BI: Inteligencia aplicada al negocio

Artículo escrito por Oscar Alonso Llombart en exclusiva para CMS-Spain.com

Autor: *Óscar Alonso Llombart (oalonsol@uoc.edu)*

Introducción

En una empresa es necesario tomar decisiones día a día (algunas muy estratégicas) que están basadas en información generada en base a datos reales. Generar esa información es la labor de los sistemas de Inteligencia en el Negocio (en inglés, *Business Intelligence*), en adelante BI (ver glosario).

Toda toma de decisiones implica aceptar un riesgo, lo que es indudable es que el objetivo es minimizar ese riesgo. Aquí es donde entran en juego las herramientas de BI. Ellas son las encargadas de transformar los datos corporativos de nuestro sistema de *backoffice* en información.

El objetivo de este artículo es dar una introducción al concepto de BI y a algunas de sus fases a través de un proyecto de este tipo en el que se optó por las técnicas OLAP (*Online Analitical Processing*) para el análisis de la información y para el que se utilizó la metodología CRISP-DM. Las fases cubiertas en el presente documento son:

- Preparación de los datos.
- Modelado.
- Evaluación.

Los sistemas de información se dividen de acuerdo al siguiente esquema, representado en la Figura 1:

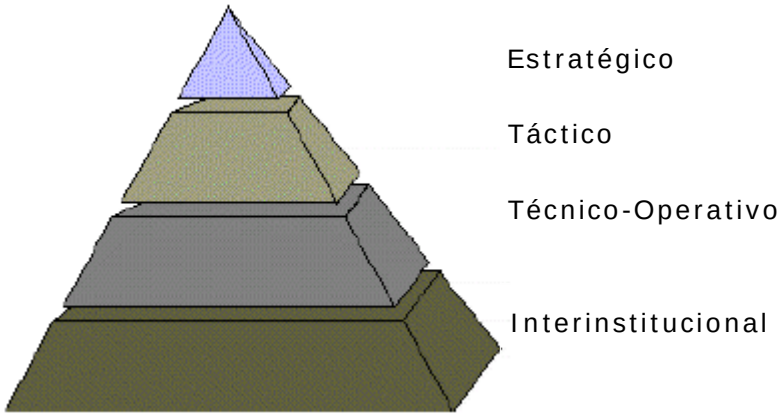

FIGURA 1 - TIPOS DE SISTEMAS DE INFORMACIÓN

- Sistemas interinstitucionales, este nivel de sistemas de información está surgiendo recientemente, es consecuencia del desarrollo organizacional orientado a un mercado de carácter global, el cual obliga a pensar e implementar estructuras de comunicación más estrechas entre la organización y el mercado (empresa extendida, organización inteligente e integración organizacional), todo esto a partir de la generalización de las redes informáticas de alcance nacional y global (internet), que se convierten en vehículo de comunicación entre la organización y el mercando, no importa dónde esté la organización (intranet), el mercado de la institución (extranet) y el mercado (internet).

- Sistemas técnico-operativos, que cubren el núcleo de operaciones tradicionales de captura masiva de datos (*data entry*) y servicios básicos de tratamiento de datos, con tareas predefinidas (contabilidad, facturación, almacén, presupuesto, personal y otros sistemas administrativos). Estos sistemas están evolucionando con la irrupción de censores, autómatas, sistemas multimedia, bases de datos relacionales más avanzadas y *datawarehousing*.

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

- Sistemas estratégicos, facilitan la labor de la dirección, proporcionándole un soporte básico, en forma de mejor información, para la toma de decisiones. Se caracterizan porque son sistemas sin carga periódica de trabajo, es decir, su utilización no es predecible, al contrario de casos anteriores, cuya utilización es periódica.

El objetivo es generar información que pueda ser trabajada y analizada de forma intuitiva, más o menos en tiempo real, y con la posibilidad de integrar diferentes fuentes de datos para ofrecer una visión global que puede ser compartida y distribuida por todos los departamentos de la empresa.
No es que todo el mundo tenga acceso a toda la información, se trata de que todo el mundo tenga acceso y genere la información que necesita para que su trabajo sea los más eficiente posible.
Aunque este tipo de herramientas llevan existiendo desde hace años, la mayoría han sufrido importantes procesos de reingeniería para permitir el despliegue de sus funcionalidades a través de internet.

La evolución última de estos productos ha hecho que estén preparados para poder generar un portal como acceso a la información desde donde todos los usuarios de la empresa podrán realizar procesos de consulta (*query)*, análisis, informes, gráficos, entre otros. Estamos hablando de un portal, que puede estar totalmente integrado con sistemas de gestión del conocimiento.
De esta forma, la organización podrá crear un modelo común de información que soporte fuentes de datos heterogéneas y múltiples aplicaciones, tanto del interior como del exterior de la organización.

El concepto de BI es lo suficientemente amplio como para que entremos a determinar en detalle los componentes que en él se engloban. En este sentido aquí debemos considerar desde entornos operacionales para la captación de datos (y esto implica la recogida de datos de diferentes canales), pasando por contenedores de estos datos, y acabando con las herramientas que trabajan para generar la información a partir de los datos.

En la Figura 2 se muestra un diagrama de cada uno de estos componentes.

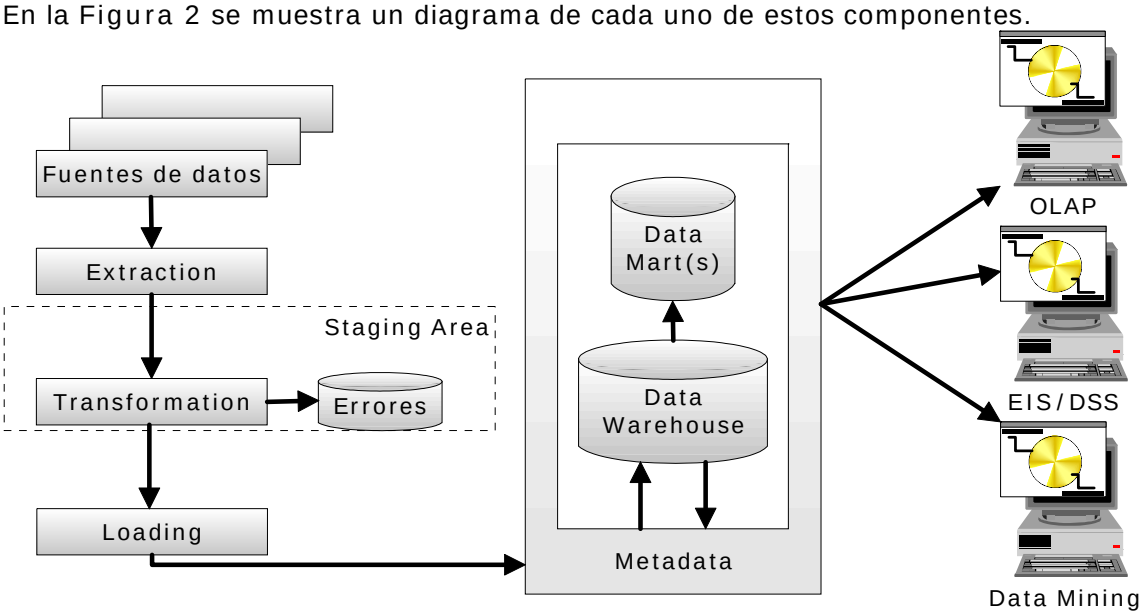

FIGURA 2 - COMPONENTES DEL ENTORNO BI

A partir de los sistemas operacionales de una compañía se extrae la información relevante y significativa, se limpia, ordena e introduce en un centro de información especial, a partir del cual los responsables departamentales y ejecutivos analizan la marcha del negocio utilizando alertas, semáforos o mecanismos de análisis interactivo.

Es necesario limpiar e integrar los datos que proceden de distintas fuentes. Esta función, soportada por las herramientas de Extracción, Transformación y Carga (en inglés *Extraction, Tranformation and Loading*), en adelante ETL (ver glosario). Ver Figura 2.

Una vez que se han obtenido los datos, es necesario guardarlos en un contenedor eficaz Esta función está cubierta por los gestores de bases de datos que realizan la carga de los mismos, los aceptan comprobando su integridad, los interrelaciona y posteriormente procesa las peticiones realizadas por las consultas del usuario obteniendo como resultado el *datawarehouse ó un datamart* (ver glosario).

Es el usuario final, quien dependiendo de sus requerimientos, define el tipo de herramientas de análisis a utilizar, desde el simple *query & reporting,* que realiza informes predefinidos, hasta la posibilidad de analizar información mediante *queries ad hoc* (OLAP), la implementación de técnicas de minería de datos que permiten encontrar normas de comportamiento en los conjuntos de datos analizados, para realizar, por ejemplo, clasificaciones o predicciones, o generar información para los EIS (*Executive Information Systems*), DSS (*Decision Support Systems*, Sistemas de ayuda a la toma de decisiones) ó el BSC (cuadro de mandos integral).

Datos, información y conocimiento

La tecnología de los almacenes de datos ó *datawarehouses*, se encuadra dentro de la línea de evolución de las bases de datos hacia una mayor funcionalidad e inteligencia.
Las empresas han visto aumentada su capacidad de generar y recoger datos (introducción de internet en las empresas, tecnologías de entrada de datos...).
Estas grandes cantidades de datos (obtenidas a un coste relativamente bajo) no aportan información a las organizaciones.

"Una organización puede ser rica en datos y pobre en información, sino sabe como identificar, resumir y categorizar los datos" (Madnick, 1993)

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o construyen un modelo, haciendo que la interpretación del confronto entre la información y ese modelo represente un valor añadido, entonces nos referimos al conocimiento. En la Figura 3 se ilustra la jerarquía que existe en una base de datos entre dato, información y conocimiento [MOL02]. Se observa igualmente el volumen que representa en cada nivel y el valor que los responsables de las decisiones le dan en esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento.

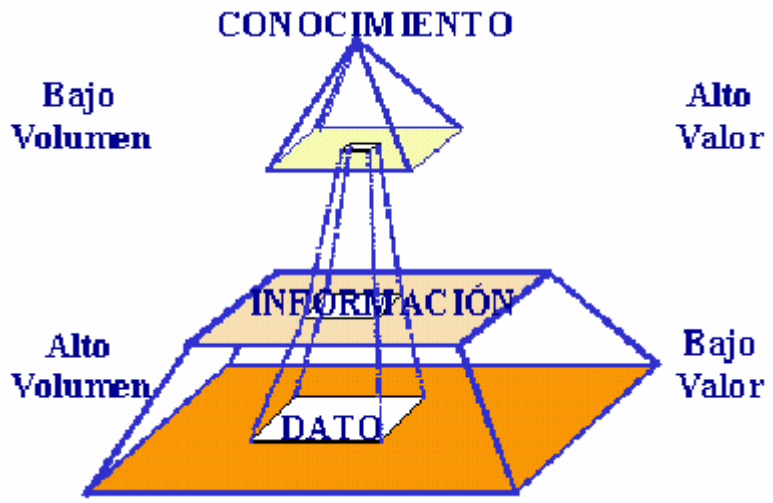

FIGURA 3 – RELACIÓN ENTRE DATO, INFORMACIÓN Y CONOCIMIENTO

El objetivo de este tipo de proyectos es transformar los datos en información que nos pueda resultar útil para el conocimiento del negocio que se encuentra disperso en la organización y ofrecer las herramientas necesarias para el análisis de datos de la información.
El almacén de datos pretende dar un soporte a la organización para proporcionarle una buena gestión de sus datos, que le ayude en la toma de decisiones estratégicas y tácticas.

Lo que se busca es poder gestionar la información para que en una primera instancia mejore el proceso actual, transformando los datos orientados a las aplicaciones en información orientada a la toma de decisiones.

Con el uso de estos almacenes se ahorra mucho tiempo en el preprocesamiento de los datos al construirse mediante la integración de fuentes de datos múltiples y heterogéneas (como puedan ser las bases de datos relacionales, ficheros planos, registros de transacciones *online*...) y al aplicarse sobre él técnicas de limpieza e integración se asegura la consistencia en el nombrado, estructuras codificadas, medidas de los atributos y demás aspectos entre las múltiples bases de datos que lo integran se asegura la calidad de la información.

¿Por qué datawarehouses y datamarts?

El diseño de las bases de datos transaccionales sobre las que corren los aplicativos no está orientado a la extracción de la información [FER02].
Sería posible trabajar con la base de datos transaccional, aunque nada recomendable, el tiempo de respuesta incidiría sobre el rendimiento del sistema puesto que estas bases de datos han sido diseñadas para una escritura y modificación intensiva, no para su lectura.
Así el diseño de las bases de datos transaccionales y las desarrolladas para abordar proyectos de BI son totalmente distintos.
Las bases de datos de BI (*datawarehouses ó datamarts*) están específicamente diseñadas para abordar consultas, por lo que son capaces de recoger datos de diferentes aplicativos y homologarlos en un repositorio central, todo ello con la ayuda de las herramientas ETL.

Tomar los datos desde varias bases de datos operacionales y transformarlos en datos requeridos para el depósito, se refiere a la transformación o a la integración de datos. Las bases de datos operacionales, diseñadas para el soporte de varias aplicaciones de producción, frecuentemente difieren en el formato.
Los mismos elementos de datos, si son usados por aplicaciones diferentes o administrados por diferentes software DBMS, pueden definirse al usar nombres de elementos inconsistentes y/o ser codificados de manera diferente. Todas estas inconsistencias deben resolverse antes que los elementos de datos sean almacenados en el *datamart*.

"Se llama datawarehouse al almacén de datos que reúne la información histórica generada por todos los distintos departamentos de una organización, orientada a consultas complejas y de alto rendimiento. Un datawarehouse tiene una orientación corporativa que pretende conseguir que cualquier departamento pueda acceder a la información de cualquiera de los otros mediante un único medio, así como obligar a que los mismos términos tengan el mismo significado para todos. Un datamart es un almacén de datos históricos relativos a un departamento de una organización, así que puede ser simplemente una copia de parte de un datawarehouse para uso departamental. Tanto el datawarehouse como el datamart son sistemas orientados a la consulta, en los que se producen procesos en lote (batch) de carga de datos (altas) con una frecuencia baja y conocida." [HC03].

Muchos almacenes de datos comienzan siendo *datamarts* (para minimizar riesgos) y se va ampliando su ámbito ya que estos están centrados en un tema concreto y están diseñados para una unidad de negocio específica. La implementación incremental reduce riesgos y asegura que el tamaño del proyecto permanezca manejable en cada fase.

Otro paso necesario es crear los metadatos (es decir, datos acerca de datos que describen los contenidos del almacén de datos). Los metadatos consisten en definiciones de los elementos de datos en el depósito, sistema(s) del (os) elemento(s) fuente. Cómo los datos se integran y transforman antes de ser almacenados en información similar, es necesario contar con un diccionario donde se explique el contexto y su procedencia.

Con un pequeño entrenamiento en la estructura del almacén y en las posteriores técnicas que se aplicarán sobre él, el usuario tendrá un único canal de dónde extraer la información y no cómo ahora en que el personal dedicado a estas tareas se ve saturado de diferentes reportes y consultas, a menudo realizados por personal que ya no está en la empresa y del que tienen que ir cuadrando y extrayendo información “artesanalmente”. Permitiendo así un análisis inmediato de los resultados y conectando departamentos empresariales (que antes formaban islas).

Preparación de los datos

La *staging area* ó preparación de los datos, es una colección de procesos que limpian, transforman, combinan, y preparan los datos originales para su utilización en el *datawarehouse*. En la *staging area*, los datos originales son transformados a formatos comunes, comprobada su consistencia y su integridad referencial, y preparados para cargar en la base de datos del *datawarehouse*.

Una vez localizadas las fuentes de datos, estos se han de preparar para que se les puedan aplicar los métodos o herramientas que construirán el modelo deseado. Esta fase aunque parezca sencilla conlleva aproximadamente el 70% del esfuerzo en los proyectos de *data mining* de nueva implantación.

En este punto hay que asegurarse de unas cuantas cosas:

- Que los datos tengan la calidad suficiente: es decir, que no contengan errores, redundancias o que presenten otro tipo de problemas.
- Que los datos sean los necesarios, quizás haya que no nos harán falta y quizás tendremos que añadir.
- Que están en la forma adecuada: muchos métodos de construcción de modelos requieren que los datos estén en un formato determinado que no ha de coincidir necesariamente con el que están almacenados.

Las técnicas utilizadas para asegurar los tres aspectos comentados son la limpieza de datos, la transformación de los datos y la reducción de la dimensionalidad.

- La limpieza de datos, consiste en procesar los datos eliminando los atributos que sean erróneos o redundantes, siendo los factores de distorsión más importantes:
    1. Datos incompletos, puede pasar especialmente en aquellos atributos en que cuando se diseñó el proceso correspondiente a la entrada de datos se decidió que no eran obligatorios o que tenían formato libre.
    2. Datos redundantes, repetición de tuplas.
    3. Datos incorrectos o inconsistentes, muy común cuando el tipo de valores que puede recibir un atributo no está controlado porque está declarado como "texto libre".
- Transformación de datos, no siempre los datos están en la forma más adecuada para poder aplicar los métodos que hacen falta para la tarea que se ha de llevar a cabo y el modelo que se quiere obtener.
- Reducción de la dimensionalidad, una de las justificaciones más frecuentes para la utilización de técnicas de *data mining* es su capacidad para trabajar con grandes conjuntos de datos. Ahora bien el tamaño de un conjunto de datos, o de un problema de *data mining*, la da tanto la cantidad de registros que tiene como el número de atributos que se manejan.

Lo que pasa es que a partir de ciertos niveles de registros y atributos, la eficiencia de los algoritmos de *data mining* comienza a reducirse. Lo que se busca es reducir el número de atributos y registros sin que el modelo sufra una degradación en su predicción o representación.

Modelado del almacén

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas y sus parámetros son calibrados a valores óptimos.
Existen varias técnicas para los mismos problemas, algunas de estas técnicas tienen requerimientos específicos en el formato de los datos, por lo que puede ser necesario el paso atrás a la fase de preparación de los datos.

Como primer paso se selecciona en función del problema la técnica a utilizar, en el caso que nos ocupa sobre el almacén construido se aplicarán técnicas de análisis en línea de la información u OLAP.

Antes de construir el modelo, necesitamos generar un procedimiento o mecanismo para probar la calidad y validez del modelo.
Una vez construido el modelo el diseñador interpreta el modelo de acuerdo a sus conocimientos del dominio y al criterio de éxito del proyecto.
Esta tarea interfiere con la fase de evaluación siguiente, el diseñador contacto con los analistas del negocio y expertos del dominio para discutir los resultados en el contexto del negocio.

Métodos OLAP

Los métodos OLAP [CCS93] surgieron para analizar los datos de ventas y marketing, así como para procesar datos administrativos y consolidar datos procedentes de diversas fuentes de cara a efectuar un análisis de rentabilidad, mantenimiento de la calidad y otros tipos de aplicaciones que se caracterizan porque redefinen de manera continua y flexible el tipo de información que hay que extraer, analizar y sintetizar (en comparación con las bases de datos tradicionales, dirigidas a responder consultas muy prefijadas y rutinarias).

Los sistemas OLAP se alimentan de los datos generados por los sistemas transaccionales (facturación, ventas, producción,...). Herramientas típicas de OLAP son las que permiten un análisis multidimensional de los datos en contra de las típicas facilidades de creación de resúmenes e informes propios de los sistemas de bases de datos tradicionales.

Desde un punto de vista conceptual, OLAP tiene las cuatro siguientes funcionalidades [GL96]:

- *Querying*: posibilidad de generar consultas potentes a través de una interfase simple y declarativa.
- *Restructuring*: capacidad de reestructurar información en una base de datos multidimensional explotando la dimensionalidad de los datos y mostrando diferentes perspectivas de los datos.
- *Classification*: capacidad de clasificar ó agrupar conjuntos de datos en una manera apropiada para la consiguiente sumarización.
- *Summarization/consolidation*: esta es una generalización de los operadores de agregación en SQL estándar. En resumen, la sumarización mapea conjuntos de valores de tipo numérico a un único valor “consolidado”.

La unidad de datos de OLAP es el cubo, que es una representación de los datos de interés para el análisis. Posteriormente, hablaremos en mayor detalle de cada una

de las operaciones que permiten “cortar” y mirar los cubos desde la perspectiva de muchos grupos diferentes de usuarios.

La característica principal de los cubos es que optimizan las consultas. Normalmente se guardan en forma de tabla relacional especial que facilita ciertos tipos de consultas. Por ejemplo, hay columnas de las tablas que se llaman columnas de dimensión que facilitan y proveen datos para resúmenes e informes. Las columnas llamadas columnas agregadas permiten precalcular cantidades como conteos, sumas y medias.

Construir un cubo requiere un análisis detallado de las necesidades de datos del grupo de usuarios a los cuales va dirigido y puede requerir mucho tiempo, tanto de diseño como de instalación por primera vez. Compensa por el hecho que facilita extraordinariamente las tareas de análisis de datos de los diversos grupos de usuarios y, una vez establecido, resulta más sencillo de modificar que las tablas relacionales tradicionales.

Análisis y modelado multidimensional

El objetivo del análisis multidimensional es ganar comprensión en el conocimiento contenido en las bases de datos. Su principal ventaja es que facilitan los análisis complejos (al estar muy próximos a la manera de pensar del analista) y la visualización de los datos en el *datamart* para procesos de toma de decisiones, reduciendo la confusión y disminuyendo las interpretaciones erróneas.
Además, ya que los datos están almacenados físicamente en una estructura multidimensional ó base de datos n-dimensional, la velocidad de estas operaciones es varias veces superior y más consistente de lo que es posible en otras estructuras de bases de datos. La combinación de simplicidad y velocidad es uno de los principales beneficios del análisis multidimensional.

El modelo está basado en la noción de dimensión que permite especificar diferentes maneras de estudiar la información, de acuerdo con las perspectivas del negocio bajo las cuales el análisis puede ser realizado. Cada dimensión se organiza en una jerarquía de niveles, correspondiendo a dominios de datos en diferentes niveles ó granularidades. Un esquema multidimensional consiste en un conjunto de tablas de hechos (también llamadas *f-tables*) que se definen respecto a combinaciones particulares de niveles.
Una instancia multidimensional asocia medidas, que corresponden a los datos a ser estudiados, con coordenadas simbólicas a las tablas de hechos.
Finalmente, en una dimensión, los valores con un gran nivel de detalle pueden hacer *roll-up* (agruparse) a valores más generales [ASS01].
En la Figura 4 se muestra un ejemplo de esquema multidimensional.

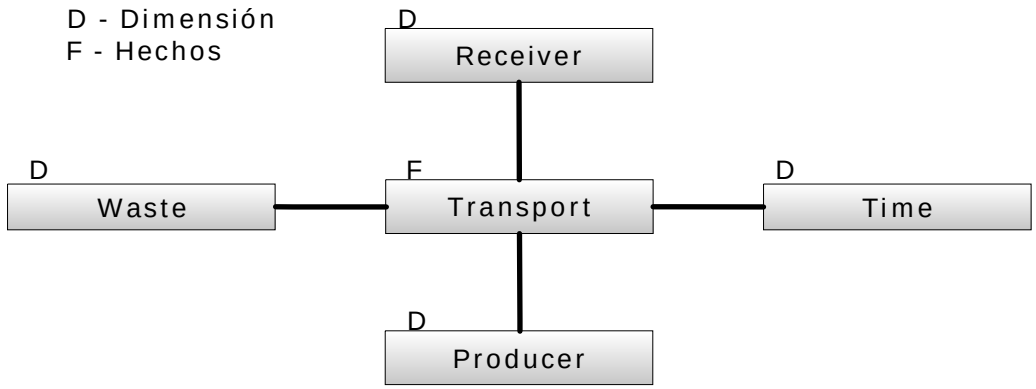

FIGURA 4 - EJEMPLO DE UN ESQUEMA MULTIDIMENSIONAL

Conceptualmente, en un modelo multidimensional, distinguimos 3 niveles diferentes de detalle [ASS01]:

- *Upper Level* (UL): a este nivel, encontramos dimensiones (D) y hechos (F). Las dimensiones se utilizan para caracterizar los hechos, y muestran los puntos de vista desde los que analizaremos los hechos. Relacionando un conjunto de dimensiones a un hecho, obtenemos un esquema en estrella. La posibilidad de navegar de una estrella a otra se observa por su compartición de dimensiones.
- *Intermediate Level* (IL): las dimensiones y los hechos se descomponen en niveles de dimensiones (DL), y celdas de hechos (FC) respectivamente. Los diferentes niveles de una dimensión forman una jerarquía. Cada FC contiene datos (un conjunto de medidas) para un DL determinado para cada dimensión al que el hecho esté relacionado.
- *Lower Level* (LL): el último nivel muestra los atributos de los DL y los FC. Estos son, respectivamente, atributos de clasificación (CA), y medidas (M).

En la Figura 5, se detalla el esquema multidimensional presentado en la Figura 4 con los 3 niveles de detalle UL, IL, LL.

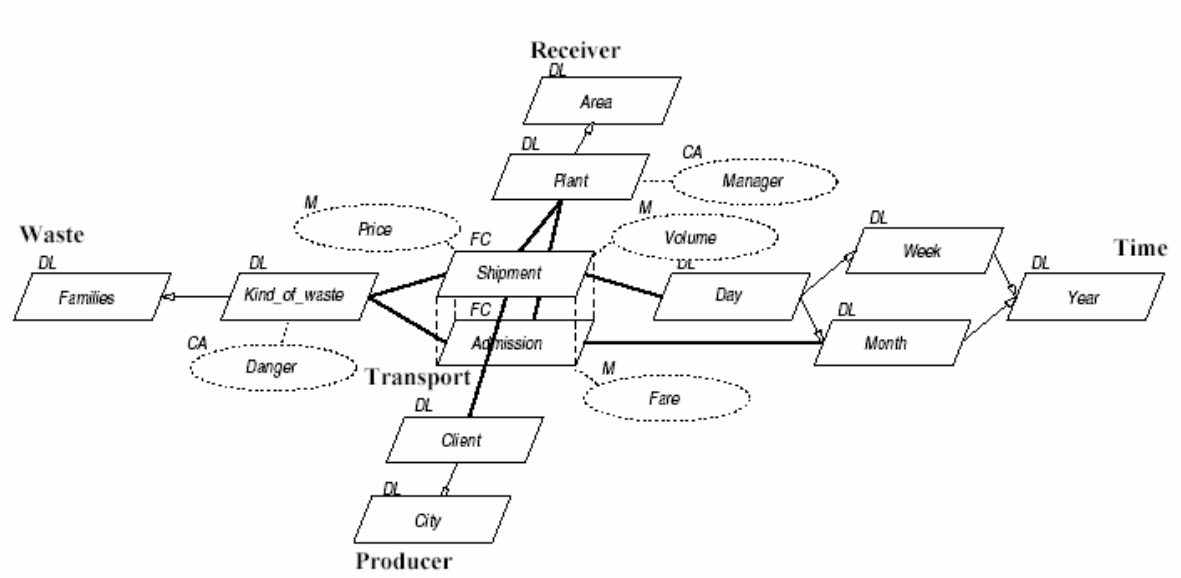

FIGURA 5 - EJEMPLO DE UN ESQUEMA MULTIDIMENSIONAL DETALLADO

El uso de dimensiones es una forma de mostrar (y a veces almacenar) datos muy útil en sistemas con grandes cantidades de información. Las dimensiones son ejes de análisis o criterios de clasificación de la información que ofrecen un índice a los datos mediante una lista de valores [HC03]. Por ejemplo son dimensiones <Tiempo>, <Geografía> y <Producto>.

A continuación se muestra una representación espacial de una variable multidimensional con una, dos y tres dimensiones. En la Figura 6 los cubitos representan valores de dimensión, y las esferas son datos.

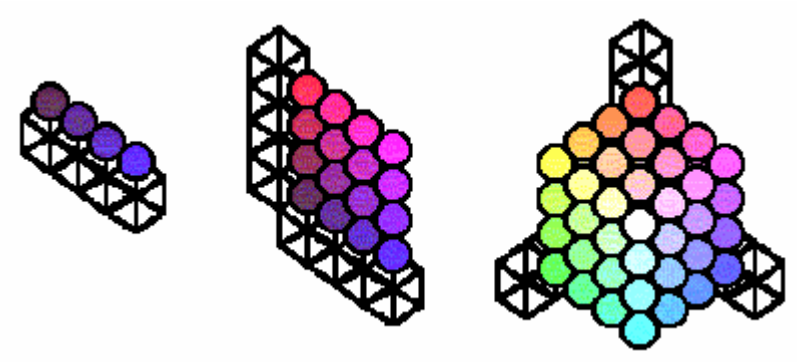

FIGURA 6 - VARIABLES CON UNA, DOS Y TRES DIMENSIONES

Una variable unidimensional podría ser el cambio del euro con el dólar, que sólo varía en la dimensión <Tiempo>. Los cubitos serían, por ejemplo, los meses del año y las esferas serían los valores numéricos correspondientes al cambio monetario en cada momento. Un ejemplo de variable de dos dimensiones es el número de habitantes, que se mueve por las dimensiones <Geografía> y <Tiempo>. Finalmente, los ingresos de una organización podrían almacenarse mediante una variable de tres dimensiones: <Producto>, <Geografía> y <Tiempo>.

Para los desarrolladores de aplicaciones acostumbrados a trabajar con bases de datos relacionales, el diseño de una base de datos multidimensional puede ser complejo o al menos, extraño. Pero la experiencia nos dice que el diseño de dimensiones y variables es mucho más sencillo e intuitivo que un diseño relacional. Esto es debido a que las dimensiones y variables son reflejo directo de los informes en papel utilizados por la organización.
Normalmente los usuarios no pueden “digerir” más de seis o siete dimensiones en un informe OLAP. Por esta razón, los diseñadores restringen el número de dimensiones un cubo basándose en las limitaciones de los usuarios más que en las del sistema.
Los cubos contienen típicamente de cuatro a diez dimensiones en sistemas en producción.

Operaciones

Un analista quizás solo quiera ver un subconjunto de los datos y algunos atributos y dentro de cada atributo seleccionado quiera restringir los valores de interés. En la terminología de bases de datos multidimensionales, estas operaciones se llaman *pivoting* (rotar el cubo para mostrar una cara determinada) y *slicing-dicing* (seleccionar algún subconjunto del cubo). Las vistas multidimensionales permiten que las jerarquías asociadas con cada dimensión se vean de una manera lógica [HC03].

La agregación de la dimensión <Producto> desde el <Producto> al <Tipo de producto> se llama una operación del *roll-up*. La operación inversa es el *drill-down* que muestra información detallada para cada punto agregado, esta operación es

esencial porque a menudo los usuarios sólo quieren ver datos agregados primero y seleccionando ver datos más detallados [HC03].

Normalmente los elementos de una dimensión forman una jerarquía, con lo que algunos son padres de otros. Cuando las variables multidimensionales de un *datamart* son cargadas con nueva información (por ejemplo, mensualmente a partir de ficheros de texto), ésta se refiere a los nodos hoja del árbol jerárquico de cada una de las dimensiones [HC03]. Por ejemplo, la información de ventas llega detallada por producto, por provincia y por mes. Pero si queremos obtener el total de ventas de todos los productos, el total de ventas de todas las provincias, el de todos los meses del año, o alguna combinación de estos, deberemos realizar un proceso de agregación de la información.
Por ejemplo, en la dimensión <Producto> incluiremos un valor llamado "Total Productos" que será padre de todos los demás productos y que contendrá el acumulado de todos ellos. En la dimensión tiempo podremos tener, por ejemplo, el año 2002 descompuesto en trimestres, y estos a su vez en meses. La información llega detallada por producto y por mes, y posteriormente a la carga de datos, se realiza un proceso de agregación que calcula estos acumulados.

Estructura del cubo

La estructura del cubo se define por medio de sus medidas y dimensiones. Derivan de tablas del origen de datos del cubo. El conjunto de tablas del que derivan las dimensiones y medidas de un cubo se denomina esquema del cubo. Las medidas del cubo derivan de las columnas de la tabla de hechos. Las dimensiones del cubo derivan de columnas de las tablas de dimensiones.

Hay dos tipos comunes de esquemas de cubo: en estrella y en copo de nieve [Mic01]. En un esquema en estrella, cada tabla de dimensión se combina con la tabla de hechos. En un esquema en copo de nieve, una o más tablas de dimensiones que no se combinan con la tabla de hechos son para dimensiones con varias tablas de dimensiones.

Esquema en estrella

Para facilitar el acceso a los datos y el análisis, un *datamart* organiza físicamente en estructuras llamadas esquemas en estrella. Un *datawarehouse* construido sobre *data marts* utiliza uno o más esquemas en estrella para representar eventos o procesos específicos del negocio.

Un esquema en estrella se caracteriza por tener una tabla central de hechos rodeada por tablas de dimensiones que contienen información desnormalizada de los hechos [Mic01].

Los siguientes elementos son característicos en un esquema en estrella:

- El centro del esquema es la tabla de hechos (DWH001), contiene las métricas o medidas del negocio.
- Las puntas de la estrella son las tablas de dimensiones. Estas puntas son utilizadas para describir la información existente en un proceso específico del negocio y proveen el contexto a los datos numéricos.

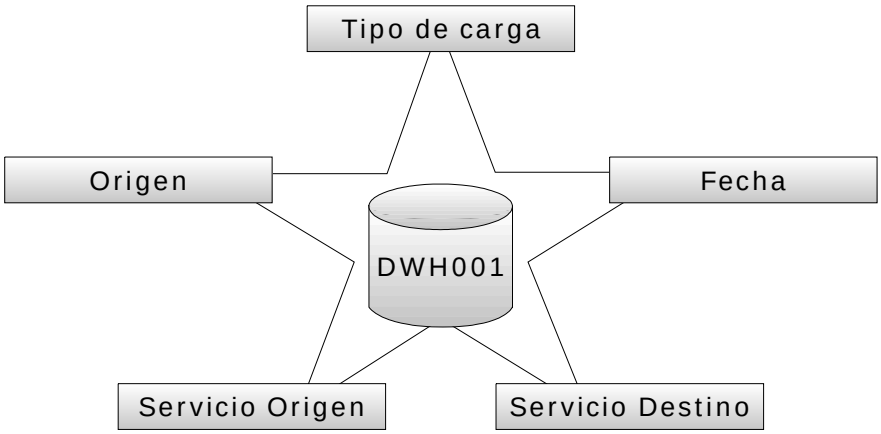

FIGURA 7 - DIAGRAMA DE UN ESQUEMA EN ESTRELLA

Los cubos OLAP utilizan los esquemas en estrella como sus orígenes de datos, ver Figura 7.

Tabla de hechos

La tabla de hechos es la tabla central del esquema en estrella que representa datos numéricos en el contexto de las entidades del negocio. La tabla de hechos está constituida por medidas y por *foreign keys*.

Medidas

Una medida, es una columna numérica de la tabla de hechos. Las medidas representan los valores que se van a analizar, tales como las unidades vendidas o el número de empleados.

Foreign keys

Una *foreign key* es la representación de la *primary key* de una dimensión en la tabla de hechos. Las *foreign keys* son tomadas de las *primary keys* para cada dimensión de la tabla. La combinación de estas claves es el identificador para cada registro de la tabla de hechos.

Tablas de dimensiones

Una tabla de dimensión representa una entidad del negocio, dotando de contexto a los datos numéricos de la tabla de hechos. El diseño de las tablas de dimensiones apunta a las necesidades analíticas del usuario presentando información descriptiva que es fácil de utilizar para los usuarios.
Una tabla de dimensiones:

- Describe las entidades del negocio, en los esquemas en estrella representan una única entidad del negocio, tales como un producto o un cliente.
- Contienen atributos que proveen contexto a los datos numéricos.
- Presentan los datos organizados en jerarquías, en cada dimensión, se pueden organizar los datos en una o varias jerarquías. En la Figura 8 se muestra un ejemplo de las jerarquías, la dimensión <Fecha> se puede

desglosar en <Año>, <Trimestre>, <Mes> y <Día> permitiendo a los usuarios ver los datos detallados y sumarizados.

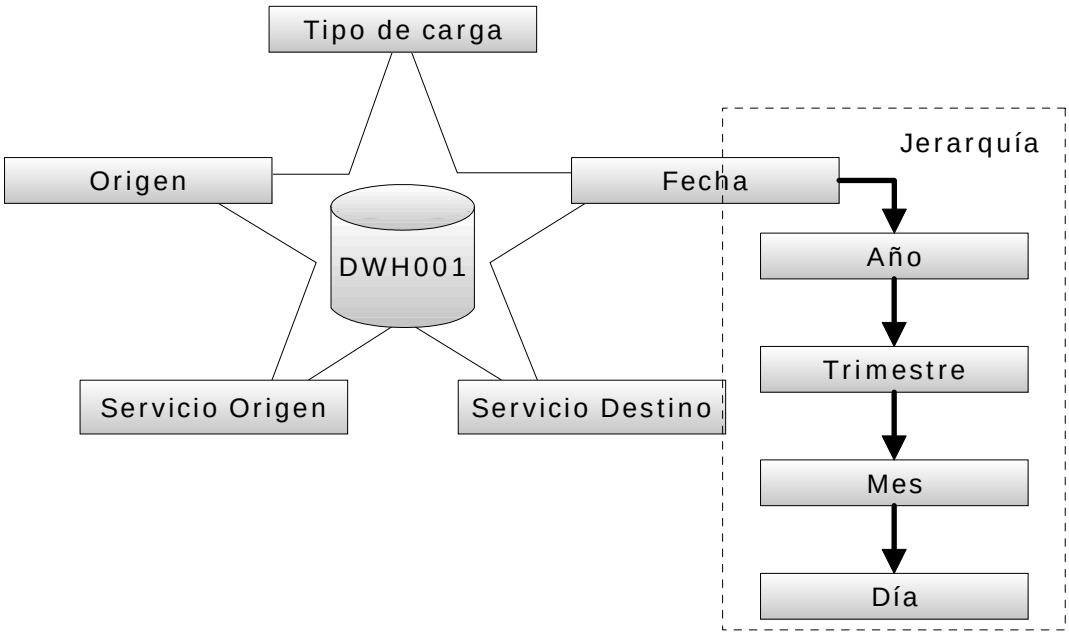

FIGURA 8 - DIAGRAMA DE UNA JERARQUÍA

Dimensiones del cubo

Cuando se diseña un *datawarehouse* es una buena práctica crear una tabla de dimensión separada para las fechas más que utilizar una columna de fecha/hora en la tabla de hechos [Mic01].
Como con otras tablas de dimensiones, se utiliza una clave entera para ligar la tabla de dimensión de fechas con la tabla de hechos. Crear una tabla de dimensión de fechas separada tiene ciertas ventajas sobre almacenar una columna de fecha/hora en la tabla de hechos:

- Contiene propiedades adicionales de la fecha, tales como la estación para un mes ó una marca de festivo para un día.
- Reduce el espacio de almacenamiento y diseñar y procesar una dimensión basada en una tabla con sólo unos pocos registros es mucho más rápido que extraer los valores de la dimensión de una gran tabla de hechos.
- Se utiliza en múltiples tablas de hechos, esto permite crear una dimensión única y compartida en múltiples cubos.

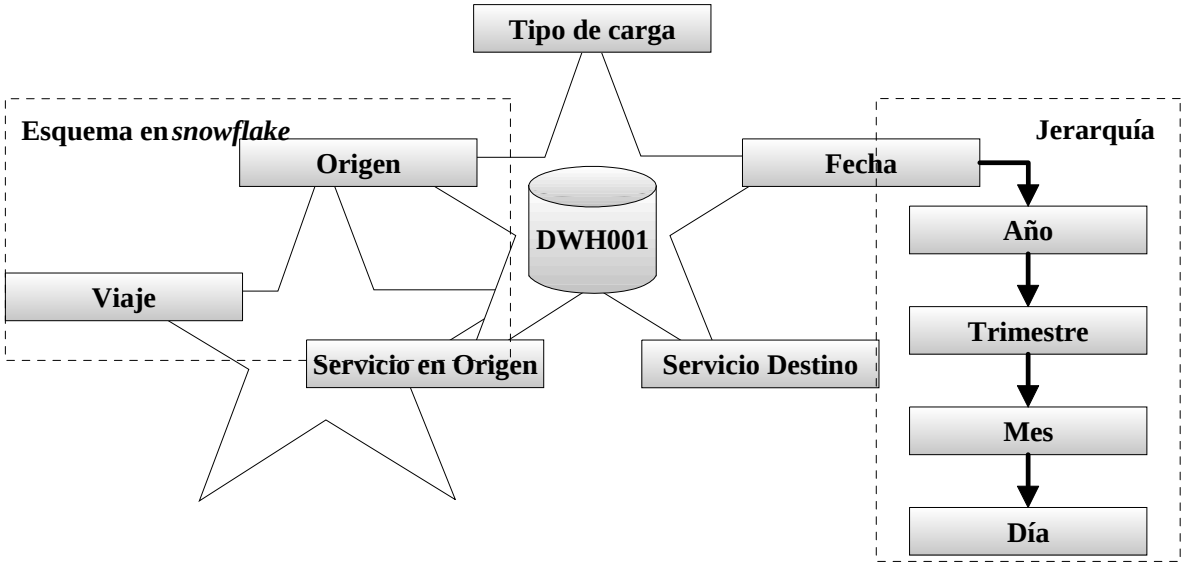

FIGURA 9 - DIAGRAMA DE UN ESQUEMA SNOWFLAKE

Almacenamiento de cubos

Una vez decidido emplear un entorno de consulta es necesario decidir su modo de almacenamiento.

Las opciones de almacenamiento físico afectan al rendimiento, los cubos pueden necesitar una cantidad considerable de espacio para contener los datos y las agregaciones (resúmenes de datos precalculados que mejoran el tiempo de respuesta de consultas al tener preparadas las respuestas antes de que se planteen las preguntas) en estructuras multidimensionales.
Los cubos pueden dividirse en particiones y cada una de ellas se puede almacenar mediante un modo de almacenamiento distinto.

Los modos de almacenamiento son:

- OLAP multidimensional (MOLAP).
- OLAP relacional (ROLAP).
- OLAP híbrido (HOLAP).

ROLAP es la arquitectura de bases de datos multidimensional en la que los datos se encuentran almacenados en una base de datos relacional, la cual tiene forma de estrella (también llamada copo de nieve ó araña). En ROLAP, en principio la base de datos sólo almacena información relativa a los datos en detalle, evitando acumulados (evitando redundancias).
Con el modo de almacenamiento ROLAP, la respuesta de consultas suele ser más lenta que la de los otros dos modos de almacenamiento. ROLAP se usa normalmente para el acceso a grandes conjuntos de datos que se consultan con poca frecuencia, tales como datos históricos de años no recientes.

En un sistema MOLAP, en cambio, los datos se encuentran almacenados en ficheros con estructura multidimensional, los cuales reservan espacio para todas las

combinaciones de todos los posibles valores de dimensión que representan acumulados.
Es decir, un sistema MOLAP contiene precalculados (almacenados) los resultados de todas las posibles consultas a la base de datos. MOLAP consigue consultas muy rápidas a costa de mayores necesidades de almacenamiento, y retardos en las modificaciones (que no deberían producirse salvo excepcionalmente), y largos procesos *batch* de carga y cálculo de acumulados. En ROLAP, al contener sólo las combinaciones de valores de dimensión que representan detalle, es decir, al no haber redundancia, el fichero de base de datos es pequeño. Los procesos *batch* de carga son rápidos (ya que no se requiere agregación), y sin embargo, las consultas pueden ser muy lentas, por lo que se aplica la solución de tener al menos algunas consultas precalculadas.
En general, el modo MOLAP es más apropiado para las particiones de cubos que se utilizan con frecuencia y donde se necesita una rápida respuesta de consultas.

HOLAP, esta es la solución recomendada, combina atributos de los modos MOLAP y ROLAP. La idea que yace detrás de HOLAP es que obtengamos los beneficios de una arquitectura MOLAP para la información resumida, pero necesitamos el detalle, el sistema “retrocederá” a una base de datos relacional. Por lo tanto el rendimiento aumenta aunque solo a nivel de información resumida, y la escalabilidad puede ser resuelta porque todo el detalle está almacenado en tablas relacionales convencionales.
Las particiones almacenadas como HOLAP son más pequeñas que sus equivalentes MOLAP y responden más rápidamente que las particiones ROLAP a las consultas que implican datos de resumen. El modo de almacenamiento HOLAP suele ser más adecuado para particiones en cubos que requieren una respuesta de consultas rápida basada en una gran cantidad de datos de origen.

Evaluación

Introducción

Antes de proceder a la fase de despliegue final del modelo construido, es importante evaluar el modelo y revisar la construcción con el fin de comprobar que se cumplen los objetivos del negocio. Aquí es crítico determinar si partes importantes del negocio no han sido lo suficientemente consideradas. Al final de esta fase, el líder del proyecto debe decidir exactamente cómo utilizar los resultados del proceso.

Es en esta fase y con la ayuda del analista de los datos que gracias a la naturaleza iterativa de este tipo de proyectos, pueden surgir nuevas preguntas a responder que hagan que el proyecto retorne a la fase de conocimiento del negocio a fin de poder responderlas.

Una buena manera de definir los resultados de un proyecto de este tipo es utilizar la ecuación:

RESULTADOS = MODELOS + DESCUBRIMIENTOS

En esta ecuación estamos definiendo que el resultado del proyecto no son sólo los modelos (aunque, por supuesto, son importantes), también los descubrimientos, que se definen como cualquier cosa (aparte del modelo) que son importantes para alcanzar los objetivos del negocio (o importantes para conducir a nuevas preguntas o a los efectos secundarios, por ejemplo problemas en la calidad de los datos no cubiertos en el ejercicio de *data mining*).
Las fases de este paso son: la evaluación de los resultados, la revisión del procesos y determinar los siguientes pasos a seguir.

- Evaluación de los resultados, pasos de evaluación previos trabajan con factores tales como la precisión y la generalidad del modelo. Este paso evalúa el nivel en que el modelo se encuentra respecto a los requerimientos del negocio e intenta determina si existen requerimientos que no cumple el modelo.
- Revisión del proceso, en este punto el modelo parece ser satisfactorio y cubrir los requerimientos del negocio. Es ahora cuando se realiza una revisión del proceso para determinar si hay algún factor o tarea importante que de alguna manera no se ha tenido en cuenta.
  Esta revisión cubre además cuestiones de calidad, p.e., ¿hemos construido correctamente el modelo?, ¿hemos utilizado atributos que podíamos utilizar y estarán disponibles para análisis futuros?.
- Determinar los siguientes pasos, de acuerdo con los resultados de la revisión, se decide como proceder al siguiente paso. Se decide si finalizar el proyecto y pasarlo a desplegar o iniciar nuevas iteraciones ó nuevos proyectos. Esta tarea incluye evaluar los recursos existentes y presupuestarlos para influir en la decisión.

Evaluación del modelo

Este capítulo contiene un resumen de las principales ventajas que supone la implantación de la solución en la empresa, algunos comentarios son resultado de entrevistas realizadas con un grupo de los potenciales usuarios del modelo.

Si consideramos la suma de ventajas, podemos ver que adoptar esta solución comporta incrementar el potencial global de análisis de la información y mejorar la capacidad y control de gestión de la empresa, características cada vez más necesarias para seguir siendo competitivos.

Las ventajas se analizan desde 4 ópticas diferentes para ver su equilibrio global:

- Óptica de negocio, es decir, los beneficios directos al negocio que aporta la solución.
- Óptica económica, el factor económico de la solución: ahorro de costes y rentabilidad de la inversión.
- Óptica del usuario, desde este punto de vista se analizan las ventajas y el impacto tecnológico de la solución para el usuario.
- Óptica técnica, factores técnicos que hacen también muy coherente y ventajosa la solución presentada.

Óptica de negocio

Los beneficios más importantes de la solución son el impacto que causa en el negocio.

Los listados de papel se sustituyen por hojas de cálculo formateadas y enriquecidas con todas las posibilidades de *Microsoft Excel*. Las siguientes funciones permiten analizar mejor los datos para una toma de decisiones más efectiva y certera:

- Análisis dinámico e interactivo, el análisis multidimensional de la información permite analizarla desde múltiples visiones, con el fin de ver más información en los mismos datos.
- Resaltados y gráficos, los datos se pueden resaltar de manera que los incrementos se muestren en verde y los descensos en rojo. Así se puede centrar la atención en los puntos que la merecen. Los gráficos también cumplen su función de resumir la información a alto nivel.
- Datos enriquecidos, los datos se pueden insertar dentro de hojas de cálculo ya definidas que pueden contener fórmulas para calcular más datos, como previsiones e indicadores estadísticos.

Otro punto fuerte del modelo es su velocidad, ya que para este caso se ha decidido la utilización de la técnica de almacenamiento MOLAP que consigue consultas muy rápidas ya que las agregaciones están precalculadas. Los datos serán refrescados de las fuentes operacionales una vez a la semana, no es necesario hacerlo más a menudo pero en función de cómo evolucione el modelo y su explotación, esta técnica de almacenamiento puede ser cambiada sin problemas.

Óptica económica

No hay un coste adicional por licencia, la solución se apoya en la infraestructura ofimática ya implementada en la empresa.

El trabajo derivado de la creación de informes puede verse como un coste variable dependiente de la cantidad de informes que se quieran desarrollar que a posteriori repercute en un ahorro del tiempo de diseño de informes, un único informe satisface decenas de informes convencionales.

Se ahorra en papel y en tiempo de impresión.

Óptica del usuario

Desde este punto de vista analizamos cómo verá el usuario la implantación de esta solución y cómo le afectará.
Observamos que con el *datamart* y la aplicación de técnicas OLAP sobre él se facilita tremendamente lo que antes era muy costoso y dependía en gran medida de la habilidad técnica de los usuarios, la creación de informes para los directivos.
El usuario tiene cierta familiaridad con Excel, luego no hay un "nuevo sistema" que implantar ya que todo el *reporting* lo recibe en la herramienta que más conoce: *Microsoft Excel*. No es necesaria formación (o es muy mínima).
No es necesario ninguna instalación en el PC, no hay que molestar al usuario para instalar algo y el coste de formación es bajísimo debido a que la herramienta de consulta es *Microsoft Excel*.

El usuario tiene más potencia de análisis. Ahora tiene una herramienta flexible que le permite manipular fácilmente la información.

Óptica técnica

Existen factores técnicos que hacen también muy coherente y ventajosa la solución presentada. Esta se integra con los sistemas informáticos de la organización, con *SQL Server* y se reaprovecha la inversión ofimática, la solución se apoya en la infraestructura que ya tiene actualmente la empresa.

Revisión del proceso

Se decide revisar el proceso de carga y limpieza de datos, incidiendo especialmente en la calidad de los datos de la tabla de clientes. A veces se introduce un mismo cliente con diferentes códigos y a menudo el personal encargado de introducir los datos en las aplicaciones dan de alta un cliente y utilizan siempre ese código de cliente, cuando posiblemente ya esté dado de alta en el sistema.

Siguientes pasos

Los datos consensuados en una única fuente abren la posibilidad de nuevos análisis, por ejemplo se ha observado que sería interesante ampliar el modelo incluyendo datos procedentes de las operativas de carga y descarga de los buques, lo que podría desembocar en una optimización de estas operativas tan costosas y cruciales para la naviera. Se desea además incluir información procedente de las unidades organizativas de transporte terrestre y distribución a fin de poder brindar al cliente un mejor servicio, pudiendo informar con facilidad de cuáles son los tiempos de entrega previstos de su mercancía en función de datos históricos. A fin de mejorar este punto se propone tener los metadatos dentro de la estructura del *datamart* a fin observar fácilmente cómo los datos se integran y transforman antes de ser almacenados.
Otras preguntas que han surgido es si sería posible realizar clasificaciones de los clientes y de la carga por puertos, preguntas para las que el modelo está preparado para responder aplicando sobre él técnicas de *clustering* y árboles de decisión.
Se ha conseguido una recomendación a seguir en los proyectos de *datawarehousing* y *datamining*, y es que en las siguientes iteraciones el proyecto pase de ser una iniciativa de IT a ser una iniciativa del negocio [SF00], permitiendo el desarrollo incremental y manejable para asegurar la flexibilidad del modelo y haciendo que los cambios en los procesos del negocio se reflejen en el diseño del modelo.

Adicionalmente se plantea la posibilidad de la adquisición de un *software* generador y distribuidor de informes por correo electrónico para que la información circule apropiadamente desde los sistemas hasta los usuarios, permitiendo además la gestión por excepción avisando de la ocurrencia de un evento importante en el negocio como la superación de umbrales y ratios y la comparación de ventas con previsiones y gastos con presupuestos. De esta forma los directivos y mandos pueden centrarse en sus tareas productivas sabiendo que el sistema de *reporting* les alertará antes situaciones alarmantes.

GLOSARIO

Agregación, son resúmenes de datos precalculados que mejoran el tiempo de respuesta de consultas al tener preparadas las respuestas antes de que se planteen las preguntas.

BSC (Cuadro de mandos integral), dota a las empresas de una estructura de indicadores que permiten medir de forma coherente la evolución de la empresa.

Business Intelligence, conjunto de metodologías y tecnologías orientadas a potenciar la gestión inteligente de la empresa que permitan a los equipos directivos controlar los negocios.

Cubo, unidad de datos del análisis en línea de la información, es la representación de los datos de interés para el análisis.

CRISP-DM metodología, el término general para todos los conceptos desarrollados y definidos en CRISP-DM.

Data mining, proceso no trivial de análisis de grandes cantidades de datos con el objetivo de extraer información útil, por ejemplo para realizar clasificaciones o predicciones.

Datamart, almacén de datos departamental, no son más que datos históricos pero tratados para evitar datos duplicados, atributos no existentes, etc.

Datawarehouse, almacén de datos que reúne la información histórica generada por todos los distintos departamentos de una organización, orientada a consultas complejas y de alto rendimiento.

Dimensión, atributos de los datos a analizar, no son más que los filtros que podemos aplicar a nuestros datos, tanto filas como columnas.

Drill down, operación que muestra información detallada para cada punto agregado.

DSS (Decision Support Systems), sistemas de ayuda a la toma de decisiones.

EIS (Executive Information Systems), sistemas de información para ejecutivos, independientes de aplicaciones convencionales, ergonomía de presentación y manipulación de datos y alta disponibilidad de información y análisis.

Esquema en estrella, organización física de los *datamarts* que facilita el acceso a los datos y al análisis. Se caracteriza por tener una tabla central de hechos rodeada por tablas de dimensiones que contienen información desnormalizada de los hechos.

ETL (Extraction, Transformation and Loading), herramientas dedicadas a la extracción de los datos desde las fuentes donde estos se encuentren a los *datamatarts*.

Foreign key, representación de la *primary key* de una dimensión en la tabla de hechos.

HOLAP (Hybrid OLAP), los datos son almacenados y recuperados de una combinación multidimensional de una estructura de cubos y tablas de bases de datos relacionales.

Medida, columna numérica de la tabla de hechos que representa los valores que se van a analizar, tales como las unidades vendidas o el número de empleados.

Metadatos, datos acerca de datos que describen los contenidos del almacén de datos.

MOLAP (Multidimensional OLAP), los datos son almacenados y recuperados de cubos que están separados de la base de datos relacional que es el origen de los datos.

OLAP, *Online Analitical Processing*, análisis on-line de información, análisis de datos con los que se trabaja día a día.

Pivoting, operación sobre un cubo que lo permite rotar para mostrar una cara determinada.

ROLAP (Relational OLAP), los datos son almacenados y recuperados de una base de datos relacional.

Slicing-dicing, operación que permite seleccionar un subconjunto del cubo.
Staging area, ó preparación de los datos, es una colección de procesos que limpian, transforman, combinan, y preparan los datos originales para su utilización en el *datawarehouse* ó el *datamart*

Tipo de problema de data mining, clasificación de problemas típicos de *data mining* tales como la descripción de datos y la sumarización, segmentación, clasificación, predicción y análisis de dependencias.

BIBLIOGRAFÍA

[AGS96**]** Rakesh Agrawal, Ashish Gupta, Sunita Sarawagi. *Modeling multidimensional databases, Research Report*, IBM *Almaden Research Center*, San Jose, California, 1996.

[ASS01**]** Alberto Abelló (UPC), José Samos (UGR), Félix Saltor (UPC). *A datawarehouse multidimensional data models classification*, 2001.

[CCKKRSW00**]** Peter Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinatz, Colin Shearer, Rudiger Wirth. CRISP-DM 1.0, *Step by step data mining guide*, www.spss.com, 2000.

[CCS93**]** E.F. Codd, S.B. Codd, C.T. Salley. *Beyond decisión support, Computerworld*, 27, *July* 1993.

[CT97**]** Luca Cabibbo, Riccardo Torlone. *Querying multidimensional databases, In Proc. of the 6th Int. Workshop on Database Programming Languages*, 1997.

[FER02**]** Montse Fernández. Inteligencia aplicada al negocio, Datamation, n° 191, septiembre 2002.

[GL96**]** Marc Gyssens (LUC), Laks V.S. Lakshmanan (CUM). *A foundation for multidimensional databases, In Proc. Of the 22nd VLDB Conference*, Mumbai (Bombay), India, 1996.

[GOR02**]** Davor Gornik. *Data Modeling for Data Warehouses, Rational Software White Paper*, 2002.

[HC03**]** Manuel de la Herrán Gascón, Vicent Castellar-Busó. Cómo diseñar grandes variables en bases de datos multidimensionales. REDcientífica, 2003.

[HHD01**]** Mark W. Humphries, Michael W. Hawkins, Michelle C. Dy. *Data Warehousing Architecture and Implementation, Harris Kern's Enterprise Computing Institute,* 2001.

[Mic01**]** *Microsoft Training and Certification. Designing and implementing OLAP solutions with MS SQL Server 2000. Microsoft Technical Report*, 2001.

[Mic02**]** *Microsoft Training and Certification. Data Warehousing with SQL Server, Technical Report. Microsoft*, 2002.

[MOL02] Luis Carlos Molina Félix. *Data mining*: torturando a los datos hasta que confiesen. Universitat Oberta de Catalunya, 2002.

[SF00] David Sammon, Pat Finnegan. *The ten commandments of data warehousing. The DATABASE for Advances in Information Systems, volume* 32, *number* 4, *fall* 2000.

[SHE00**]** Colin Shearer. *The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, volume* 5, *number* 4, *fall* 2000.

[SM00**]** Ramon Sangüesa i Solé, Luis Carlos Molina Félix. Postgrau en *Data Mining*. Universitat Oberta de Catalunya, 2000.

[TOD00**]** Chris Todman. *Designing a Data Warehouse: Supporting Customer Relationship Management, Prentice Hall*, 2000.

[Two99**]** *Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery, Third Edition. Two Crows Corporation*, 1999.

Autor: Óscar Alonso Llombart (oalonsol@uoc.edu)