Sistemas de Bases de Datos 2

2024 Ing. Luis Alberto Arias Solórzano

Unidad 7

Inteligencia de Negocio

- 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"

- Los orígenes de dichos sistemas están en la investigación de operaciones, teorías de administración científicas o basadas en comportamiento, y control de procesos estadísticos.

- 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 de sólo lectura.**

DATA WAREHOUSES

- Un data warehouse es un tipo especial de base de datos, o diseño especial, se define como **"un almacén de datos orientado a un tema, integrado, no volátil y variante en el tiempo, que soporta decisiones de administración";** donde el término *no volátil* significa que una vez que los datos han sido insertados, no pueden ser cambiados, aunque sí pueden ser 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 la toma de decisiones; segundo, la necesidad de hacerlo sin afectar a los sistemas operacionales.

- Por definición, las cargas de trabajo del data warehouse están destinadas para el apoyo a la toma de decisiones y por lo tanto, tienen consultas intensivas (con actividades ocasionales de inserción por lotes); asimismo, los propios data warehouses tienden a ser bastante grandes y con un alta tasa de crecimiento. Por consecuencia, es difícil —aunque no imposible— perfeccionar el rendimiento. También puede ser un problema la escalabilidad.

DATA MARTS

- Es un tipo de "almacén" limitado que esta hecho a la medida de algún propósito. Proporcionaba un acceso más rápido a los datos, que si tuvieran que ser sincronizados con los demás datos cargados en todo el data warehouse.

- Existe alguna controversia sobre la definición precisa del término data mart. Para nuestros propósitos podemos definirlo como **"un almacén de datos especializado, orientado aun tema, integrado, volátil y variante en el tiempo para apoyar un subconjunto específico de decisiones de administración".**

- La principal diferencia entre un data mart y un data warehouse es que el data mart es *especializado* y *volátil.* Por *especializado* queremos decir que contiene datos para dar apoyo (solamente) a un área específica de análisis de negocios; por *volátil* queremos decir que los usuarios pueden actualizar los datos e incluso, posiblemente, crear nuevos datos (es decir, nuevas tablas) para algún propósito.

DATA MARTS

Hay tres enfoques principales para la creación de un data mart:

- ■ Los datos pueden ser simplemente extraídos del data warehouse; de hecho, sigue un enfoque de "**divide y vencerás**" sobre la carga de trabajo general de apoyo para la toma de decisiones, a fin de lograr un mejor rendimiento y escalabilidad. Por lo general, los datos extraídos son cargados en una base de datos que tiene un esquema físico que se parece mucho al subconjunto aplicable del data warehouse; sin embargo, puede ser simplificado de alguna manera gracias a la naturaleza especializada del data mart.

- ■ A pesar del hecho de que el data warehouse pretende proporcionar un "**punto de control único**", un data mart puede ser creado todavía en forma independiente (es decir, *no* por medio dela extracción a partir del data warehouse). Dicho enfoque puede ser adecuado si el data warehouse es inaccesible por alguna razón, digamos razones financieras, operacionales o incluso políticas (o puede ser que ni siquiera exista todavía el data warehouse; vea el siguiente punto).

- ■ Algunas instalaciones han seguido un enfoque de "**primero el data mart**", donde los data marts son creados conforme van siendo necesarios y el data warehouse general es creado, finalmente, como una consolidación de los diversos data marts

PREPARACIÓN DE LOS DATOS

Extracción

- La **extracción** es el proceso de capturar datos de las bases de datos operacionales y otras fuentes. Hay muchas herramientas disponibles para ayudar en esta tarea, incluyendo herramientas proporcionadas por el sistema, programas de extracción personalizados y productos de extracción comerciales (de propósito general).

- El proceso de extracción tiende a ser intensivo en E/S y por lo tanto, puede interferir con las operaciones de misión crucial; por esta razón, este proceso a menudo es realizado en paralelo (es decir, como un conjunto de subprocesos paralelos) y en un nivel físico.

- Sin embargo, dichas "extracciones físicas" pueden ocasionar problemas para el procesamiento subsecuente, ya que pueden perder información —en especial información de vínculos— que está representada de alguna manera física (por ejemplo, por apuntadores o por contigüidad física). Por esta razón, los programas de extracción proporcionan en ocasiones un medio para preservar dicha información introduciendo números de registro secuenciales y reemplazando apuntadores por lo que en realidad son valores de clave externa.

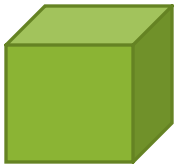
PREPARACIÓN DE LOS DATOS

Transformación y consolidación

- Además, las fechas y horas asociadas con el significado que tienen los datos en los negocios, necesitan ser mantenidas y correlacionadas entre fuentes; un proceso llamado "sincronización en el tiempo". Por razones de rendimiento, las operaciones de transformación se realizan frecuentemente en paralelo. Pueden ser intensivas tanto en E/S como en CPU.

EJEMPLO:

- La sincronización en el tiempo puede ser un problema difícil. Por ejemplo, suponga que queremos encontrar el promedio de las ganancias por cliente y por vendedor en cada trimestre. Suponga que los datos del cliente contra las ganancias son mantenidos por trimestre fiscal en una base de datos de contabilidad, y en cambio, los datos del vendedor contra el cliente son mantenidos por trimestre de calendario en una base de datos de ventas. De manera clara, necesitamos fusionar los datos de las dos bases de datos. La consolidación de clientes es fácil, involucra simplemente la coincidencia de IDs de clientes. Sin embargo, la cuestión de la sincronización de tiempo es mucho más difícil. Podemos encontrar las ganancias de cliente por trimestre *fiscal* (a partir de la base de datos de contabilidad), pero no podemos decir qué vendedores fueron responsables de cuáles clientes en ese momento y, a fin de cuentas, no podemos encontrar las ganancias de clientes por trimestre de *calendario.*



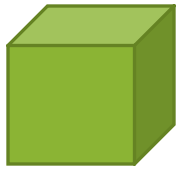
PREPARACIÓN DE LOS DATOS

Limpieza

- Pocas fuentes de datos controlan adecuadamente la calidad de los datos. Por consecuencia, los datos requieren frecuentemente de una **limpieza** (por lo general, 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 establecimiento de abreviaturas y formatos estándares, el reemplazo de sinónimos por identificadores estándares, etcétera. Los datos que son erróneos y que no pueden ser limpiados, serán reemplazados.

- En ocasiones, la información obtenida durante el proceso de limpieza puede ser usada para identificar la causa de los errores en el origen y por lo tanto, mejorar la calidad de los datos a través del tiempo.



PREPARACIÓN DE LOS DATOS

Transformación y consolidación

- Aun después haber sido limpiados, es probable que los datos todavía no estén en 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 será un conjunto de archivos, uno por cada tabla identificada en el esquema físico; como resultado, la transformación de los datos puede involucrar la división o la combinación de registros fuente.

- A veces, los errores de datos que no fueron corregidos durante la limpieza son encontrados durante el proceso de transformación. Por lo general cualquier dato incorrecto es rechazado. La información obtenida como parte de este proceso puede ser usada, en ocasiones, para mejorar la calidad de la fuente de datos.

- La transformación es particularmente importante cuando necesitan mezclarse varias fuentes de datos, un proceso al que se llama **consolidación.** En estos casos, cualquier vínculo implícito entre datos de distintas fuentes necesita volverse explícito (introduciendo valores de datos explícitos).

PREPARACIÓN DE LOS DATOS

Carga

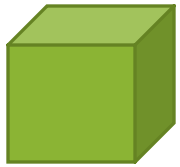
- Los fabricantes de DBMS han puesto considerable importancia en la eficiencia de las operaciones de **carga.** Para los propósitos actuales, consideramos que las "operaciones de carga"
- a. *Movimiento de datos.* Por lo general, los sistemas modernos proporcionan herramientas de carga en paralelo. En ocasiones formatearán previamente los datos para darles el formato físico interno requerido por el DBMS de destino antes de la carga real. (Una técnica alterna que proporciona gran parte de la eficiencia de las cargas pre-formateadas es cargar los datos en tablas de trabajo que se asemejan al esquema de destino. La verificación de la integridad necesaria puede ser realizada en esas tablas de trabajo y luego usar los INSERTS en el nivel de conjunto para mover los datos desde las tablas de trabajo hacia las tablas de destino.)
- b. *Verificación de integridad.* La mayor parte de la verificación de integridad de los datos a ser cargados puede ser realizada antes de la carga real, sin hacer referencia a los datos que ya están en la base de datos. Sin embargo, ciertas restricciones no pueden verificarse sin examinar la base de datos existente; por ejemplo, una restricción de unicidad tendrá que ser verificada, por lo general, durante la carga real (o por lotes después de que se haya terminado la carga).

PREPARACIÓN DE LOS DATOS

Carga

- c. *Construcción de índices.* La presencia de índices puede hacer significativamente lento el proceso de carga, debido a que la mayoría de los productos actualiza los índices conforme cada fila es insertada en la tabla subyacente. Por esta razón, en ocasiones es buena idea eliminar los índices antes de la carga y luego volverlos a crear. Sin embargo, este enfoque no vale la pena cuando la proporción de los nuevos datos (con respecto a los existentes) es pequeña, ya que el costo de crear un índice no se compensa con el tamaño de la tabla a indexar. Además, la creación de un índice grande puede estar sujeta a errores de asignación irrecuperables, y entre más grande sea el índice, es más probable que ocurran tales errores.

- La mayoría de los productos DMBS soportan la creación de índices en paralelo en un esfuerzo para agilizar los procesos de carga y de construcción de índices.



PREPARACIÓN DE LOS DATOS

Actualización

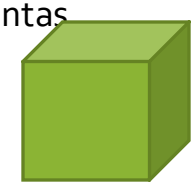
- La mayoría de las bases de datos de apoyo para la toma de decisiones, aunque no todas, requieren una **actualización** periódica de los datos para mantenerlos razonablemente vigentes.
- La actualización involucra por lo general una carga parcial, aunque algunas aplicaciones de apoyo para la toma de decisiones requieren la eliminación de lo que hay en la base de datos y una recarga completa. La actualización involucra todos los problemas que están asociados con la carga, pero también es probable que deba realizarse mientras los usuarios están accediendo a la base de datos.

PROCESAMIENTO ANALÍTICO EN LINEA (OLAP)

- El término **OLAP** ("procesamiento analítico en línea") fue acuñado en un artículo escrito por Arbor Software Corp. en 1993, aunque (como sucede con el término "data warehouse") el concepto es mucho más antiguo.

- 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 general en muchas formas diferentes (es decir, de acuerdo con muchos agrupamientos diferentes). De hecho, uno de los problemas fundamentales del procesamiento analítico es que la cantidad de agrupamientos posibles llega rápidamente a ser muy grande y los usuarios deben considerarlos todos o casi todos. Ahora bien, por supuesto que los lenguajes relaciónales soportan tal agregación, pero cada consulta individual en un lenguaje de éstos produce como resultado una sola tabla (y todas las filas de esa tabla tienen la misma forma y el mismo tipo de interpretación). Por lo tanto, para obtener *n* agrupamientos distintos se requieren *n* consultas distintas y *n* tablas de resultados distintas.



PROCESAMIENTO ANALÍTICO EN LINEA (OLAP)

- El término CUBO poco útil, se deriva del hecho de que en la terminología OLAP (o al menos multidimensional), los valores de datos pueden ser percibidos como si estuvieran almacenados en las celdas de un arreglo multidimensional o *hipercubo.* En el caso que estamos viendo (a) los valores de datos son cantidades, (b) el "cubo" es de dos dimensiones: una dimensión de proveedores y una dimensión de partes (¡y el "cubo" es bastante plano!) y por supuesto, (c) esas dos dimensiones son de tamaños desiguales (por lo que el "cubo" ni siquiera es un cuadrado, sino un rectángulo general). De cualquier forma.

BD Multidimensionales

Tabulaciones cruzadas

- Los productos OLAP despliegan los resultados de las consultas no como tablas estilo SQL sino como **tabulaciones cruzadas.**

MOLAP

- El MOLAP involucra una **base de datos multidimensional,** que es una base de datos en la cual los datos están almacenados conceptualmente en las celdas de un arreglo multidimensional.
- Como un ejemplo simple, los datos podrían estar representados como un arreglo de tres dimensiones que corresponden a productos, clientes y periodos, respectivamente; cada valor individual de celda podría representar la cantidad total del producto indicado, vendido al cliente indicado en el periodo indicado. Como ya dijimos, las tabulaciones cruzadas también pueden ser consideradas como dichos arreglos.

MINERÍA DE DATOS

- La **minería de datos** puede describirse como "análisis de datos exploratorio". El propósito es buscar patrones interesantes en los datos, patrones que pueden usarse para especificar la estrategia del negocio o para identificar comportamientos fuera de lo común.

- Las herramientas de minería de datos aplican técnicas estadísticas a una gran cantidad de datos almacenados para buscar tales patrones.

- Las bases de datos para minería de datos frecuentemente son *extremadamente* grandes, y es importante que los algoritmos sean escalables.

Gracias