Fundamentos de Data Warehouse

Mendez, A., Mártire, A., Britos, P. Y Garcia-Martínez, R.

Centro de Actualización Permanente en Ingeniería del Software
Escuela de Postgrado
Instituto Tecnológico de Buenos Aires
Av. Eduardo Madero 399 – (C1106ACD) – Buenos Aires - ARGENTINA
mendez_andrea@yahoo.com.ar, ariel_martire@yahoo.com

1. Introducción

El Data Warehouse es una tecnología para el manejo de la información construido sobre la base de optimizar el uso y análisis de la misma utilizado por las organizaciones para adaptarse a los vertiginosos cambios en los mercados. Su función esencial es ser la base de un sistema de información gerencial, es decir, debe cumplir el rol de integrador de información proveniente de fuentes funcionalmente distintas (Bases Corporativas, Bases propias, de Sistemas Externos, etc.) y brindar una visión integrada de dicha información, especialmente enfocada hacia la toma de decisiones por parte del personal jerárquico de la organización.

Es un sitio donde se almacena de manera integrada toda la información resultante de la operatoria diaria de la organización. Además, se almacenan datos estratégicos y tácticos con el objetivo de obtener información estratégica y táctica que pueden ser de gran ayuda para aplicar sobre los mismos técnicas de análisis de datos encaminadas a obtener información oculta (Data Mining).

Esta información incluye movimientos que modifican el estado del negocio, cualquier interacción que se tenga con los clientes y proveedores, y cualquier dato adicional que ayude a comprender la evolución del negocio.

Esta tecnología ayuda a la organización a responder preguntas esenciales para la toma de decisiones que le permitan obtener ventajas competitivas y mejorar su posición en el mercado en el que operan. Algunas de las preguntas podrían ser:

- Cuál es el perfil de mis clientes?
- Cómo es su comportamiento?
- Cuál es la rentabilidad que me deja?
- Cuál es el riesgo que corro con él?
- Qué servicios y productos utiliza y cómo puedo incrementarlos?
- Etc.

Además, se aplican técnicas de limpieza e integración de datos, esto asegura la existencia de estructuras homogéneas persistentes en el tiempo.

Para comprender mejor el funcionamiento de ésta tecnología explicaremos su arquitectura y los sistemas OLTP y OLAP.

2. Arquitectura del Data Warehouse

La arquitectura (Figura 2) de esta tecnología está integrada por los siguientes componentes:

2.1. OLTP (On-Line Transaction Processing)

Son aplicaciones que definen el comportamiento habitual de un entorno operacional de gestión y ejecutan las operaciones del día a día. Algunas de las características más comunes de este tipo de transacciones podrían ser:

-Altas/Bajas/Modificaciones
- Consultas rápidas, escuetas y predecibles
- Poco volumen de información e información disgregada
- Transacciones rápidas
- Gran nivel de concurrencia
- Modo de actualización on-line
- Baja redundancia de datos

Algunos ejemplos de este tipo de aplicaciones son:

- Compras
- Ventas
- Inventario
- Sueldos

2.2. Consolidación

Es la parte del proceso de Data Warehouse que se encarga de producir el cambio de los sistemas OLTP a las Bases de Datos OLAP. Consolidan datos de aplicaciones no integradas, sumarizan datos

disgregados y los transforman. Este proceso está compuesto por tres pasos

Validación de Consistencia de los datos

- Comprueba la validez de los datos en el entorno operacional
- Inconsistencia entre distintas aplicaciones dentro del sistema

Mecanismos de Consolidación

- Refresco de datos: Volcado completo de los datos procedentes del sistema operacional
- Actualización de datos: Volcado incremental tomando como criterio la fecha de operación
- Propagación de datos

Factores técnicos

- Mecanismo de transporte
- Tiempos de carga
- Reformateo de datos

2.3. Middleware

Es un software que reside físicamente en un Cliente y en un Servidor de Comunicaciones, localizado entre el Cliente y el Servidor. Actúa como traductor entre distintas tecnologías. Permite que dos o más sistemas trabajen juntos aunque no estén preparados para ello. (Figura 1).

Algunas de sus características más relevantes son:

- Un mismo middleware puede poseer más de una máquina virtual para soportar diferentes entornos de desarrollo
- Gestiona las comunicaciones con el Data Warehouse
- Controla la concurrencia y controla los procesos Batch
- Posee diversos controladores de Bases de Datos para acceder a las distintas fuentes, por ejemplo, Oracle, Sybase, AS400, etc.

Ejemplos:

- Monitores de procesamiento de transacciones
- Convertidores de datos
- Replicación de datos
- Controladores de comunicación

2.4. OLAP (On-Line Analytical Process)

Son aplicaciones que se encargan de analizar datos del negocio para generar información táctica y estratégica que sirve de soporte para la toma de decisiones. Mientras que las transacciones OLTP utilizan Bases de Datos Relacionales u otro tipo de archivos, OLAP logra su máxima eficiencia y flexibilidad operando sobre Bases de datos Multidimensionales.

Podemos nombrar las siguientes características como las más sobresalientes de estas aplicaciones:

- Estructura de datos transparente al usuario
- Solo Consulta, trabajan sobre la información operacional generada por los sistemas OLTP
- Consultas sobre grandes volúmenes de datos no predecibles
- Información histórica
- Modo de actualización Batch
- Alta redundancia de datos para facilitar la generación de consultas y obtener buenos tiempos de respuesta

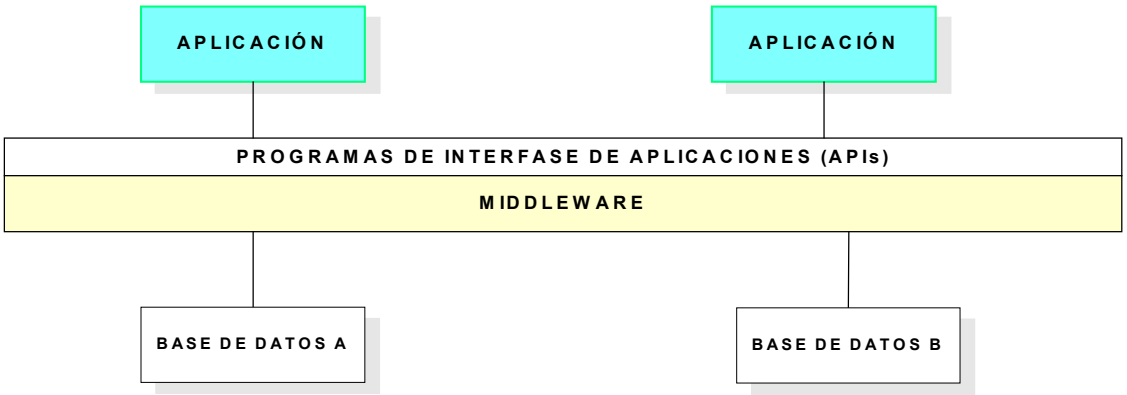

Figura 1 – Middleware

- Poderoso Back-end analítico para múltiples aplicaciones de usuarios

- Trabaja con resúmenes de miles de registros condensados en una sola respuesta

2.5. Data Marts

Una vez contando con la base de información empresarial integrada y, a partir de esta, se crean subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones. Los datos existentes en este contexto pueden ser sumarizados, agrupados, explorados y reportados de múltiples formas para que diversos grupos de usuarios realicen la explotación de los mismos.

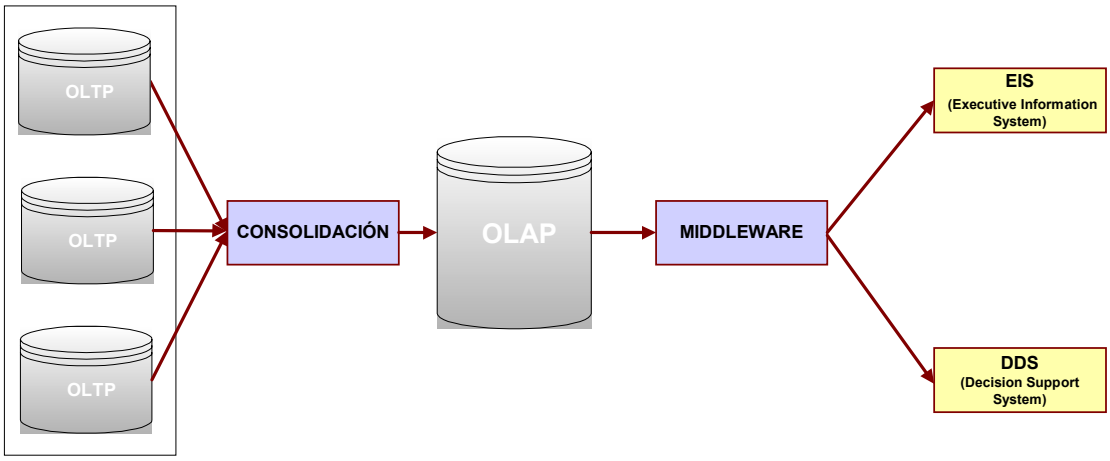

Figura 2 – Arquitectura del Data Warehouse

Es un modelo multidimensional basado en tecnología OLAP, incluyendo variables claves y los indicadores claves para el proceso de toma de decisiones.

Algunas ventajas de la construcción del Data Mart:

- Son más simples de implementar que un Data Warehouse
- Pequeños conjuntos de datos y, en consecuencia, menor necesidad de recursos
- Se encuentran más rápidamente las necesidades de las Unidades de Negocio
- Queries más rápidos por menor volumen de datos

Como desventaja se puede decir que, en algunos casos, añaden tiempo al proceso de actualización.

En síntesis, son pequeños Data Warehouse centrados en un tema o un área de negocio específico. En muchos casos, los Data Warehouse comienzan siendo Data Marts con el objetivo de minimizar los riesgos para luego ir ampliando su espectro gradualmente.

3. Aplicaciones

3.1. EIS (Executive Information System)

Son herramientas para proveer información estratégica a los ejecutivos mediante informes, comparativas y cuadros de mando multidimensionales.

3.2. DSS (Decission Support System)

Herramienta de soporte para la toma de decisiones. Incorpora reglas de decisión y análisis de datos no predefinidos en las posibilidades de un EIS.

- Sistemas de presentación
- Sistemas Interrogativos
- Sistemas de Simulación
- Sistemas funcionales
- Sistemas Expertos

4. Diferencias entre OLTP y OLAP

Mientras que las aplicaciones OLTP se caracterizan por estar actualizadas constantemente por varios usuarios a través de transacciones operacionales sobre datos individuales, las aplicaciones OLAP son utilizadas por personal de niveles ejecutivos que requieren datos con alto grado de agregación y desde distintas perspectivas (dimensiones), como ser: totales de venta por región, por producto, por período de tiempo,..., etc. (Ver figura 3).

OLTP	OLAP
Atomizado	Sumarizado
Datos Históricos	Datos Actuales
Un registro a la vez	Muchos registros a la vez
Orientado a la información operativa	Orientado a la información estratégica
Datos relacionales	Datos Multidimensionales
Consultas simples predefinidas	Consultas ad-hoc
Volumen de datos acotados	Grandes volúmenes de datos

5. Concepto de datos multidimensionales

En el análisis multidimensional, los datos se representan mediante dimensiones como producto, territorio y cliente. En general, las dimensiones se relacionan en jerarquías, por ejemplo, ciudad, estado, región, país y continente. El tiempo es también una dimensión estándar con sus propias jerarquías tales como: día, semana, mes, trimestre y año. (Figura 4).

No es común que, por ejemplo, alguien dentro de la organización se pregunte: "¿cuánto vendí?".

Figura 4 – Estructura multidimensional de los datos

En general, un Gerente de Ventas podría preguntarse: ¿Cuánto vendí del producto "A" en el períodos "X" en la región "Y"? (Figura 5).

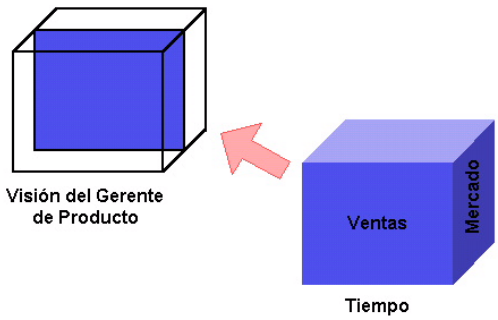

Figura 5 – Análisis de los datos desde el punto de viste del gerente de producto

En cambio, para un gerente de Finanzas la necesidad es diferente y su pregunta sería: ¿A cuánto ascendieron las ventas de todos los productos en todas las regiones al cierre del mes "M"? y para el caso de un gerente regional: ¿Cuánto fueron las ventas de todos los productos en el período J ó K en mi región? **Figura 6**.

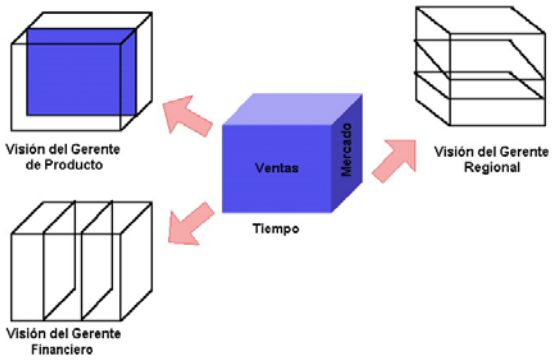

Figura 6 – Visión de los gerentes financiero y regional

6. Implementación de un Data Warehouse

La estructura adoptada para el almacén de datos se debe realizar de tal modo que satisfaga las necesidades de la empresa, dicha elección es clave en la efectividad del Data Warehouse. Existen tres formas básicas de estructura del almacén:

Data Warehouse central
La implementación consta de un solo nivel con un solo almacén que soporta los requerimientos de información de toda la empresa.

Data Warehouse distribuido
Es una estructura de un solo nivel que se particiona para distribuirlo a nivel departamental.

Data Warehouse de dos niveles
Es una combinación de los anteriores que soporta requerimientos de información tanto a nivel empresarial como departamental.

6.1. Costos del Data Warehouse

Uno de los puntos más importantes a tener en cuenta en el momento de decidir implementar un Data Warehouse es el costo que trae aparejado. A grandes rasgos los costos asociados a un proyecto Data Warehouse son el Costo de Construcción y el costo de Mantenimiento y Operación una vez construido.

6.1.1. Costo de Construcción

Es similar a al Costo de Construcción de cualquier sistema de Tecnología. Se pueden clasificar en tres tipos:

RECURSOS HUMANOS: Es necesario contar con conocimiento sobre el perfil y cualidades del personal ya que el desarrollo de esta tecnología requiere de la participación tanto del personal técnico como de los especialistas de negocios, estos dos grupos trabajarán juntos durante todo el desarrollo del Data Warehouse.

TIEMPO: Además de los tiempos de construcción y entrega del Data Warehouse, se debe tener en cuenta los tiempos de planificación del proyecto y de definición de la Arquitectura.

TECNOLOGÍA: El costo de la nueva tecnología introducida por el Data Warehouse se debe considerar solo como el costo inicial de la implementación.

6.1.2. Costo de Operación y Mantenimiento

Es necesario, una vez que se ha finalizado la construcción y se ha entregado el producto se debe dar soporte que es una fuente continua de costos. Los costos de operación se dividen en:

6.1.3. Costo de Evolución

Es necesario realizar ajustes continuos del Data Warehouse a través del tiempo, muchas veces estos cambios se deben al aprendizaje mediante el uso.

6.1.4. Costo de Crecimiento

Incrementos de volúmenes de datos, de cantidad de usuarios accediendo al Data Warehouse desembocará en un aumento en los recursos necesarios para que los tiempos de respuesta y recuperación de datos, principalmente, sigan siendo óptimos.

6.1.5. Costo producido por cambios

El Data Warehouse necesita soportar los cambios en el origen de datos que utiliza como así también soportar los cambios de la información que produce. Por ejemplo, si el cambio se produce en el ambiente empresarial, seguramente, cambiarán las necesidades de información de los usuarios serán necesarios, entonces, cambios en las Aplicaciones DSS y EIS. Si por el contrario cambio viene dado por el sector tecnológico y éste afecta el modo de almacenamiento de los datos, implicaría ajustes en los procesos de Extracción, Soporte y Carga para adaptarse a las variaciones.

6.2. Impactos de implementación del Data Warehouse

El éxito del Data Warehouse no está en la construcción sino en utilizarlo para mejorar los procesos empresariales, operacionales y de toma de decisiones, para que esto suceda se deben tener en cuenta los impactos producidos en los siguientes ámbitos:

6.2.1. Impacto en la gente

La construcción requiere de la participación activa de quienes utilizarán el Data Warehouse, depende tanto de la realidad de la empresa como de las condiciones que existan en ese momento, las cuales determinarán cual será su contenido.

El Data Warehouse provee los datos que posibilitará a los usuarios a acceder a su propia información en el momento que la necesitan. Esta posibilidad para entregar información presenta varias implicancias:

Los usuarios deberán adquirir nuevas destrezas.
Se eliminará los largos tiempos de análisis y programación para obtener información. Como la información estará lista para ser utilizada, probablemente, aumenten las expectativas. Pueden existir nuevas oportunidades en la comunidad empresarial para los especialistas de información. Se reducirá hasta casi eliminarse la gran cantidad de reportes en papel.

La madurez del Data Warehouse dependerá del uso activo y retroalimentación de sus usuarios.

6.2.2. Impactos en los procesos empresariales y de toma de decisiones

Mejora del proceso de toma de decisiones por medio de la disponibilidad de la información. Las decisiones se toman más rápidamente por gente más informada.

Los procesos empresariales pueden ser optimizados, se elimina el tiempo de espera de información que, generalmente, es incorrecta o no se encuentra.

Se reducen los costos de los procesos y muchas veces se aclaran sus conexiones y dependencias, aumentando así la eficiencia en dichos procesos.

El Data Warehouse permite que los datos de los sistemas operaciones sean utilizados y examinados, cuando estos datos se organizan para tener significado para la empresa la gente comienza a

aprender de los sistemas y pueden quedar expuestos posibles defectos de las aplicaciones actuales.

Aumenta la confianza de las decisiones tomadas en base a la información del Data Warehouse, debido a que tanto los responsables de la toma de decisiones como los afectados conocen que están basadas en información de buena calidad.

La información compartida conduce a un lenguaje común, conocimiento común y mejora de la comunicación en la empresa.

Teniendo en cuenta las etapas de construcción, soporte del Data Warehouse y soporte de los sistemas operacionales, algunos de los impactos técnicos son los siguientes.
En el momento de construcción de un Data Warehouse el impacto más grande sobre la gente técnica está dado por la curva de aprendizaje, algunas de las nuevas destrezas a adquirir son:

- Conceptos y estructura del Data Warehouse
- Nuevas de demandas de soporte técnico debido a la utilización de nuevas tecnologías, nuevas demandas de recursos.
- Es necesario adquirir destrezas de desarrollo incremental evolutivo.
- Trabajo en equipo con gente del área de negocios como participantes activos del desarrollo del proyecto.

Por último, podemos decir que un Proyecto de Data Warehouse se considera exitoso cuando la gente de la empresa lo utiliza para satisfacer sus necesidades operacionales y de negocio.

7. El Data Mining y su relación con el Data Warehouse

Las técnicas de Data Mining son el resultado de un largo proceso de investigación y desarrollo de productos orientados al almacenamiento, extracción análisis de datos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. Data Mining está soportado por las siguientes tecnologías:

- Soportes de almacenamiento masivo de datos
- Potentes computadoras con multiprocesadores
- Data Warehouse
- Algoritmos de Data Mining

Data Mining es la extracción de información oculta y predecible de grandes bases de datos.

Un sistema Data Mining es una tecnología de soporte para usuario final cuyo objetivo es extraer conocimiento útil y utilizable a partir de la información contenida en las bases de datos de las empresas.

Las herramientas de Data Mining sirven para predecir tendencias y comportamientos, de esta manera permiten a las organizaciones tomar decisiones proactivas para adaptarse rápidamente a los cambios del mercado obteniendo así ventajas

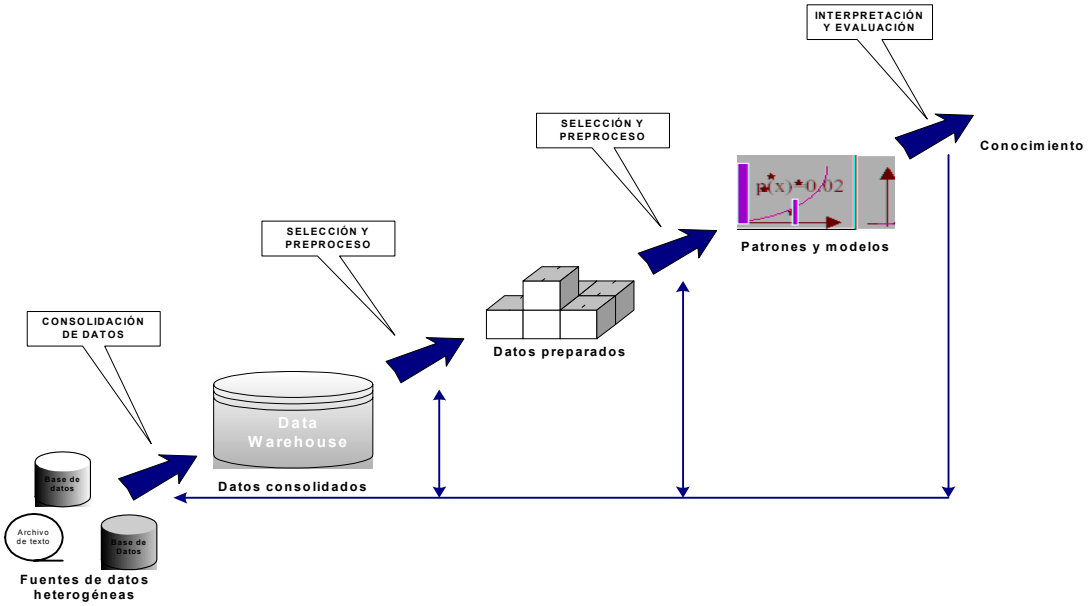

Figura 7 – Data Warehouse y la relación con el Data Mining

competitivas.

Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas por consultas en un sistema tradicional de soporte operacional. La potencialidad de estas herramientas reside en la capacidad de explorar las bases de datos en busca de patrones ocultos, encontrando información predecible que para un experto sería casi imposible debido al gran volumen de información.

heterogéneas Bases de Datos relacionales, ficheros planos y registros de transacciones en línea. (Figura 7).

El Data Warehouse dota a las organizaciones de memoria, y el Data Mining de inteligencia.

La mejor forma de aplicar las técnicas de Data Mining es que éstas se encuentren totalmente integradas con el Data Warehouse así como también con herramientas flexibles e interactivas para el análisis de negocios. Varias herramientas de Data Mining actualmente operan fuera del Data

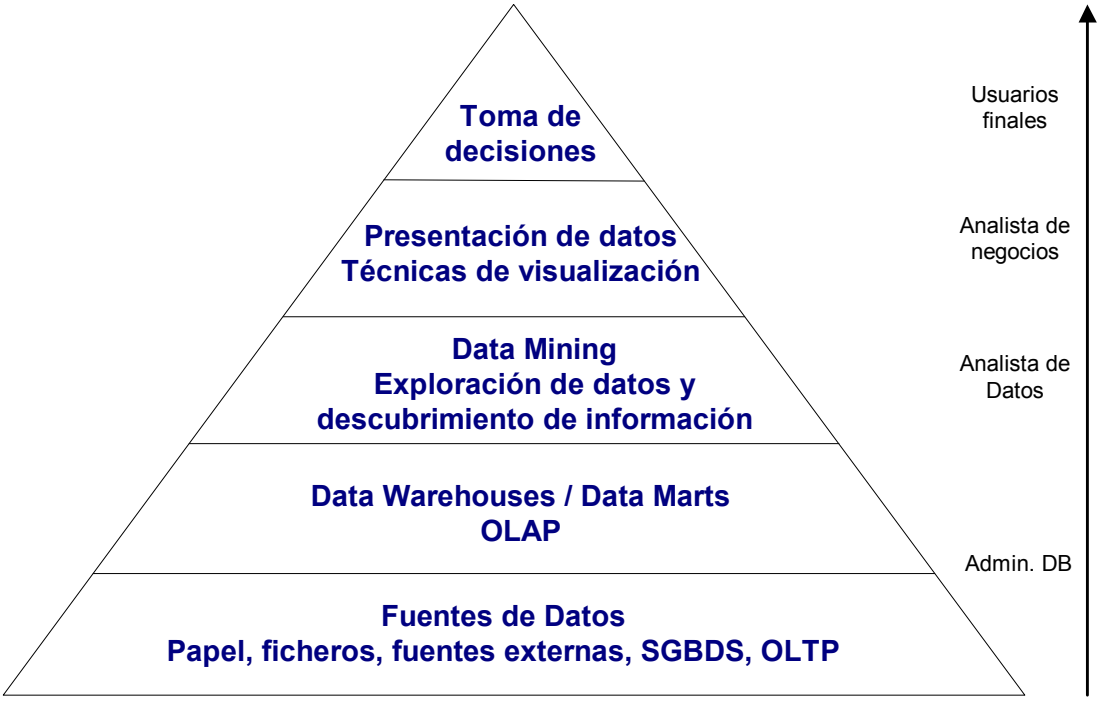

Figura 8 – Evolución desde los datos operacionales hasta la información para la toma de decisiones

Una vez que las herramientas de Data Mining fueron implementadas en computadoras cliente servidor de alto performance o de procesamiento paralelo, pueden analizar bases de datos masivas para brindar respuesta a preguntas tales como, "¿Cuáles clientes tienen más probabilidad de responder al próximo mailing promocional, y por qué? y presentar los resultados en formas de tablas, con gráficos, reportes, texto, hipertexto, etc.
El origen de la información que utilizan los algoritmos de Data Mining, por lo general, son datos históricos que se encuentran almacenados en un Data Warehouse. El partir de un Data Warehouse simplifica la etapa previa a la etapa de preparación de los datos ya que se construye en base a la integración de fuentes de datos múltiples y

Warehouse, requiriendo pasos extra para extraer, importar y analizar los datos. Además la integración con el Data Warehouse permite que ni bien los cambios originados en las bases de datos operacionales son replicados al Data Warehouse pueden ser analizados directamente y monitoreados mediante las técnicas de Data Mining.
El server de Data Mining debe estar integrado con el Data Warehouse y el server OLAP para insertar el análisis de negocios directamente en esta infraestructura. Un avanzado, metadata centrado en procesos define los objetivos del Data Mining para resultados específicos tales como manejos de campañas promocionales, optimización de promociones, etc. A medida que el Data Warehouse crece con nuevas decisiones y resultados, la organización puede aplicar Data Mining para

obtener las mejores prácticas y aplicarlas en futuras decisiones.

Este diseño representa una transferencia fundamental desde los sistemas de soporte de decisión convencionales. Más que simplemente proveer datos a los usuarios finales a través de software de consultas y reportes, el server de Data Mining aplica los modelos de negocios del usuario directamente al Data Warehouse y devuelve un análisis proactivo de la información más relevante. Estos resultados mejoran los metadatos en el server OLAP proveyendo un estrato de metadatos que representa una vista fraccionada de los datos. Generadores de reportes, visualizadores y otras herramientas de análisis pueden ser aplicadas para planificar futuras acciones y confirmar el impacto de esos planes. (Figura 8).

9. Referencias

1- Rubinstein Jacobo, 2000. The Data Warehouse. Cambridge Technology Partners.

2- David Friend, 1995. Introducción al procesamiento analítico on-line (OLAP). Chairman Pilot Software Inc.

3- Gabriel Buades, 1990. Data Warehouse.

4- Ernestina Mensalvas Ruiz, José María Peña Sanchez, 2000. Data Mining: Técnicas y herramientas. Universidad Politécnica de Madrid, departamento de Lenguajes y sistemas informáticos e Ingeniería del Software.

5- José Martín Arevalillo, 2000. Data Mining, una herramienta para la toma de decisions. U.N.E.D. Departamento de estadística e investigación operativa.

6- Areas de Investigación, Data Warehousing y Tecnología OLAP en
http://gplsi.dlsi.ua.es/gplsi/areas.htm

7- Javier Cantoral Justo, 2002. Data Mining Conceptos y Técnicas. Universidad de Alicante, Grupo de investigación de sistemas de información en la empresa.

8- Data Mining y Data Warehousing en www.kdnuggets.com

9- Data Warehouse Terminology, 2003. En http://www.credata.com/research/terminology.html

10- On Line Analitical Processing en http://altaplana.com/olap/