Del Query al Data Mining

por Jorge Gros

La posibilidad de que los usuarios exploten la información almacenada en sus sistemas transaccionales ha dado lugar a una serie de nuevas tecnologías que van desde el simple Query hasta las más sofisticadas herramientas de minería de datos

OTROS ARTICULOS RELACIONADOS

El poder de la replicación de datos
(Nigel Stokes, Abril 1996)

Benefíciese de las ventajas del Data Warehousing
(Nigel Stokes, Marzo 1997)

Guía práctica para la minería de datos
(Mark Wulf, Abril 1997)

En la informática, tanto en hardware como en software se pueden distinguir claramente distintas generaciones de productos; según surgen, las distintas tecnologías tienen su auge y posteriormente su declive. Al principio, en la prehistoria de la informática, nos colmaba de satisfacción el haber impreso los recibos de nómina o las facturas. Más adelante, los usuarios empezaron a interesarse por la posibilidad de explotar para su trabajo la información almacenada en los sistemas transaccionales. En este proceso evolutivo de los Sistemas de Información empresariales, surgieron dos tecnologías aún hoy muy importantes, como son los EIS (Sistemas de Información para la Dirección) y los queries (software de consultas).

Los proyectos EIS

Un típico proyecto EIS, comienza por una fase previa de análisis muy extensa, en la cual se define la información que necesita tener la alta dirección en su "tablero de mandos". A continuación se desarrolla un proyecto, cuyo objetivo es extraer de las bases de datos operativas la información necesaria, sintetizarla y presentarla, generalmente de una manera muy espectacular, tal como se lo merecen "sus consumidores", los más altos ejecutivos de una gran empresa. Los resultados suelen ser muy satisfactorios, queda un solo problema, el muy alto coste de este tipo de proyectos. Pero peor aún, la evolución del mercado y de la empresa demanda que el EIS de la empresa también evolucione y el coste del mantenimiento puede llegar a ser igual de alto como el del proyecto inicial. Por este mismo motivo el EIS no es una herramienta apropiada para el análisis de datos a nivel departamental.

Los productos Query

Al realizarse el análisis detallado de ventas, costes, márgenes, rotaciones, etc., las preguntas que se plantea un usuario varían constantemente. Por cierto, se tardó bastante tiempo para que los informáticos nos diéramos cuenta que esto no es precisamente un capricho del usuario, sino una necesidad objetiva del negocio. Aquí no hay un tablero de mando. No se constata el "qué está pasando", se analiza el "porqué puede estar pasando". Por esto los queries son un software con una capacidad de extraer la información sin ningún proyecto previo. La información se presenta en forma poco espectacular, muchas de las consultas son para "usar y tirar".

Los queries han sido un avance fabuloso tanto para el informático como para el usuario. El usuario puede por sí solo extraer información de la base de datos, sin esperar que los informáticos, generalmente ya bastante agobiados por otros proyectos, programen la salida en el formato solicitado por el usuario.

La evolución hacia Data Warehouse

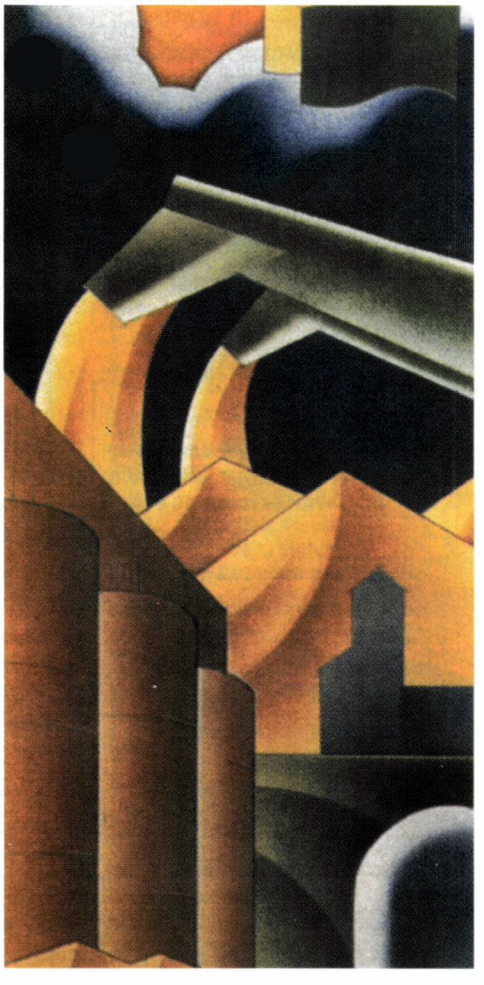
Una vez descubierta la posibilidad de obtener información sin programar, su uso aumenta constantemente. Pero el verdadero desbordamiento se produce con las hojas electrónicas bajo Windows y la posibilidad de alimentarlas con la información de la base de datos operativa. Ahora los usuarios pueden no solamente acceder a la información, sino realmente trabajar con ella. Además las hojas electrónicas permiten presentar la información en una forma muy atractiva.

Aparece un nuevo tipo de productos que permite al usuario, directamente desde su PC y trabajando con interfaz gráfica, solicitar la información a la base de datos central. La utilidad del binomio query - hoja electrónica es tan enorme, que en muchas instalaciones la ocupación del hardware por los queries representa un porcentaje elevado. Este fenómeno constituye la mejor prueba de la importancia que tiene para una empresa el trabajar con la información. Sin embargo, la masificación del uso de las consultas, llega a poner en evidencia algunas deficiencias intrínsecas de este tipo de productos.

En primer lugar, la velocidad. Los tiempos de respuesta *son muy lentos* y se convierten en un freno importante del análisis de la información, puesto que cada consulta resuelta, genera nuevas preguntas. Los tiempos de respuesta lentos, interrumpen el hilo del razonamiento del usuario, dificultando así la tarea de profundizar el análisis. Pero además de la lentitud, también la alta ocupación de los recursos de hardware por los queries constituye un problema. La sobrecarga aumenta aún más la lentitud de las consultas y baja el rendimiento de las propias aplicaciones operativas. Esta situación conduce a constantes y costosos upgrades de ordenadores.

La proliferación de queries no sólo ha llegado a colapsar muchos ordenadores, sino que también *absorbe un volumen significativo* de los recursos humanos de los departamentos de informática. Ello se debe al hecho de que la petición de la consulta se especifica en términos de la base de datos. Hay que utilizar los nombres de columnas y tablas y por tanto, hay que saber cómo están diseñadas las bases de datos. Muchas veces hay que definir joins, e inclusive, tablas especiales para poder unir información de una aplicación con otra, cuando sus tablas no son compatibles. Demasiadas veces, el resultado de una petición no es el pretendido por el usuario. Por todo esto, en las instalaciones, donde no se haya hecho una inversión considerable para que los usuarios adquieran un grado de cultura informática relativamente elevado, hay que dedicar recursos para el soporte de los queries.

Otro problema es la inconsistencia de la información obtenida. Frecuentemente, los usuarios omiten algún factor importante al formular sus peticiones de consulta. Como resultado, en el mejor de los casos hay que repetir las consultas, modificando la petición, en el peor se trabaja con información distorsionada. Y no es nada raro que en una reunión de ventas o un consejo, diferentes personas se presenten con información distinta. *Más de una reunión ha degenerado en un intento de conciliar la información que traen consigo los distintos participantes.*



Entre los datos almacenados en algún lugar de sus repletos discos, se hallan las respuestas a las preguntas más desconcertantes de sus usuarios... Eso sí, siempre que se disponga de la herramienta adecuada

Data Warehouse

Todos estos problemas no son deficiencias de productos determinados, ni mucho menos culpa de los desarrolladores. Se producen porque una tecnología llega al límite de sus posibilidades. La solución ya no se puede producir por la vía del "más de lo mismo".

Los problemas de velocidad y de sobrecarga de procesadores se deben al hecho que las bases de datos relacionales han sido diseñadas para el trabajo transaccional y no para atender consultas. El objetivo de obtener cualquier información en cualquier momento, existan o no los índices necesarios en la base de datos, es una tarea demasiado difícil para las bases de datos relacionales. Por ello, ha surgido la tecnología Data Warehouse que se apoya en otro tipo de bases de datos. Se define un nuevo entorno, el OLAP (Procesamiento Analítico en Tiempo Real) en contraste con el existente OLTP (Procesamiento de Transacciones en Tiempo Real).

Las bases de datos para OLAP tienen una estructura totalmente

distinta, orientada a alta velocidad de recuperación de la información para consultas. Se les suele llamar bases de datos multidimensionales o también "hiper-cubos". Los conceptos como Zona, Producto, Vendedor o Período de Venta se convierten en las múltiples dimensiones de estos cubos, y en las intersecciones de sus coordenadas, están almacenados los valores correspondientes. De esta manera, los Data Warehouse homogenizan la información en unas estructuras que permiten acceder más fácilmente a cualquier dimensión y de allí navegar mediante funciones específicas tales como "la rotación del cubo" (detallar determinada información siguiendo el criterio de otra dimensión, como por ejemplo, detallar zona por producto); "drill down" (bajar al siguiente nivel, como por ejemplo, detallar zona por comercial) etc.

Estas nuevas estructuras de datos, además de responder a las consultas bastante más rápidamente, lo hacen con menor utilización de recursos. Pero sobre todo, al trabajar sobre otra base de datos, se separan los procesos transaccionales de los procesos de consulta, colocando los procesos de consulta generalmente en otro ordenador, a veces con procesadores de características especiales.

La forma de plantear la distribución de datos en un Data Warehouse tiene un impacto decisivo en su rendimiento y condiciona el éxito de todo el proyecto

La idea de la coexistencia de dos bases de datos que almacenan esencialmente la misma información estructurada en formas distintas, según el tipo de proceso a que se estén sometiendo los datos, resultaba inicialmente un tanto chocante. No es de sorprenderse, puesto que los informáticos estábamos adoctrinados para perseguir con insistencia el inalcanzable objetivo de la integración total de las aplicaciones y la eliminación de todo tipo de redundancias. En la actualidad, esta idea está apoyada por la existencia de cada vez más evidencias de los buenos resultados y también por la baja de los precios de hardware. Además en muchas de las aplicaciones, la replicación de la información no tiene que ocurrir estrictamente en tiempo real y en aquellos casos donde sí sea indispensable, tenemos a mano otra nueva tecnología: la de "mirroring". Los productos que usan esta tecnología son capaces de replicar datos y objetos en tiempo real, con garantía de una sincronización exacta. En este punto también conviene hacer la observación, que la redundancia es uno de los conceptos inherentes de Data Warehouse, utilizado para aumentar las velocidades de acceso.

Lo dicho en el párrafo anterior no significa de ninguna manera que un proyecto de definición de un Data Warehouse y su alimentación es algo simple. Todo lo contrario: los grandes Data Warehouse son sistemas sumamente sofisticados y para su implantación hace falta un conocimiento muy profundo, tanto del producto como de la problemática de usuario. La forma de plantear la distribución de datos en un Data Warehouse *tiene un impacto decisivo en su rendimiento.* Hay muchos proyectos que han resultado una decepción porque se subestimaron estos factores.

La homogeneización de las estructuras de datos tiende a resolver otro gran problema de los queries: El del soporte requerido por los usuarios. La tecnología Data Warehouse facilita que el usuario especifique sus peticiones de consulta en su lenguaje propio, en vez de usar la terminología de base de datos. (Esto, cuando los diseñadores del producto logran evitar la tentación de sustituir una jerga informática "plana" por otra "multi-dimensional", que puede resultar aún más difícil para los mortales.)

Finalmente, el problema de inconsistencia de la información elaborada por los queries, también puede encontrar su solución en Data Warehouse: Cuando se define el contenido del Data Warehouse, es necesario definir con exactitud la correspondencia de conceptos entre la base de datos transaccional y el Data Warehouse. A partir de allí, cualquier usuario que seleccione una consulta como "Venta del Producto X en el mes corriente con detalle por Clientes", obtiene la misma respuesta.

El avance de Data Warehouse

Es natural que una nueva tecnología, con un potencial para resolver problemas tan importantes como los sufridos por los usuarios de los queries y sus departamentos de informática, tenga un éxito notable. Las predicciones de crecimiento del mercado hablan de 700% de incremento interanual en los próximos 5 años y textualmente cada mes aparecen nuevos productos. En grandes empresas, los productos Data Warehouse han asumido la mayoría de funciones antes realizadas con queries y también una parte de las funciones del EIS. La eliminación total de los queries normalmente no es práctica ni debe ser el objetivo, puesto que siempre existirán consultas puntuales. Como en todos los proyectos, aquí también es aplicable la regla 80 / 20 - el último 20 % cuesta 80% del esfuerzo. Por tanto, el proponerse el objetivo que todas las consultas tienen que ser satisfechas por el Data Warehouse, encarece el proyecto de una forma muy significativa.

Las funciones de los EIS que conviene que sean absorbidas por Data Warehouse son aquellas que requieren una presentación menos espectacular y tienen una variedad de posibles consultas muy grande. Absorbiendo estas funciones, un Data Warehouse puede producir muy significativos ahorros de coste de mantenimiento del EIS. De hecho están surgiendo nuevos productos que combinan la tecnología Data Warehouse con la opción de definir output tipo EIS. Es una combinación muy lógica y a la vez potente, que le da un alto valor a los productos que la usen.

Data Mart: ¿hermano menor de Data Warehouse?

A pesar de las grandes ventajas de Data Warehouse, parecen existir unas importantes barreras para su utilización en empresas de tamaño mediano. Los productos Data Warehouse han nacido para resolver problemas de análisis de grandes masas de información, en empresas donde una pequeña diferencia en el valor de una variable, puede afectar la cuenta de resultado con unas diferencias de millones de dólares.

Los productos y proyectos Data Warehouse están dimensionados para este tipo de empresas, contando con hardware muy potente (muchas veces especializado) y la masiva intervención de consultores externos, expertos en la realización de la puesta en marcha. Un proyecto de este tipo resulta en todos los aspectos excesivo para un departamento de ventas que necesita analizar la información de 500.000 - 3.000.000 de líneas de pedidos, o una cantidad equivalente de información financiera, que es lo normal para una empresa mediana.

Recientemente han surgido los denominados Data Mart, productos que utilizan la tecnología Data Warehouse adaptada a las necesidades de las empresas

Para resolver este tipo de necesidades han surgido los *Data Mart, productos que utilizan la tecnología Data Warehouse adaptada a las necesidades de las empresas medias.* Data Mart se destaca por una definición de requerimientos más fácil y rápida. También se simplifica el desarrollo de todo el mecanismo de su base de datos y con ello baja substancialmente todo el coste del proyecto, así como su duración. Normalmente, Data Mart resuelve aplicaciones a nivel departamental, aunque en ocasiones se desarrolla una aplicación que integre todas ellas y proporciona las funciones de un EIS.

Los esfuerzos de los desarrolladores de productos Data Mart, junto con los mejoras del índice precio/rendimiento del hardware, suben constantemente el límite de penetración de Data Mart, permitiendo asumir proyectos más y más importantes. La simplicidad de los proyectos de Data Mart y el menor coste en comparación con Data Warehouse, significan una ventaja competitiva muy grande a favor de Data Mart, donde el mercado de los dos tipos de productos se solapa.

Data Mining

Data Mining es, aparentemente hasta ahora, la forma más avanzada de extraer la información de las bases de datos. En su máxima expresión, ya no es el usuario quien formula las consultas. "Agentes inteligentes" recorren las bases de datos y buscan en ellas posibles relaciones. Veamos un ejemplo distinto al que casi siempre se ha visto en las revistas, (el de la relación de la hora en que se compran los pañales y la cerveza por cajas):

Si en la base de datos está la información de venta de agua mineral por días y las condiciones climatológicas, lo obvio es que existirá una relación y no se necesita data mining para cuantificarla. Sin embargo si esta relación cuantificada la comparamos con la del año pasado, seguro que no coincidirá. ¿Qué ha cambiado? Aquí es dónde empieza tener sentido utilizar data mining. Las preguntas que surgen aquí son: ¿Cuántos de los factores que han producido el cambio están reflejados en nuestra base de datos y con qué precisión? ¿Hasta qué punto se las pueden arreglar los agentes inteligentes con las deficiencias del diseño de una base de datos y la falta de su normalización? ¿Cuánto tratamiento previo hay que darle a la base de datos, para que tenga sentido empezar con data mining y cuántas posibles relaciones útiles se perderán en este tratamiento? Pero indudablemente, data mining nos puede aportar ideas muy importantes. La cuestión es qué coste de hardware, software y recursos humanos tienen.

Parece por todas las preguntas sugeridas y muchas otras que seguramente quedan en el tintero, que aunque se hacen muchas presentaciones de data mining a empresas medianas y hasta pequeñas, este tipo de tecnología, por algún tiempo, solo puede resultar rentable para empresas muy grandes.

Conclusiones

He tratado aquí de escribir en forma sumamente resumida sobre las distintas herramientas, usadas para poder trabajar con la información que se oculta en nuestras bases de datos transaccionales, sus características y su evolución. Las características de cada tipo de producto tal como se han comentado, muchas veces no se presentan en forma tan tajante como en este artículo, puesto que existen productos híbridos. Por otro lado la forma de implantar un producto y la calidad profesional de los instaladores, pueden significar una diferencia enorme entre dos instalaciones de un mismo producto. Si el artículo les ha ayudado a pensar qué tipo de preguntas se debe hacer a los proveedores de este tipo de herramientas, considero que ha cumplido su misión.

Aparte de esto nunca sobra volver a recordar las reglas básicas, tales como la regla 80 / 20, la que distintos tipos de tarea requieren herramientas distintas y que para cada tamaño del problema hay que calibrar el tamaño de la solución (mejor un martillo y un destornillador que un martillo más grande). Ni el uso de las más avanzadas tecnologías se escapa a las reglas básicas del sentido común. ■

Jorge Gros *es socio y director de Software Greenhouse, S.A. Si desea contactar con él, su E-Mail es jgros@swgreenhouse.com*