DATA WAREHOUSE (BODEGA DE DATOS). HERRAMIENTA PARA LA TOMA DE DECISIONES (Parte II)

ALONSO TAMAYO ALZATE* NÉSTOR DARÍO DUQUE MÉNDEZ*

PC: Base de datos, Bodega de datos, Datamarts, Cliente-servidor.

RESUMEN

Este artículo es el complemento de la primera parte y enfoca desde una perspectiva más técnica la importancia, los conocimientos requeridos, las posibles arquitecturas, las fases de desarrollo, las dificultades y retos que se encuentran involucradas en un proceso de Data Warehouse; también esboza las herramientas de análisis y brevemente explica algunas técnicas de minería de datos.

Introducción

En la parte I de este artículo se presentó un esbozo de la importancia que reviste para la organización, contar con herramientas de gestión de la información para soportar las decisiones, igualmente mostraba al Data Warehouse como la combinación de las tecnologías de almacenamiento de datos y el proceso de información que convierte esos datos fríos en valioso recurso para la dirección empresarial.

Enfrentar un proyecto de Data Warehouse implica apoyarse en diferentes técnicas de campos tales como el administrativo, diseño e implementación de bases de datos, administración y análisis de información, entre otros, lo que exige definir con base a una situación concreta, la arquitectura de la bodega de datos.

Como todo proyecto informático, su desarrollo recorre varias etapas; éstas son similares a las de cualquier otro proyecto, pero con particularidades muy específicas para el Data Warehouse. Es claro, que el Data Warehouse tiene como objetivo, partiendo de datos almacenados, extraer información relevante para el negocio. Dado que mucha

* Profesores Universidad Nacional de Colombia Sede Manizales. Departamento Administración y Sistemas.

de esta información está oculta, no es de fácil obtención y tiene el agravante de presentar grandes volúmenes de datos almacenados, por lo que es menester apoyarse o construir agentes o herramientas de análisis de gran potencia y versatilidad.

Todo lo anterior deja ver la importancia de abordar los proyectos de bodegas de datos, con la conciencia de los retos y riesgos que se corren, además de los recursos humanos, técnicos, económicos y de tiempo que se involucran.

Técnicas requeridas para el Data Warehouse

- **Técnicas Administrativas.** La información del Data Warehouse es propia para cada empresa, está estrechamente ligada con el negocio que se está sistematizando, por lo tanto, el diseño e implementación deben apoyar la solución a las necesidades planteadas. Se debe partir de los requerimientos funcionales de información que generen una ventaja competitiva para la empresa y faciliten la toma de decisiones por parte de la administración. Como plantean Gill y Rao "Con frecuencia, el reto reside en transformar los enunciados estratégicos generales de la empresa en indagaciones empresariales precisas y después convertirlos en solicitudes y reportes del Data Warehouse".

- **Técnicas de almacenamiento y extracción de datos.** Recordemos que varios son los procesos asociados con ésta tecnología: Población inicial y actualizaciones, almacenamiento y análisis de datos.

Como se explicó en el artículo anterior, en ocasiones los datos que poblarán la bodega de datos provienen de diferentes orígenes, por lo tanto se requiere definir una estructura y esquema eficiente; además, consolidar esos datos implica conocer y manejar diferentes sistemas, diferentes motores de bases de datos y eventualmente varios lenguajes de programación, que permitan la extracción de datos desde las fuentes. Las extracciones iniciales implicarán generalmente una conversión de tipo de datos y el manejo de datos ausentes o inconsistentes, que garantice la integridad.

Las actualizaciones implican la extracción de datos desde sistemas en operación, que se harán periódica y cíclicamente. Se requiere, de acuerdo al conocimiento de cada situación en particular, definir si se hará semanal, mensualmente o en otro período establecido. Actualizaciones muy constantes normalmente no benefician el análisis de datos, puesto que rara vez cambian las tendencias y/o comparaciones. Se recomienda, en caso de extracciones voluminosas hacerlo hacia un archivo, esto facilita reiniciar desde distintos puntos, repetir el cargue y preprocesar antes de enviar a la

red. Los cargues deben ser masivos, aprovechando los utilitarios de las bases de datos o las rutinas desarrolladas para esto y no una simple instrucción insert, que generalmente es ineficiente. Es usual y conveniente eliminar índices en este proceso y posteriormente volverlos a crear.

En el caso de los refrescos es preferible manejar la detección y propagación de cambios. Eventualmente usar triggers (disparadores: Acciones especiales definidas por el usuario que son automáticamente ejecutadas por el servidor de bases de datos a partir de eventos sucedidos: insert, update, delete) o aplicaciones propias. También es permitido la comparación de versiones, que algunos sistemas operativos apoyan a través de breves comandos.

En el almacenamiento se deben usar estrategias para lograr eficiencia. En las bodegas de datos es posible manejar diversos niveles de granularidad, a menor granularidad mayor es la cantidad de detalle. Para aumentar la granularidad, los datos operacionales deben resumirse y acumularse. Entre mayor sea la granularidad mas procesamiento se tendrá para convertir y resumir los datos desde las fuentes pero, al mismo tiempo, menor será el volumen de almacenamiento y mayor la facilidad de las consultas. Como se nota algunos datos se pueden almacenar como agregados, eso implica un especial cuidado al momento de los refrescos, para que estos datos sumarios también sean actualizados.

Otro elemento importante son las dimensiones de categorización. Un especial interés al momento del análisis es el tiempo, que permite determinar tendencias e información por períodos. También se usan las siguientes dimensiones: grupos de clientes, líneas de productos, ubicación geográfica, grupo industrial, área en la organización, estrato social y las especificas del negocio a modelar.

No obstante lo dicho hasta ahora, existen varios enfoques de la arquitectura del Data Warehouse y en algún caso podría optarse por no generar copias de los datos de las aplicaciones en producción, sino utilizar los datos operacionales usando aplicaciones que los consulten directamente.

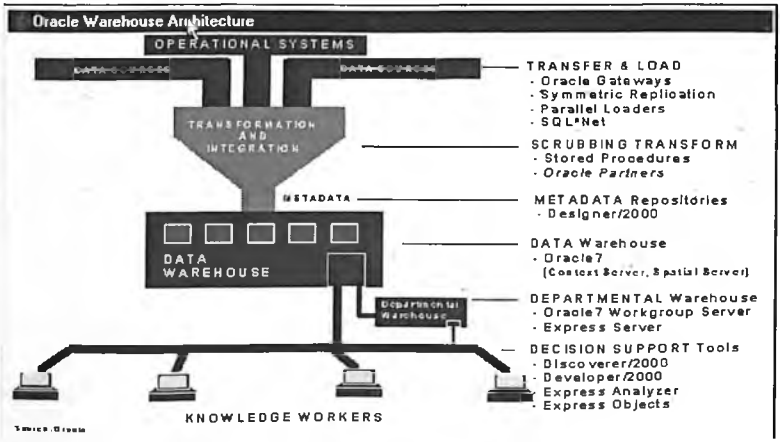
- **Técnicas de administración del programa y análisis de datos.** No es suficiente con almacenar un volumen alto de datos; el Data Warehousing implica la gestión de los mismos para convertirse en vital herramienta de soporte a las decisiones, derivar conclusiones a partir de la historia. Esto incluye el descubrimiento de patrones y tendencias que puedan ser extrapoladas, e intentar predecir comportamientos futuros. Estas técnicas se basan en las matemáticas, estadísticas, en la psicología, algoritmos genéticos, redes neuronales e incluso en la experiencia.

Partiendo de datos almacenados, es posible obtener consultas sencillas y descriptivas de datos independientes; también se pueden obtener reportes que manejen varias dimensiones y permitan crecer o bajar en la granularidad, dando una visión de los diferentes valores combinados cuando se requiera. Pero, como se planteó en el párrafo anterior, se puede ser más exigentes y a partir de software especializado, optar por la predicción.

Selección de Arquitectura del Data Warehouse

Otro elemento que reviste importancia al momento de implementar una bodega de datos, es la selección de la arquitectura. La arquitectura enfoca el proyecto como componentes (Fuente de datos, bodega de datos, datamart, su acceso y uso).

El diagrama siguiente explica como se organizan estos elementos. Este es un caso particular propuesto por la firma Oracle.



Otros proveeedores ofrecen diferentes modelos, de los cuales varios enfoques son elegibles:

- Consultas desde un esquema virtual hacia los datos operacionales. Normalmente una bodega de datos se asocia con un almacén donde se hacen copias de datos de aplicaciones en producción y de carácter histórico. En esta arquitectura se elimina la copia y actualización y se usan los datos de las bases de datos operacionales, a partir del metamodelo del Data Warehouse, los cuales se accesarán al momento de la consulta.

• Almacenamiento propio a partir de varias fuentes. Bodega de datos empresarial, no necesariamente centralizada. Se apoya en la normal necesidad de preprocesar los datos desde las fuentes en operación y aboga por realizar esta tarea una vez y almacenar los datos en bases propias, que serán actualizadas periódicamente; a partir de éstas se aplican las herramientas de análisis. Esta estrategia asegura la consistencia, pero es compleja de crear.

• Datamarts o mercado de datos únicamente. Plantea y reconoce las particularidades de cada área o departamento de una organización y la imposibilidad de ser satisfechos sus requerimientos por un solo Data Warehouse. El concepto de datamarts es una analogía a tiendas de vecindario que sirven a la población del sector, en lugar de un gran supermercado que abastece toda la ciudad. Los Datamarts son sub-bodegas organizadas por temas a nivel de departamentos. Esta arquitectura solo usa datamart.

• Data Warehouse y mercado de datos. Es una combinación de las dos anteriores, el Data Warehouse corporativo es un recopilador y distribuidor de la información sin desconocer las particularidades especificas de cada área. Esta estrategia permite posibles inconsistencias en los datos.

• Cliente Servidor en dos capas. Solo existen servidores de datos y clientes que los usan. En el servidor (o servidores) residen las fuentes de datos, el Data Warehouse y los datamarts. En los clientes, se ejecutan las herramientas de acceso del usuario final; éstas son generalmente aplicaciones gráficas.

• Cliente Servidor en tres capas. Las tareas se dividen en tres niveles:

- Un servidor de datos, que contiene las fuentes de los datos.
- Un servidor de aplicaciones, que contienen los datos de la bodega de datos y manejan el software de Data Warehouse y datamarts.
- La porción cliente, que manejan las aplicaciones de consulta y reporte.

Construcción del Data Warehouse

El ciclo del desarrollo del Data Warehouse no difiere en mucho de las fases de perfeccionamiento de todos los desarrollos de software. Las fases y las secuencias son las mismas, pero existen variantes únicas asociadas al Data Warehouse.

Comprende :

Planeación. En esta fase se determina:

- El enfoque que se optará para la implementación: Top-Down (De Arriba abajo), Bottom-up (De abajo a arriba) o una combinación de estos.
- La metodología de desarrollo: Las más usuales son el método de análisis y diseño estructurado y el método del desarrollo en espiral.
- El alcance inicial de proyecto.
- Selección del enfoque arquitectónico.
- Programa y presupuesto.
- Definir las expectativas del usuario final.
- Recopilación de metadatos.

Requerimientos

Especificación clara y precisa de las funciones que se esperan obtener del Data Warehouse. Estos deben definirse desde varias perspectivas: propietario, arquitecto/ desarrollador del Data Warehouse y desde la visión del usuario. Se definen las áreas tema que apoyará la bodega de datos, el nivel de detalle de la información requerida (nivel de granularidad), las dimensiones de categorización (tiempo, geografía, industria, grupo de clientes, línea de producto, etc.).

Análisis

Consiste en convertir todos los requerimientos conseguidos en la fase anterior en especificaciones concretas que sirvan de base para el diseño. Se definen los modelos lógicos de los datos para el Data Warehouse, los mercados de datos, definir los procedimientos de conexión con las fuentes de datos y el Data Warehouse y las herramientas de acceso del usuario final.

Diseño

Los modelos lógicos conseguidos en la anterior fase se convierten en modelos físicos. Se generan los diseños para programas y procesos que se requieren según la arquitectura, tanto a nivel de los datos como de aplicación.

Construcción

Se conoce también como diseño físico y consiste en plasmar en la práctica, los diseños lógicos de la fase anterior. Incluye la construcción de programas que creen y

modifiquen las bases de datos, que extraigan datos de las fuentes, programas para transformación de datos tales como integración, resumen y adición, programas para la actualización de los datos y programas para búsquedas en bases de datos muy grandes.

Montaje

Relacionados con la instalación, puesta en marcha y uso del Data Warehouse. Un elemento importante consiste en concientizar a los usuarios sobre la disponibilidad, beneficios y presentación de Data Warehouse, esto se conoce como comercialización de la información.

Retos de la Implementación

Como se aprecia, enfrentar un proyecto de Data Warehouse exige el conocimiento de la empresa, capacidades administrativas y fortalezas técnicas. Estos proyectos deben ser asumidos por equipos de trabajo multidisciplinarios, que logren que las ventajas potenciales se lleven a la práctica.

Estas son algunas de las tareas que deben ser sorteadas por el equipo:

1. La integración de datos y metadatos de diferentes fuentes y épocas. Esto conlleva la necesidad de generar datos a almacenar en forma consistente partiendo de datos símiles, sin perder información importante.

2. Limpieza, filtrado y refinación de los datos. Para el proceso de análisis de los datos es problemático la ausencia de valores de atributos y la existencia de valores ilógicos o inconsistentes.

3. En los sistemas de procesamiento en línea (OLTP) los detalles de las operaciones son muy importantes mientras que en el Data Warehouse se busca almacenar datos en forma condensada y agrupada.

4. Siendo la bodega de datos el resultado de la importación de datos de diferentes fuentes, las cuales son dinámicas porque cambian con el tiempo, se requiere generar mecanismos que garanticen la sincronización y aseguren la actualización a partir de los cambios en las fuentes.

5. Para una correcta operación de la bodega de datos es necesario tener la correcta información sobre los datos que se tienen almacenados, la administración de metadatos toma importancia.

Herramientas de Análisis

Recordemos que no es suficiente con almacenar datos, es necesario procesarlos para convertirlos en información importante para la organización.

Los sistemas de apoyo a las decisiones (DSS), conecta a las personas con las bodegas de datos. De la calidad de estas herramientas depende el grado de aprovechamiento de las mismas. Pueden ser:

- Herramientas de consultas/reportes, con interfaz gráfica, que facilitan sin usar sentencias SQL, realizar queries complejos.
- Herramientas OLAP (Online Analytical Processing), generando consultas multidimensionales, con columnas y filas móviles y diversos grados de agrupamiento, para diferentes parámetros.
- Minería de Datos (Datamining). En octubre de 1995 Edmun DeJesús, editor de la famosa revista BYTE Magazine, escribió al respecto: "Gracias a la minería de datos, las computadoras se encargan de seleccionar vastos almacenes de datos. Con una incansable e incesante búsqueda, será posible encontrar la diminuta pepita de oro en una montaña de datos de desperdicio". En datamining las búsquedas se hacen sobre datos dispersos, con poca o ninguna intervención del usuario. No se requiere formular un requerimiento estricto para que la herramienta entregue algunas relaciones ocultas y patrones interesantes, conseguidos a través de clasificación y predicción.

Algunas aplicaciones de estas técnicas están directamente relacionadas con el mercadeo de producto, pudiendo predecir el comportamiento de los clientes ante una oferta o un producto en particular, de acuerdo a su ubicación geográfica. También para conocer las preferencias de los consumidores y tomar medidas que los acerquen a los productos que se distribuyen.

En estas tareas de minería de datos se encuentran inconvenientes inherentes a las bodegas de datos, ellos pueden ser:

- Grandes volúmenes de información y altamente dimensionales, lo que dificulta el hallazgo de patrones.
- Valores inconsistentes o no existentes en algunos atributos importantes. Estas situaciones deberían haberse corregido en la fase de población y actualización, pero en caso de presentarse se debe tener una política para su manejo.

- La representación de los resultados no siempre es comprensible para todos los usuarios.
- Valor estadístico de los patrones hallados.

Hoy existe una buena cantidad de productos de diversos fabricantes para la minería de datos, varios impulsados por universidades reconocidas; destacándose entre ellos: Intelligent Miner (IBM), KDD Project (GTE laboratories), Datamind (Datamind Inc), Saxon (PMSI). Algunos se pueden conseguir en sitios Internet, para las diferentes plataformas: Data Surveyor (www.ddi.nl, IDIS (http, VisDB (http.) Este último producto tiene una versión para el sistema operativo Linux. El VisDB se ha desarrollado para apoyar la exploración de bancos de datos grandes. Los instrumentos de VisDB implementan severas técnicas visuales, permitiendo trabajar con bodegas de datos de aproximadamente un millón de valores de datos. Las técnicas apoyadas por el sistema son: Técnicas orientadas a pixel (espirales, Ejes y Técnicas de Agrupación), Coordenadas Paralelas y figuras de madera.

Estos productos, en forma integrada o separada se basan en: Redes neuronales, algoritmos genéticos, arboles de decisión, algoritmos estadísticos, funciones de visualización gráfica, técnica de K-vecinos, reglas de producción.

Algunos pasos deben seguirse para lograr resultados provechosos:

- Qué se espera?. Qué se quiere descubrir?
- Conjuntos de datos que se analizarán.
- Pre-procesamiento. Buscan desechar los valores con desviaciones muy altas, generados por ausencia o datos incorrectos.
- Limpieza. A partir de un previo conocimiento obtenido en los pasos anteriores se determinan las variables y registros que realmente representarán importancia.
- Elegir la función de la minería y sus algoritmos.

Técnicas de Minería de Datos

Como las enumeradas en el artículo anterior, las más utilizadas son redes neuronales, árboles de decisión, algoritmos genéticos, análisis de correlaciones y k-vecinos. Todas estas técnicas pueden ser mezcladas para obtener los resultados esperados. Se presenta a continuación una breve descripción de cada una de ellas:

Redes Neuronales Artificiales (RNA). Como su nombre lo indica, simula el sistema nervioso real en forma abstracta. Estas deben ser entrenadas para que brinden solución a los problemas. Esta enseñanza se realiza repitiendo sistemáticamente entradas clásicas, con sus respectivas salidas o respuestas. Son usadas para reconocimiento de patrones, clasificaciones de voz e imagen, procesamiento de lenguaje natural, predicción y optimización.

Arboles de Decisión (AD). Representan reglas donde atributos independientes determinan los valores finales. En estos árboles cada nodo representa una propiedad que puede tomar diversos valores, cada uno de los cuales genera una rama. Los nodos hojas representan las clasificaciones finales, usadas donde se deben tomar decisiones a partir de varias alternativas combinadas y con pesos diferentes. Son útiles en problemas de alta dimensionalidad y pequeño número de valores para cada atributo. Se usan, por enumerar unos, en dominios médicos y en simulaciones de juegos de ajedrez.

K-Vecinos. Usa razonamiento basado en memoria (MBR) para las predicciones. Identifica los vecinos mas cercanos (valores similares para igual atributo) y observa como se comporta la variable de salida. Parte de un conjunto de datos modelo que representa el mecanismo de clasificación y se determina la cantidad de vecinos que participan en la clasificación (K). Es permitido ponderar atributos para expresar su importancia en la técnica.

Reglas de Producción. Generalmente son transformaciones de árboles de decisión que han crecido mucho, llevándolos al plano proposicional, lo cual facilita el entendimiento.

BIBLIOGRAFÍA

GILL, Harjinder; RAO Prakash. Data Warehousing. La integración de información para la mejor toma de decisiones. Prentice Hall, 1996.

PECH ESCALANTE, Ivan. Soluciones Avanzadas No.34. Data Warehouse.

ORFALI, Robert; HARKEY Dan; EDWARDS, Jeri. Cliente/Servidor Guía de Supervivencia. McGraw-Hill. 1997

RUBLE, David. Análisis y Diseño Práctico de Sistemas Cliente/Servidor con GUI. Prentice Hall. 1997

VILLAMIL, Omar. Herramientas de Minería de Datos. Conferencias Internet.

Artículos de Internet. Búsquedas con WebFerret de la casa FerretSoft.