APOYO PARA

LA TOMA DE DECISIONES

Cátedra: Gestión de Datos
Profesor: Santiago Pérez
Año: 2006
Bibliografía: Introducción a las Bases de Datos. DATE

Prof. Santiago Perez
Nestor Fernandeze

"APOYO PARA LA TOMA DE DECISIONES"

1. INTRODUCCION

Los sistemas de apoyo para la toma de decisiones son sistemas que ayudan en el análisis de información de negocios. Su propósito es ayudar a la administración para que "marque tendencias, señale problemas y tome decisiones inteligentes".
La idea básica es recolectar datos operacionales del negocio y reducirlos a una forma que pudiera ser usada para analizar el comportamiento del mismo y modificarlos de una manera inteligente.

1.1 ASPECTOS DEL APOYO PARA LA TOMA DE DECISIONES

Las bases de datos de apoyo para la toma de decisiones muestran determinadas características especiales, de las cuales sobresale ésta: la base de datos es principalmente (aunque no totalmente) de sólo lectura. Por lo general, la actualización que se da está limitada a operaciones de carga o actualizaciones periódicas y esas operaciones están dominadas a su vez por INSERTs – los DELETEs se realizan muy ocasionalmente y los UPDATEs casi nunca.
También vale la pena señalar las siguientes características adicionales de las bases de datos de apoyo para la toma de decisiones.

- Se tiende a usar las columnas en combinación.
- Por lo general, no preocupa la integridad.
- Las claves incluyen frecuentemente un componente temporal.
- La base de datos tiende a ser grande (especialmente cuando se acumulan los detalles de las transacciones de negocios a lo largo del tiempo, y con frecuencia así sucede).
- La base de datos tiende a estar muy indexada.
- La base de datos involucra frecuentemente varios tipos de redundancia controlada.

Las consultas de apoyo para la toma de decisiones tienden a ser bastante complejas. Éstos son algunos de los tipos de complejidades que pueden presentarse:

- Complejidad de expresiones lógicas: las consultas de apoyo para la toma de decisiones involucran expresiones complejas en la cláusula WHERE; las cuales son difíciles de escribir, difíciles de comprender y difíciles de manejar adecuadamente por el sistema.
- Complejidad de juntas: las consultas de apoyo para la toma de decisiones requieren frecuentemente acceso a muchas clases de hecho. Por consecuencia, en una base de datos diseñada adecuadamente dichas consultas involucran, por lo general a muchas juntas.
- Complejidad de función: las consultas de apoyo para la toma de decisiones involucran frecuentemente funciones estadísticas y matemáticas. Pocos productos soportan tales funciones.
- Complejidad analítica: las preguntas de negocios rara vez son respondidas con una sola consulta. No sólo es difícil para los usuarios escribir consultas de gran complejidad, sino que las limitaciones que tienen las implementaciones de SQL pueden impedir el procesamiento de una de las consultas.

1.2 DISEÑO DE BASE DE DATOS DE APOYO PARA LA TOMA DE DECISIONES

El diseño de base de datos debe ser realizado en al menos dos etapas, primero la lógica y luego la física:

a- El diseño lógico debe ser analizado primero. En esta etapa, el enfoque está en la corrección relacional: las tablas deben representar relaciones adecuadas y por lo tanto garantizar que las operaciones relacionales funcionen tal como se indica y no produzcan resultados sorprendentes.

b- Segundo, el diseño físico debe surgir a partir del diseño lógico. Por supuesto, en esta etapa el punto de atención está puesto sobre la eficiencia y el rendimiento del almacenamiento. En principio es permisible cualquier acomodo físico de los datos, siempre y cuando exista una transformación que conserve la información entre los esquemas lógico y físico, y que pueda ser expresada en el álgebra relacional.

1.2.1 Diseño lógico

Las reglas de diseño lógico no dependen del uso que se pretenda dar a la base de datos, ya que se aplican las mismas reglas sin tomar en cuenta los tipos de aplicaciones. Por lo tanto, no debe haber diferencia si esas aplicaciones son operacionales (OLTP) o de apoyo para la toma de decisiones; de cualquier forma, es necesario seguir el mismo procedimiento de diseño. Entonces, volvamos a ver las tres características lógicas de las bases de datos de apoyo para la toma de decisiones.

- *Combinaciones de columnas y muy pocas dependencias*
  Las consultas de apoyo para la toma de decisiones con frecuencia tratan a las combinaciones de columnas como una unidad, lo que significa que las columnas componentes nunca son acreditadas en forma individual. El efecto neto es que la cantidad total de dependencias se reduce y el diseño se vuelve más sencillo, con menos columnas y posiblemente hasta menos tablas.
- *Las restricciones de integridad en general*
  Como ya explicamos, (a) las bases de datos de apoyo para la toma de decisiones son principalmente de sólo lectura y (b) la integridad de los datos se verifica al cargar (o actualizar) la base de datos.
  Declarar las restricciones proporciona un medio para decirles a los usuarios lo que significa los datos y por lo tanto, les ayuda en sus tareas de formular consultas. Además, la declaración de restricciones también puede proporcionar información crucial al optimizador.
- *Claves temporales*
  Por lo general, las bases de datos operacionales involucran sólo datos actuales. Por el contrario, las bases de datos de apoyo para la toma de decisiones involucran por lo general datos históricos y por lo tanto, tienden a poner marcas de tiempo en la mayoría o en todos los datos. Por consecuencia, las claves de dichas bases de datos incluyen frecuentemente columnas de marca de tiempo.

1.2.2 Diseño físico

Las bases de datos de apoyo para la toma de decisiones tienden a ser grandes y fuertemente indexadas, e involucran diversos tipos de redundancia controlada.

Primero consideramos el partido (también conocido como fragmentación). El partido representa un ataque al problema de la " base de datos grande"; divide una tabla dada en un conjunto de particiones o fragmentos separados para efectos de almacenamiento físico. Dichas particiones pueden mejorar significativamente el manejo y la accesibilidad de la tabla en cuestión.
Ahora pasemos al indexado. Por supuesto, es bien conocido que el uso del tipo adecuado de índice puede reducir drásticamente la E/S. La mayoría de los primeros productos SQL proporcionaban solamente un tipo de índice, el árbol B, pero a través de los años se han tenido otros tipos disponibles, en especial en conexión con las bases de datos de apoyo para la toma de decisiones; estos incluyen a los índices de mapa de bits, dispersión, multitabla, lógicos y funcionales, así como a los de árbol B en sí.

- Índice de árbol B. Los índices de árbol B proporcionan acceso eficiente para consultas de alcance. La actualización de árboles B es relativamente eficiente.
- Índices de mapas de bits. Estos índices son eficientes para las consultas que involucran conjuntos de valores, aunque llegan a ser menos eficientes cuando los conjuntos se vuelven demasiado grandes.
- Índices de dispersión. Los índices de dispersión son eficientes para acceder a filas específicas (no rangos).
- Índices multitabla. Una entrada de índice multitabla contiene esencialmente apuntadores hacia filas de varias tablas, en lugar de sólo hacia filas de una tabla. Dichos índices pueden mejorar el rendimiento de las juntas.
- Índices lógicos. Un índice lógico indica para qué filas de una tabla específica, una expresión lógica específica da como resultado verdadero.
- Índices funcionales. Un índice funcional indexa las filas de una tabla no con base en los valores de esas filas, sino con base en el resultado del llamado a alguna función especificada sobre esos valores.

Por último, pasemos al asunto de la redundancia controlada. La redundancia controlada es una herramienta importante para reducir E/S y minimizar la contienda. La redundancia está controlada cuando es administrada por el DBMS y está oculta para los usuarios. Hay dos tipos amplios de esta redundancia:

- El primero implica mantener copias exactas, o réplicas, de los datos básicos.
- El segundo implica mantener datos derivados además de los datos básicos, muy frecuentemente en la forma de tablas de resumen o de columnas calculadas o derivadas.

1.2.3 Errores comunes de diseño

- *Filas duplicadas*. Los diseñadores de apoyo para la toma de decisiones dicen con frecuencia que sus datos simplemente no tienen un identificador único y que por lo tanto, tienen que permitir duplicados. Esto surge debido a que el esquema físico no deriva a partir de un esquema lógico (el cual probablemente nunca se creó).
- *Esquemas de estrella*: son el resultado de intentar "tomar atajos" en una técnica adecuada de diseño. Es podo lo que se puede ganar con esos atajos. Con frecuencia afectan el rendimiento y la flexibilidad conforme crece la base de datos.
- *Nulos*. Los diseñadores tratan frecuentemente de ahorrar espacio permitiendo nulos en las columnas. Sin embargo, por lo general dichos intentos son erróneos.

- *Diseño de tablas de* resumen. La cuestión del diseño lógico de tablas de resumen es con frecuencia ignorada, lo que da lugar a una redundancia no controlada y a dificultades para mantener la consistencia.
- *Varias rutas de navegación*. A menudo, los diseñadores de apoyo para la toma de decisiones y los usuarios dicen (incorrectamente) que hay una "multiplicidad de rutas de navegación" hacia algún dato deseado, cuando en realidad quieren decir que los mismos datos pueden ser alcanzados por medio de varias expresiones relacionales diferentes.

Es claro que los usuarios pueden llegar a confundirse en tales casos y no estar seguros de que expresión usar o de si habrá alguna diferencia o no en el resultado. Por supuesto, parte de este problema solo puede ser resuelta mediante una enseñanza adecuada para el usuario. Otra parte pude ser resuelta si el optimizador hace su trabajo adecuadamente.

1.3 PREPARACION DE LOS DATOS

Los datos deben ser extraídos de diversas fuentes, limpiados, transformados y consolidados, cargados en la base de datos de apoyo para la toma de decisiones y luego actualizados periódicamente. Cada una de estas operaciones involucra sus propias consideraciones especiales.

1.3.1 Extracción

La extracción es el proceso de capturar datos de la base de datos operacionales y otras fuentes. Hay muchas herramientas disponibles para ayudar en esta tarea, incluyendo herramienta proporcionadas por el sistema, programas de extracción personalizados y productos de extracción comerciales.

1.3.2 Limpieza

Pocas fuentes de datos controlan adecuadamente la calidad de los datos. Por consecuencia, los datos requieren frecuentemente de una limpieza (por lo gral. por lote) antes de que puedan ser introducidos en la base de datos de apoyo para la toma de decisiones. Las operaciones de limpieza típicas incluyen el llenado de valores faltantes, la corrección de errores tipográficos y otros de captura de datos, el reemplazo de sinónimos por identificadores estándares etc.

1.3.3 Transformación y consolidación

Aun después de haber sido limpiados , es probable que los datos todavía no estén de la forma en que se requieren para el sistema de apoyo para la toma de decisiones y por lo tanto, deberán ser transformados adecuadamente. Por lo general, la forma requerida sera un conjunto de archivos, uno por cada tabla identificada en el esquema físico.
La transformación es importante cuando necesitan mezclarse varias fuentes de datos, un proceso al que se llama consolidación.

1.3.4 Carga

Los fabricantes de DBMS han puesto considerable importancia en la eficiencia de las operaciones de carga. Para los propósitos actuales, consideramos que las operaciones de carga incluyen (a) el movimientos de los datos transformados y consolidados hacia la base

de datos de apoyo para la toma de decisiones, (b) la verificación de su consistencia (es decir, verificación de integridad) y (c) la construcción de cualquier índice necesario.

1.3.5 Actualización

La mayoría de las bases de datos de apoyo para la toma de decisiones (aunque no todas) requieren una actualización periódica de los datos para mantenerlos vigentes. La actualización involucra por lo general una carga parcial, aunque algunas aplicaciones de apoyo para la toma de decisiones requieren la eliminación de lo que hay en la base de datos y una recarga completa.

1.3.6 Almacenes de datos operacionales

Un ODS (almacén de datos operacionales) es una colección de datos actuales integrados y volátiles (actualizables) que están orientados a un tema. El termino orientado a un tema significa que los datos en cuestión tienen que ver con alguna área temática especifica (por ejem. clientes, productos etc.). Un almacén de datos operacionales puede ser usado (a) como un área transitoria para la reorganización física de los datos operacionales extraídos, (b) para proporcionar informen operacionales y (c) para apoyar la toma de decisiones operacionales.

2. DATA WAREHOUSES Y DATA MARTS

2.1 Data warehouses

Al igual que los almacenes de datos operacionales, una data warehouse es un tipo especial de base de datos: "un almacén de datos orientado a un tema, integrado, no volátil y variante en el tiempo, que soporta decisiones de administración" (donde el termino no volátil significan que una vez que los datos han sido insertados, no pueden ser cambiados, aunque sí borrados). Los data warehouses surgieron por dos razones: primero, la necesidad de proporcionar una fuente única de datos limpia y consistente para propósitos de apoyo para lo toma de decisiones; segundo, la necesidad de hacerlo sin afectar a los sistemas operacionales.

2.2 Data marts

Se define como un almacén de datos especializado, orientado a un tema, integrado, volátil, y variante en el tiempo para apoyar un subconjunto especifico de decisiones de administración. La principal diferencia entre una data marts y una data warehouse es que la primera es especializada y volátil.
Hay tres enfoques para la creación de una data mart:

- Los datos pueden ser simplemente extraídos de la data warehouse.
- A pesar del hecho de que la data warehouse pretende proporcionar un punto de control único una data mart puede ser creado todavía en forma independiente (es decir, no por medio de la extracción a partir de la data warehouse.
- Algunas instalaciones han seguido un enfoque de "primero la data mart" donde estos son creados conforme van siendo necesarios y la data warehouse gral. es creada, como una consolidación de los diversos data mart.

3. PROCESAMIENTO ANALITICO EN LINEA (OLAP)

El término OLAP puede ser definido como el proceso interactivo de crear, mantener, analizar y elaborar informes sobre datos y es usual añadir que los datos en cuestión son percibidos y manejados como si estuvieran almacenados en un arreglo multidimensional.
El primer punto, es que el procesamiento analítico requiere invariablemente, algún tipo de agregación de datos, por lo gral. en muchas formas diferentes.
Las desventajas de este enfoque son obvias: la formulación de tantas consultas similares pero distintas, es tediosa para el usuario y la ejecución de todas esas consultas es probablemente bastante costosa en tiempo de ejecución. Por lo tanto, debemos encontrar una forma de solicitar varios niveles de agregación en una sola consulta y ofrecer a la implementacion la oportunidad de calcular todas esas agregaciones de manera más eficiente.
Dichas consideraciones son la motivación que hay tras las opciones GROUPING SETS, ROLLUP y CUBE de la cláusula GROUP BY.
La opción GROUPING SETS permite al usuario especificar con exactitud qué agrupamientos específicos van a ser realizados. La opción ROLLUP se encarga de agrupar todas las combinaciones que se requieren, y la opción CUBE (poco útil) se deriva del hecho de que en la tecnología OLAP, los valores de datos pueden ser percibidos como si estuvieran almacenados en las celdas de un arreglo multidimensional o hipercubo.
Una cláusula GROUP BY dada puede incluir cualquier mezcla de especificaciones GROUPING SETS, ROLLUP y CUBE.

4. MINERIA DE DATOS

La minería de datos puede describirse como el análisis de datos exploratorio.
El propósito de buscar patrones interesantes en los datos, patrones que pueden usarse para especificar la estrategia del negocio o para identificar comportamientos fuera de lo común (por ejemplo un incremento súbito de una tarjeta de crédito puede indicar que la misma ha sido robada).
Podemos descubrir reglas de asociación más generales a partir de agregaciones adecuadas de los datos dados: por ejemplo el agrupamiento por cliente nos permitirá probar la validez de reglas tales como “si un cliente compra zapatos, es probable que también compre calcetines, aunque no necesariamente en la misma transacción”.
También podemos definir otro tipo de reglas: por ejemplo, una regla de correlación de secuencia podría ser usada para identificar patrones de compra a lo largo del tiempo (“sí un cliente compra zapatos hoy, es probable que compre calcetines dentro de los cinco días siguientes”). Una regla de clasificación podría ser usada para ayudar a decidir si se otorga un crédito (“sí un cliente tiene ingresos a $75.000 anuales es probablemente un buen sujeto de créditos”).