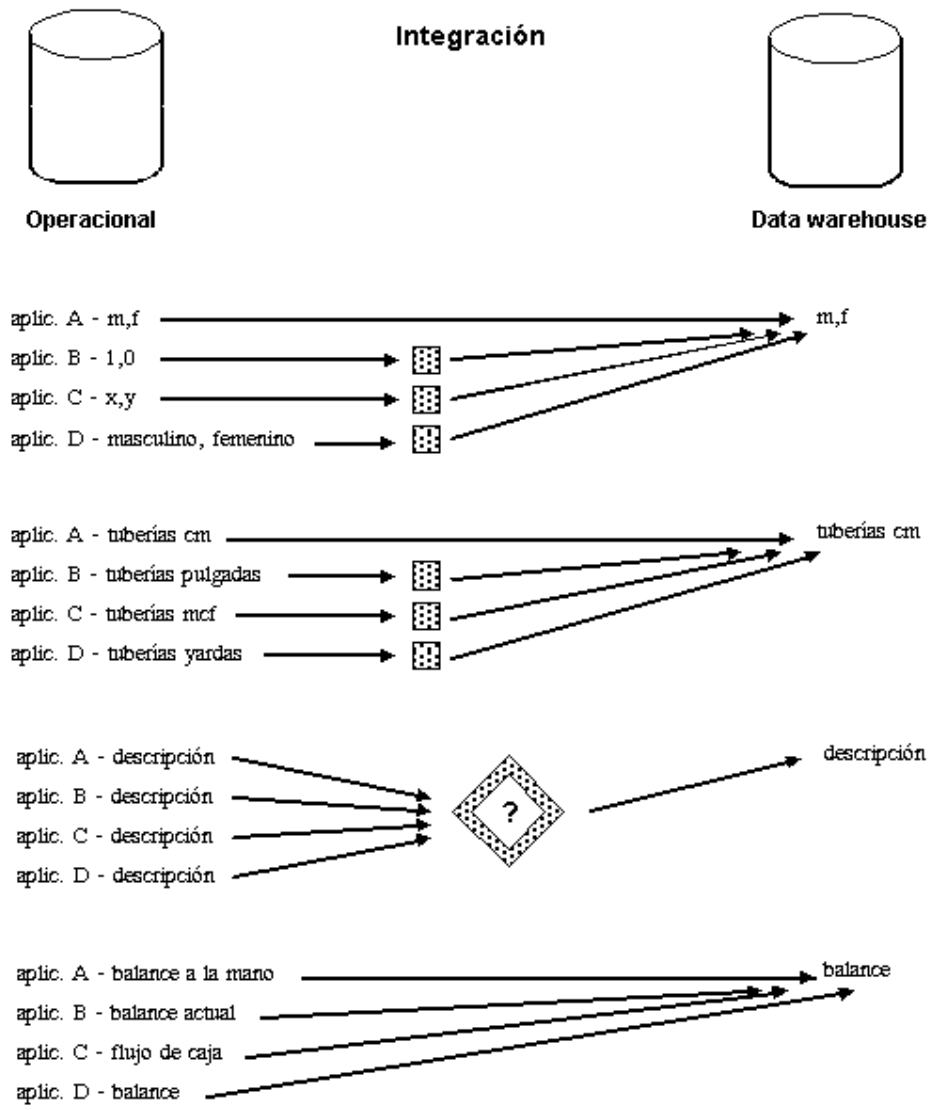
El aspecto más importante del ambiente data warehousing es que la información encontrada al interior está siempre integrada.

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 data warehouse con la carencia de integración del ambiente de aplicaciones, se muestran en la Figura 3, 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 3 mencionada, muestra algunas de las diferencias más importantes en las formas en que se diseñan las aplicaciones [4].



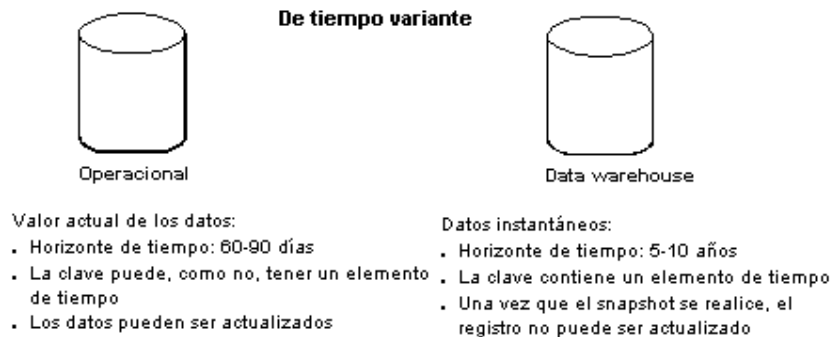
**Figura 3. Características del Datawarehouse. Integración.**

### De Tiempo Variante

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

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

Los datos históricos son de poco uso en el procesamiento operacional. La información del depósito por el contraste, debe incluir los datos históricos para usarse en la identificación y evaluación de tendencias [4].



**Figura 4. Características del Datawarehouse. De Tiempo Variante.**

El tiempo variante se muestra de varias maneras:

**1°:** La más simple es que la información representa los datos sobre un horizonte largo de tiempo - desde cinco a diez años. El horizonte de tiempo representado para el ambiente operacional es mucho más corto - desde valores actuales hasta sesenta a noventa días. Las aplicaciones que tienen un buen rendimiento y están disponibles para el procesamiento de transacciones, deben llevar una cantidad mínima de datos si tienen cualquier grado de flexibilidad. Por ello, las aplicaciones operacionales tienen un corto horizonte de tiempo, debido al diseño de aplicaciones rígidas.

**2°:** La segunda manera en la que se muestra el tiempo variante en el data warehouse está en la estructura clave. Cada estructura clave en el data warehouse contiene, implícita o explícitamente, un elemento de tiempo como día, semana, mes, etc. El elemento de tiempo está casi siempre al pie de la clave concatenada, encontrada en el data warehouse. En ocasiones, el elemento de tiempo existirá implícitamente, como el caso en que un archivo completo se duplica al final del mes, o al cuarto.

**3°:** La tercera manera en que aparece el tiempo variante es cuando la información del data warehouse, una vez registrada correctamente, no puede ser actualizada. La información del data warehouse es, para todos los propósitos prácticos, una serie larga de "snapshots" (vistas instantáneas). Por supuesto, si los snapshots de los datos se han tomado incorrectamente, entonces pueden ser cambiados. Asumiendo que los snapshots se han tomado adecuadamente, ellos no son alterados una vez hechos. En algunos casos puede ser no ético, e incluso ilegal, alterar los snapshots en el data warehouse. Los datos operacionales, siendo requeridos a partir del momento de acceso, pueden actualizarse de acuerdo a la necesidad [4].

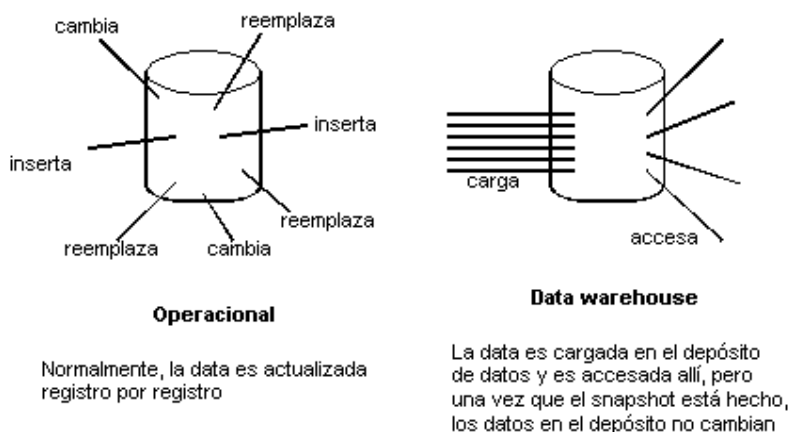
## **No Volátil**

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

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

Otra consecuencia de la simplicidad de la operación del data warehouse está en la tecnología subyacente, utilizada para correr los datos en el depósito. Teniendo que soportar la actualización de registro por registro en modo on-line (como es frecuente en el caso del procesamiento operacional) requiere que la tecnología tenga un fundamento muy complejo debajo de una fachada de simplicidad [4].



**Figura 5. Características del Datawarehouse. No Volátil.**

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

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

Se debe considerar lo siguiente:

Los datos se filtran cuando pasan desde el ambiente operacional al de depósito. Existe mucha data que nunca sale del ambiente operacional. Sólo los datos que realmente se necesitan ingresarán al ambiente de data warehouse.

El horizonte de tiempo de los datos es muy diferente de un ambiente al otro. La información en el ambiente operacional es más reciente con respecto a la del data warehouse. Desde la perspectiva de los horizontes de tiempo únicos, hay poca superposición entre los ambientes operacional y de data warehouse.

El data warehouse contiene un resumen de la información que no se encuentra en el ambiente operacional.

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

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

## **Organización de un Proyecto**

Antes de plantear el datawarehouse, debemos encontrar en la empresa nuestro sistema nervioso digital. ¿Y que es el sistema nervioso digital?, es como el sistema nervioso humano, en el cual todas sus partes se interrelacionan haciendo que la empresa funcione productivamente controlando sus procesos. Por ello un sistema nervioso digital transforma 3 elementos de un negocio:

- Su relación con los clientes y asociado (comercio).
- El flujo de información y la relación entre sus empleados (administración del conocimiento).
- Procesos de negocios internos (operaciones de negocios).

Se considera al Sistema Nervioso Digital como punto de partida para plantear que data warehouse es la tecnología que en la actualidad necesitará la empresa. La construcción del DW empieza con el Planeamiento, Requerimiento, Análisis, Diseño, seguido de la Construcción, Despliegue y Expansión que este puede tener en la empresa donde se desearía implementar, siguiendo con especificaciones en cada uno de los procesos [2].

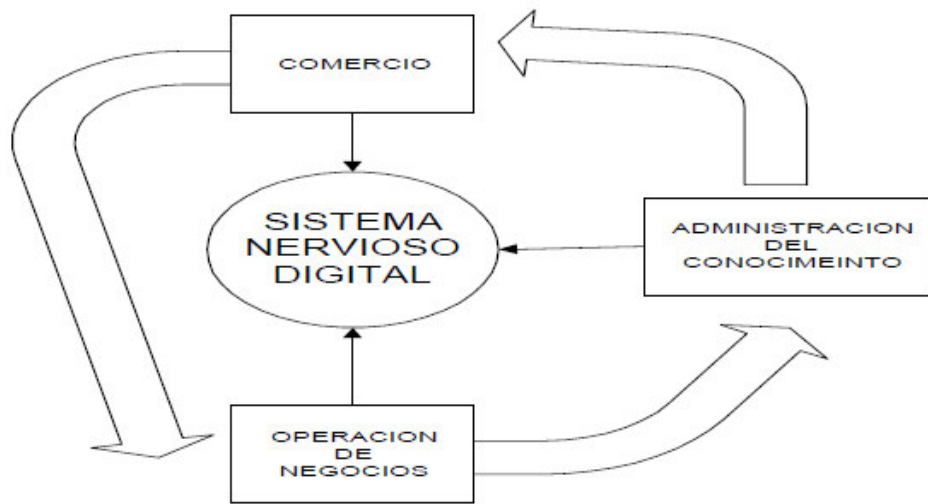


Figura 6. Sistema Nervioso Digital. Transformación.

## Planeamiento

La Figura 7 muestra la planeación que se tiene que realizar en un datawarehouse. Algunos de los pasos se pueden efectuar al mismo tiempo (en paralelo), lo cual acorta la duración de esta fase.

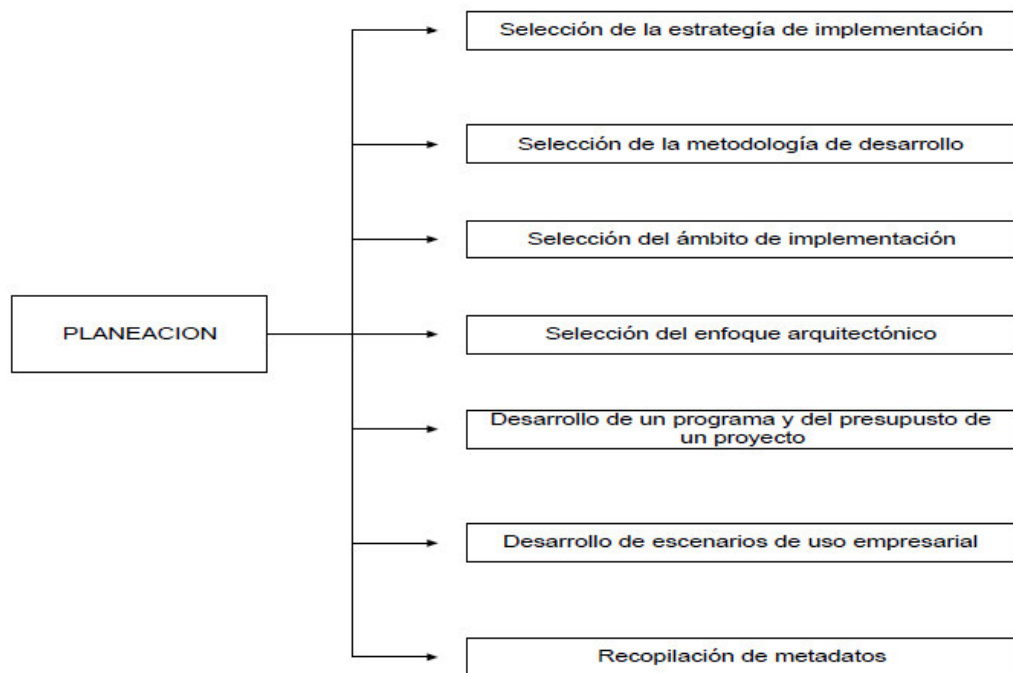


Figura 7. Planeación necesaria para el sistema de Data Warehouse.

Uno de los primeros pasos más importantes consiste en decidir la estrategia general de implementación. La decisión tiene mucho que ver con la cultura de la organización y se basa en cómo se llevan a cabo las tareas dentro de la organización.

Se debe tener en cuenta la metodología a utilizar, las más conocidas son: Método en Cascada y Método Espiral, se define el método arquitectónico, el desarrollo del programa y los escenarios que la empresa va a tener cuando se implemente el data warehouse, para ello se define claramente los metadatos, que son elementos que se va a utilizar para la planeación efectiva del datawarehouse.

## Requerimiento

La fase en mención es una especificación precisa de las funciones que se obtendrán del data warehouse, para ello se debe definir los requerimientos que se necesitará, como se muestra en la Figura 8.

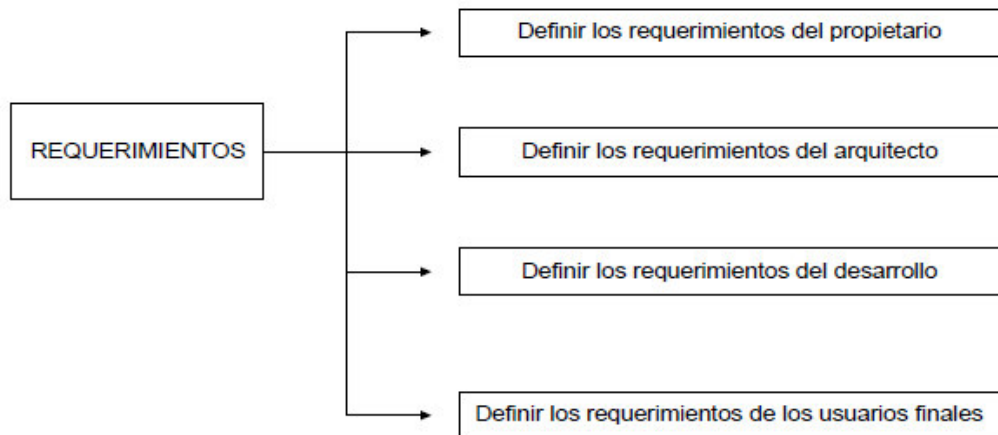


Figura 8. Requerimientos para la solución del Data Warehouse.

## Análisis

Esta fase significa convertir los requerimientos agrupados en un conjunto de especificaciones que puedan apoyar el diseño. En este análisis debe considerarse 3 tipos de especificaciones:

- Requerimientos de enfoque empresarial que delinear las fronteras de la información que debe comprender el data warehouse. El enfoque empresarial determinará también la audiencia y sus requerimientos de información.
- Especificación de requerimientos de fuentes de datos que delinear las fronteras de información disponible en las fuentes de datos actuales.
- Especificaciones de requerimientos de usuario final y acceso, las cuales definen cómo se utilizará la información del data warehouse. Junto con éstas se encuentra la especificación de los tipos de herramientas y técnicas de exhibición que se usarán.

## Diseño

En la fase de diseño se encuentran las siguientes dos actividades principales:

- Diseño detallado de la arquitectura de datos: Es el desarrollo del modelo físico de datos para la base de datos de almacenamiento del datawarehouse y mercado de datos.
- Diseño detallado de la arquitectura de aplicaciones: Es la Correspondencia de los modelos físicos de datos de la fuente de datos con los modelos físicos data warehouse y mercado de datos.

## **Construcción**

En esta fase se realiza la implementación física de los diseños desarrollados durante la fase de diseño. Las aplicaciones que se necesitan construir son las siguientes:

- Programas que creen y modifiquen la base de datos para el datawarehouse.
- Programas que traigan datos de fuentes relacionadas y no relacionadas.
- Programas que realicen transformación de datos.
- Programas que realicen actualización de base de datos.
- Programas que efectúen búsquedas en base de datos muy grandes.

## **Despliegue**

Los requerimientos de despliegue para un data warehouse son:

- La información contenida en el data warehouse debe estar en términos y lenguajes que comprendan los usuarios ya que ellos no son técnicos.
- Debe existir una necesidad de que la información que proporcione el data warehouse debe de ser precisa para los usuarios finales.

## **Expansión**

En esta etapa se prevé algunas de las siguientes áreas de mejora:

- Consultas empresariales que no pueden formularse o satisfacerse debido a la limitación del data warehouse.
- Consultas empresariales que comprenden fuente de datos externas que no formaron parte de la implementación inicial.
- Desempeño no satisfactorio de componentes del data warehouse.

## **Impactos Data Warehouse**

El éxito de DW no está en su construcción, sino en usarlo para mejorar procesos empresariales, operaciones y decisiones. Posesionar un DW para que sea usado efectivamente, requiere entender los impactos de implementación en los siguientes ámbitos [3]:

- Impactos Humanos.

Efectos sobre la gente de la empresa:

- *Construcción del DW*: Construir un DW requiere la participación activa de quienes usarán el DW. A diferencia del desarrollo de aplicaciones, donde los requerimientos de la empresa logran ser relativamente bien definidos producto de la estabilidad de las reglas de negocio a través del tiempo, construir un DW depende de la realidad de la empresa como de las condiciones que en ese momento existan, las cuales determinan qué debe contener el DW. La gente de negocios debe participar activamente durante el desarrollo del DW, desde una perspectiva de construcción y creación.
  - *Accesando el DW*: El DW intenta proveer los datos que posibilitan a los usuarios acceder su propia información cuando ellos la necesitan. Esta aproximación para entregar información tiene varias implicancias:
    - o La gente de la empresa puede necesitar aprender nuevas destrezas.
    - o Análisis extensos y demoras de programación para obtener información será eliminada. Como la información estará lista para ser accesada, las expectativas probablemente aumentarán.
    - o Nuevas oportunidades pueden existir en la comunidad empresarial para los especialistas de información.
    - o La gran cantidad de reportes en papel serán reducidas o eliminadas.
    - o La madurez del DW dependerá del uso activo y retroalimentación de sus usuarios.
  - *Usando aplicaciones DSS/EIS*: Usuarios de aplicaciones DSS y EIS necesitarán menos experiencia para construir su propia información y desarrollar nuevas destrezas.
- Impactos Empresariales.

#### Procesos Empresariales y Decisiones Empresariales.

Se deben considerar los beneficios empresariales potenciales de los siguientes impactos:

- Los Procesos de Toma de Decisiones pueden ser mejorados mediante la disponibilidad de información. Decisiones empresariales se hacen más rápidas por gente más informada.
  - Los procesos empresariales pueden ser optimizados. El tiempo perdido esperando por información que finalmente es incorrecta o no encontrada, es eliminado.
  - Conexiones y dependencias entre procesos empresariales se vuelven más claros y entendibles. Secuencias de procesos empresariales pueden ser optimizadas para ganar eficiencia y reducir costos.
  - Procesos y datos de los sistemas operacionales, así como los datos en el DW, son usados y examinados. Cuando los datos son organizados y estructurados para tener significado empresarial, la gente aprende mucho de los sistemas de información. Pueden quedar expuestos posibles defectos en aplicaciones actuales, siendo posible entonces mejorar la calidad de nuevas aplicaciones.
- Comunicación e Impactos Organizacionales.

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

- La gente tiene mayor confianza en las decisiones empresariales que se toman. Ambos, quienes toman las decisiones como los afectados conocen que está basada en buena información.
  - Las organizaciones empresariales y la gente de la cual ella se compone queda determinada por el acceso a la información. De esta manera, la gente queda mejor habilitada para entender su propio rol y responsabilidades como también los efectos de sus contribuciones; a la vez, desarrollan un mejor entendimiento y apreciación con las contribuciones de otros.
  - La información compartida conduce a un lenguaje común, conocimiento común, y mejoramiento de la comunicación en la empresa. Se mejora la confianza y cooperación entre distintos sectores de la empresa, viéndose reducida la sectorización de funciones.
  - Visibilidad, accesibilidad, y conocimiento de los datos producen mayor confianza en los sistemas operacionales.
- Impactos Técnicos de DW.

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

- Nuevas destrezas de desarrollo: Cuando se construye el DW, el impacto más grande sobre la gente técnica está dada por la curva de aprendizaje, muchas destrezas nuevas se deben aprender, incluyendo:
  - Conceptos y estructura DW.
  - El DW introduce muchas tecnologías nuevas (ETT, Carga, Acceso de Datos, Catálogo de Metadatos, Implementación de DSS/EIS), y cambia la manera en que se usa la tecnología existente. Nuevas responsabilidades de soporte, nuevas demandas de recursos y nuevas expectativas, son los efectos de estos cambios.
  - Destrezas de diseño y análisis donde los requerimientos empresariales no son posibles de definir de una forma estable a través del tiempo.
  - Técnicas de desarrollo incremental y evolutivo.
  - Trabajo en equipo cooperativo con gente de negocios como participantes activos en el desarrollo del proyecto.
- Nuevas responsabilidades de operación: Cambios sobre los sistemas y datos operacionales deben ser examinados más cuidadosamente para determinar el impacto que estos cambios tienen sobre ellos, y sobre el DW.

### **¿Cómo trabaja el Data Warehouse?**

- Extrae la información operacional.
- Transforma la operación a formatos consistentes.
- Automatiza las tareas de la información para prepararla a un análisis eficiente.

### **¿En qué se puede usarlo?**

- Manejo de relaciones de marketing.
- Análisis de rentabilidad.
- Reducción de costos.

## **Estructura de un Datawarehouse**

Los data warehouses tienen una estructura distinta. Hay niveles diferentes de esquematización y detalle que delimitan el data warehouse. La estructura de un data warehouse se muestra en la Figura 9 [5].

En la figura 9, se muestran los diferentes componentes del data warehouse que 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:

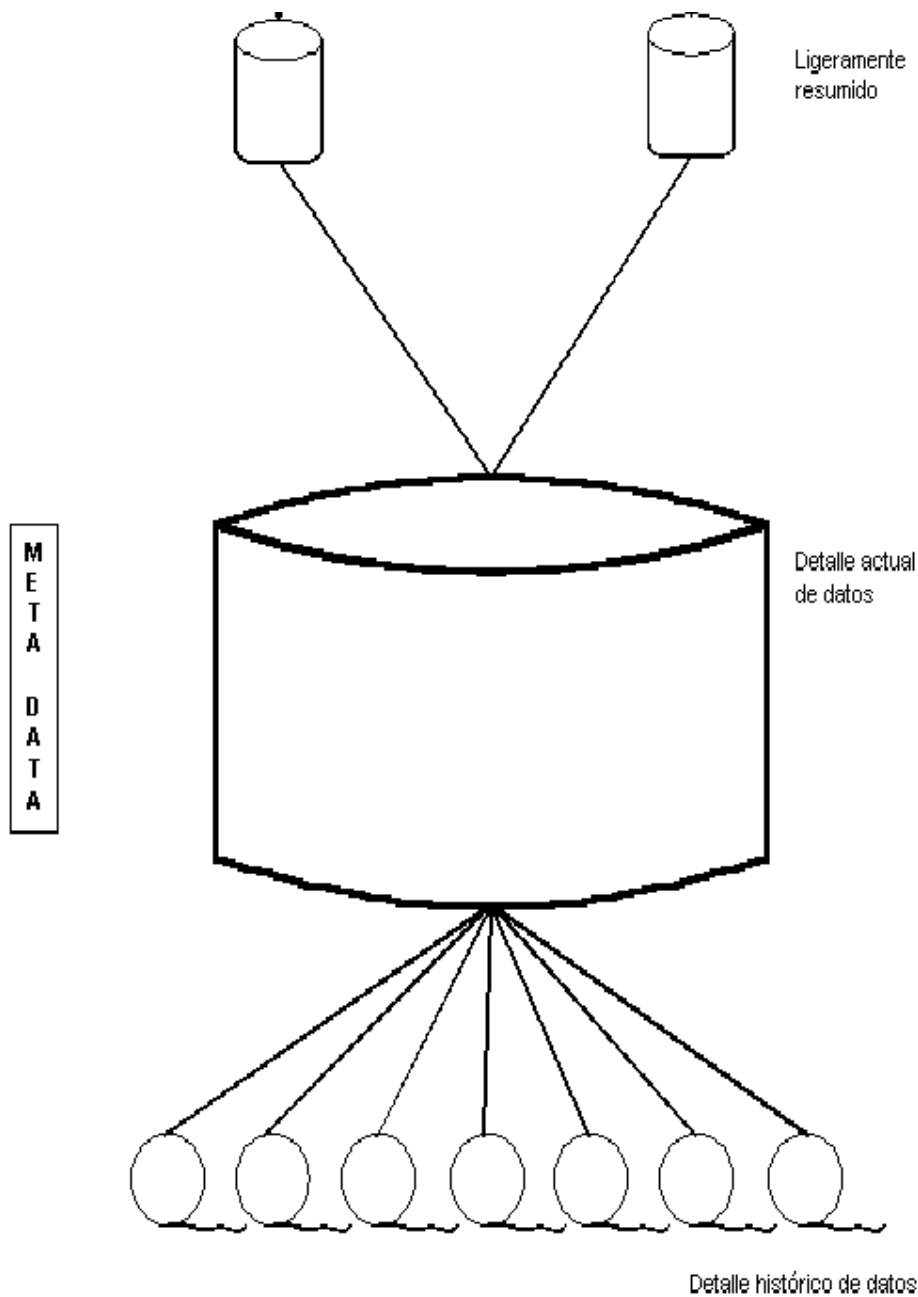
- Refleja las ocurrencias más recientes, las cuales son de gran interés.
- Es voluminoso, ya que se almacena al más bajo nivel de granularidad.
- Casi siempre se almacena en disco, el cual es de fácil acceso, aunque su administración sea costosa y compleja.

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 data warehouse casi siempre se almacena en disco. Los puntos en los que se basa el diseñador para construirlo son:

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

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



**Figura 9. Estructura de los Datos de un Data Warehouse.**

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

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

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

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

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

A fin de recordar los diferentes niveles de los datos encontrados en el data warehouse, se considera el ejemplo mostrado en la Figura10.

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.

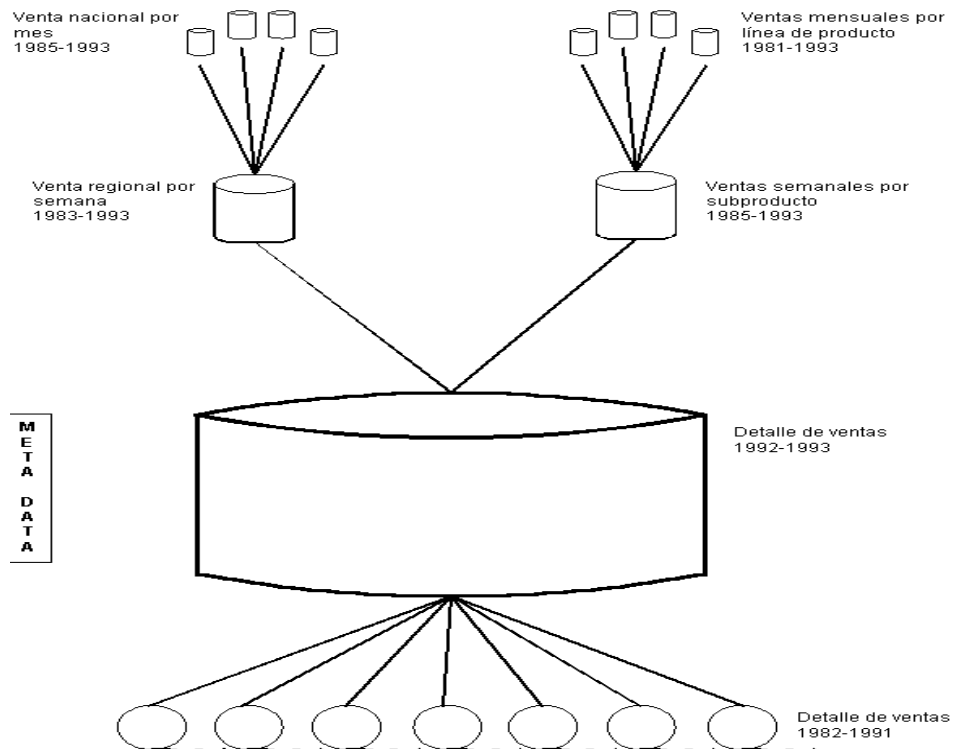


Figura 10. Ejemplo de Niveles de Esquematización que podría encontrarse en un DW.

## Inteligencia de Negocio

El término inteligencia empresarial se refiere al uso de los datos de una empresa para facilitar la toma de decisiones a las personas que deciden, es decir, la comprensión del funcionamiento actual y la anticipación de acciones para dar una dirección bien informada a la empresa.

Las herramientas de inteligencia se basan en la utilización de un sistema de información de inteligencia que se forma con distintos datos extraídos de los datos de producción, con información relacionada con la empresa o sus ámbitos y con datos económicos.

Mediante las herramientas y técnicas ETL (extraer, transformar y cargar) se extraen los datos de distintas fuentes, se depuran y preparan (homogeneización de los datos) para cargarlos en un almacén de datos.

La vida o el periodo de éxito de un software de inteligencia de negocios dependerá únicamente del nivel de éxito del cual haga en beneficio de la empresa que lo usa, si esta empresa es capaz de incrementar su nivel financiero, administrativo y sus decisiones mejoran el accionar de la empresa, la inteligencia de negocios usada estará presente por mucho tiempo, de lo contrario será sustituido por otro que aporte mejores resultados y más precisos.

Por último, las herramientas de inteligencia analítica posibilitan el modelado de las representaciones en base a consultas para crear tablas de bordes; lo cual se conoce como presentación de informes [6].

## Características

Este conjunto de herramientas y metodologías tienen en común las siguientes características:

- *Accesibilidad a la información.* Los datos son la fuente principal de este concepto. Lo primero que debe garantizar este tipo de herramientas y técnicas será el acceso de los usuarios a los datos con independencia de la procedencia de estos.
- *Apoyo en la toma de decisiones.* Se busca ir más allá en la presentación de la información, de manera que los usuarios tengan acceso a herramientas de análisis que les permitan seleccionar y manipular sólo aquellos datos que les interesen.
- *Orientación al usuario final.* Se busca independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

## Niveles de realización de BI

De acuerdo a su nivel de complejidad se pueden clasificar las soluciones de Business Intelligence en:

- Consultas e informes simples (Querys y reports).
- Cubos OLAP (On-Line Analytic Processing).
- Data Mining o minería de datos.
- Sistemas de previsión empresarial; predicción mediante estudio de series temporales (ejemplo: Previsión de ventas).

## **Extracción, transformación y carga (ETL)**

Para poblar el Data Warehouse se deben mover bloques de datos, muchas veces desde diferentes sistemas operativos, estructuras de archivos y bases de datos, mediante procesos programados que se ejecutan frecuentemente fuera del horario de trabajo para no insumir tiempo de procesamiento del hardware de la empresa, entorpeciendo la operatoria de la misma.

Los subsistemas para poblar el Data Warehouse se pueden construir utilizando herramientas y productos disponibles en el mercado, programas y procesos codificados desde cero, o combinaciones de estos elementos.

Al construir los sistemas para poblar el Data Warehouse, se debe considerar la posibilidad de que estos permitan regular el crecimiento evolutivo del Data Warehouse, brindando escalabilidad y soporte para grandes cantidades de datos y consultas complejas. Se pueden encontrar dificultades adicionales dependiendo de las fuentes de datos que se tengan disponibles, que implican el uso de diferentes herramientas y tecnologías para acceder a cada uno de ellos.

### **Extracción**

El propósito principal de la fase de *extracción* es capturar y copiar los datos requeridos de uno o más sistemas operacionales o fuentes de datos. Los datos que se extraen son colocados en un archivo intermedio con un formato definido, que luego será utilizado por la siguiente fase del proceso.

Los registros que sean rechazados en el proceso deben ser registrados en un archivo o *log* de rechazos para que puedan ser analizados posteriormente y así tener la posibilidad de cargarlos en el Data Warehouse correctamente. Además, esto permite descubrir los errores que han ocurrido en los procesos de creación de los datos operacionales.

Ejemplos de estos errores son violaciones de integridad, claves duplicadas, formatos de datos incorrectos y datos inválidos como campos vacíos, fechas futuras e importes negativos cuando estos no correspondan.

Además, esto permite descubrir los errores que han ocurrido en los procesos de creación de los datos operacionales. Ejemplos de estos errores son violaciones de integridad, claves duplicadas, formatos de datos incorrectos y datos inválidos como campos vacíos, fechas futuras e importes negativos cuando estos no correspondan.

## Transformación

Las funciones básicas a ser realizadas en esta fase consisten en leer los archivos intermedios generados por la fase de extracción, realizar las transformaciones necesarias, construir los registros en el formato del Data Warehouse y crear un archivo de salida conteniendo todos los registros nuevos a ser cargados en el Data Warehouse.

La mayor parte del trabajo en esta fase involucra el efectuar las transformaciones necesarias. Estas transformaciones incluyen:

- Combinar campos múltiples de nombres y apellidos en un solo campo.
- Fusionar campos o datos homónimos.
- Separar un campo de fecha en campos de año, mes y día.
- Cambiar la representación de los datos, como TRUE (verdadero) a 1, y FALSE (falso) a 0, o códigos postales numéricos a alfanuméricos, respetando los estándares de la empresa.
- Cambiar un dato que tiene múltiples representaciones a una sola representación, como por ejemplo definir un formato común para números telefónicos, o establecer un término común para los nombres de los campos o los valores de los datos que sean sinónimos.

## Carga

El objetivo de esta fase consiste en tomar los registros formateados por la fase de transformación y cargarlos en el Data Warehouse, que es el contenedor para todos los datos informativos (actuales e históricos) requeridos por las operaciones del Data Warehouse.

Generalmente los datos son insertados en el Data Warehouse, rara vez son actualizados o eliminados.

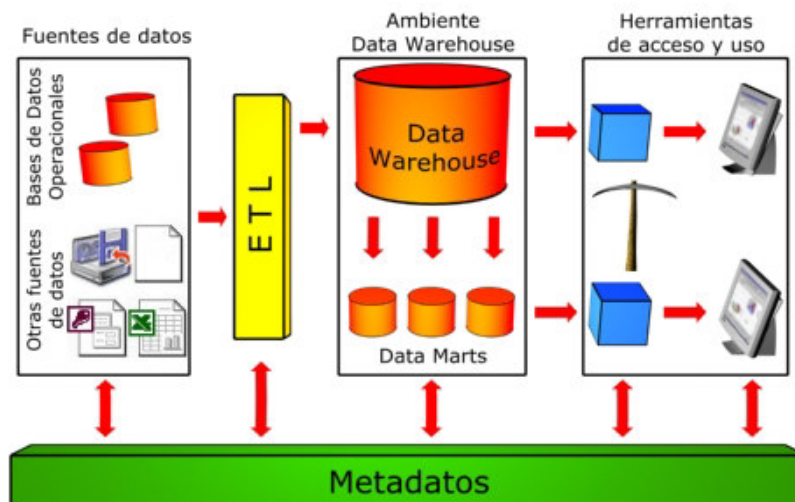


Figura 11. Extracción, transformación y carga. ETL.

## Ambiente Data Warehouse

Es el bloque donde se almacenan los datos informativos, utilizado principalmente para usos estratégicos. No obstante, existen herramientas que no hacen uso de este bloque, realizando las consultas multidimensionales directamente sobre la base operacional. En este caso se puede pensar en el Data Warehouse simplemente como una vista lógica o virtual de datos.

## Data Mart

Un *Data Mart* es una implementación de un Data Warehouse con un determinado alcance de información y un soporte limitado para procesos analíticos, que sirve a un sólo departamento de una organización o para el análisis de problemas de un tema particular.

El Data Mart es un subconjunto de información corporativa con formato adicional a la medida de un usuario específico del negocio. Un Data Mart será siempre menor en complejidad y alcance de los datos.

Un Data Warehouse tiene más usuarios y más temas que un Data Mart, brindando una vista más amplia entre múltiples áreas.

Existen dos grandes filosofías con respecto a la relación entre los conceptos de Data Warehouse y Data Mart.

Bill Inmon, quien es considerado el padre del Data Warehouse, propuso la idea de que los Data Marts se sirven del Data Warehouse para extraer información. La misma está almacenada en tercera forma normal, en un modelo relacional.

Por su parte, Ralph Kimball, el principal propulsor del enfoque dimensional para el diseño del Data Warehouse, sostiene que el Data Warehouse es el resultado de la unión de los Data Marts de la empresa.

## Metadatos

Los metadatos son datos acerca de los datos. En una base de datos los metadatos son la representación de los diversos objetos que definen una base de datos, por ejemplo, ubicación y descripción de base de datos, tablas, nombres y resúmenes. También podemos mencionar las descripciones lógicas y físicas de tablas, columnas y atributos.

Uno de los problemas con el que pueden encontrarse los usuarios de un data warehouse es saber lo que hay en él y cómo pueden acceder a lo que quieren. A fin de proveer el acceso a los datos universales, es absolutamente necesario mantener los metadatos. Un componente llamado repositorio les ayuda a conseguirlo.

Los metadatos son sólo una de las utilidades del repositorio, pero éste tiene muchas funcionalidades: catalogar y describir la información disponible, especificar el propósito de la misma, indicar las relaciones entre los distintos datos, establecer

quién es el propietario de la información, relacionar las estructuras técnicas de datos con la información de negocio, establecer las relaciones con los datos operacionales y las reglas de transformación, y limitar la validez de la información.

## Modelado de Datos

Para comprender uno de los aspectos más relevantes de la arquitectura del Data Warehouse, como es el modelado de datos, es necesario establecer primero las diferencias sustanciales entre los dos mundos de modelado existentes: entidad-relación (ER) y dimensional.

El *modelado entidad-relación* se utiliza habitualmente para crear un único modelo complejo de todos los procesos de una organización. Este enfoque ha demostrado ser efectivo para crear sistemas eficientes de procesamiento transaccional en línea (OLTP).

Por otra parte, el *modelado dimensional* crea modelos individuales para reflejar procesos discretos de negocio. Este modelado organiza la información en estructuras que usualmente corresponden a la forma en que los analistas realizan sus consultas de los datos del Data Warehouse.

## El modelo relacional

En la mayoría de los sistemas transaccionales el objetivo del modelo es garantizar la integridad de los datos, además de eliminar cualquier tipo de redundancia en los datos. Este enfoque es perfecto para los entornos de procesamiento transaccional, porque las transacciones son muy simples y deterministas.

El éxito del procesamiento transaccional en un entorno de bases de datos relacionales se debe básicamente a esta disciplina de modelado.

Sin embargo, para el propósito de un Data Warehouse, el modelo relacional (ER) presenta los siguientes problemas:

- Legibilidad limitada. Los usuarios finales no son capaces de entender el modelo ER. Por tanto, no pueden “navegar” por dicho modelo en busca de información.
- Dificultad para las herramientas de consulta en el acceso a un modelo ER general. Las herramientas de consulta a menudo poseen prestaciones mediocres o inaceptables cuando se trabaja en entornos relacionales de grandes volúmenes de información.
- La utilización de la técnica de modelado ER frustra el principal atractivo del Data Warehouse. Al utilizar el modelado ER queda frustrada la recuperación de información intuitiva y con alto rendimiento.

## El modelo dimensional

El modelado dimensional es una técnica de diseño lógico que busca presentar la información en un marco estándar e intuitivo que permita un acceso de alto rendimiento. Este modelado se vale de los principios de la disciplina que emplea el modelo relacional con algunas importantes restricciones.

El modelado dimensional es esencialmente útil para resumir y organizar los datos y la presentación de información para soportar el análisis de la misma. Existen algunos conceptos básicos para comprender la filosofía de este tipo de modelado: áreas temas, medidas, dimensiones y hechos.

Un *área tema* es una cuestión de interés de una función empresarial. Las áreas tema en conjunto constituyen el ámbito de implementación del Data Warehouse. Por ejemplo, el departamento de Comercialización de una empresa puede estar interesado en las áreas tema de Pedidos, Promociones, Mercados y Ventas.

Para especificar las áreas tema se deben identificar las medidas. Una *medida* o *indicador* es un cuantificador del desempeño de un ítem o una actividad del negocio. La información que brinda una medida es usada por los usuarios en sus consultas para evaluar el desempeño de un área tema.

El Data Warehouse organiza un gran conjunto de datos operacionales mediante múltiples dimensiones. Una *dimensión* es una colección de miembros o entidades del mismo tipo y constituye un calificador conceptual que provee el contexto o significado para una medida.

La forma de representar la organización de los datos en un modelo dimensional es a través de un *cubo* (el cual no necesariamente debe tener tres dimensiones). Los miembros de una dimensión pueden estar organizados en una o más jerarquías.

Una *jerarquía* es un conjunto de miembros de una dimensión, los cuales se definen por su posición relativa con respecto a los otros miembros de la misma dimensión, y forman en su totalidad una estructura de árbol. Partiendo de la raíz del árbol, los miembros son progresivamente más detallados hasta llegar a las hojas, donde se obtiene el mayor nivel de detalle.

Puede darse el caso en que una dimensión no necesite jerarquizarse debido a que ninguno de sus miembros posee una posición relativa con respecto a los otros miembros. Por ejemplo, una dimensión Cliente que tiene como miembros nombre, sexo y fecha de nacimiento, no necesita organizar estos miembros porque todos están al mismo nivel de detalle, a menos que se desee agruparlos por alguno de ellos para visualizar los datos.

Existen principalmente dos *esquemas para el modelo dimensional*: el esquema *estrella* (star), y el esquema *copo de nieve* (snowflake).

En el esquema estrella, cada modelo dimensional está compuesto de una tabla central con una clave primaria compuesta, denominada tabla de *hechos*, y un conjunto de tablas periféricas denominadas tablas de dimensiones.

Cada una de las tablas de dimensiones tiene una clave primaria que corresponde exactamente con uno de los componentes de la clave compuesta de la tabla de hechos. Las tablas de hechos, además de sus campos clave, contienen una o más medidas, indicadores o “hechos”. Las medidas más útiles en una tabla de hechos son numéricas y aditivas. La aditividad es crucial porque las aplicaciones Data Warehouse casi nunca recuperan un solo registro de la tabla de hechos, sino que acceden a cientos, miles o incluso millones de registros a la vez.

Las tablas de dimensiones, por el contrario, contienen información textual descriptiva. Los atributos de las dimensiones se emplean como fuente de las restricciones en las consultas al Data Warehouse.

En el modelo estrella las dimensiones no se normalizan. Con ello se logra minimizar el número de uniones y, por consiguiente, incrementar el rendimiento de las consultas (una tabla de hechos está relacionada con numerosas tablas de dimensiones).

Una variante del modelo en estrella es el modelo copo de nieve o snowflake. En este modelado se normalizan las dimensiones reflejando las jerarquías en las mismas y conservando lo esencial del modelo en estrella: las tablas de hechos. La ventaja del modelo copo de nieve es eliminar la redundancia de datos y por lo tanto ocupar menos espacio en disco.

En las bases de datos usadas para data warehousing, un esquema en estrella es un modelo de datos que tiene una tabla de hechos (o tabla fact) que contiene los datos para el análisis, rodeada de las tablas de dimensiones. Este aspecto, de tabla de hechos (o central) más grande rodeada de radios o tablas más pequeñas es lo que asemeja a una estrella, dándole nombre a este tipo de construcciones.

Las tablas de dimensiones tendrán siempre una clave primaria simple, mientras que en la tabla de hechos, la clave principal estará compuesta por las claves principales de las tablas dimensionales [7].



**Figura 12. Modelado de Datos. Esquema en Estrella.**

Esquema en copo de nieve (bola de nieve) es una variedad más compleja del esquema estrella. El afinamiento está orientado a facilitar mantenimiento de dimensiones.

Lo que distingue a la arquitectura en copo de nieve de la arquitectura en esquema estrella, es que las tablas de dimensiones en este modelo representan relaciones normalizadas (3NF) y forman parte de un modelo relacional de base de datos.

Con varios usos del esquema en bola de nieve, el más común es cuando las tablas de dimensiones están muy grandes o complejas y es muy difícil representar los datos en esquema estrella.

El problema es que para extraer datos de las tablas en esquema de copo de nieve, a veces hay que vincular muchas tablas en las sentencias SQL que puede llegar a ser muy complejo y difícil para mantener [8].

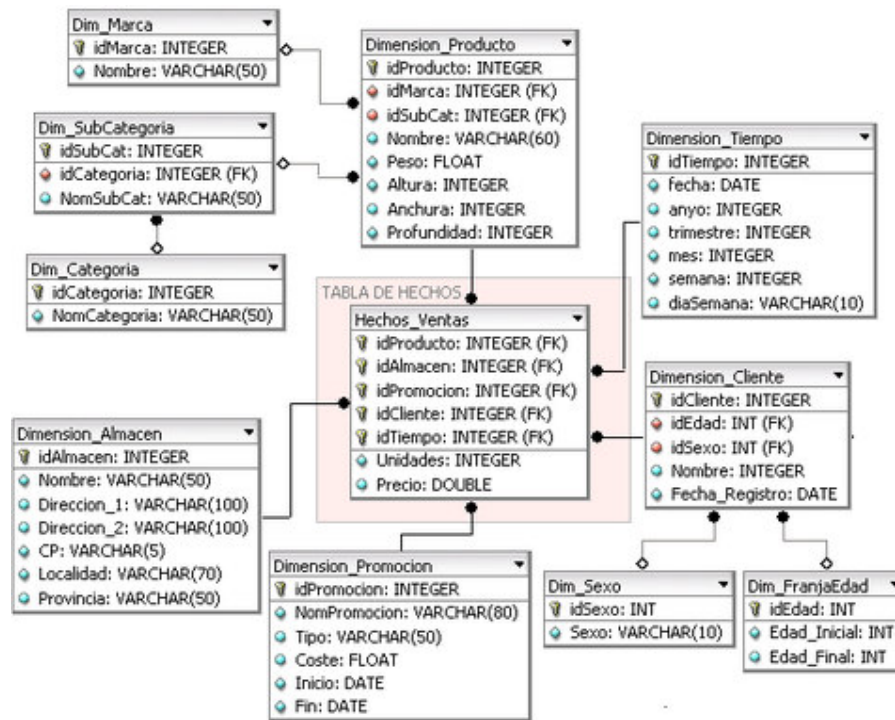


Figura 13. Modelado de Datos. Esquema Copo de Nieve.

## Ventajas del modelo dimensional

El modelo dimensional presenta importantes ventajas de las que carece el modelo relacional. Uno de los puntos fuertes del modelo dimensional es que el marco predecible del esquema estrella resiste a los cambios inesperados en el comportamiento del usuario.

Cada dimensión es equivalente a las demás y todas las dimensiones pueden ser concebidas como puntos de entrada hacia la tabla de hechos. El diseño lógico puede realizarse independientemente de los patrones de consulta esperados, siendo consideradas de la misma forma tanto las interfaces de usuario como las estrategias de consulta, así como el lenguaje de consulta generado contra el modelo dimensional.

Otra cualidad del modelo dimensional es la flexibilidad. Los nuevos elementos de datos y las nuevas decisiones de diseño son fácilmente adaptables. Todas las tablas pueden modificarse simplemente agregando nuevos registros de datos o se pueden incluir nuevas dimensiones al modelo sin necesidad de volver a cargar los datos posteriormente. Además no es necesario volver a programar las herramientas de consulta o de informes para adaptarse a los cambios, y las aplicaciones existentes pueden continuar su ejecución brindando los mismos resultados.

Las modificaciones ante las cuales el modelo dimensional es flexible incluyen:

- Agregar medidas a la tabla de hechos, siempre que sean aditivas y consistentes con el mayor nivel de detalle de las dimensiones.
- Agregar atributos a las dimensiones.

- Agregar nuevas dimensiones, siempre que exista un único valor de dicha dimensión definido para cada registro de la tabla de hechos.
- Particionar los registros de una dimensión a un mayor nivel de detalle a partir de un determinado punto en el tiempo. Los registros anteriores permanecerán sin cambios mientras que los futuros registros se almacenarán de acuerdo al nuevo modelo.

Una ventaja adicional del modelo dimensional es el creciente número de utilidades administrativas y aplicaciones que gestionan y utilizan los agregados. Los *agregados* son registros resumidos que son lógicamente redundantes con la información ya existente en el Data Warehouse y son empleados para mejorar el rendimiento de las consultas.

Cualquier implementación de tamaño mediano o grande del Data Warehouse requiere la creación de una estrategia de agregados. Todas las aplicaciones software de gestión de agregados, así como las utilidades de navegación de agregados, dependen de una estructura específica de las tablas de hechos y dimensiones que es absolutamente dependiente del modelo dimensional. Si no se emplea el enfoque del modelo dimensional no es posible beneficiarse de tales aplicaciones.

## **Herramientas de acceso y uso**

Sin las herramientas adecuadas de acceso y análisis el Data Warehouse se puede convertir en una mezcla de datos sin ninguna utilidad. Es necesario poseer técnicas que capturen los datos importantes de manera rápida y puedan ser analizados desde diferentes puntos de vista.

También deben transformar los datos capturados en información útil para el negocio. Actualmente a este tipo de herramientas se las conocen como Herramientas de Inteligencia de Negocio (Business Intelligence Tools, BIT) y están situadas conceptualmente sobre el Data Warehouse.

Cada usuario final debe seleccionar la herramienta que mejor se ajusta a sus necesidades y a su Data Warehouse. Entre ellas podemos citar las Consultas SQL (Structured Query Language), las Herramientas MDA (Multidimensional Analysis) y OLAP (On-Line Analytical Processing) y las herramientas Data Mining.

Este bloque también incluye el hardware y software involucrados en mostrar la información en pantalla y emitir reportes de impresión, hojas de cálculo, gráficos y diagramas para el análisis y presentación.

## **OLAP (On Line Analytical Processing)**

En un Data Warehouse se depositan datos para consulta y análisis, a diferencia del Procesamiento Transaccional en Línea (OLTP), en donde los datos se almacenan para operación y control.

El Procesamiento Analítico en Línea (OLAP) es una tecnología de análisis de datos que crea nueva información empresarial a partir de los datos existentes. Posee las siguientes características:

- Presenta una visión multidimensional lógica de los datos del Data Warehouse, independiente de su forma de almacenamiento.
- Crea resúmenes, adiciones y jerarquías.
- Comprende consultas interactivas y análisis de los datos. Permite una profundización hacia niveles más detallados o un ascenso a niveles superiores de resumen y adición.
- Ofrece opciones de modelado analítico, incluyendo un motor de cálculo para medir datos numéricos a través de muchas dimensiones, así como también provee modelos estadísticos básicos.
- Responde con rapidez a las consultas, de modo que el proceso de análisis no se interrumpa y la información no se desactualiza.
- Recupera y exhibe datos tabulares en dos o tres dimensiones, cuadros o gráficos, con un fácil pivoteo de los ejes.

Esta tecnología es independiente de la implementación y permite el empleo de cualquier base de datos, ya sea relacional (ROLAP, cuando se aplica el modelo dimensional a una base de datos relacional), dimensional (MOLAP, modelo dimensional sobre base de datos dimensional), de objetos, etc.

## **Drill Down y Roll Up**

Una de las características del acceso a la información es la posibilidad de representarla a diferentes niveles de agregación. Esto se logra mediante las características conocidas como Drill Down y Roll Up.

Estas son técnicas para navegar a través de distintos niveles de detalle de una jerarquía de datos, desde los de mayor nivel de agregación (también llamados datos sumariados) hasta los más detallados.

Drill Down explora los hechos hacia los niveles más detallados de la jerarquía de dimensiones, mientras que Roll Up explora los hechos iterativamente hacia el nivel más alto de agregación.

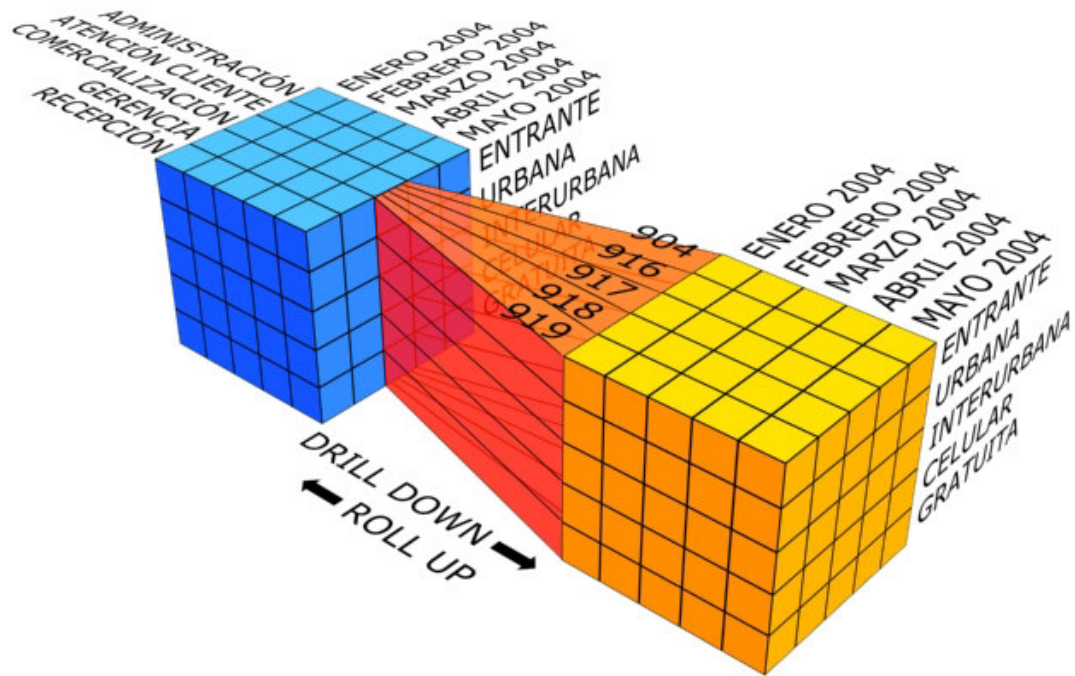


Figura 14. Herramientas de Acceso y Uso. Drill Down y Roll Up.

## Slice y Dice

Estos términos son utilizados para describir cómo los datos multidimensionales pueden ser mostrados aplicando diferentes filtros a los cubos.

Slice (Rebanar) es la acción de conformar un subconjunto de los datos de un cubo aplicándole una única restricción a una sola dimensión, mediante la elección de un miembro en particular.

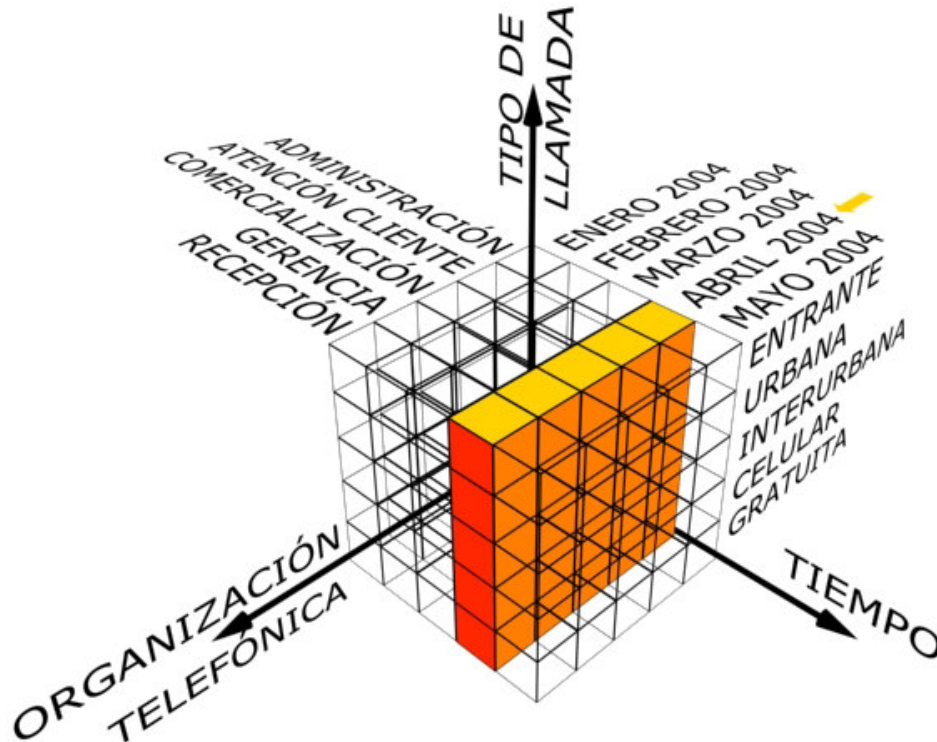


Figura 15. Herramientas de Acceso y Uso. Slice y Dice.

## Data Mining (Minería de Datos)

El Data Mining consiste en el análisis y modelización estadística de datos. Es una poderosa tecnología con gran potencial para ayudar a las compañías a concentrarse en la información más importante de su Data Warehouse.

Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información. Los análisis prospectivos automatizados ofrecidos por un producto así van más allá de los eventos pasados provistos por herramientas retrospectivas típicas de Sistemas de Soporte a las Decisiones.

Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

Las *técnicas* que se pueden utilizar en el proceso de Data Mining se clasifican en:

- *Análisis estadístico.* Se utiliza para detectar patrones no usuales de datos. Estos patrones se describen mediante modelos estadísticos y matemáticos. Algunas técnicas son análisis lineal y no lineal, análisis de regresión, análisis de

univariación y multivariación, análisis de series históricas, pronósticos, lógica difusa y clustering.

- *Descubrimiento de conocimientos.* Tiene sus raíces en la inteligencia artificial y el aprendizaje con máquinas. Permite extraer de los datos información implícita, no trivial, desconocida y potencialmente útil. Como proceso, consiste en buscar en los datos sin establecer por adelantado una hipótesis, e incluso de esa forma encontrar información inesperada e interesante de relaciones y patrones entre los elementos de datos o reglas empresariales de los datos investigados. Tiene como resultado el descubrimiento de hechos empresariales ocultos en el Data Warehouse o en los Data Marts. Algunos usos de esta técnica incluyen la clasificación empresarial, la cual consiste en detectar reglas empresariales que dividan los registros de datos en grupos inconexos; las redes neuronales, para descubrir grupos preferenciales de clientes, detectar fraudes financieros y planear la logística y transporte; descubrimiento de reglas y representación mediante árboles de decisión; y las asociaciones, para describir la afinidad entre un conjunto de elementos de datos con un nivel de confianza.
- *Sistemas de visualización.* Permiten a los analistas hacer descubrimientos mediante el análisis gráfico de muchas variables para luego ver patrones y relaciones que serían difíciles de determinar por medio de algoritmos automatizados.
- *Sistemas de información geográfica.* Relaciona los datos de Data Warehouse de diferentes ubicaciones físicas con representaciones geográficas.

## Sitios de Internet consultados

- [1] <http://www.lawebdelprogramador.com/cursos/mostrar.php?id=278&texto=Data+Warehouse>
- [2] <http://www.emagister.com/datawarehouse-cursos-1107904.htm>
- [3] <http://es.geocities.com/cibercero/mtd/foro/datawarehouse11.htm#data>
- [4] <http://www.sqlmax.com/dataw1.asp>
- [5] <http://www.ongei.gob.pe/publica/metodologias/Lib5084/14.HTM>
- [6] [http://es.wikipedia.org/wiki/BI\\_\(inform%C3%A1tica\)](http://es.wikipedia.org/wiki/BI_(inform%C3%A1tica))
- [7] [http://es.wikipedia.org/wiki/Esquema\\_en\\_estrella](http://es.wikipedia.org/wiki/Esquema_en_estrella)
- [8] [http://etl-tools.info/es/bi/almacenedatos\\_esquema-copo-de-nieve.htm](http://etl-tools.info/es/bi/almacenedatos_esquema-copo-de-nieve.htm)