Universidad Tecnológica Nacional
Facultad Regional Buenos Aires

Gestión de Datos

Data Warehouse

Data Mining

Tecnologías OLAP

Ing. Enrique Reinosa
Octubre 2007

INTRODUCCION

¿Que es una Base de Datos?

La pregunta sobre que es exactamente una base de datos, podría abrir en si mismo muchos debates entre los profesionales de sistemas. Dependiendo de cuan experimentada es una persona y que es lo que se está tratando de realizar, se puede interpretar a una base de datos de varias maneras.

En un sentido amplio, una base de datos se puede considerar como la suma total de todos los datos que guarda una organización. Esta definición general podría incluir los datos almacenados en las bases de datos definidas tradicionalmente, los datos almacenados en diferentes archivos y cintas guardadas en toda la empresa, y podría aún incluir los datos tipeados o escritos a mano en formularios, memos y otras fuentes no relacionadas con computadoras.

Una *Base de Datos* es una colección de datos organizados lógicamente y administrados por un conjunto unificado de principios, procedimientos, y funcionalidades que ayudan a garantizar la aplicación e interpretación consistente de los datos a través de toda una organización.

Un *Producto de Base de Datos* es un producto de software para la computadora (un programa o colección de programas) que administra el almacenamiento y la recuperación de un conjunto de datos, los organiza lógicamente, y le provee al usuario ciertas funcionalidades para garantizar que los datos se organizarán lógicamente y se aplicarán consistentemente. Existen muchas formas de categorizar a los productos de bases de datos. Se los puede clasificar por la plataforma en la que corren, por las funciones que ejecutan, o por su arquitectura.

¿Cual es la diferencia entre un *Producto de Base de Datos* y un *Sistema de Administración de Bases de Datos?*

Otra manera de comercializar un producto es referirse a él como un Sistema de Administración de Base de Datos (DBMS). No existe una definición clara sobre cual es la diferencia entre uno y otro, pero en general, cuando a un producto se lo denomina Sistema de Administración de Bases de Datos, generalmente se lo ha construido sobre sus capacidades en vez de las simples funciones de recuperación y almacenamiento de datos. Estos productos comúnmente incluyen grandes capacidades de back-up y recuperación de datos, ingreso al sistema, control de concurrencia y bloqueo, y mecanismos de seguridad que los Productos de Base de Datos más livianos no poseen.

Arquitecturas

La arquitectura de una base de datos describe la organización lógica y operacional básica que una base de datos particular tomará en la ejecución de las tareas que se le requieren. Algunas de las principales arquitecturas incluyen ***archivo plano, Xbase, jerárquica, red, relacional y orientada a objetos***.
La arquitectura de las base de datos es la que manda como será percibida la base de datos por los usuarios y programadores. Las decisiones de diseño físico y lógico de bases de datos se toman basándose en la arquitectura.

¿Que es la Arquitectura de Base de Datos?

En su forma fundamental, la palabra arquitectura se refiere al estilo ó método de construir algo. Se usa el término arquitectura de base de datos con dos significados distintos.
Primero, de manera singular, la arquitectura de un producto de base de datos se refiere al modo específico en que se han puesto juntos y el método de uso y el estilo al que los usuarios se deben adherir.
Segundo, y más común, se usa el término arquitectura para referirse al agrupamiento de productos de bases de datos definidos en forma aislada en familias o en conjuntos relacionados de productos de bases de datos que comparten algunas características en común. Comparar las diferentes arquitecturas de bases de datos no es una tarea simple.
El término arquitectura de base de datos se refiere al método y al estilo de construir un sistema que un producto particular de base de datos soportará. Aunque existen cientos de productos de base de datos en el mercado hoy en día, hay menos de una docena de arquitecturas que se usan. La arquitectura dice algo sobre la manera en que un producto de base de datos realiza su objetivo primario, que es administrar los datos. También indica algo sobre el medio en el cual corre, la plataforma que usa, y el resto de las funciones que proveerá.

Bases de Datos Relacionales

El término *relacional* se ha usado para describir más tipos de productos que cualquier otro término en la historia de las bases de datos. Por lejos la contribución más sustancial que ha hecho el modelo relacional a la industria de base de datos ha sido la creación del SQL (structured query language – lenguaje estructurado de consulta). Este lenguaje se ha establecido en un tiempo relativamente corto como el Standard de los lenguajes de bases de datos. Al hacer esto, se les dio a los vendedores de bases de datos un Standard común para adherirse. Es interés del vendedor adherirse a un

lenguaje de acceso Standard debido que le garantiza que el perfil de programación es consistente con los de otros vendedores. Esto significa que uno puede comprar cualquier sistema para la administración de interfase de usuario o lenguaje de programación y hacerlo funcionar con distintas bases de datos.

La palabra *Relacional*, cuando se la aplica a los sistemas de bases de datos, tiene diversas definiciones;

> *Es una teoría de las Ciencias de la Computación/Matemática desarrollada por Dr. Codd. Esta visión submite una representación de datos y un esquema de almacenamiento que intenta utilizar algebra relacional avanzada y sus correspondientes propiedades lógicas y matemáticas como un medio óptimo de almacenamiento y acceso a los datos es un sistema de base de datos.*

Hoy en la práctica, la teoría relacional como la propone Dr. Codd se la aplica para formar lo que se puede identificar estrechamente como un intento relacional práctico para la administración de datos.
El modelo relacional de Codd direcciona muchas áreas y facetas de la administración de datos, siendo las más importantes:

- Estructura de Datos (Como deberían ser almacenados los datos)
- Integridad de Datos (Como deberían ser clasificados los datos)
- Manejo de Datos (Como deberían se accedido los datos)

Reglas de Estructuras de Datos

Las reglas de estructuras de datos definen una terminología y un conjunto de reglas para la creación de estructuras de datos. Los datos se ven en términos de elementos de datos individuales. Estos elementos son *atómicos*, es decir, no se pueden dividir en partes más pequeñas. Cada elemento de datos se dice que tiene un *dominio*, que es un conjunto de valores considerados válidos para ese elemento.
Cualquiera de los elementos de datos que comparten el mismo dominio pueden utilizar operadores de comparación matemáticos (tale como mayor que, menor que, igual, distinto, etc.), para generar la información.
Una *relación* es una colección de *atributos* (tipos de elementos de datos) que están relacionados de alguna manera. Por ejemplo, los atributos Nombre, Dirección, y Teléfono podrían formar una relación denominada Información del Cliente.
Una *tupla* es una simple ocurrencia de una relación.

Integridad de Datos

Teniendo definidas las reglas para la creación de estructuras de datos, el papel elaborado por el Dr. Codd continúa definiendo guías de referencia para determinar que campos deberían pertenecer a que tablas.

Para lograr la eficiencia relacional en el almacenamiento y en la relación de elementos de datos, el diseñador comienza identificando las tablas *base*. Las tablas base son representaciones almacenadas de relaciones del mundo real. Por ejemplo, una compañía podría tener tablas base para empleados, clientes y productos.
Estas tablas deberían estar todas clasificadas según las siguientes reglas de integridad:

1. Cada relación tiene una serie de claves candidatas. Una clave candidata es cualquier atributo o combinación de atributos que identifican las tuplas.

2. Cada relación tiene una y solo una clave candidata que sirve como su clave primaria. Una clave primaria es aquella clave candidata que identifica unívocamente la tupla y la distingue de todas las otras tuplas.

3. Cualquier atributo de una relación puede ser una clave foraña. Una clave foraña es un atributo de una relación que es principal de alguna otra tabla.

4. Cada relación puede tener muchas claves alternativas que son claves candidatas no calificadas como claves primarias pero que identifican tuplas de alguna manera.

Manejo de datos

Con la construcción y relacionamiento entre las estructuras de datos establecidos, el papel elaborado por el Dr. Codd continúa definiendo un conjunto de operaciones que pueden ser utilizadas con la base de datos para extraer la información requerida. Teniendo creadas tablas que representan las relaciones, los datos pueden ser accedidos usando los ocho operadores relacionales. Los operadores tradicionales, basados en la teoría de conjuntos algebraicos, incluyen la unión, intersección, diferencia, producto cartesiano, select, project, join y división. Estas operaciones luego son definidas usando un lenguaje de acceso a datos denominado SQL.
Esta colección de reglas y guías constituyen la base sobre la cual se ha construido toda la actual tecnología relacional.

Conclusión de Bases de Datos Relacionales

Hay varios aspectos de la arquitectura relacional y de las bases de datos que la utilizan que son de particular importancia. Primero y principal, es crítico que la persona que trata de entender como funciona una base de datos relacional comprenda el lenguaje SQL y el único estilo de navegación que provee. En segundo lugar, también es importante tener en mente la manera en que las bases de datos relacionales intentan simplificar el proceso de administrar grandes sistemas con el transcurso del tiempo con independencia de los datos. En tercer lugar, la performance de una base de datos relacional, como también del resto, se debería basar en todos los perfiles del producto (programación, arquitectura, almacenamiento de datos, y administración), no solo en la arquitectura misma.

Aunque la aproximación relacional a las arquitecturas de bases de datos representaron un cambio radical y significante en las formas que una organización hace uso de las bases de datos, estos cambios aún no son tan fácilmente comprendidos según las demandas reales de procesamiento de datos en el día a día.

Data Warehousing

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

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

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

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

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

El concepto DataMart es una extensión natural del DataWarehouse, y está enfocado a un departamento o área especifica, como por ejemplo los departamentos de Finanzas o Marketing. Permitiendo así un mejor control de la información que se está abarcando.

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

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

En esta nueva tecnología cada eslabón de la cadena de valor será representado por una base de datos multidimensional, la cual permite potencialmente administrar la etapa productiva que representa. La cadena de valor total será representada entonces por el conjunto de bases de datos multidimensionales asociadas a cada eslabón.

En primer lugar, DW no es un producto que pueda ser comprado en el mercado, sino más bien un concepto que debe ser construido. DW es una combinación de conceptos y tecnología que cambian significativamente la manera en que es entregada la información a la gente de negocios. El objetivo principal es satisfacer los requerimientos de información internos de la empresa para una mejor gestión, con eficiencia y facilidad de acceso.
La manera tradicional hasta ahora de entregar la información es a través de emisión de reportes impresos desde los sistemas operacionales, con consultas a nivel de cliente y extracción ocasional de datos para suplir actividades basadas en papel. Los problemas con la entrega de la información actual son muchos, incluyendo inconsistencia, inflexibilidad y carencia de integración a través de la empresa.
El DW puede verse como una bodega donde están almacenados todos los datos necesarios para realizar las funciones de gestión de la empresa, de manera que puedan utilizarse fácilmente según se necesiten. El contenido de los datos, la organización y estructura son dirigidos a satisfacer las necesidades de información de analistas.
Los sistemas transaccionales son dinámicos, en el sentido que constantemente se encuentran actualizando datos.
Analizar esta información puede presentar resultados distintos en cuestión de minutos, por lo que se deben extraer y almacenar fotografías de datos (snapshots), para estos efectos, con la implicancia de un consumo adicional de recursos de cómputo. Llevar a cabo un análisis complejo sobre un sistema transaccional, puede resultar en la degradación del sistema, con el consiguiente impacto en la operación del negocio.
El datawarehouse intenta responder a la compleja necesidad de obtención de información útil sin el sacrificio del rendimiento de las aplicaciones operacionales, debido a lo cual se ha convertido actualmente en una de las tendencias tecnológicas más significativas en la administración de información.
Los almacenes de datos (o Datawarehouse) generan bases de datos tangibles con una perspectiva histórica, utilizando datos de múltiples fuentes que se fusionan en forma congruente. Estos datos se mantienen actualizados, pero no cambian al ritmo de los sistemas transaccionales. Muchos datawarehouses se diseñan para contener un nivel de detalle hasta el nivel de transacción, con la intención de hacer disponible todo tipo de datos y características, para reportar y analizar. Así un datawarehouse resulta ser un recipiente de datos transaccionales para proporcionar consultas operativas, y la información para poder llevar a cabo análisis multidimensional. De esta forma, dentro de un almacén de datos existen dos tecnologías complementarias, una relacional para consultas y una multidimensional para análisis.
Existen muchas definiciones para el DW, la más conocida fue propuesta por Inmon (considerado el padre de las Bases de Datos) en 1992: “Un DW es una colección de datos orientados a temas, integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales”. En 1993, Susan Osterfeldt [MicroSt96] publica una definición que sin duda acierta en la clave del DW: “Yo considero al DW como algo que provee dos beneficios empresariales reales: Integración y Acceso de datos. DW elimina una gran cantidad de datos inútiles y no deseados, como también el procesamiento desde el ambiente operacional clásico”.

Esta última definición refleja claramente el principal beneficio que el datawarehouse aporta a la empresa, eliminar aquellos datos que obstaculizan la labor de análisis de información y entregar la información que se requiere en la forma más apropiada, facilitando así el proceso de gestión.

1. DATA WAREHOUSE

1.1. Sistemas de Data Warehousing.

Los sistemas de Data Warehousing son el centro de la arquitectura de los Sistemas de Información de los 90's. Han surgido como respuesta a la problemática de *extraer información sintética a partir de datos atómicos almacenados en bases de datos de producción.* Uno de los objetivos principales de este tipo de sistemas es servir como base de información para la toma de decisiones. Los beneficios obtenidos por la utilización de este tipo de sistemas se basan en el acceso interactivo e inmediato a información estratégica de un área de negocios. Este acercamiento de la información al usuario final permite una toma de decisiones rápida y basada en datos objetivos obtenidos a partir de las bases de datos (eventualmente heterogéneas) de la empresa. Estos beneficios aumentan cuanto más importantes son las decisiones a tomar y cuanto más crítico es el factor tiempo. [1]

Un *Sistemas de Data Warehousing* incluye funcionalidades tales como:

1. *Integración de bases de datos* heterogéneas (relacionales, documentales, geográficas, archivos, etc.).
2. *Ejecución de consultas complejas no predefinidas* visualizando el resultado en forma de gráfica y en diferentes niveles de agrupamiento y totalización de datos.
3. *Agrupamiento y desagrupamiento* de datos en forma interactiva.
4. *Análisis de problema en términos de dimensiones*. Por ejemplo, permite analizar datos históricos a través de una dimensión tiempo.
5. *Control de calidad de datos* para asegurar, no solo la consistencia de la base, sino también la relevancia de los datos en base a los cuales se toman las decisiones.

1.2. ¿Qué es un Data Warehouse?

Un Data Warehouse es una colección de datos

- orientada a sujetos
- integrada
- variante en el tiempo
- no volátil

Que soporta el proceso de toma de decisiones.
Un Data Warehouse soporta procesamiento informático, brindando una sólida plataforma de datos históricos, integrados, de los cuales hacer análisis.

Es orientado a sujetos:

Un primer aspecto de un data warehousing es que esta orientado a los mayores sujetos de la empresa. El mundo operacional esta diseñado alrededor de aplicaciones y funciones, como por ejemplo pagos, ventas, entregas de mercadería, para una institución comercial. Un data warehouse esta organizado alrededor de los mayores sujetos, como cliente, vendedor, producto y actividades El mundo operacional concierne al diseño de la base de datos y al diseño de procesos. Un data warehousing está enfocado en la modelización de los datos y el diseño de la base de datos, exclusivamente. El diseño de procesos (en su forma clásica) no es parte del data warehouse.

Los datos son integrados:

El aspecto más importante del ambiente de un data warehouse es que sus datos están integrados. Cuando los datos son movidos del ambiente operacional, son integrados antes de entrar en el warehouse. Por ejemplo, un diseñador puede representar el sexo como "M" y "F", otro puede representarlo como "0" y "1", o "x" e "y", y otro usar las palabras completas "masculino" y "femenino". No importa la fuente de la cual el sexo llegue al data warehouse, debe ser guardado en forma consistente; los datos deben ser integrados.

Es variante en el tiempo

Los datos en el warehouse son precisos para un cierto momento, no necesariamente ahora; por eso se dice que los datos en el warehouse son variantes en el tiempo. La varianza en el tiempo de los datos de un warehouse se manifiesta de muchas maneras. El data warehouse contiene datos de un largo horizonte de tiempo. Las aplicaciones operacionales, sin embargo, contienen datos de intervalos de tiempo pequeños, por cuestiones de performance (tamaño chico de las tablas). Toda estructura clave en un warehouse contiene implícita o explícitamente un elemento del tiempo. Esto no necesariamente pasa en el ambiente operacional. Los datos de un warehouse, una vez almacenados, no pueden ser modificados (no se permiten updates). En el ambiente operacional, los datos, precisos al momento de acceso, pueden ser actualizados, según sea necesario.

Es simple de manejar

Updates, inserts y deletes son efectuados regularmente, en una base de record-por-record, a los datos operacionales. La manipulación de datos en un warehouse, es mucho más sencilla. Solo ocurren dos operaciones, la carga inicial, y el acceso a los datos. No hay necesidad de updates (en su sentido general). Hay consecuencias muy importantes de esta diferencia de procesos

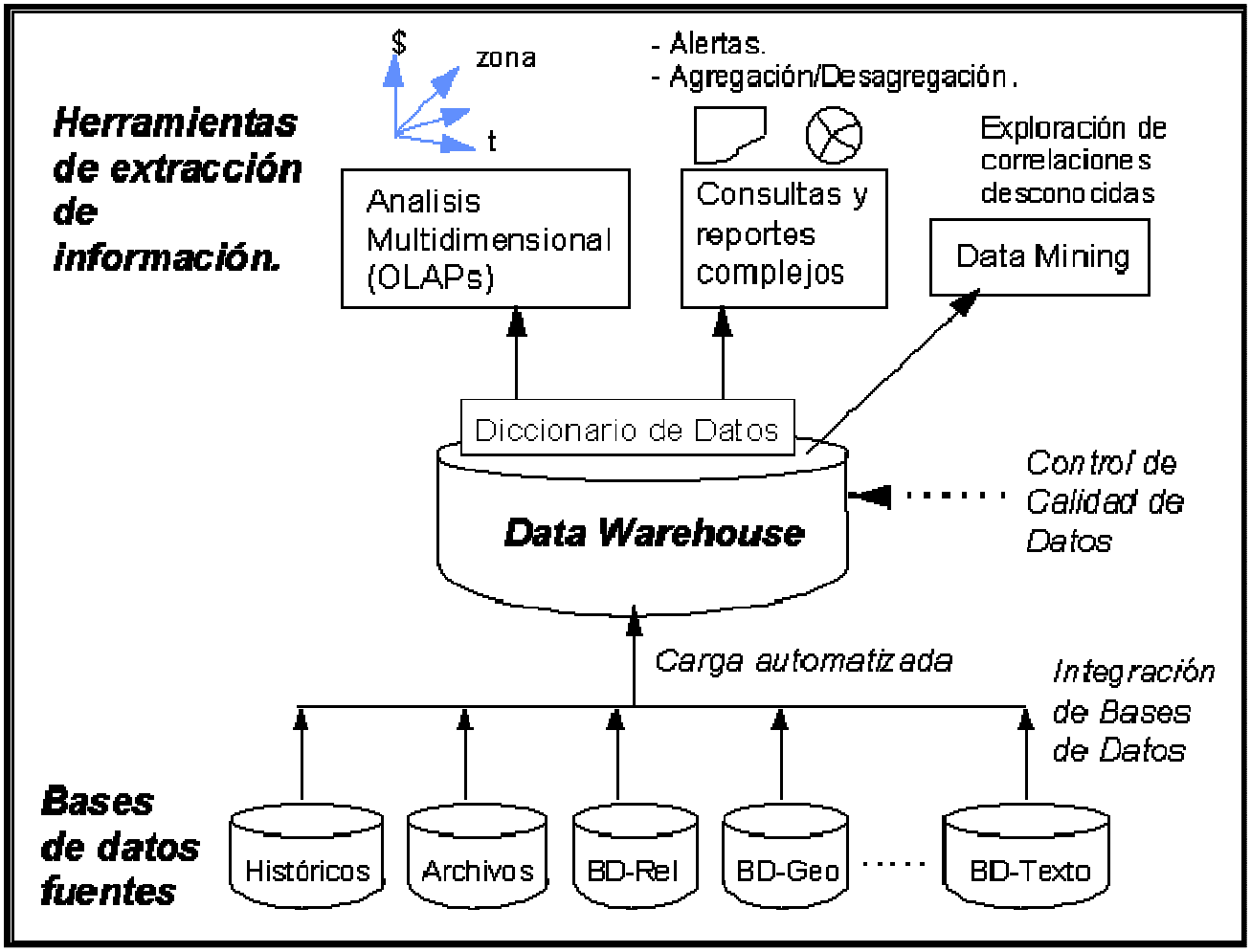
con un sistema operacional: A nivel de diseño, en un warehouse, no hay que controlar anomalías producidas por los updates, ya que no hay updates. Se pueden tomar libertades de diseño físico como optimizar el acceso a los datos, y de normalización física. Otra consecuencia es la simplicidad de la tecnología del warehouse, en lo que respecta a backups, recuperación, locks, integridad, etc.

1.3. La arquitectura del Data Warehouse

La arquitectura lógica de un sistema de Data Warehousing es del tipo mostrado en la Figura 1. Un Sistema de Data Warehousing consta de tres niveles: (1) bases de datos fuentes (de producción e históricos), (2) una base de datos con datos resumidos extraídos de las bases de producción (el Data Warehouse), y (3) interfaces orientadas a usuarios que extraen información para la toma de decisiones. Las clásicas son: Análisis Multidimensional, consultas y reportes y Data Mining.

Las bases de datos fuentes consisten en bases de datos de producción así como en históricos de dichas bases. Estas bases de datos pueden estar implementadas en diferentes tipos de sistemas: BD-Relacionales, BD-geográficas, BD-textos, archivos, etc. Una característica común es que almacenan ítems de datos atómicos, los cuales son relevantes como datos de producción, pero pueden ser demasiado finos como base para la toma de decisiones. Además, la noción de calidad de los datos en estas bases se basa en la consistencia de dichos registros, independientemente de la relevancia que estos tengan dentro del problema.

El Data Warehouse es una base de datos que incluye los datos relevantes para la toma de decisiones en un área de negocios o globalmente en la empresa. Los datos almacenados en el Data Warehouse son, fundamentalmente, agrupamientos y totalizaciones de los datos relevantes que se encuentran en las bases de producción y en los históricos. Una componente importante en el Data Warehouse es el **Diccionario de Datos** (o Meta-Data), el cual describe los datos almacenados con el objetivo de facilitar el acceso a los mismos a través de las herramientas de explotación del Data Warehouse. El Diccionario de Datos establece correspondencias entre los datos almacenados y los conceptos que estos representan de forma de facilitar la extracción de información por parte del usuario final.



Arquitectura lógica de un Sistema de Data Warehousing

El objetivo de un ambiente de Data Warehousing es principalmente convertir los datos de aplicaciones del ambiente transaccional (OLTP), en datos integrados de gran calidad. Luego se los debe almacenar en una estructura que optimice el acceso por parte de usuarios finales en un ambiente decisional (OLAP). Durante este proceso, datos totalizados son agregados al Warehouse. Los datos son transferidos desde el ambiente operacional al Warehouse, en una base periódica, apropiada al tipo de análisis de negocios necesario.
Podemos dividir las funcionalidades del Data Warehouse en cinco grandes grupos, cada uno de las cuales es responsable de un conjunto de procesos específicos, esenciales para el ambiente de soporte decisional:

- Acceso a Fuentes (Source)
- Carga (Load)
- Almacenamiento (Storage)
- Consultas (Query)
- Meta Datos (Meta Data)

Las funcionalidades de acceso a fuentes, carga y almacenamiento soportan la migración de los datos operacionales al Warehouse. La funcionalidad de consultas maneja los procesos que soportan el acceso y análisis de los datos para toma de decisiones. La funcionalidad de meta datos sirve como base para las otras cuatro, ya que provee los datos que controlan sus procesos e interacciones.

Acceso a Fuentes

La funcionalidad de acceso a fuentes, incluye los procesos que se aplican en las bases de datos fuentes, a los datos que serán transferidos. Las bases de datos fuentes son típicamente las bases de datos operacionales de la organización; sin embargo, se están integrando cada vez más, a bases de distribución pública sobre industria, demografía y clientes potenciales. Los datos pueden provenir de fuentes muy diversas. Determinar la mejor fuente de datos, evitando redundancias, es una de las tareas más largas y difíciles.
Muchos de los procesos asociados con la función de acceso a fuentes, como **mapeo, integración, análisis y calidad de los datos**, ocurren durante la fase de análisis y diseño del Data Warehouse. En realidad entre un 75 y 80% del tiempo de desarrollo del Warehouse está destinado a estas actividades.
Desafortunadamente, automatizar estas tareas, no es nada fácil. Algunas herramientas pueden ayudar a detectar problemas en la calidad de los datos, y generar programas de extracción; pero la mayor parte de la información requerida para el desarrollo, está en la mente de los analistas que trabajan con las bases de datos fuentes.

Los factores que impactan directamente sobre el tiempo destinado a estas actividades son: el número de aplicativos fuentes que serán mapeados al Data Warehouse, la calidad de los meta datos mantenidos en esas aplicaciones, y las reglas de empresa que las gobiernan.

Carga

La funcionalidad de carga comprende los procesos asociados con la migración de los datos desde los aplicativos fuentes a las bases del Warehouse. Incluyen **extracción, depuración, conversión y carga de datos**.
La **extracción** involucra acceder a los datos de los aplicativos. Es el primer paso la preparación de los datos. Hay varias alternativas de extracción que balancean la performance y las restricciones de tiempo y almacenamiento. Si las aplicaciones fuentes mantienen una base de datos en línea, se puede hacer una consulta que cree directamente los archivos de extracción. Hay que asegurarse que no se actualicen los datos mientras se hace la extracción para no generar inconsistencias. La performance puede caer si las transacciones en línea compiten con la extracción. Una solución alternativa es crear una vista, desde la cual extraer los datos. El inconveniente aquí, es el espacio de disco adicional para guardar esa copia de la base. El tiempo es un factor crucial; muchos aplicativos de extracción tienen un ciclo batch, en el cual transacciones fuera de línea son aplicadas a la base de datos.
Luego de la extracción, los datos son accedidos para determinar si hay problemas de calidad. La **depuración** de los datos puede ser manejada de muchas maneras. Si los errores son inherentes a los aplicativos fuentes, los datos pueden ser limpiados sistemáticamente como parte del proceso de conversión. Desafortunadamente, muchos errores ocurren porque los aplicativos fuentes sólo tienen una mínima validación de dominio, que permite la aparición de datos inválidos. La única manera de solucionarlos es corriendo rutinas pesadas de validación a nivel de fuentes. Los errores que surgen de tipeos incorrectos, son muy difíciles de detectar y corregir.
El paso final en la preparación de los datos para ser cargados en el Warehouse, es la **conversión**. Este proceso invoca reglas de conversión, de valores de aplicativos locales, a valores globales, integrados.
Cuando este proceso es completado, se **cargan** los datos al Data Warehouse.

Almacenamiento

La funcionalidad de almacenamiento comprende la arquitectura necesaria para integrar las vistas varias, al Data Warehouse. Aunque a menudo hablamos del Warehouse como si fuera un único almacén de datos, sus datos pueden estar distribuidos en múltiples bases manejadas por diferentes DBMSs. Dos tipos de manejadores se ajustan bien a esta tarea: relacionales (RDBMSs) y multidimensionales (MDDBMSs).
Un MDDBMS organiza los datos en un array de n dimensiones. Cada dimensión representa algún aspecto de los negocios a ser analizado. Las

bases multidimensionales presentan los datos de manera que los usuarios puedan entenderlos y accederlos fácilmente.
Cada área de la empresa puede necesitar que su propia visión de los negocios sea organizada como un array multidimensional, de manera de optimizar sus requerimientos específicos. Generalmente no es deseable que la misma base multidimensional soporte los requerimientos de todas las áreas de la empresa. Una RDBMS usualmente se ajusta más, al manejo de la base integrada.

Mientras que las vistas multidimensionales son diseñadas para optimizar el acceso de usuarios finales de cada área, la base de datos integrada del Warehouse es diseñada para optimizar el acceso de todas las áreas.

Se le llama Data Warehouse a la base integrada, y Data Marts a las vistas multidimensionales de cada área.
La separación entre el Data Warehouse corporativo y sus Data Marts satélites, introduce la necesidad de una estrategia que coordine la distribución de los datos hacia los Data Marts. Se debe considerar la incorporación de un servidor de replicación, que entregue los datos correctos, al Data Mart correcto en el momento correcto.
Los datos son almacenados en varios niveles. Los más actuales se guardan en un medio de fácil acceso en línea. Datos más viejos se pueden guardar en un medio seguro, pero más barato. Y los datos históricos pueden ser guardados en otros medios, o eliminados si ya no tienen más valor decisional.

Consultas

El ambiente de consultas permite a los usuarios conducir el análisis y producir reportes a través de sus herramientas OLAPs multidimensionales. Nuevas tecnologías prometen soportar la nueva generación de herramientas de análisis: **data mining y simulación de negocios**.
Las herramientas de **data mining** analizan los datos para identificar correlaciones inesperadas entre ellos.
Uno de los principales propósitos de estas tecnologías es chequear la efectividad de las reglas de empresa. Las herramientas de **simulación de negocios** crean modelos para testear el impacto de cambios en el ambiente de negocios. Se pueden establecer nuevas reglas de empresa. Luego hay que realimentar los aplicativos operacionales.
El arquitecto del Data Warehouse, debe determinar como totalizar los datos. Existen varios enfoques viables: la sumarización puede ser hecha durante la carga, y almacenada en el Data Warehouse; durante la replicación a los Data Marts; o a demanda, por las herramientas de consulta y simulación.

Meta Datos

El conocimiento de los meta datos es tan esencial como el conocimiento de los datos del Data Warehouse.

Deben incluir dominio, reglas de validación, derivación y conversión de los datos extraídos. También describen las bases de datos del Warehouse,

incluyendo reglas de distribución y control de la migración hacia los Data Marts. Los procesos que monitorean los procesos del Warehouse (como extracción, carga, y uso) crean meta datos que son usados para determinar que tan bien se comporta el sistema.
Los meta datos, deberían estar disponibles para los usuarios, para ser usados en sus análisis. Los administradores pueden manejar y proveer el acceso a través de los servicios del repositorio.

Las cinco funcionalidades del Warehouse proveen un marco de trabajo para controlar la arquitectura de los componentes. Este marco, describe las conversiones de los datos desde un ambiente OLTP, a un ambiente OLAP. [2]

1.4. Data Warehousing vs. Data Mart

Por muchos años, los sistemas que extraían y almacenaban datos de diversas fuentes para que ayudaran a la toma de decisiones, se llamaron *Data Warehouses*. En fechas recientes se ha hecho una distinción entre los grandes sistemas para almacenar datos (*data warehouses*) y los sistemas más pequeños (*data marts*), aún cuando el concepto general sigue nombrándose Data Warehousing.

Información proveniente de la industria indica que aproximadamente el 75% del data warehouse actual es, de hecho, data mart. En la Conferencia Mundial de Data Warehousing de META Group/DCI 1997 en febrero de 1997, se observó que **"el objetivo de las empresas se ha desplazado de la justificación del costo del *data warehousing* a la aplicación interna de la emisión de *data marts*."**

Los *data marts* se ajustan mejor a las necesidades que tiene una parte específica de un negocio, más que a las de toda una empresa. Optimizan la distribución de información útil para la toma de decisiones y se enfocan al manejo de datos resumidos o de muestras, más que a la historia presentada con detalle. De igual forma, no necesitan ser administrados centralmente por el departamento de sistemas de una organización, sino que pueden estar a cargo de un grupo específico dentro del área de la empresa que los utilice.

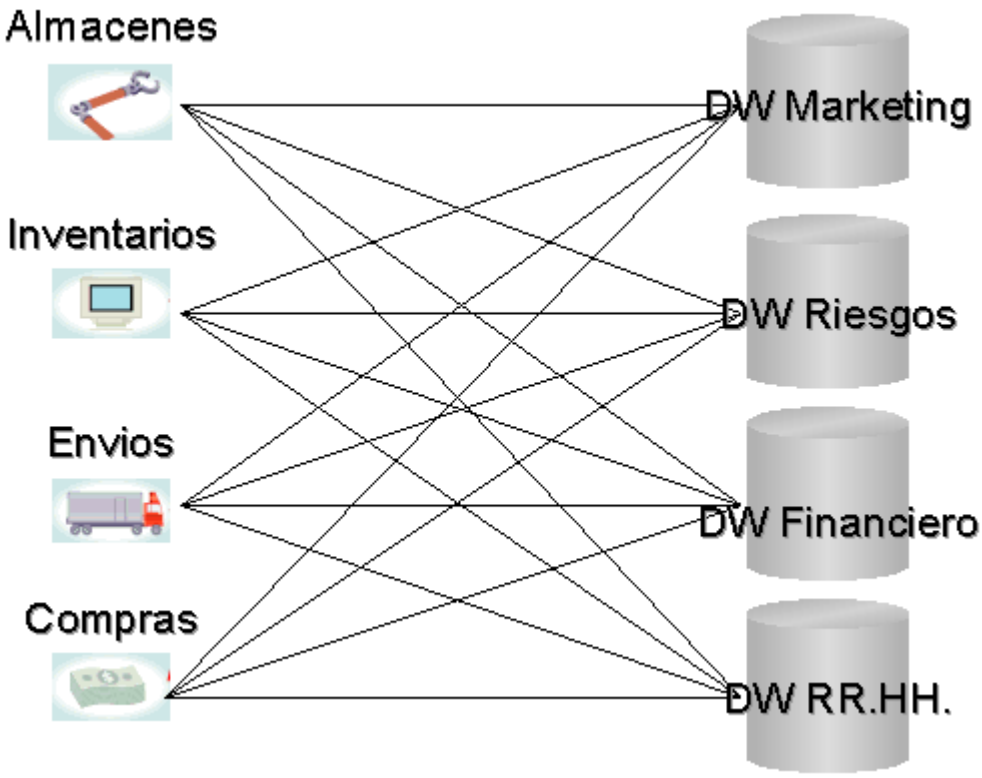
La creciente popularidad de los data marts se basa en varias buenas razones. Por un lado, **disminuyen significativamente el costo de creación y de operación, lo cual los pone al alcance de muchas compañías. Con los data marts se puede llegar a prototipos más rápidamente y obtener sistemas completamente desarrollados e implementados dentro de tres a seis meses.** El problema es que tienen un alcance más limitado que el *data warehousing*, ya que se enfocan a un conjunto concreto de necesidades, por lo mismo, son ideales para trabajar con objetivos y equipos de trabajo precisos.

A menudo, las pequeñas empresas y los departamentos autónomos de una organización prefieren utilizar éstos para construir su propio mecanismo para toma de decisiones. M**uchos departamentos de sistemas aprovechan la eficacia de esta aproximación y actualmente construyen un *data warehouses* para un solo tema o un data mart cada vez que se necesita, lo que les permite ganar experiencia y el apoyo de los administradores, los cuales ven los beneficios constantemente.** El hecho de comenzar con un plan modesto e ir creciendo conforme se aprende más sobre la fuente de los datos y sobre las necesidades finales del usuario, permite que las organizaciones justifiquen el uso de los *data marts* conforme estos avanzan.

En ocasiones, los proyectos que comienzan como *data warehouses* evolucionan a *data marts*. Cuando las organizaciones acumulan grandes cantidades de datos históricos para el apoyo de decisiones, que rara vez o nunca usan, pueden reducir la información guardada y convertir su *data warehouse* en un *data mart* mejor enfocado. O bien, pueden dividir el data Warehouse en diferentes data marts que ofrecen un tiempo de respuesta más rápido, un acceso más fácil y una menor complejidad para los usuarios finales.

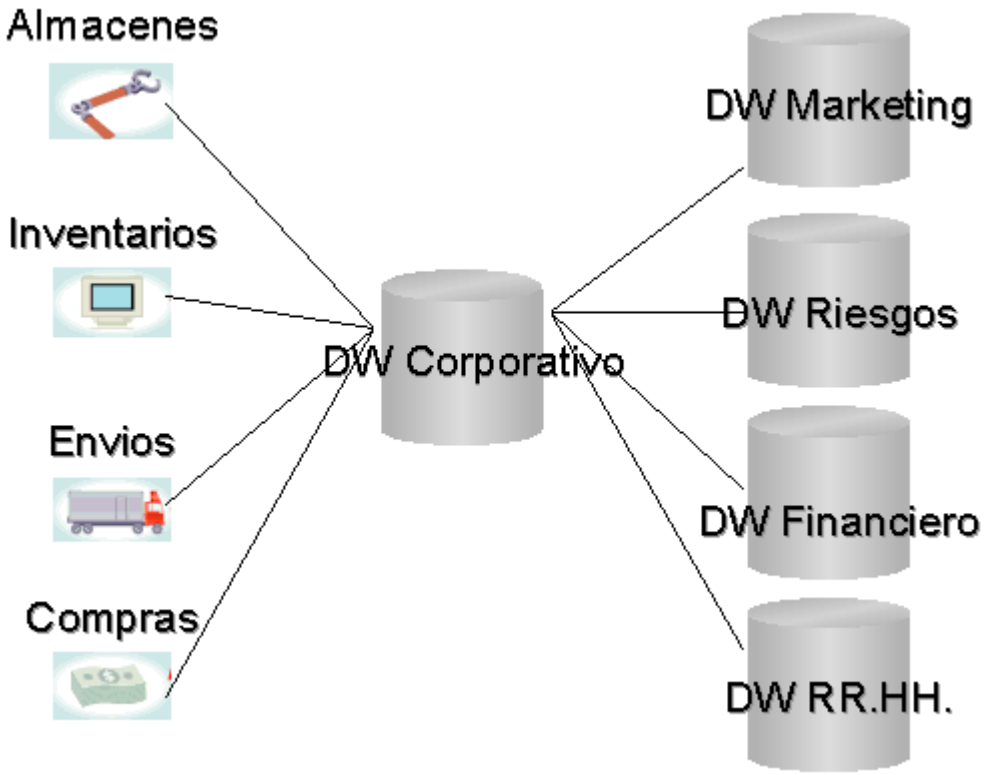
La duplicación en otro entorno de datos es un término que suele ser mal interpretado e incomprendido. Así es usado por los fabricantes de SGBD en el sentido de simple réplica de los datos de un sistema operacional centralizado en sistemas distribuidos. En un contexto de Data Warehouse, el término duplicación se refiere a la creación de Data Marts locales o departamentales basados en subconjuntos de la información contenida en el Data Warehouse central o maestro.
Según define Meta Group, "un Data Mart es una aplicación de Data Warehouse, construida rápidamente para soportar una línea de negocio simple". Los Data Marts, tienen las mismas características de integración, no volatilidad, orientación temática y no volatilidad que el Data Warehouse. Representan una estrategia de "divide y vencerás" para ámbitos muy genéricos de un Data Warehouse.

Esta estrategia es particularmente apropiada cuando el Data Warehouse central crece muy rápidamente y los distintos departamentos requieren sólo una pequeña porción de los datos contenidos en él. La creación de estos Data Marts requiere algo más que una simple réplica de los datos: se necesitarán tanto la segmentación como algunos métodos adicionales de consolidación.
La primera aproximación a una arquitectura descentralizada de Data Mart, podría ser venir originada de una situación como la descrita a continuación.

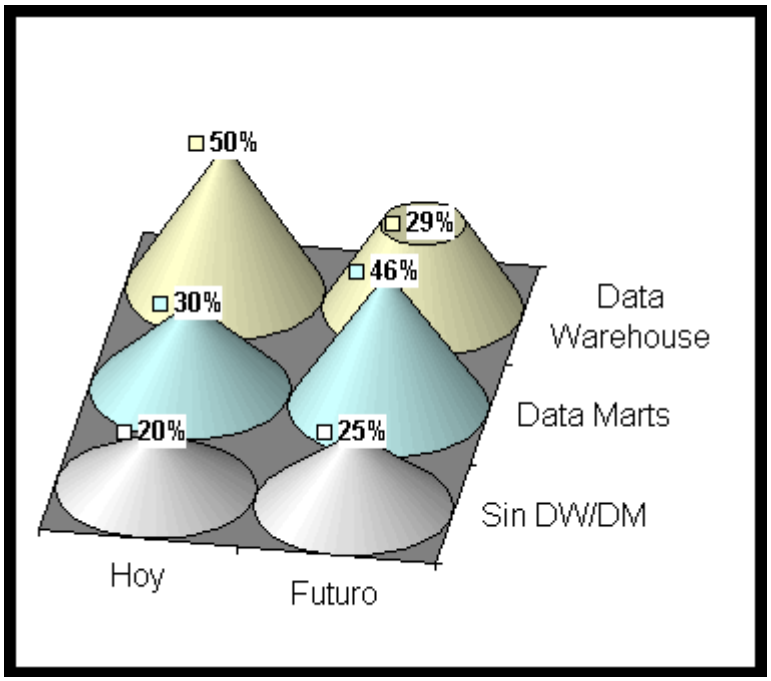


El departamento de Marketing, emprende el primer proyecto de Data Warehouse como una solución departamental, creando el primer Data Mart de la empresa.
Visto el éxito del proyecto, otros departamentos, como el de Riesgos, o el Financiero se lanzan a crear sus Data Marts. Marketing, comienza a usar otros datos que también usan los Data Marts de Riesgos y Financiero, y estos hacen lo propio.

Esto parece ser una decisión normal, puesto que las necesidades de información de todos los Data Marts crecen conforme el tiempo avanza. Cuando esta situación evoluciona, el esquema general de integración entre los Data Marts pasa a ser, la del gráfico de la derecha.
En esta situación, es fácil observar cómo este esquema de integración de información de los Data Marts, pasa a convertirse en un rompecabezas en el que la gestión se ha complicado hasta convertir esta ansia de información en un auténtico quebradero de cabeza. No obstante, lo que ha fallado no es la integración de Data Marts, sino su forma de integración.



En efecto, un enfoque más adecuado sería la coordinación de la gestión de información de todos los Data Marts en un Data Warehouse centralizado.
En esta situación los Data Marts obtendrían la información necesaria, ya previamente cargada y depurada en el Data Warehouse corporativo, simplificando el crecimiento de una base de conocimientos a nivel de toda la empresa.
Esta simplificación provendría de la centralización de las labores de gestión de los Data Marts, en el Data Warehouse corporativo, generando economías de escala en la gestión de los Data Marts implicados.
Según un estudio de IDC (*International Data Corporation*) tras analizar 541 empresas, la distribución de las implantaciones de Data Warehouse y Data Marts en la actualidad, y sus opiniones respecto a esta distribución en el futuro, nos muestra los siguientes datos:
En la gráfica, observamos, cómo en la actualidad, de las empresas consultadas, un 80% de ellas cuentan con implantaciones de Data Warehouse o Data Marts.



La proporción actual de implantaciones de Data Warehouse es casi el doble que el de Data Mart.
No obstante, seguramente tras la andadura inicial de alguno de estos proyectos de Data Mart, se ve como más adecuado para el futuro este enfoque "divide y vencerás", previéndose una inversión de estos papeles y duplicando la implantación de Data Marts a los Data Warehouse.
Probablemente, el 5% de usuarios que disponen de tecnología de Data Warehouse y piensan renunciar a ella en el futuro, no han realizado previamente un estudio de factores implicados en un Data Warehouse, o han pasado por la situación inicial de partida, y no se han planteado una reorganización del mismo.

1.5. Costos v/s Valor del DW

En todo proyecto es importante e inevitable realizar un análisis desde la perspectiva Costo/Valor.
A grandes rasgos, los costos asociados a un proyecto DW incluyen el costo de construcción y, la manutención y operación una vez que está construido. En cuanto al valor, éste considera, el valor de mejorar la entrega de información, el valor de mejorar el proceso de toma de decisiones y el valor agregado para los procesos empresariales.

1.5.1. Costos De Un DW

Costos De Construcción

Los costos de construir un DW son similares para cualquier proyecto de tecnología de información. Estos pueden ser clasificados en tres categorías:
RRHH: la gente necesita contar con un enfoque fuerte sobre el conocimiento del área de la empresa y de los procesos empresariales. Además es muy importante considerar las cualidades de la gente, ya que el desarrollo del DW requiere participación de la gente de negocios como de los especialistas tecnológicos; estos dos grupos de gente deben trabajar juntos, compartiendo su conocimiento y destrezas en un espíritu de equipo de trabajo, para enfrentar los desafíos de desarrollo del DW.

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

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

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

Evolutivos: ajustes continuos del DW a través del tiempo, como cambios de expectativas y, cambios producto del aprendizaje del RRHH del proyecto mediante su experiencia usando el DW.

Crecimiento: Incrementos en el tiempo en volúmenes de datos, del número de usuarios del DW, lo cual conllevará a un incremento de los recursos necesarios

como a la demanda de monitoreo, administración y sintonización del DW (evitando así, un incremento en los tiempos de respuesta y de recuperación de datos, principalmente).

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

Los dos primeros tipos de costos de operación, son básicos en la manutención de cualquier sistema de información, por lo cual no nos resultan ajenos; sin embargo, se debe tener especial cuidado con los costos de operación por cambios, ya que ellos consideran el impacto producto de la relación del OLTP y del Ambiente Empresarial, con el DW.
Resulta esencial para llevar a cabo un proyecto DW, tener claridad en la forma que éste se ve afectado por medio de cambios a nivel de OLTP como del Ambiente Empresarial; por ello entonces, a continuación se analiza más en detalle este tipo de costos de operación.

1.5.2. Cambios y el DW.

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

- Cambios en el ambiente empresarial: Un cambio en el ambiente empresarial puede cambiar las necesidades de información de los usuarios. Así, el contenido del DW se puede ver afectado y las aplicaciones DSS y EIS pueden requerir cambios.
- Cambios en la tecnología: Un cambio en la tecnología puede afectar la manera que los datos operacionales son almacenados, lo cual implicaría un ajuste en los procesos de Extracción, Transporte y Carga para adaptar las variaciones presentadas.

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

1.5.3. Valor Del DW

El valor de un DW queda descrito en tres dimensiones:

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

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

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

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

1.5.4. Balance de Costos v/s Valor.

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

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

1.5.5. Impactos DW

El éxito de DW no está en su construcción, sino en usarlo para mejorar procesos empresariales, operaciones y decisiones. Posicionar un DW para que

sea usado efectivamente, requiere entender los impactos de implementación en los siguientes ámbitos:

Impactos Humanos.

Efectos sobre la gente de la empresa:

- Construcción del 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 accesar su propia información cuando ellos la necesitan. Esta aproximación para entregar información tiene varias implicancias :

    a) La gente de la empresa puede necesitar aprender nuevas destrezas.

    b) Análisis extensos y demoras de programación para obtener información será eliminada. Como la información estará lista para ser accesada, las expectativas probablemente aumentarán.

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

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

    e) 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:

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

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

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

d) Procesos y datos de los sistemas operacionales, así como los datos en el 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:

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

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

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

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

Impactos Técnicos De 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:

  a) Conceptos y estructura DW.

  b) El DW introduce muchas tecnologías nuevas (ETT, Carga, Acceso de Datos, Catálogo de Metadatos, Implementación de DSS/EIS), y cambia la manera que nosotros usamos la tecnología existente. Nuevas responsabilidades de soporte, nuevas demandas de recursos y nuevas expectativas, son los efectos de estos cambios.

  c) Destrezas de diseño y análisis donde los requerimientos empresariales no son posibles de definir de una forma estable a través del tiempo.

  d) Técnicas de desarrollo incremental y evolutivo.

  e) 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.

1.5.6. Consideraciones finales del análisis

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

Como ya hemos visto, son variados los cambios que comenzarán a producirse al implementar un DW. Es importante entonces anticiparse a estos cambios, considerar sus implicancias y planificarlos en la empresa. Las siguientes situaciones, gatillan el comienzo de estos cambios:

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

1.6. Data Warehouse: Estrategia Recomendada

Prototipo

La meta del prototipo de Data Warehouse es proveer a los usuarios finales con una aproximación de lo que el Data Warehouse les puede proporcionar en un período de tiempo tan corto como sea posible tal que el grupo de Data Warehouse pueda demostrar los beneficios del Data Warehouse a los usuarios y recolectar lo más pronto la retroalimentación crítica de los usuarios. Un prototipo es un esfuerzo designado a simular tan cerca como sea posible del ver y sentir del producto que será entregado a los usuarios. En el Data Warehouse esto quiere decir que los datos deben ser llevados e integrados y cargados en estructuras de Datos apropiadas. Deben ser distribuidas herramientas de acceso de datos a usuario finales y aplicaciones para realizar queries. Deben ser creadas herramientas de soporte en la Decisión si es aplicable. Sin embargo el proceso de integrar y convertir los datos no será completamente automatizado. En la mayoría de los casos el prototipo contemplará una cargada no repetible (de una sola vez) de los datos de las estructuras del Data Warehouse. La plataforma y la Base de Datos para el almacenamiento puede también diferir de aquellas para la arquitectura definitiva del Data Warehouse, Lo que es importante también, es que la presentación de los Datos al usuario final sea tan fiel como sea posible para que sea igualmente presentada en posteriores etapas del Data Warehouse. En la mayoría de los casos la herramienta que será utilizada en el desarrollo es la misma que la que se ha utilizado para el prototipo.

Piloto

El piloto, simplemente es el primero de muchos esfuerzos iterativos que se harán para llegar a la construcción de un Data Warehouse. Se debe observar especial cuidado porque es la primera fase del proyecto en el cual el equipo de trabajo utilizará los métodos, técnicas y herramientas que será la base para un Data Warehouse completo. Por esta razón el proyecto piloto debe tener un pequeño alcance y tiempo adicional comparativamente con los esfuerzos sucesivos.

Prueba del concepto tecnológico

La prueba del concepto tecnológico es un paso opcional que se puede necesitar para definir si la arquitectura especificada para el Data Warehouse funcionará finalmente como se intenta. En la mayoría de los proyectos el esfuerzo del piloto ha servido también como la prueba del concepto para la arquitectura técnica. Es crítico que la prueba del concepto tecnológico no esté cercana al prototipo, dado que la meta del prototipo es poner datos en las manos de los usuarios tan pronto como sea posible. Dada esta meta el hecho

de proveer factibilidad técnica durante el prototipo podría adicionar enormes e inaceptables riesgos al prototipo.

Arquitectura del Data Warehouse

Dependiendo de la estructura interna de los datos y especialmente del tipo de consultas a realizar, se diseña la arquitectura del Data Warehouse.
Con este criterio los datos deben ser repartidos entre numerosos Data Marts.
Datos de los sistemas de Aplicación y de otras fuentes de Data Warehouse deben ser periódicamente extraídos y alimentados en la capa de Data Scrubbing. La extracción debe ser realizada en muchos casos utilizando los programas para acompañar esta tareas. El Data scrubbing debe ser hecho bien sea con ayuda de programas desarrollados para esto, o con ayuda de herramientas de scrubbing.
El corazón del Data Warehouse debe ser organizado desde un punto de vista de negocio. Se puede utilizar una estructura de Datos para el corazón del Data Warehouse, ligeramente normalizada. Esta estructura de Datos parece estar normalizada cuando se vé al nivel de Entidad - relación. Cuando se miran los atributos sin embargo, la estructura de datos puede estar desnormalizada. Esta aproximación asegura que el corazón del Data Warehouse siempre representa una visión de negocio de la Compañía y sus datos, independientemente de cuántos usuarios están mirando a esos datos en un momento en particular. Esto es importante debido a que la forma en que la información es usada, cambiará frecuentemente y se necesita una Base de Datos estable para soportar el cambio. El corazón del Data Warehouse sirve como este conjunto estable de datos. Mientras que el corazón del Data Warehouse preserva las reglas lógicas del negocio implícitas en los datos, no es siempre fácilmente accesado por los usuarios finales.
Por esta razón se debe utilizar una serie de Data Marts para proveer a los usuarios finales con fácil acceso a sus datos. Los data Marts deben consistir en Datos extraídos del corazón del Data Warehouse y reorganizados y/o reformateados para hacer más fácil su uso para diferentes propósitos. Pero dado que esos propósitos específicos pueden cambiar en el tiempo, los Data Marts deben ser concebidos con estructuras de Datos temporales. Cuando los usuarios no ven más los datos como están presentados por un Data Mart en particular, este Data Mart debe ser removido. Y mientras los usuarios desarrollan nuevas formas de hacer búsquedas y mirar los datos, deben ser creados nuevos Data Marts para hacer sus búsquedas más simples y con un mejor desempeño.
Los Data Mart pueden incluir una gran variedad de estilos de tablas. Algunas pueden ser simplemente un subconjunto de datos en el Data Warehouse, conteniendo solamente datos para una particular zona geográfica, un período específico de tiempo, una unidad de negocios. Otros Data marts pueden ser el resultado de reunir información proveniente de diferentes tablas del corazón del Data Warehouse en una tabla Data Mart desnormalizada .O tal vez los Data Marts serán construidos para contener elementos de datos calculados y derivados que no están explícitamente almacenados en el corazón del Data Warehouse. Lógicamente algunas tablas de Data Marts serán combinación de

estas técnicas. También es posible mencionar que el uso de estructuras de datos multidimensionales debería estar reservado para Data Marts. Esto es, datos que están en el corazón del Data Warehouse deberían almacenarse en forma relacional y luego ser extraídos en un Data Mart multidimensional si es requerido.
Las herramientas de acceso de usuario final son un componente crítico de los Data Warehouses. Para la mayoría de los usuarios finales, la herramienta de acceso es el Data Warehouse. Ellos no son (ni deberían serlo), conocedores de la compleja arquitectura de Datos y análisis que está detrás de la información que ellos ven en sus pantallas de computador. Por lo tanto es crítico que los usuarios sean provistos de un método apropiado para utilizar la información de los Data Warehouses. No se debe esperar que un usuario novato negocie una compleja y poderosa herramienta sólo para hacer una simple pregunta del Data Warehouse. Similarmente un usuario adelantado rápidamente quedará frustrado si el o ella esperan hacer un complejo análisis de negocio usando una herramienta de acceso con menos poder del que se necesita. Es importante reconocer que hay diferentes estilos de usuarios finales cada uno con su propio nivel de conocimiento y necesidades, para así proveer de apropiados mecanismos de acceso para cada clase de usuarios.

Acceso a datos de usuario finales

Las herramienta de acceso para usuarios finales deben ser cuidadosamente seleccionadas. Los usuarios han demostrado a través de su experiencia con los sistemas de manejo de información existentes que tendrán una baja tolerancia a herramientas que no sean fáciles de usar. La mayoría de los usuarios finales para esta etapa inicial serán ejecutivos y gerentes y como tales pondrán énfasis en una herramienta intuitiva y fácil de utilizar.
Eventualmente pueden surgir una clase de usuarios poderosos que necesita hacer análisis complejos de datos, excediendo las capacidades de SQL Nativo. Para este caso de usuarios se debe considerar herramientas OLAP aunque las mejores de esta herramientas pueden ser difíciles de usar y envuelven una curva de aprendizaje. Estudios recientes demuestran que sólo 1 de 10 usuarios a los que se les ha dado herramientas OLAP realmente hacen el tipo de análisis complejos que hacen que esas herramientas sean necesarias. Para otro tipo de usuarios herramientas OLAP pueden ser inútiles y de frustrante complejidad.

Factores de riesgo

Es importante conocerlos para poder monitorearlos. Son estos

- **Expectativas de los usuarios.** Se debe trabajar en las expectativas de los usuarios. Muchas veces el éxito depende de la diferencia entre lo que los usuarios esperan y lo que ellos perciben que les es entregado. Es crítico que el equipo de trabajo comunique las expectativas acerca de lo que será entregado muy claramente y ayude al usuario final a entender la naturaleza iterativa de construir un Data Warehouse.

- **Experiencia con Data Warehouses**. Este riesgo se puede reducir con el uso juicioso de experiencias de proveedores y consultores.

- **Dirección estratégica.** Es relativamente lógico definir un punto de inicio lógico para el Data Warehouse. Sin embargo cuando esta primera área se haya completado, es más difícil identificar áreas para esfuerzos futuros y asegurar que esos esfuerzos están alineados con los objetivos y necesidades del negocio. El riesgo se puede mitigar siguiendo la estrategia recomendada, para entender las necesidades y prioridades de la información del negocio y desarrollar una implementación de Data Warehouses a largo plazo que cumpla con estas prioridades. [3]

2. DATA MINING

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

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

2.1. Los Fundamentos del Data Mining

Las técnicas de Data Mining son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. Data Mining toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. Data Mining está listo para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

- Recolección masiva de datos

- Potentes computadoras con multiprocesadores
- Algoritmos de Data Mining

Las bases de datos comerciales están creciendo a un ritmo sin precedentes. Un reciente estudio del META GROUP sobre los proyectos de Data Warehouse encontró que el 19% de los que contestaron están por encima del nivel de los 50 Gigabytes, mientras que el 59% espera alcanzarlo en el segundo trimestre de 1997. En algunas industrias, tales como ventas al por menor (retail), estos números pueden ser aún mayores. MCI Telecommunications Corp. cuenta con una base de datos de 3 terabytes + 1 terabyte de índices y overhead corriendo en MVS sobre IBM SP2. La necesidad paralela de motores computacionales mejorados puede ahora alcanzarse de forma más costo - efectiva con tecnología de computadoras con multiprocesamiento paralelo. Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que consistentemente son más performantes que métodos estadísticos clásicos.

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

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

2.2. El Alcance de Data Mining

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

- **Predicción automatizada de tendencias y comportamientos**. Data Mining automatiza el proceso de encontrar información

predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos. Un típico ejemplo de problema predecible es el marketing apuntado a objetivos (targeted marketing). Data Mining usa datos en mailing promociónales anteriores para identificar posibles objetivos para maximizar los resultados de la inversión en futuros mailing. Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.

- **Descubrimiento automatizado de modelos previamente desconocidos.** Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar *datos anormales* que pueden representar errores de tipeado en la carga de datos.

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

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

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

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

Las técnicas más comúnmente usadas en Data Mining son:

- **Redes neuronales artificiales:** modelos predecibles no-lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.

- **Árboles de decisión:** estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Métodos específicos de árboles de decisión incluyen Árboles de Clasificación y Regresión (CART: Classification And Regression Tree) y Detección de Interacción Automática de Chi Cuadrado (CHAI: Chi Square Automatic Interaction Detection)

- **Algoritmos genéticos:** técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y selección natural en un diseño basado en los conceptos de evolución.

- **Método del vecino más cercano:** una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases del/de los *k* registro (s) más similar/es a él en un conjunto de datos históricos (donde *k* □ 1). Algunas veces se llama la técnica del vecino *k*-más cercano.

- **Regla de inducción:** la extracción de reglas if-then de datos basados en significado estadístico.

Muchas de estas tecnologías han estado en uso por más de una década en herramientas de análisis especializadas que trabajan con volúmenes de datos relativamente pequeños. Estas capacidades están ahora evolucionando para integrarse directamente con herramientas OLAP y de Data Warehousing.

2.3. ¿Cómo Trabaja el Data Mining?

¿Cuán exactamente es capaz Data Mining de decirle cosas importantes que usted desconoce o que van a pasar? La técnica usada para realizar estas hazañas en Data Mining se llama *Modelado*. Modelado es simplemente el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta. Por ejemplo, si busca un galeón español hundido en los mares lo primero que podría hacer es investigar otros tesoros españoles que ya fueron encontrados en el pasado. Notaría que esos barcos frecuentemente fueron encontrados fuera de las costas de Bermuda y que hay ciertas características respecto de las corrientes oceánicas y ciertas rutas que probablemente tomara el capitán del barco en esa época. Usted nota esas similitudes y arma un modelo que incluye las características comunes a todos los sitios de estos tesoros hundidos. Con estos modelos en mano sale a buscar el tesoro donde el modelo indica que en el pasado hubo más probabilidad de darse una situación similar. Con un poco de esperanza, si tiene un buen modelo, probablemente encontrará el tesoro.

Este acto de construcción de un modelo es algo que la gente ha estado haciendo desde hace mucho tiempo, seguramente desde antes del auge de las computadoras y de la tecnología de Data Mining. Lo que ocurre en las computadoras, no es muy diferente de la manera en que la gente construye modelos. Las computadoras son cargadas con mucha información acerca de una variedad de situaciones donde una respuesta es conocida y luego el software de Data Mining en la computadora debe correr a través de los datos y distinguir las características de los datos que llevarán al modelo. Una vez que el modelo se construyó, puede ser usado en situaciones similares donde usted no conoce la respuesta.

Si alguien le dice que tiene un modelo que puede predecir el uso de los clientes, ¿Cómo puede saber si es realmente un buen modelo? La primera cosa que puede probar es pedirle que aplique el modelo a su base de clientes - donde usted ya conoce la respuesta. Con Data Mining, la mejor manera para realizar esto es dejando de lado ciertos datos para aislarlos del proceso de Data Mining. Una vez que el proceso está completo, los resultados pueden ser testeados contra los datos excluidos para confirmar la validez del modelo. Si el modelo funciona, las observaciones deben mantenerse para los datos excluidos.

2.4. Una arquitectura para Data Mining

Para aplicar mejor estas técnicas avanzadas, éstas deben estar totalmente integradas con el data warehouse así como con herramientas flexibles e interactivas para el análisis de negocios. Varias herramientas de Data Mining actualmente operan fuera del warehouse, requiriendo pasos

extra para extraer, importar y analizar los datos. Además, cuando nuevos conceptos requieren implementación operacional, la integración con el warehouse simplifica la aplicación de los resultados desde Data Mining. El Data warehouse analítico resultante puede ser aplicado para mejorar procesos de negocios en toda la organización, en áreas tales como manejo de campañas promocionales, detección de fraudes, lanzamiento de nuevos productos, etc.

El punto de inicio ideal es un data warehouse que contenga una combinación de datos de seguimiento interno de todos los clientes junto con datos externos de mercado acerca de la actividad de los competidores. Información histórica sobre potenciales clientes también provee una excelente base para prospecting. Este warehouse puede ser implementado en una variedad de sistemas de bases relacionales y debe ser optimizado para un acceso a los datos flexible y rápido.

Un server multidimensional OLAP permite que un modelo de negocios más sofisticado pueda ser aplicado cuando se navega por el data warehouse. Las estructuras multidimensionales permiten que el usuario analice los datos de acuerdo a como quiera mirar el negocio - resumido por línea de producto, u otras perspectivas claves para su negocio. 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 campaña, prospecting, y optimización de promociones. La integración con el data warehouse permite que decisiones operacionales sean implementadas directamente y monitoreadas. A medida que el data warehouse crece con nuevas decisiones y resultados, la organización puede "minar" 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 Análisis Avanzado aplica los modelos de negocios del usuario directamente al 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.

3. OLAP

Diariamente los gerentes empresariales se enfrentan con dos retos fundamentales: operar la empresa de manera eficiente para maximizar la recuperación de la inversión y planear el futuro. El procesamiento informático y el procesamiento analítico son dos formas básicas de aprovechar el Data Warehouse para abordar esos dos retos.
Las empresas de hoy día que operan en una economía global con competidores globales, necesitan buscar mercados en donde sus productos y servicios tengan claras ventajas competitivas y sean diferentes. Un requerimiento fundamental es buscar nuevas oportunidades de mercados y segmentos de micromercados y crear programas de comercialización detallados. Para lograr esto, es un requisito el análisis multidimensional.

3.1. Análisis Multidimensional

En el análisis multidimensional, los datos se presentan mediante dimensiones como producto, territorio y cliente. Por lo regular las dimensiones se relacionan en jerarquías, por ejemplo, ciudad, estado, región, país y continente, o estado, territorio y región. El tiempo es también una dimensión estándar con su propia jerarquía como día, semana, mes trimestre y año, o día y año calendario.
Un usuario empresarial puede acceder los ingresos por departamento y tiende para los últimos cuatro trimestres, para un conjunto dado de productos. Los resultados se pueden pivotear o girar para cambiar los ejes y la perspectiva. Además, los usuarios empresariales pueden navegar por las dimensiones profundizando u obteniendo resúmenes a lo largo de los elementos de una dimensión, o pueden penetrar a través de las dimensiones para ver otras perspectivas.
Al procesamiento analítico o análisis multidimensional se le conoce también como procesamiento analítico en línea (OLAP). Se procesa en una visión multidimensional de los datos empresariales en el Data Warehouse y puede tener un motor de depósito de base de datos multidimensional.
De esta forma, dentro de un Data Warehouse existen dos tecnologías complementarias, una relacional para consultas y una multidimensional para análisis. Esta distinción se ha establecido como respuesta a los requerimientos del usuario (quienes marcan las tendencias en los negocios) y por la debilidad fundamental de las bases de datos relacionales para proveer capacidad de análisis y sistemas de soporte a las decisiones (quienes marcan las tendencias tecnológicas).
Los RDBMS´s y sus herramientas asociadas para generar aplicaciones, pueden ser limitativas cuando se requieren tareas de análisis, debido a que se cuenta con un manejo ineficiente de relaciones multidimensionales, habilidad limitada de análisis y consolidación. Esto se debe principalmente a que el modelo relacional se basa en un esquema bidimensional (tablas) y el manejo de datos mediante SQL no es adecuado para los cálculos que requiere una aplicación OLAP. Sí existen herramientas propias del usuario en su PC, orientadas a generar funcionalidad

multidimensional (investigación a detalle, consolidaciones y rotaciones), éstas tienen la limitante del manejo masivo de datos.

Como E.F. Codd lo establece:
"De hecho, los sistemas transaccionales nunca fueron diseñados con la intención de proveer funciones poderosas para la síntesis, el análisis y la consolidación de datos, que define el análisis multidimensional de datos. Este tipo de funciones tuvieron la intensión de ser proporcionadas por herramientas de usuarios distintas y complementarias de los productos transaccionales".

Los requerimientos de un sistema OLAP y el deseo de aventajar con las bases de datos relacionales, ha conducido a la creación de productos referidos como OLAP relacionales (ROLAP). Estas herramientas, extienden la capacidad de las bases de datos relacionales en un intento de hacerlas también adecuadas para aplicaciones OLAP.

3.2. Definición de OLAP

El procesamiento analítico en línea es una tecnología de análisis de datos que hace lo siguiente:

- Presenta una visión multidimensional lógica de los datos en el Data Warehouse. La visión es independiente de como se almacenan los datos.

- Comprende siempre la consulta interactiva y el análisis de los datos. Por lo regular la interacción es de varias pasadas, por lo que incluye la profundización en niveles cada vez más detallados o el ascenso a niveles superiores de resumen y adición.

- Ofrece opciones de modelado analítico, incluyendo un motor de cálculo para obtener proporciones, desviaciones, etc., que comprende mediciones de datos numéricos a través de muchas dimensiones.

- Crea resúmenes y adiciones (también conocidas como consolidaciones), jerarquías y cuestiona todos los niveles de adición y resumen de cada intersección de las dimensiones.

- Maneja modelos funcionales de pronóstico, análisis de tendencias y análisis estadísticos.

- Recupera y exhibe datos tabulares en dos o tres dimensiones, cuadros y gráficas, con un pivoteo fácil de los ejes. El pivoteo es fundamental ya que los usuarios empresariales necesitan analizar los datos desde perspectivas diferentes; y el análisis desde una perspectiva conduce a otra cuestión empresarial que se va a examinar desde otra perspectiva.

- Responde con rapidez a las consultas, de modo que el proceso de análisis no se interrumpe y la información no se desactualiza.
- Tiene un motor de depósito de datos multidimensional, que almacena los datos en arreglos. Esos arreglos son una representación lógica de las dimensiones empresariales.

La tecnología OLAP se aplica en muchas áreas funcionales de una empresa, tales como producción, ventas y análisis de rentabilidad de la comercialización, mezcla de manufacturas y análisis de logística; consolidaciones financieras, presupuestos y pronósticos, planeación de impuestos y contabilidad de costos.

OLAP surge como un proceso para ser usado en el análisis de ventas y mercadotecnia, para elaborar reportes administrativos y consolidaciones, para presupuestación y planeación, para análisis de rentabilidad, reportes de calidad y otras aplicaciones que requieren una visión flexible, de arriba a abajo, del negocio. OLAP provee de reportes sumarios que los ejecutivos requieren para tomar decisiones, así como la facilidad de elaborar cálculos complejos, enfoques a detalles operativos y consultas no programadas. OLAP se alimenta principalmente de los sistemas transaccionales y como tales, debe considerar una eficiente administración de la base de datos y proveer un nivel adecuado de seguridad.

Prendse y Creeth, en su “OLAP report”, establecen que mientras la definición actual de OLAP ha justificado su aceptación, poco se ha recorrido en la definición de un estándar computacional, similar al que SQL es para los sistemas de bases de datos relacionales (RDBMS). Ellos proponen el concepto FASMI (Fast Analysis Shared Multidimensional Information) para definir un OLAP. Dentro de este contexto:

Fast (Rápido), significa que el sistema está orientado a tener tiempos de respuesta de 5 segundos, con respuestas de 1 segundo a consultas sencillas y, algunas más complejas, con respuestas de hasta 30 segundos . En situaciones típicas, si las respuestas toman más tiempo, el usuario perderá la concentración en lo que busca analizar.

Analysis (Análisis), lo que indica que al usuario debe proporcionársele suficiente funcionalidad para resolver sus problemas, sin la necesidad de contar con el apoyo de sistemas o de una pre-programación. De esta forma, se podrán realizar consultas no definidas, cálculos de diferencias, variaciones y tendencias, consolidar y llevar a cabo análisis de sensibilidad y de búsqueda de metas.

Shared (Compartida), significa que debe haber facilidad de acceso simultáneo, tanto de lectura como de escritura, aunado a un esquema de seguridad adecuado, con el objeto de guardar la confidencialidad (probablemente a nivel de celda).

Multidimensional, ha sido y continuará siendo un requisito, con la habilidad de manejar múltiples jerarquías y dimensiones.

Information (Información), son todos los datos y la información derivada, cuando y donde sea necesaria, en su contexto, tanto de información "suave" como información "dura", interna o externa.

3.3. OLAP: Multidimensional vs. Relacional

Las dos opciones para almacenar los datos son el depósito multimensional y el depósito relacional.

Desde la perspectiva del usuario, los aspectos a comparar son:

- ¿Cuántos datos empresariales están almacenados y disponibles?
- ¿Es adecuada la capacidad de almacenamiento?
- ¿Es aceptable el desempeño?
- ¿Se justifica el costo?
- ¿La consulta de los mismos es manejable por los usuarios?
- ¿Qué tipos de vínculos se puede asociar?

La siguiente figura resume los aspectos de selección de uno u otro.

	Base de Datos Relacional	Base de Datos Multidimensional
Depósito de datos, acceso y visión	▪ Relacional ▪ Tablas de filas y columnas ▪ SQL con ampliaciones ▪ Herramientas API	▪ Dimensional ▪ Arreglos: Hipercubo/Multicubo ▪ Tecnología de matriz dispersa ▪ Propietario de hoja de cálculo
Utilización e incorporación	▪ OLTP ▪ Motor RDBMS ▪ Profundización a nivel de detalle ▪ Desempeño de consultas: rango amplio	▪ OLAP ▪ Motor multidimensional ▪ Profundización a nivel de resumen/adición ▪ Desempeño de consultas: rápido
Tamaño y actualización de la base de datos	▪ Gigabytes a Terabytes ▪ El depósito de índices y el retiro de normas incrementan el tamaño	▪ Gigabytes ▪ Compresión y adición de datos dispersos ▪ Difícil actualización durante el

	▪ Consulta y carga paralelas ▪ Actualización durante el año	uso; los cambios pequeños podrían requerir reorganización
Alcance de interacción del usuario	Transaccional	Base de datos completa
Datos afectado por interacción del usuario	Registros individuales	Grupos de registros
Utilización de la máquina	Estática	Dinámica
Prioridades (fuente:IBM Consulting Group)	Alto desempeño Alta disponibilidad	Ala flexibilidad Alta facilidad de consulta por usuarios

Depósito de datos multidimensional vs. Relacional

Basado en la tabla anterior un administrador RDBMS es una mejor opción cuando los datos sean voluminosos (5 – 10 GB en adelante) o se estime crecimiento constante del volumen de datos; al igual es conveniente cuando el número de usuarios concurrentes esperados sea mayor a 50 (considere por ejemplo que una misma bodega de datos puede ser utilizada por usuarios de varias aplicaciones, Data Marts, simultáneamente).
Similarmente, un MDD es una buena opción en el caso de tener volúmenes alrededor de 10 GB, con relativamente pocos usuarios concurrentes. La integración natural de herramientas de explotación, con la estructura de la representación dimensional de los datos, hace que los MDD sean una opción viable en el desarrollo de aplicaciones con funcionalidad superior, a realizar en períodos cortos de tiempo.

Una característica no técnica, pero no menos importante a considerar, es el costo del administrador en sí. Si ya cuenta con un RDBMS (común en empresas grandes), una estrategia adecuada sería desarrollar su Data Warehouse con su RDBMS familiar, ya que el costo sería solamente por licencias adicionales. Lo anterior, sumado al importante factor que es el contar en casa con diseñadores y DBA con experiencia en optimización de aplicaciones y afinación de bases de datos es conocimiento difícil de adquirir en el corto plazo trabajando con MDD, favorece la implementación de Data Warehouses basados en RDBMS.

3.4. Depósito de datos multidimensional y servicios OLAP

El diseño inicial y la actividad de configuración son conducidas por el “diseño lógico” o el “modelo de información”. Los pasos básicos son los siguientes:

- Seleccionar la función empresarial, como análisis de ingresos por ventas y reportes financieros.
- Identificar los valores numéricos, es decir, las mediciones que se van a almacenar, tales como ingresos por ventas y clientes.
- Determinar las dimensiones (tiempo, geografía y producto), por ejemplo tiempo por mes y trimestre, y geografía por estado o región.
- Definir el "modelo lógico" y cargar el depósito de datos multidimensionales, ya sea directamente de la fuente de datos o filtrando y ajustando el contenido seleccionado del Data Warehouse.

Las funciones principales que se ofrecen a usuarios empresariales comprenden las siguientes:

- Rápida respuesta a consulta de cómputos intensiva, tales como escenarios "¿qué pasa si…?.
- Actualización interactiva de la base de datos multidimensional, para permitir aplicaciones de pronósticos, planeación a futuro y presupuestación.
- Explotación de relaciones ricas entre los elementos o valores de las dimensiones para descubrir relaciones insospechadas.
- Un poderoso motor de cálculo y análisis comparativo: posiciones, comparaciones, porcentaje de clase, máximo, mínimo, promedios, promedios móviles, comparaciones entre períodos, y otros.
- Cálculos cruzados entre dimensiones, como asignación de costos y eliminaciones dentro de la compañía, o cálculos a nivel de filas para aplicaciones orientadas a hojas de cálculo como las declaraciones de pérdidas y ganancias .
- Ampliación de las funciones básicas con funciones definidas por el usuario.
- Potentes funciones estadísticas y financieras.
- Inteligencia de tiempo.
- Pivoteo, tabulación cruzada, profundización, niveles de resumen para una o varias dimensiones, y otras funciones poderosas de navegación.

La administración general y la administración de sistemas requieren del siguiente:

- El modelado inicial de los datos en donde son consideraciones clave elegir las dimensiones correctas, prever como se accederán los datos y seleccionar los filtros apropiados para cargar los datos desde el Data Warehouse.

- Transferencias periódicas y actualizaciones en bloque, debido a que las actualizaciones en incrementos son un reto y casi imposibles mientras que la base de datos está en uso.

- Adición, resúmenes y precálculo durante el proceso de cálculo.

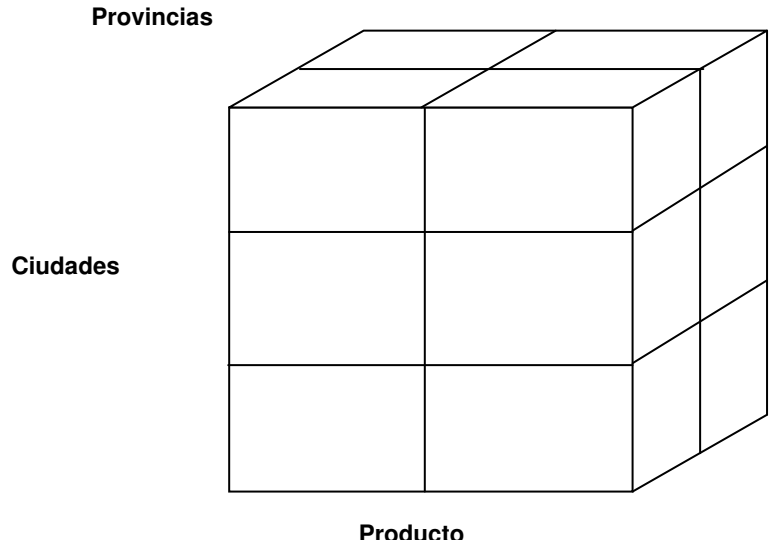
- Capacitación en una tecnología diferente y uso de nuevas habilidades.

- Escritura de nuevas aplicaciones en lenguaje propietario para ampliar y mejorar los procesos frontales comunes de la base de datos.

La razón por la que se necesitan varias dimensiones es que no se pueden mezclar informaciones distintas (Ciudades con Provincias, Provincias con regiones, etc.).

En las BDM sin jerarquías, la solución puede ser por dimensiones separadas.

A continuación, se representa gráficamente como debe visualizarse.



Base de Datos Multidimensional (BDM)

3.5. OLAP relacional (ROLAP)

Aunque los datos se almacenan en forma relacional (fila, columna), se presentan al usuario empresarial en forma de dimensiones empresariales. A fin de "ocultar" la forma de depósito, se crea una capa semántica de

metadatos. Esta capa ubica las dimensiones para las tablas relacionales. También se crean metadatos adicionales para cualquier resumen o adición, con el fin de mejorar el tiempo de respuesta. Todos estos datos se almacenan en la base de datos relacional para su mantenimiento y administración.

La actividad inicial de diseño y configuración se conduce mediante el "diseño técnico de una base de datos". Se deben seguir estos pasos básicos:

- Construir el "diseño dimensional" utilizando técnicas como el retiro de normas, el esquema de estrella.

- Incorporar los datos adecuados de adición y resumen.

- Dividir los grandes conjuntos de datos en segmentos más pequeños y manejables para mejorar el desempeño. Por ejemplo, se dividen las unidades de tiempo u organizacionales.

- Agregar índices creativos o de mapa de bits para mejorar el desempeño.

- Crear y almacenar los metadatos. Los metadatos incluyen las definiciones de las dimensiones, la ubicación de las dimensiones para las tablas relacionales, las relaciones jerárquicas entre dimensiones, la información de segmentos, las definiciones y descripciones de resúmenes y adiciones, las fórmulas y cálculos, la v vigencia de uso y muchas otras .

Desde la perspectiva operacional, los pasos para ejecutar una consulta son los siguientes:

- Construir la herramienta cliente utilizando una visión dimensional o empresarial de los datos.

- Consultar el servidor OLAP desde la herramienta cliente, y examinar los metadatos en tiempo real.

- Crear declaraciones SELECT de pasos múltiples y/o subconsultas correlacionadas y someterlas a la base de datos relacional.

- Sobre los resultados de la consulta a la base de datos, realizar funciones multidimensionales, como cálculos y fórmulas, traducción de bits para descripciones empresariales, etc.

- Devolver los resultados a la herramienta cliente para un mayor procesamiento y exhibición. O para su exhibición inmediata.

Las herramientas ROLAP mapean una estructura multidimensional en una capa por encima de la estructura de tablas original, proveen capacidad de generar cálculos analíticos que limitan a las bases relacionales y al usuario de herramientas para poder observar datos bajo un esquema multidimensional. Esta capa multidimensional es esencial, puesto que una base de datos relacional inherentemente no entiende estructuras multidimensionales.

Con el objeto de optimizar el desempeño de un ROLAP, se hace necesaria la creación de una tabla de sumarización que debe pasar a un servidor de cálculos, que se encuentra fuera de la base de datos relacional. Un ROLAP requiere de la creación y mantenimiento de un índice. Para mantener un sistema eficiente, es necesario que todos los renglones se ordenen (o indexen, como también se le conoce), con la consecuencia de que este índice adquiere un tamaño considerablemente grande, que al no poder "subir" a la memoria, afecta el desempeño de la aplicación.

3.6. Evaluación de servidores y herramientas OLAP

Los servidores y herramientas OLAP se evalúan utilizando cuatro conjuntos de criterios:

3.6.1. Características y Funciones.

La interfaz del usuario y el acceso a lo servicios OLAP deben ofrecer diversas opciones que confieran mayor poder a las habilidades que ya posee el usuario y al conocimiento incorporado en los modelos analíticos OLAP. Las opciones potenciales deben incluir las siguientes:

- Hoja de cálculo
- Herramientas cliente de propietario
- Herramientas de otros fabricantes
- Ambiente 4GL
- Interfase con lenguajes estándar
- Navegadores de cubo para clientes

3.6.2. Motores de servicios OLAP.

El motor de servicios OLAP, en cualquiera de sus configuraciones, depósito dimensional ó relacional, debe satisfacer la capacidad, la viabilidad de cambio de escala y las características tecnológicas del modelo analítico y la aplicación planeados. Las características de tecnología requeridas dependen del modelo analítico y del uso contemplado.

3.6.3. Administración.

Las funciones necesarias de administración para fines de preparación inicial, configuración y operación continua incluyen lo siguiente:

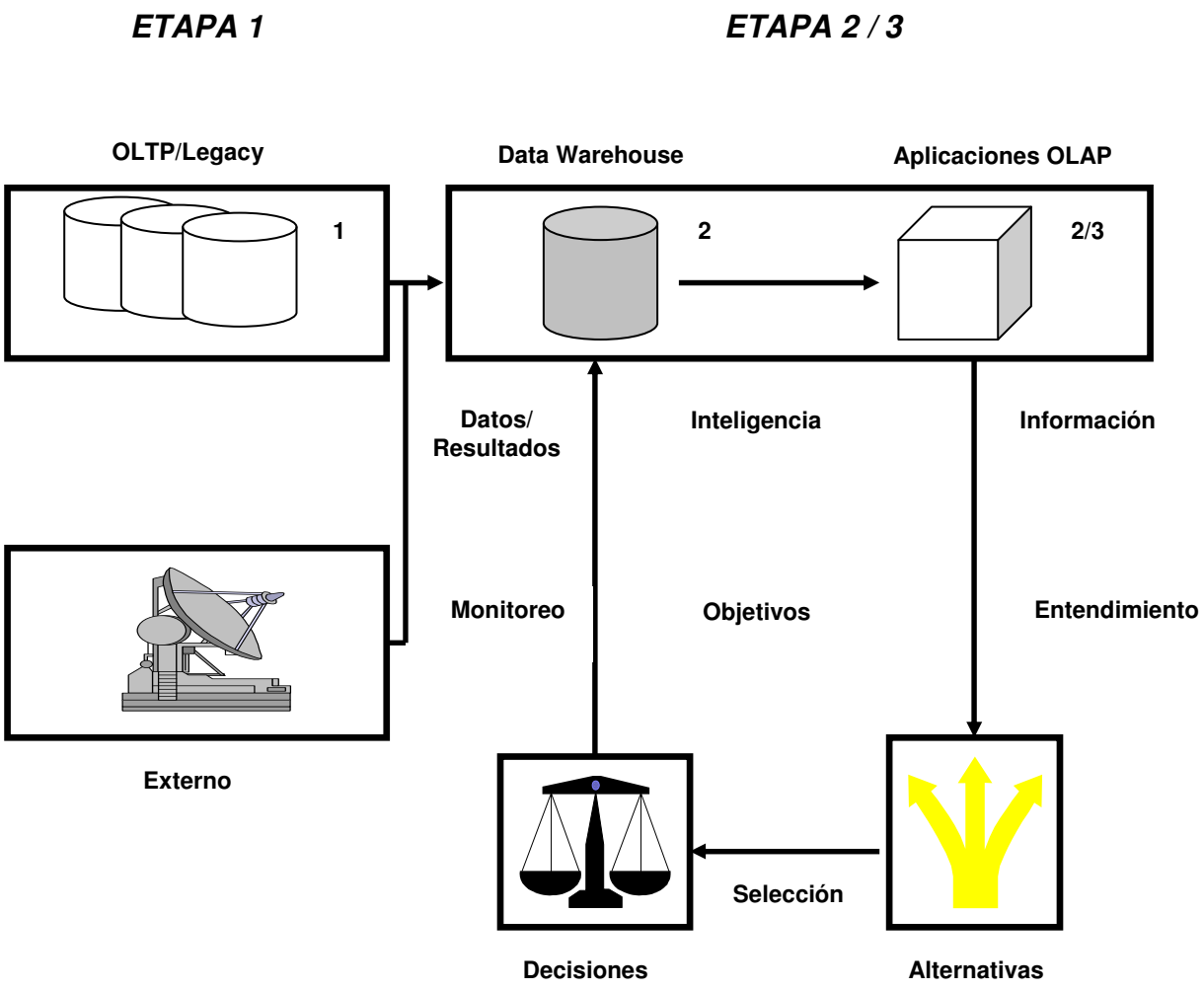
- Definición del modelo analítico dimensional
- Creación y mantenimiento del depósito de metadatos
- Control de acceso y privilegios con base en el uso
- Carga del modelo analítico desde el Data Warehouse
- Afinación del desempeño a niveles aceptables para permitir un análisis que no desorganice
- Reorganización de la base de datos
- Administración de todas las partes del sistema, incluyendo el Middleware
- Distribución de los datos al cliente para el análisis adicional y local

3.6.4. Arquitectura global.

Desde una perspectiva de arquitectura global no es sencillo elegir entre el depósito de datos multidimensionales o el relacional para el OLAP. La empresa necesita proporcionar los criterios para hacer la elección adecuada.

Por fortuna, la tendencia de la industria es ofrecer los servicios OLAP con una combinación de un proceso frontal de servidor OLAP, con un depósito multidimensional incorporado para datos gruesos y un proceso posterior de depósito relacional. En esta configuración de arquitectura, la información y consultas de acceso frecuente se precalculan, se resumen, se agregan y luego se almacenan en depósitos de datos multidimensionales relacionales de Data Warehouse. Las consultas complejas o de cómputo intensivo, o los datos complejos basados en cálculos también se procesan y almacenan.

Pasaje de OLTP a OLAP



3.7. Desarrollo de Aplicaciones OLAP

El éxito en las empresas que pueden implementar aplicaciones OLAP, dependerá frecuentemente de sus experiencias y de la metodología empleada. Puesto que las aplicaciones de OLAP son principalmente desarrolladas y mantenidas por el usuario, no requieren de un soporte amplio por las áreas de sistemas. La innovación en el proceso de la toma de decisiones (IDP), se basa en facilitar un entorno mediante herramientas analíticas bien probadas y de técnicas que ayudan a determinar perspectivas claras sobre la dinámica del negocio. IDP es una metodología completa que se enfoca en la integración de tres elementos básicos: el proceso, la tecnología y la cultura de la organización. El objetivo consiste en mejorar la productividad y eficiencia

de análisis y del proceso de toma de decisiones a lo largo de la organización.

Tanto IDP, como los mapas de decisión (indicado en el diagrama), ayudan a definir las mejores prácticas desde la definición de objetivos y medidas de desempeño, hasta el desarrollo de planes, tendencias claves de decisión, información requerida para verificar supuestos y la selección de los mejores cursos de acción. Como resultado de este proceso, se genera una lista de aplicaciones junto con sus prioridades correspondientes. Mediante los mapas de decisión, es posible definir el modelo mental que los ejecutivos utilizan para conceptualizar y controlar su negocio.

Las aplicaciones OLAP brindan la inteligencia de negocios requerida para tomar decisiones, y por consiguiente, definen la estructura y el contenido del almacén de datos, el nivel histórico y de detalle requerido de la información. Finalmente, el proceso de colección de datos, filtrado e integración de los mismos, obtenida de los OLTP, permite generar información valiosa propia del almacén de datos y los OLAP.

3.8. OLTP v/s OLAP: Dos Mundos Diferentes

De acuerdo como se entiendan las diferencias entre estos dos tipos de sistemas uno gana un mejor entendimiento de OLAP. Esto es muy importante en especial para diseñadores, ya que ellos necesitan ver estas diferencias para poder llevar a cabo de mejor manera un proyecto de esta naturaleza.

Las diferencias entre ambos procesamientos se establecen en distintos ámbitos; el siguiente es un paralelo entre ambas filosofías:

3.8.1. Orientación o Alineación de Datos.

OLTP	OLAP
Alineación por aplicación	Alineación por dimensión
Datos son organizados inherentemente por aplicación.	Datos son organizados por dimensiones definidas del negocio.
Focalizado en encontrar requerimientos de aplicaciones específicas para tareas específicas.	Focalizado en encontrar requerimientos de análisis empresarial.

3.8.2. Integración

OLTP	OLAP
Típicamente no integradas	Debe ser integrada
Cada tema de negocios puede tener información en diferentes sistemas.	Toda información de un tema, alimentado de varios sistemas, reunidos en una sola B.D.
Diferentes sistemas contienen diferentes tipos de datos.	Todos los tipos de datos integrados en un sistema.
Diferentes convenciones de nomenclatura	Convenciones de nomenclatura estandarizadas
Diferentes formatos de archivos	Formato de archivo standard
Diferentes plataformas Hardware	Un sólo servidor (lógico) warehouse

3.8.3. Acceso y Manipulación de datos por parte de Usuarios finales.

OLTP	OLAP
Los usuarios son los que giran las ruedas de la organización: ingresan datos nuevos, abren y cierran registros, corrigen datos antiguos. (Es decir, selección, inserción, modificación y eliminación de datos.)	Los usuarios miran las ruedas de la organización.(Sólo selección)
Se ejecutan muchas veces las mismas acciones.	El usuario continuamente cambia el tipo de preguntas que hace a la base de datos.

3.8.4. Administradores

OLTP	OLAP
Manipulación de datos registro a registro.	Carga y Acceso de datos en forma Masiva.
Transacciones y/o rutinas de validación a nivel de registro.	Validación realizada antes o después de cada carga (nunca a nivel de registro o de transacción).

3.8.5. Transacción

OLTP	OLAP
Se manejan cientos de transacciones por día.	Se maneja sólo una transacción que contiene cientos de registros. Esto se hace a través de una carga de datos desde OLTP al repositorio del DW, dependiendo de la estructura que se adopte.
Si está bien la transacción (se realiza exitosamente) se asegura la consistencia de ese único pedazo de datos.	Si la carga termina bien se tiene la consistencia asegurada de TODO el conjunto de datos.

3.8.6. La dimensión Tiempo

OLTP	OLAP
Hay una falta de soporte explícito para reconstruir la historia previa. Si se reescribe sobre los datos, no se puede recuperar un estado anterior. Si se mantiene un historial, al aumentar los cambios es casi imposible reconstruir rápidamente hasta un punto pasado.	Similar a las capas geológicas, la base de datos dimensional puede verse como una serie de capas de datos, compuestas cada una por una snapshot del OLTP tomadas a intervalos regulares (mientras más se cava en las capas más se ahonda en el pasado).
Datos operacionales son altamente volátiles, cambian en la medida que opera la empresa y sus sistemas computacionales reflejan la operación.	Los datos del DW son altamente estables, son insertados en intervalos definidos, y no son modificados.
Pueden haber cambios en la base de datos mientras se está consultando en ella. Estas bases que sufren continuos cambios reciben el nombre de bases de datos parpadeantes.	No hay parpadeo.

3.9. Conclusiones

La necesidad de análisis multidimensional oportuno, como soporte para la toma de decisiones, es cada vez más creciente. Como respuesta, los departamentos de sistemas se han inclinado por la tecnología de almacenes de datos para satisfacer la demanda de los usuarios. Sin embargo, es importante tomar en cuenta que existen diversas tecnologías orientadas a consultar, almacenar datos y analizar resultados. Aunado a lo anterior, se conoce de pocos almacenes de datos exitosos, que satisfacen las demandas del usuario. En este sentido, los mercados de datos a base de OLAP, dentro o fuera de la arquitectura de almacenes de datos, ha demostrado ser una solución práctica para el usuario final. En este sentido, la selección de la arquitectura correcta para el desarrollo de la aplicación es de vital importancia.

Mediante el uso de IDP y mapas de decisiones, o con metodologías similares, se ayuda a definir el proceso de la toma de decisiones dentro de la organización. Esto a su vez, define los requerimientos analíticos de la aplicación OLAP. Puesto que el usuario final es la clave para la definición y el desarrollo de la aplicación, su involucramiento es necesario para alcanzar el éxito en la primera aplicación. La aplicación OLAP, dependiendo del volumen y del nivel de detalle requerido, debe definir el contenido del almacén de datos y los accesos a los sistemas transaccionales y no lo contrario.

CONCLUSION

Una de las principales limitaciones del manejo de información ha sido la complejidad que representa su integración, puesta en marcha y, sobre todo, aprovechamiento de datos contenida en ese tipo de soluciones, la llegada de Internet empieza a modificar esto para hacerlo más accesible al usuario.

Desentrañar una molécula de ADN (ácido desoxirribonucleico), la cual contiene toda la información que compone a un ser humano, puede llevarle al científico toda su vida si no cuenta con las herramientas adecuadas y si no sabe cuáles son las bases de organización y componentes de la propia molécula. Algo similar sucede en las grandes empresas cuyos datos se encuentran recluidos en enormes bases de datos pero no hay forma de aprovecharlos porque se desconoce cómo está organizada o porque es demasiado difícil acceder a ella.

Los Datawarehouses, o almacenes de datos, fueron una de las propuestas modernas de organización y administración de datos centralizados que permitió a la gente desentrañar todo ese historial de información que compone y ha compuesto la compañía durante años, a fin de usarlo en beneficio propio.

Organizar y administrar los datos fue un proceso difícil y lento que viajó de lo más complicado a lo más sencillo. Así los Data Warehouses han visto su crecimiento lleno de conceptos complejos que el cliente no puede asimilar fácilmente y que no le sirven de nada a la hora de mejorar su productividad, lo cual es el meollo del concepto.

El hecho de que el número de instalaciones de Datawarehouses en el país sea muy bajo y los que han sido exitosos sean menos, mucho se debe a la falta de la filosofía que debe rodear un sistema de ese tipo, pues el simple DWH no puede cargar con todo el paquete. A través de tantos nombres raros como MetaData, Data Marts, Data Cartridges, servidores universales, entre otros, se perdió de vista que se trataba con personas, no con máquinas, y la gente no pudo identificarse con ese nuevo sistema como parte de él, por lo que no cambió su forma de actuar para adaptarse a él.

Ahora ya se sabe cómo funciona y de qué forma se pueden organizar los datos, pero ese es problema principalmente de los administradores de base de datos, la meta es saber qué tan fácil y eficiente puede ser, para quienes manejan la empresa, en cuanto a mejorar la situación y posición competitiva de la misma, y sobre todo, qué tanto se adapta la solución a la forma de actuar de la empresa.

De tal suerte, el Datawarehousing debió dejar de ser complejo para formar parte de una solución donde la interactividad y el cliente son la parte más importante. Ello quiere decir darle al usuario u hombre de negocios una interfase conocida con herramientas flexibles y suficientes para la información que requiere y, ahora sí, mostrarle todo lo que se había estado perdiendo por no consultar su propia historia de datos.

Lo mismo que los genes no necesitan saber cómo está compuesto el ADN para actuar sobre él a fin modificarlo, multiplicarlo u obtener información valiosa de quien lo porta, los usuarios de un Data Warehouse no necesita conocer todo lo que está atrás de una interfase para alcanzar una meta.

La experiencia que tiene toda la gente de la empresa no debe verse limitada por el desconocimiento de un sistema o porque no entiende cómo operar dicho sistema. Cuando se piensa que el Datawarehouse dará mejores resultados si logra llegar hasta el cliente la facilidad toma mayor valor, pues la interfase debe ser tan amigable que cualquier persona en Internet logra acceder sus datos y llamar información para diseñar su propio producto u orden, por ejemplo.

APENDICE I: "*ASEGURANDO LA INTEGRIDAD DE DATOS"*

Introducción

Existen numerosos factores que pueden provocar que un proyecto de Data Warehouse fracase. Desde requerimientos funcionales de negocio juntados en forma pobre hasta la falta de miembros del equipo de trabajo capacitados adecuadamente, ninguna causa es más devastadora para la organización que la falla de los usuarios de aceptar el Data Warehouse debido a que la información que contiene es de calidad cuestionable. Entender porque hay problemas con la información en el Data Warehouse puede ser una tarea desmoralizante para los responsables de resolverlo. El problema puede ser debido a uno o una combinación de los siguientes ítems:

- Corrupción de los datos o falta de datos dentro del sistema fuente de origen,
- La rutina de migración de datos falla al mover los datos que son necesarios por los usuarios finales o el Data Warehouse,
- El proceso de depuración de datos no resuelve las inconsistencias de datos o no verifica correctamente la información seleccionada,
- El proceso de conversión de datos no es construido adecuadamente o contiene reglas de cambio incorrectas, y/o
- El proceso de carga de datos en el Warehouse no es construido apropiadamente o falla.

A continuación se trata el tema de enfocar la integridad de los datos a través del establecimiento de un medio de control que identifica y modera los problemas en los datos dentro de un Data Warehouse.

Integridad de Datos

¿Qué es la integridad de datos?, La integridad de datos se define como la información que se adhiere a un Standard estricto de valor y completitud. Es decir, los datos son precisos y toda la población de datos relevantes está contenida dentro del Data Warehouse.
La credibilidad del Data Warehouse solamente descansa en la integridad de sus datos. De esta manera, entender el impacto de las percepciones es importante para apreciar el requisito de integridad de datos dentro del Data

Warehouse. El uso habitual del Data Warehouse solamente se apoya en la percepción del usuario del Data Warehouse. Las percepciones negativas de los usuarios conducirán a una disminución de su uso y del valor de la organización. Primero y principal, el Data Warehouse debería estar compaginado para establecer certeza tanto al usuario como a los sistemas de información, que los datos tienen integridad. Segundo,
Compaginar significa tratar los temas de usuarios relacionados a la extracción de sus datos del Warehouse.

La perspectiva del usuario final

Durante toda la vida de un Data Warehouse, los usuarios se cuestionarán la información que ellos obtienen de él a menos que el Warehouse haya establecido credibilidad dentro de los usuarios. Durante las semanas posteriores a la puesta en producción de un Data Warehouse, los usuarios preguntarán "¿Puedo confiar en la información del Data Warehouse?". Inclusive, la pregunta de usuario más común que se ha encontrado es "La información en mi query o reporte, ¿es correcta? ". Las respuestas a estas preguntas tendrán un impacto inmenso en las percepciones de los usuarios sobre el Data Warehouse.

La perspectiva IS

Luego de haber terminado la carga inicial o una actualización incremental del Data Warehouse, ¿Cómo se sabe que se cargó toda la información requerida? Se puede ver si el proceso de carga terminó exitosamente o si existen datos en las tablas del Warehouse. Suponiendo que el proceso de carga terminó bien, ¿Cómo se sabe que los datos fueron correctamente cargados? Si uno no tiene la habilidad para consolidar la reacción de que todo esta "OK", ¿Cómo se puede tratar con el usuario del Data Warehouse? Recordar que las decisiones y respuestas a sus preguntas impactarán directamente sus percepciones sobre el Data Warehouse.

Controles de Integridad de Datos

¿Como se puede prevenir el ingreso de datos redundantes, corruptos o sin sentido al Data Warehouse? ¿Como se sabe que se cargó toda la información solicitada? Al establecer un medio que incorpora controles en el flujo de proceso del Data Warehouse se asegura la integridad de los datos. Los controles de integridad de datos se pueden agrupar en dos categorías:

a) Controles de prevención.

b) Controles de detección.

Cada categoría trata a los datos en diferentes etapas del proceso.

a) Controles de prevención.

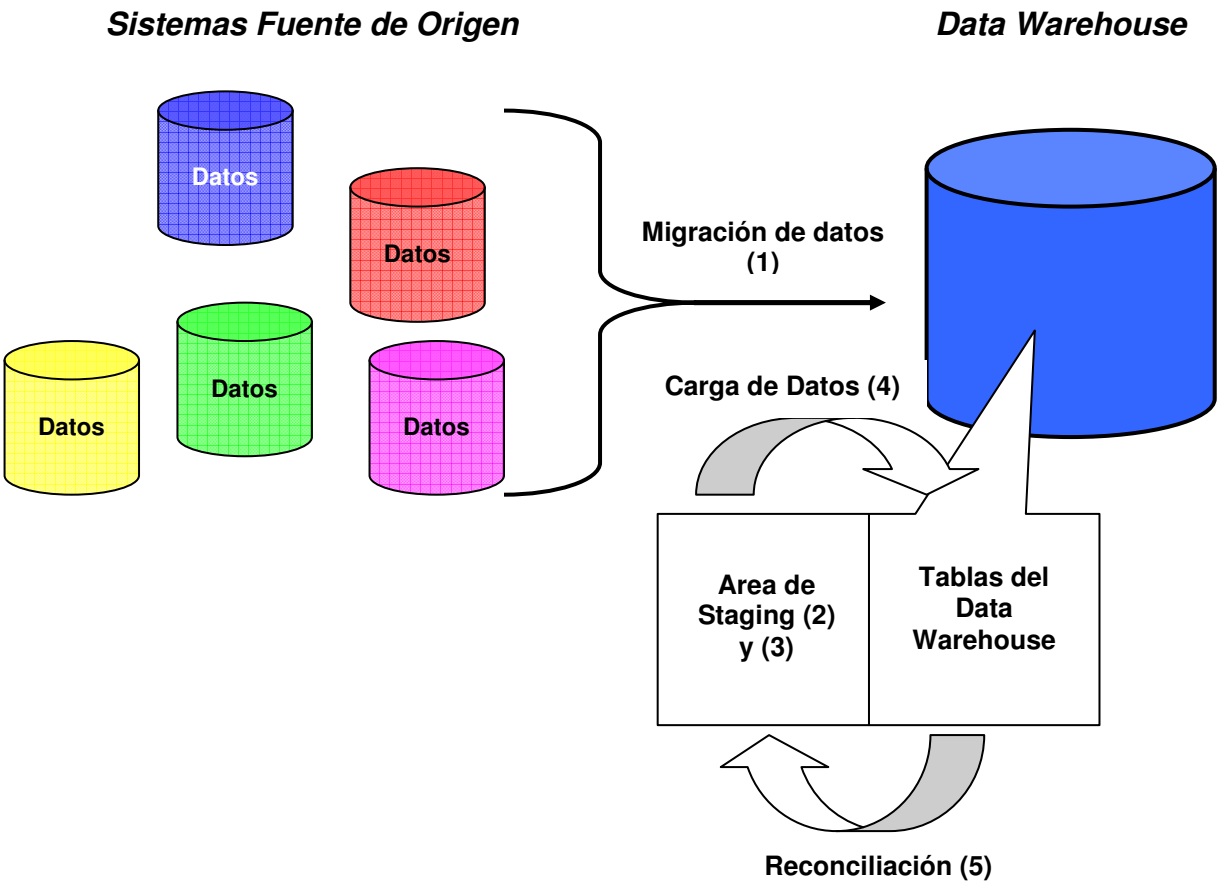
Los controles de prevención ayudan en la integridad de los datos **antes** que se carguen en el Data Warehouse. Estos controles se incorporan en los procesos de migración de datos, depuración, conversión y carga y son los principales medios para prevenir el ingreso de datos redundantes, corruptos o sin sentido al Data Warehouse.

b) Controles de detección.

Los controles de detección son aquellos controles que evalúan la precisión y completitud de los datos **después** que se cargan en el Data Warehouse o en cada etapa del proceso. Estos controles se incorporan en el proceso de reconciliación y son el principal medio de detectar información insuficiente o incorrecta dentro del Data Warehouse.

Puntos de control de los datos

En cada etapa de los procesos de migración de datos, depuración, conversión y carga, existe la oportunidad de asegurar la integridad de los datos **antes** que ingresen al Data Warehouse. Además, el proceso de reconciliación identifica cualquier discrepancia en los datos **luego** que han entrado al Data Warehouse o en cada etapa del proceso.



Etapas del proceso de datos

- Migración de datos
- Depuración
- Conversión
- Conciliación

William H. Inmon, también conocido como el Padre del Data Warehousing, ha estimado que, en promedio, el 80 % del esfuerzo en la construcción de un Data Warehouse se aplica en las etapas del proceso.

Migración de Datos (Control de Prevención)

El propósito de la migración de datos es mover los datos desde los sistemas seleccionados de origen al área de staging del Data Warehouse. Sólo se deberían mover los datos solicitados por los usuarios para la emisión de reportes o aquellos que se utilizan durante los procesos de conversión y carga. Se puede prevenir la información insignificante no ingresando datos innecesarios al Data Warehouse desde los sistemas de origen hacia el área de staging del Data Warehouse.
Los datos que se deberían mover al área de staging del Data Warehouse incluirían datos referenciales y transaccionales. Por ejemplo, en un Data Warehouse de ventas, los datos referenciales se relacionarían con la información del cliente y los datos transaccionales serían la información asociada con la venta a un cliente. La información que típicamente **no** debería moverse al área de staging del Data Warehouse incluye las tablas del administrador del sistema RDBMS y tablas de aplicación de sistema origen que contengan metadatos ó datos de proceso temporarios . Por ejemplo, en una base de datos Oracle, las tablas de diccionario de datos, tales como ALL_TABLES, contienen información que es insignificante para los usuarios finales y no se usaría en los procesos de conversión y carga del Data Warehouse . La siguiente tabla destaca los tipos de datos que deberían migrarse y los que no al área de staging del Data Warehouse.

Contenido de la Tabla	Migrar los datos	No migrar los datos
Datos referenciales del sistema origen : ▪ Requeridos para propósitos de reporte		

▪ No requeridos para propósitos de reporte	*	*
Datos transaccionales del sistema origen : ▪ Requeridos para propósitos de reporte ▪ No requeridos para propósitos de reporte	*	*
Datos del sistema RDBMS		*
Metadatos de aplicación del sistema origen ó datos de proceso temporarios		*

- *Identificar y entender donde se ubican los datos y que datos mover es una tarea difícil. No subestimar los desafíos de esta tarea.*
- *Cuando exista duda, dejarlo afuera a menos que se sepa que se quiere hacer con los datos.*

Depuración de Datos (Control de Prevención)

El fin de la depuración de datos es corregir, para estandarizar el formato, y completar cualquier valor requerido por el Data Warehouse. Este proceso también ayuda a identificar los datos redundantes los que serán prevenidos al ingresarlos en el Data Warehouse durante el proceso de carga. El siguiente es un ejemplo simple de depuración de datos.

Antes de la depuración:

La tabla de información de cliente que aparece abajo contiene datos que están en un formato no Standard. Los elementos de datos FIRST_NAME, LAST_NAME y COMPANY_NAME contienen información que están en un formato inconsistente. Inclusive, el elemento de datos STATE tiene información faltante.

TABLA DE INFORMACION DE CLIENTES

FIRST_NAME	LAST_NAME	COMPANY_NAME	AREA_CODE	PHONE	STATE
SAM	Adams	boston beer co.	617	3685000	MA
Sam	Adams	Boston beer co,.	617	3685000	MA
Samuel	Adams	Boston Beer Co.	617	3685000	
SAMUEL	ADAMS	BOSTON BEER	617	3685000	MA
Martin	Zweig	Zweig Funds	800	2722700	NY

Después de la depuración:

El proceso de depuración estandarizó los datos de la Tabla de Información de Clientes a un formato que ha identificado información redundante, que será excluida del Data Warehouse durante el proceso de carga .

TABLA DE INFORMACION DE CLIENTES

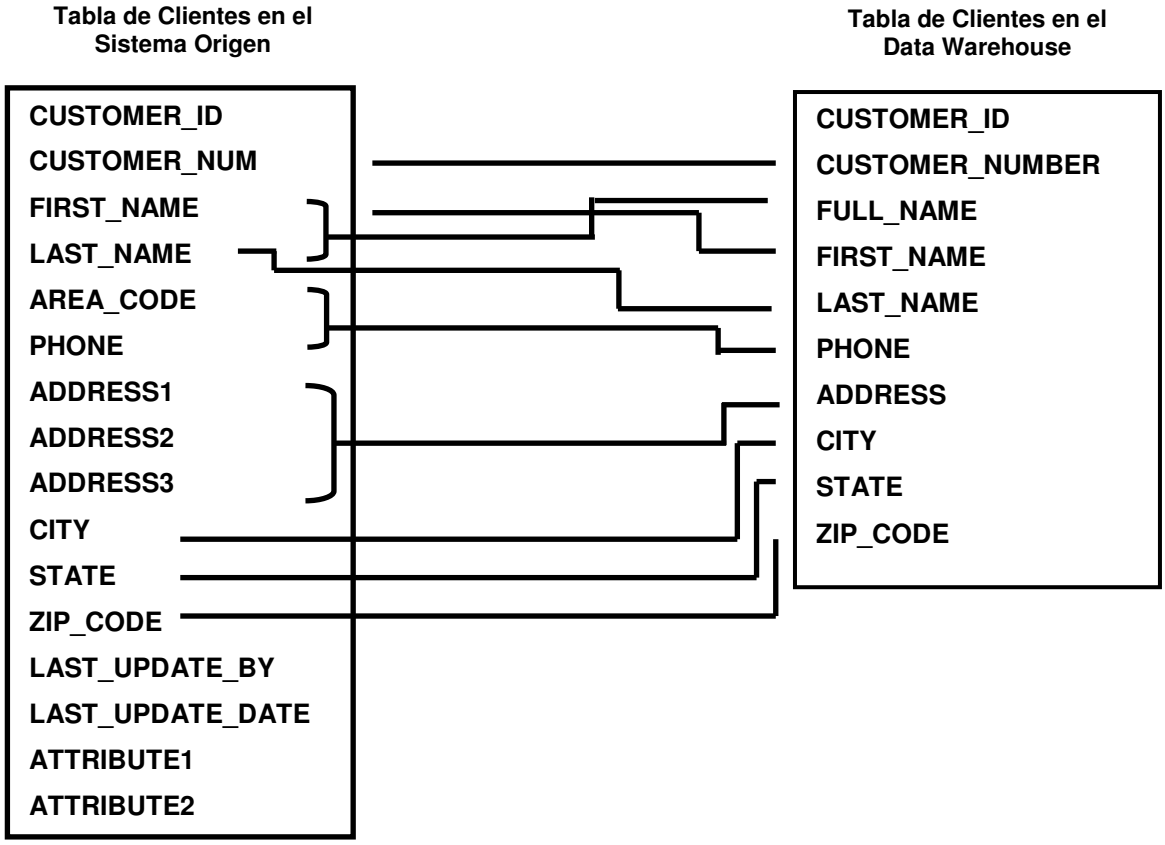
FIRST_NAME	LAST_NAME	COMPANY_NAME	AREA_CODE	PHONE	STATE
Samuel	Adams	Boston Beer Co.	617	3685000	MA
Samuel	Adams	Boston Beer Co.	617	3685000	MA
Samuel	Adams	Boston Beer Co.	617	3685000	MA
Samuel	Adams	Boston Beer Co.	617	3685000	MA

Martin	Zweig	Zweig Funds	800	2722700	NY

- *Utilizar herramientas de software que migren, depuren y conviertan los datos . El retorno de la inversión justifica comprar herramientas de software en vez de desarrollar scripts en SQL . Los costos asociados en mantener y ampliar desarrollos propios de scripts SQL excederá significativamente al de comprar herramientas de software desarrolladas por terceros .*

Conversión de Datos (Control de Prevención)

El objetivo de la conversión de los datos es convertir los datos con el formato y la estructura requeridos por el Data Warehouse . El proceso de conversión debería reducir el número de elementos de datos que se cargan desde el área de staging al Data Warehouse . Al desarrollar las reglas de conversión para este proceso, se utilizan sólo aquellos elementos de datos que se requieren para el Data Warehouse . Si existen elementos de datos que no son necesarios para reportes o carga del Data Warehouse, hay que prevenir el ingreso al Data Warehouse no incorporándolos en las sentencias de conversión ó carga . El siguiente es un simple ejemplo de la correspondencia de datos entre el formato del sistema origen y el del Data Warehouse .



Las reglas de conversión incorporadas en el ejemplo de correspondencia de datos citado anteriormente son como sigue :

Elemento de Datos del Sistema Origen	Regla de Conversión	Elemento de Datos del Data Warehouse
	Secuencia creada en el Data Warehouse .	**CUSTOMER_ID**
CUSTOMER_ID	El elemento de datos no se carga en el Warehouse .	
CUSTOMER_NUM	Correspondencia uno a uno .	**CUSTOMER_NUM**
FIRST_NAME	Correspondencia uno a uno y concatenación con **LAST_NAME** .	**FIRST_NAME** **FULL_NAME**
LAST_NAME	Correspondencia uno a uno y concatenación con **FIRST_NAME** .	**LAST_NAME** **FULL_NAME**
AREA_CODE	Concatenación con **PHONE** .	**PHONE**
PHONE	Concatenación con AREA_CODE .	**PHONE**
ADDRESS1	Concatenación con **ADDRESS2** y **ADDRESS3** .	**ADDRESS**
ADDRESS2	Concatenación con **ADDRESS1** y **ADDRESS3** .	**ADRESS**
ADDRESS3	Concatenación con **ADDRESS1** y **ADDRESS2** .	**ADDRESS**
CITY	Correspondencia uno a uno .	**CITY**
STATE	Correspondencia uno a uno	**STATE**
ZIP_CODE	Correspondencia uno a uno	**ZIP_CODE**
LAST_UPDATE_BY	El elemento de datos no se carga en el Warehouse .	
LAST_UPDATE_DATE	El elemento de datos no se carga en el Warehouse .	
ATTRIBUTE1	El elemento de datos no se carga en el Warehouse .	
ATTRIBUTE2	El elemento de datos no se carga en el Warehouse .	

- *Validar las reglas de conversión con los usuarios .*

Existen dos principales maneras de cargar un Data Warehouse: 1) Renovación completa ó , 2) Renovación incremental . La renovación completa comienza truncando

las tablas en el Data Warehouse y luego cargándolas con todos los datos requeridos . La renovación incremental identifica los cambios que se produjeron en los datos origen desde la última vez que se cargó el Data Warehouse y luego inserta, actualiza ó borra registros de datos en cada tabla del Data Warehouse como se solicite . Ambas alternativas de carga pueden prevenir el ingreso de datos indeseables al Data Warehouse, abarcando : a) condiciones en las sentencias de carga, y 2) índices únicos sobre las tablas destino dentro del Data Warehouse .

Renovación completa

La alternativa de carga con renovación completa puede prevenir que datos no deseados ingresen al Data Warehouse abarcando condiciones en las sentencias de carga . Colocando un amplio rango de condiciones en cada una de las sentencias de carga, se previene que los datos no buscados ingresen al Data Warehouse . Por ejemplo, el Data Warehouse contiene información financiera sólo en dólares estadounidenses . La sentencia de carga debería contener una condición que filtre por información financiera sólo grabada en dólares estadounidenses y prevendría que cualquier información que no sea en dólares estadounidenses ingrese al Data Warehouse . Luego de finalizado el proceso de carga, se colocan índices únicos en cada una de las tablas del Data Warehouse . En caso que existan valores redundantes en los elementos de datos que se definen por los índices únicos, el RDBMS mostrará un mensaje de error ó no se creará el índice único . En tal ocasión, los registros de datos necesitan ser identificados y eliminados dentro del Data Warehouse . Nota: Con propósitos de performance en la sentencia de carga, no deberían existir índices en las tablas destino durante la carga . Luego de terminado el proceso de carga, crear todos los índices necesarios .

Renovación incremental

La carga con renovación incremental puede prevenir que datos no deseados ingresen al Data Warehouse abarcando : 1) condiciones en las sentencias de carga y, 2) índices únicos en las tablas destino dentro del Data Warehouse . De la misma manera que con renovación completa, las condiciones en cada una de las sentencias de carga previenen que datos no buscados ingresen al Data Warehouse . Inclusive, los índices que se colocan en las tablas destino dentro del Data Warehouse prevendrán que los datos redundantes ingresen al Data Warehouse . En cada caso el RDBMS cargará sólo los registros únicos de datos en el Data Warehouse ó visualizara un mensaje de error .

- *A medida que crezca el Data Warehouse, la carga con renovación completa se vuelve intensiva en cuanto a recursos y tiempo . Planificar usar la carga con renovación incremental .*

Conciliación del Data Warehouse (Control de detección)

El proceso de conciliación identifica problemas de datos que al no darles importancia, pasarían los controles de prevención . El proceso de conciliación está diseñado para proveer veracidad así como también identificar los datos que no concuerdan con la información que contiene el sistema origen . Al conciliar los datos se determina la precisión e integridad de la información .

- **Calidad de datos**

La exactitud es evaluada con el uso de totales de control sobre los elementos de datos seleccionados, los que luego son comparados con los resultados anticipados .

- **Cantidad de datos**

La integridad se determina cuantificando el número de registros y comparando los resultados con el número de registros anticipados .
Existen dos principales enfoques para conciliar el Data Warehouse : 1) Conciliación por Fase, ó 2) Conciliación completa . La conciliación de datos por fase tiene lugar luego de cada etapa del flujo del proceso de datos, mientras que la conciliación completa ocurre sólo cuando finaliza el proceso de carga .
Con cualquier enfoque que se use, la conciliación del Data Warehouse proveerá una red segura para identificar las excepciones de datos así como también asistir con preguntas y perspectivas de direccionamiento en todas las partes interesadas dentro de la organización .

Conciliación Completa

Al finalizar cada proceso de carga, se realiza una conciliación completa que compara la información del Data Warehouse con la del sistema origen correspondiente . El reporte que se muestra más abajo es un ejemplo de un reporte que se produce como resultado de una conciliación completa . Notar que el reporte identifica un problema de integridad de datos que se necesitará investigar .

Conciliación por Fase

Se realiza la conciliación luego de cada etapa del flujo del proceso de datos. Este enfoque se utiliza cuando no es factible una conciliación completa debido al número de sistemas de origen ó a la complejidad de los procesos de depuración ó conversión. Con la conciliación por fase, se determinan la veracidad e integridad de los datos luego de cada una de las siguientes etapas :

- **Migración de datos**

  Luego que los datos del sistema origen han sido migrados al área de staging del Data Warehouse, se realiza la conciliación entre los datos del sistema origen y los datos del área de staging .

- **Depuración**

  Cuando termina el proceso de depuración, se realiza la conciliación entre los datos no depurados, el listado de excepciones y los datos depurados del área de staging .

- **Conversión**

  Una vez que finaliza el proceso de conversión, se produce la conciliación entre los datos depurados, la lista de excepciones y los datos convertidos del área de staging .

- **Carga**

  Después de terminado el proceso de carga, se hace la conciliación entre los datos convertidos del área de staging y los datos del Data Warehouse .

Se debería desarrollar un reporte similar al mostrado previamente en la conciliación completa después de cada fase de conciliación .

- *Luego de realizar la conciliación, producir un reporte que se distribuya a todos los individuos que accedan ó estén relacionados al Data Warehouse. El reporte ayuda a establecer credibilidad con el Warehouse . Una vez que se estableció la credibilidad, reducir la distribución del reporte a todas las partes interesadas ó crear un reporte basado en excepciones .*

Conclusión

A medida que los Data Warehouses se convierten en aplicaciones críticas para una organización, la necesidad de asegurar la integridad de los datos aumenta . El éxito de un Data Warehouse descansa con las percepciones de los usuarios de él . Si los datos son imprecisos ó incompletos, la confianza del usuario y el uso del Data Warehouse disminuirán . Al establecer un medio de control que incorpora controles de detección y prevención, se habrán creado los medios para proveer de confianza al usuario y detectar anormalidades en los datos .

APENDICE II : *"Casos de Estudio"*

1. Productora líder de comidas y bebidas

Este caso de estudio corresponde a una productora internacional líder en comidas y bebidas con ingresos mayores a U$S 8.5 billones en 1998. Además de su negocio central, esta compañía obtiene importantes ingresos a través de la operación de una cadena de parques temáticos .
La involucración de la consultora X con este cliente comenzó en 1995 y finalizó en Mayo de 1997, debido a problemas presupuestarios .
El proyecto incluyó la implantación de una red que vincula la compañía con sus mayoristas, la consecución de datos a través de la red y, una vez recogidos, la validación y consolidación de esos datos con los de otros clientes para crear un DataWarehouse focalizado en ventas a minoristas y entregas de productos a mayoristas .

Resumen / Contexto

Previo a la implantación de su red de datos, el cliente no tenía acceso a información sobre ventas de mayoristas a minoristas, y en consecuencia, una pequeña comprensión de la efectividad de sus campañas de marketing, promociones, etc. . Además, no existía información disponible del perfil de los consumidores finales . La implantación de la red por parte de la consultora XXX hizo disponible esta clase de información a muchos grupos funcionales dentro del cliente .

El proyecto tuvo 3 fases :

1. Fase A – Prototipo .

2. Fase B – Recolección y validación de los datos .

3. Fase C – DataWarehouse .

En este caso de estudio sólo se analizará la fase C . El proyecto de DataWarehouse fue administrado usando un plan de trabajo que incluía la extracción, consolidación y depuración de los datos, la población del DataWarehouse, la administración de los metadatos, y el acceso de los usuarios finales .

Aspectos Técnicos

Esta sección incluye lo siguiente :

El Marco de trabajo .

La Estimación del trabajo .
El uso de alguna herramienta de estimación .
El Plan de trabajo y el personal .
El proceso y los criterios para la selección de herramientas .
La arquitectura técnica : proceso de refinación, DataWarehouse y capacidades de acceso para usuarios finales .

Selección de Herramientas

La selección de las herramientas para la consecución, depuración y carga de los datos, administración de los metadatos, y componentes de acceso para usuarios finales, fue realizada en forma separada . Se eligió Prism debido a que pudo integrar la funcionalidad de consecución, depuración y carga de datos es una sola herramienta, a diferencia de otros vendedores como Platinum .
Las herramientas fueron seleccionadas por un pequeño grupo compuesto por dos consultores de XXXX y una persona del cliente. Otros miembros del equipo técnico fueron consultados por conocimientos específicos tales como bases de datos ó redes .
Los responsables por la selección de herramientas para la administración de metadatos y acceso de usuarios finales no fueron los mismos que aquellos que trabajaron con los componentes para la adquisición, depuración y carga de datos .
En todas las evaluaciones, se desarrolló una pequeña lista utilizando información de revistas y recursos de la consultora . Todos los vendedores en esta lista fueron traídos al cliente para demostrar sus productos. Luego se evaluaron las herramientas más importantes con una serie de factores de criterio y peso derivados del equipo de proyecto . La involucración de usuarios finales en esta etapa fue importante para la consideración apropiada de muchas de las comodidades de los criterios de uso . El paso final en la selección de herramientas fue una demostración y una sesión de preguntas y respuestas con personal clave de XXXX y del cliente .

Arquitectura Técnica

Prism fue elegido para la extracción, traducción y carga de los datos, como así también para la administración de los metadatos y *Business Objects* fue usado para el acceso de usuarios finales . El DataWarehouse residía sobre 21 nodos de IBM SP corriendo Oracle Parallel Server. Todos los nodos en esta configuración poseían un solo procesador . Existían 18 nodos pequeños que almacenaban los datos mientras que los otros tres restantes eran amplios para acomodar mejorar los dispositivos de cintas necesarios y la capacidad del disco .

Extracción de Datos

En los casos donde los datos residían en otros nodos de IMP SP, se usó código Cobol generado por *PRISM* para extraer los datos en archivos planos. Los datos que no se tenían que mezclar con los datos de SP también fueron extraídos utilizando código Cobol generado por *PRISM* y se enviaron al mainframe vía FTP usando un proceso construido por el cliente . Los datos del mainframe que pudieron ser cargados directamente al Warehouse serán tratados en la sección *Carga de Datos* .

Depuración de Datos

Muchos de los datos que se cargaron en el DataWarehouse venían de otras aplicaciones de red que se estaban construyendo en el mismo momento o recién terminadas . En esos casos, los datos residían en uno o más nodos del IBM SP. La

depuración de estos datos fue incorporada en las otras aplicaciones de red y no se detallarán aquí. En resumen, sin embargo, las principales verificaciones de estos datos fueron comparaciones automáticas entre los datos y reporte corporativos existentes del la misma fuente de origen (principalmente mayoristas). Las verificaciones fueron realizadas por la noche y sólo después de 3 meses consecutivos de 'limpieza' de datos por parte de un comerciante mayorista hizo que el proceso de verificación se volviera menos detallado .
Otros datos para el Warehouse venían de los sistemas de mainframe. En todos los casos, estos datos eran comparados con otros relacionados en el origen antes de que se los considere limpios . Se incorporó una regla de verificación interna de negocios en el código *PRISM ETL* . Cuando se combinaban muchos orígenes en el DataWarehouse, se designaba un origen primario. Cuando aparecía un conflicto de datos, se informaba en un archivo separado y los registros individuales no se cargaban en el Warehouse. Dado que principalmente los datos de Dimensión se llevaron al Warehouse de esta manera, la carga de datos de Hecho fue impactada por la decisión de guardar registros de Dimensión .
Un pequeño grupo de administradores de datos con conocimientos funcionales trabajó sobre temas de evaluación de consistencia de datos . Estos individuos hicieron manualmente los cambios de datos necesarios para permitir que los registros inconsistentes se cargaran la siguiente noche – un proceso requerido por menos de 10 registros por semana . Considerando que aproximadamente un millo de nuevos registros de ventas se llevaban al sistema por la noche, esto parecía aceptable. En la mayoría de los casos, se rastreaban las discrepancias en los datos , por tipeado impropio del personal de ingreso de datos – principalmente en las áreas de mantenimiento de clientes . Solo en el caso de nuevos clientes una demora en la creación de los datos de dimensión del cliente afectaría la habilidad del Warehouse para aceptar datos de venta. Dado que los clientes generalmente residían en los sistemas legacy mucho tiempo antes que se llevaran las órdenes, no tuvo impacto para el DatWarehouse de ventas .
En este sistema se dejó habilitada la integridad referencial para asegurar que los registros de hecho sean creados con referencias a dimensiones válidas .

Carga de Datos

Existían distintas maneras de cargar el DataWarehouse .
En algunos casos, los datos pudieron cargarse luego de la depuración en el sistema de origen . Cuando era posible, se usaba el hardware del sistema origen para disminuir la sobrecarga del IBM SP . Se necesitó un administrador de Oracle MVS para ejecutar operaciones SQLLOAD directamente desde el mainframe IBM hacia el DataWarehouse. Permitió que un archivo plano en el mainframe se pase con un SQL*NET directamente a la base de datos Oracle. Se lograron de esta manera las cargas directas teniendo mínimo impacto en la utilización de la CPU del IBM SP .
En la mayoría de los casos, sin embargo, los datos a colocarse en el DataWarehouse estaban en el IBM SP con propósitos de depuración o debido a que el sistema origen residía allí . Los datos se dividieron en dos variedades: altas y actualización (no se realizaban borrados en este sistema). El producto *Syncsort* se utilizó extensivamente, para mezclar los datos de múltiples orígenes en la fase de depuración, y para separar datos de alta y actualización una vez que estaban listos para cargarse en el Warehouse. Los datos de alta se cargaban en la tablas apropiadas utilizando SQLLOAD, mientras que las actualizaciones al sistema se realizaban usando código Cobol generado por *Prism* .

DataWarehouse

Como se dijo antes, el DataWarehouse mismo estaba en una base de datos Oracle Parallel Server residiendo sobre 18 nodos de un IBP SP . Un total de 600 Gb de datos se incluyeron en los planes aunque a fines del año 1999, se requerirá espacio de disco adicional .
Se usó un diseño de base de datos estrella – copo de nieve . En todos los casos que se usaban copos de nieve, los datos de dimensión se forzaban a la dimensión 'principal'. Esto permitió el acceso a las tablas de hecho utilizando consultas con un solo nivel de join . La redundancia que creo esto se consideró aceptable debido a los pequeños volúmenes comparados con los datos de hecho .
Las tablas clave de hecho (Ventas y Entregas) se construyeron utilizando vistas particionadas. En cada caso, se crearon tablas separadas para cada uno de los 25 meses (los 24 meses pasados y el actual) y se creó una vista Union All que mezcló todas las tablas . Las altas y modificaciones tenían lugar en el nivel individual de tabla, permitiendo que se realizara un back up de tablas específicas. El procesamiento nocturno monitoreaba que tablas experimentaban cambios en los datos durante la ventana batch y solo a esas tablas se les realizaba back up . Como cerca de un millón de líneas de datos de facturación se cargaban en el Warehouse por la noche, es evidente que a 25 tablas de datos de ventas que contenían 22 millones de registros cada una era difícil realizarles un back up nocturno .

Experiencia Operacional

Se comprobó que el DataWarehouse era muy adaptable . Cuando se accedía usando las opciones de consulta paralela de Oracle, se demostró que cuantos más nodos se utilizaban, se incrementaban los volúmenes de datos y la performance se mantenía muy consistente .
Se experimentaron algunos problemas de performance cuando se utilizaban opciones de StarQuery con tablas que se habían creado como vistas particionadas. En algunos casos, Oracle admitió que la consulta Star no trabajaría como lo hacía antes de la versión 8.0 . En la mayoría de los casos, sin embargo, el uso de Star Query fue muy beneficioso cuando se accedía a las tablas de hecho.
Como se mencionó antes, se dejó habilitada la integridad referencial de Oracle para el DataWarehouse. Esto se hizo principalmente para asegurar que los datos de hechos eran siempre consistentes con la información dimensional actual. La sobrecarga por usar RI fue considerada aceptable dado que las ramificaciones de datos inconsistentes eran severas .
No se permitía el mantenimiento de los datos en línea en el Warehouse. Los cambios en los datos siempre venían a través del procesamiento nocturno. Esto aseguró que los datos eran modificados en el origen y previno 'repetir' inconsistencias .

Lecciones aprendidas

El plan original para la construcción de la infraestructura del DataWarehouse fue armado considerando poco personal y poco presupuesto . El trabajo actual lleva considerablemente más de lo anticipado originalmente. Parte de esto se debe al cliente y al proceso de selección de herramientas.
El diseño de la base de datos misma es el único aspecto más importante de un proyecto como éste .

La planificación y consideración temprana de procesos de back-up es una necesidad, especialmente cuando los volúmenes de datos que se tratan son mucho más grandes que la mayoría de los sistemas .
Ninguno de los paquetes de software ETL hizo tanto como se dijo. Aún se requiere codificación por parte del cliente .
En la selección de la herramienta para el acceso de usuarios finales, es imperativo tener un usuario final involucrado. El personal tecnológico no está completamente calificado para escuchar en lugar de los usuarios finales.

2. Unión Fenosa (España)

Datos del proyecto

La actividad de Unión Fenosa en sus oficinas de Madrid, es la de una empresa eléctrica. Debido a los cambios sufridos en este sector, el departamento de Marketing tenía una problemática que resolver para adaptarse a las demandas de este mercado.
Para desarrollar la actividad del departamento de Marketing, se necesitaba disponer de una información más completa y exacta de la base de clientes (datos sociales, económicos, históricos y estadísticos) y poder obtenerla de una forma rápida y ágil.
Una de las premisas ha sido la competencia y la liberalización del mercado energético, que iba más deprisa de lo que en un principio se creía. Otra, la fidelización de los clientes, la generación de otros nuevos, así como tener que pasar de un sistema de costes estándar a otro basado en precios de mercado, lo que ha supuesto uno de los mayores cambios en el sector.
Decidieron cambiar el perfil del usuario y que éste pasara de ser un mero actor dedicado a resolver transacciones, a conseguir extraer todo el rendimiento de la información para así disponer de una eficaz ayuda en la toma de decisiones. Esta labor es muy importante para el departamento de Marketing, ya que en los últimos años el negocio eléctrico está en continuo cambio, tanto en el ámbito nacional como en el internacional y la Compañía ha de estar preparada para entrar rápidamente en un mercado más competitivo y más orientado a la demanda de los clientes.

Por todo esto, se decidió estudiar la posibilidad de implantar un proyecto de Data Warehouse a nivel departamental. Para ello, se encargó a la empresa de servicios informáticos NorSistemas, propiedad del grupo Unión Fenosa, que de acuerdo con estas necesidades, estableciera un proyecto para la creación de un Data Warehouse destinado al departamento de Marketing (Datamart).
NorSistemas identificó las necesidades que el sistema existente no podía cubrir y propuso una solución departamental en tres niveles. Esto permitía disponer de una base de datos exclusiva para el departamento, en la que se podía efectuar todo tipo de consultas para el perfecto funcionamiento del negocio. Una vez identificadas las necesidades funcionales, se procedió a la búsqueda de las soluciones que mejor podían adaptarse a la problemática del departamento.
NorSistemas se hizo cargo de la selección preliminar de las herramientas necesarias y así cubrió las diferentes áreas para la creación de un Data Warehouse: gestor de base de datos, herramientas de extracción de la información y "minería" de datos. Una vez concluida esta fase previa de selección, se implantó un proyecto piloto con los tres proveedores finalistas. Con este proyecto, se pretendía instalar toda la base de datos de Marketing con las herramientas seleccionadas a pleno rendimiento.
Tras cuatro meses, en abril de 1.997, el proyecto piloto ya estaba en funcionamiento. El proceso de evaluación de las diferentes herramientas se ha llevado a cabo con la participación de los especialistas de sistemas de Unión Fenosa, los usuarios potenciales del Data Warehouse, el personal de administración y el de Marketing. Este grupo decidió qué herramientas y qué hardware era el más idóneo para el proyecto.
La implantación a nivel departamental ha supuesto una decisión muy acertada porque la puesta en marcha del proyecto ha sido mucho más rápida, fácil de realizar y ha permitido estudiar mejor el funcionamiento de la aplicación.
El éxito en la implantación y correcta utilización del Data Warehouse ha sido obra de todos. El uso, a nivel departamental, de ese tipo de herramientas, se ha llevado a cabo de forma gradual y los usuarios finales se han ido acostumbrando a utilizarlas, consiguiendo optimizar sus procesos de trabajo. Esto ha permitido que varios departamentos de Unión Fenosa hayan comenzado a implantar su Data Warehouse.
Unión Fenosa ha sido pionera en establecer un Data Warehouse dentro de su sector de actividad. La Compañía lleva varios años teniendo en cuenta que el usuario, al que se abastece de luz, es un "cliente" y éste es un concepto difícil de mantener en una empresa eléctrica. La idea se tenía pero no la tecnología apropiada para poder llevar a cabo este cambio de rol.
Un beneficio a tener en cuenta cuando se decide la implantación de un Data Warehouse con una arquitectura de 3 niveles, es la ventaja de poder ir creciendo gradualmente en la aplicación de acuerdo con la demanda del negocio. Desde la puesta en marcha del proyecto inicial hace algunos meses, el equipo hardware de Sun se ha ampliado en varias ocasiones, tanto en capacidad de disco como en procesadores.

Solución

Los sistemas hardware instalados son dos servidores Sun Enterprise 3000 con dos procesadores. Son unos equipos potentes con una disponibilidad sin precedentes y una excelente relación precio/rendimiento. Estos servidores proporcionan el rendimiento y la fiabilidad de los grandes sistemas, siendo un equipo versátil y apropiado para aplicaciones críticas de negocio. Soporta hasta 6 procesadores, redes de banda ancha, discos de alta velocidad de entrada/salida y hasta 6 GB de memoria principal.
Todos los servidores de Sun incorporan el sistema operativo Sun Solaris, un sistema abierto basado en Unix que garantiza la compatibilidad binaria de todas las aplicaciones y soporta todas las plataformas, ya sean Intel o Risc. Tiene hasta 12.000 aplicaciones. Solaris es muy eficaz y asegura el crecimiento continuado, al ser el mismo sistema operativo para todos los sistemas de Sun, desde los de gama baja hasta los grandes servidores.
La formación se ha llevado a cabo con Sun Enterprise Services para cubrir las necesidades del manejo de las herramientas y se considera que Sun proporciona un servicio de gran calidad. Asimismo, a nivel de mantenimiento, se utiliza a Sun Enterprise Services para las aplicaciones críticas con un contra-to de 24 horas y 7 días por semana.
Las principales herramientas sobre las que el departamento de Marketing ha construido su almacén de datos inteligentes han sido:

Business Objects: herramienta integrada de consulta para la creación de informes y análisis OLAP. Combina en un único producto la funcionalidad con capacidades de consulta y gene-ración de informes. Las novedades que incluye la nueva versión son que permite a los usuarios realizar análisis sobre los informes, acceder directamente a los datos, utilizar formatos de reporte predefinido y, además, proporciona capacidad de navegación para visualizar informes a través del Web.

Data Mining: Minería de Datos de SAS Institute. El sistema dataminig para el suministro de información transforma los datos para su posterior análisis e interpretación, facilitando la toma de decisiones. Permite combinar datos y aplicaciones residentes en distintas plataformas hardware, a través de utilidades de conectividad y proporciona acceso a estos datos desde una gran variedad de plataformas operativas.

ETI-Extract Tool Suite: solución para la expansión y movimiento de los datos entre distintos sistemas de entornos informáticos heterogéneos.

El Gestor de Base de Datos y el conjunto de Herramientas de Oracle: El Gestor de Base de Datos y el conjunto de Herramientas de Oracle: las aplicaciones están integradas con la tecnología de la base de datos y responden a los requisitos de adaptabilidad y rendimiento necesario, permitiendo el crecimiento y aportan-do un diseño flexible de interfaces abiertas.

El proyecto se ha realizado con una estructura de 3 niveles federada y soluciones monotemáticas que se consolidarán en un futuro. Los niveles son:

- Mainframe como repositorio de los datos
- Plataforma Sun basada en Unix
- Herramientas de software y PC's

La decisión de elegir a Sun para este proyecto de Data Warehouse no fue consecuencia de la utilización previa de sus estaciones de trabajo en puestos de carácter técnico, sino que se efectuaron una serie de pruebas con equipos de otras compañías. Los servidores corporativos de Sun convencieron y se eligieron para el proyecto de Data Warehouse, siendo los primeros servidores que se instalaron para un entorno corporativo.

Se ha conseguido saber con qué información contaban y conocer como extraer el máximo rendimiento para la perfecta marcha del negocio. Además, en términos económicos, han recuperado la inversión realizada. En términos generales, los resultados han sido excelentes. A la vista de los mismos, la inversión en tecnología en otras áreas de la Empresa está creciendo y otros departamentos están interesados en la implantación de un proyecto similar. Según han confirmado los usuarios del sistema, se puede afirmar que los objetivos se han cumplido en su totalidad.