Introducción al almacenamiento de datos

Àngels Rius Gavídia
Montse Serra Vizern
Josep Curto Díaz

PID_00203540

UOC
Universitat Oberta de Catalunya

Índice

Introducción

Generalmente, el estudio de las bases de datos se inicia considerando las bases de datos relacionales, que son las que, de manera mayoritaria, están implantadas desde hace unos años en la industria. Este tipo de bases de datos ofrece apoyo al negocio de la organización y permite almacenar los datos y el procesamiento de la información generada día a día. Todo esto implica que están diseñadas para hacer operaciones de consulta y actualización de manera eficiente por parte de distintos usuarios. Algunos ejemplos de operaciones con estas bases de datos pueden ser la introducción de datos para hacer una factura, llenar un historial médico, gestionar un seguro de vida, etc.

Esta asignatura presenta un nuevo tipo de bases de datos que están orientadas a ofrecer apoyo a la toma de decisiones en la organización. Se trata de los denominados almacenes de datos.

Almacenes de datos

Almacén de datos es la traducción de *data warehouse*.

El objetivo principal del almacén de datos es extraer rendimiento de la información almacenada, y esto quiere decir extraer los datos para un análisis posterior que ayude a tomar decisiones. Esto significa que este tipo de bases de datos tiene un enfoque diferente respecto a las bases de datos convencionales.

A lo largo de este módulo, expondremos en qué se basan los almacenes de datos y lo haremos contraponiéndolos a las bases de datos operacionales para que se vean más claramente las diferencias entre los dos tipos de bases de datos.

Finalmente, hay que comentar que en estos últimos años han surgido con mucha fuerza lo que se denomina bases de datos informativas. Como su nombre indica, se trata de bases de datos orientadas a proporcionar información (especialmente por web). Ahora bien, este tema no lo trataremos en esta asignatura, puesto que posiblemente en un futuro podría convertirse en una nueva asignatura.

Objetivos

Los materiales didácticos incluidos en este módulo se orientan a conseguir que el estudiante alcance los objetivos siguientes:

1. Conocer la orientación y los fundamentos del almacén de datos.

2. Conocer cuál ha sido la evolución de los almacenes de datos y por qué han aparecido.

3. Comprender la importancia de los procesos de toma de decisiones en el mundo de los sistemas de información y qué papel tiene en este el almacén de datos.

4. Saber distinguir las características y los objetivos principales de los almacenes de datos y saberlos diferenciar de las bases de datos operacionales.

1. ¿Qué es un almacén de datos?

Los sistemas tradicionales siempre han tenido dificultades para satisfacer las necesidades informacionales de una organización. A raíz de este hecho, surgen soluciones para dotar a las empresas o grupos de empresas de herramientas capaces de solucionar sus necesidades informacionales globales. Una de las soluciones principales es el almacén de datos.

La ventaja principal de los almacenes de datos consiste en la capacidad de almacenar la información de manera homogénea y fiable en una estructura de datos jerárquica pensada para facilitar las consultas estratégicas de la dirección de la organización.

El término *almacén de datos* ha sido concebido por Bill Inmon y R. D. Hackathorn. La definición proporcionada por Bill Inmon es la siguiente:

Lectura recomendada

W. H. Inmon; R. D. Hackathorn (1994). *Using the Data Warehouse*. Nueva York: Wiley.

> El almacén de datos es una colección de datos orientados al tema, integrados, no volátiles e historiados, organizados para ofrecer apoyo a procesos de ayuda a la decisión.

De esta definición, se desprende el hecho de que estamos ante un nuevo tipo de bases de datos, cuya importancia reside en el apoyo que puede ofrecer a las organizaciones desde el punto de vista estratégico y que a primera vista parece que no es muy difícil de construir. La dificultad principal en el momento de llevar a cabo la creación de un almacén de datos está en el hecho de saber, *a priori*, qué datos se necesitan y de qué manera se tienen que organizar.

¿Cuántas empresas que quieren llevar a cabo proyectos de este tipo saben exactamente los datos que necesitan en el almacén de datos? La experiencia nos indica que hay muy pocas empresas que lo tengan realmente claro. Algunas de estas no saben que no tienen datos suficientemente precisos para introducir en el almacén de los que luego sea posible extraer resultados que sirvan para ofrecer apoyo a la toma de decisiones.

2. Evolución histórica

Desde hace pocos años (mediados de los años noventa), asistimos a la explosión del fenómeno de los almacenes de datos. El motivo principal que ha impulsado su aparición es la enorme cantidad de datos mal explotados.

¿Antes no había datos mal gestionados? ¿Esta investigación acerca de la rentabilidad del negocio no se llevaba a cabo antes? ¿Cómo se tomaban las decisiones estratégicas?

El primer intento importante de desarrollar un sistema de información diferente de los sistemas operacionales lo llevó a cabo IBM a mediados de década de los setenta (Art Benjamin, Canadá). El objetivo que se marcó fue la posibilidad de generar informes sin la necesidad de que los programadores de los sistemas operacionales tuvieran que llevar a cabo un nuevo desarrollo cada vez. Por este motivo, se utilizaron unas herramientas flexibles de integración y generación de informes *ad hoc* a partir de la información contenida en los ficheros de los sistemas operacionales. De este modo, se conseguían una serie de beneficios: usuarios mejor servidos y ahorro de costes.

Este primer **centro de información** conseguía desplazar el esfuerzo de mantenimiento de los sistemas de un 77% a un 33%, debido al hecho de que, al menos en aquel momento, la mayoría de los mantenimientos se centraban en corregir, mejorar o crear nuevos informes.

El 1980, Steven Alter diferenció formalmente dos tipos de sistemas de información: los EDP y los DSS. Los EDP están destinados al procesamiento de los datos operativos de la empresa y los DSS están orientados a la toma de decisiones. Estos dos tipos de sistemas tienen una clara intersección: los MIS, que permiten elaborar informes estándar para los directivos.

EDP, DSS y MIS

EDP: procesamiento electrónico de datos (*electronic data processing*).

DSS: sistemas de apoyo a la decisión (*decision support systems*).

MIS: sistemas de información para la gestión (*management information systems*).

Otro hecho significativo en la evolución de los sistemas de información es el nacimiento, a principios de los años ochenta, de la informática de usuario final, a partir de la introducción del ordenador personal (PC) y de la primera revolución ofimática: hojas de cálculo, tratamiento de textos, bases de ficheros personales, etc.

Esta revolución permitió aumentar la productividad individual frente a los desarrollos tradicionales, puesto que a partir de estas herramientas informáticas, los mismos usuarios, o personas muy cercanas a ellos, eran capaces, por medios poco "consistentes" muchas veces, de elaborar informes con más calidad y de manera más rápida que las áreas de desarrollo.

Empezaron a coexistir dos tipos de informáticos: los especialistas de la programación que se dedican al desarrollo y los usuarios finales de determinadas áreas de negocio con conocimientos mínimos de ofimática. Como consecuencia de estos dos perfiles informáticos y de la diferencia de criterios entre estos, se desencadenó la necesidad de crear los centros de información.

Un centro de información está formado por una serie de ficheros y tablas que residen, normalmente, en una partición del mismo ordenador central en el que se ejecutan los sistemas operacionales. Estas tablas suelen ser una copia de ficheros operacionales que se consideran importantes y que se han volcado en los centros de información según se ha necesitado.

El acceso a los datos de los centros de información se hace por medio de herramientas flexibles que utilizan a "seudoprogramadores", que son los responsables de elaborar y repartir los informes, así como de extraer ficheros para los usuarios finales. Estos ficheros se manipulan en los ordenadores personales de manera individualizada, por medio de herramientas informáticas, últimamente en bases de datos personales, para elaborar gráficos, simulaciones, estadísticas, previsiones, etc.

Tras la aparición de las redes locales, la manipulación de los datos que provienen de los centros de información se ha podido hacer en algunos casos en el ámbito departamental. La creciente potencia y el abaratamiento de los PC ha permitido escalar estas "manipulaciones" a auténticos sistemas de medida y potencia elevadas.

Durante los años ochenta, se imponen las bases de datos relacionales y en los noventa su despliegue llega a su máximo alcance. Como hemos comentado anteriormente, estas bases de datos presentan carencias si se quieren usar para tomar decisiones y, por este motivo, surge otro tipo de sistemas de información: los almacenes de datos.

A continuación podemos ver, desde dos puntos de vista diferentes, qué es lo que favorece que aparezca este nuevo tipo de sistema de información.

Punto de vista socioeconómico

Debido al nuevo marco mundial en las relaciones comerciales establecido por la creación de la Organización Mundial del Libre Comercio en 1995, los requerimientos de las organizaciones han cambiado.

La globalización

El hecho de la globalización no solo ha tenido incidencia en la informática, sino también en cualquier entorno de la sociedad.

A raíz del nuevo marco mundial, las relaciones comerciales tienen las características siguientes:

- Un mercado global (globalización) que se caracteriza por la supresión de barreras proteccionistas arancelarias.
- Una mayor competencia entre los diferentes sectores.
- Una disminución de los márgenes de explotación, lo cual implica menos ganancias para cada operación y, en algunos casos, las consiguientes reducciones de plantilla.
- El aumento del número de operaciones para obtener ganancias similares a las anteriores.
- La fidelización necesaria de los clientes con la consiguiente mejora del servicio posventa.

Todas estas características obligan a los sistemas de información a readaptarse:

- La necesidad no solo de conocer el día a día del comportamiento de la organización, sino también de anticiparse a los cambios que la nueva dinámica comercial genera implica la necesidad de almacenar información nueva.
- En las organizaciones se crean y destruyen departamentos y secciones de manera tan rápida que se producen importantes cambios en los modelos organizativos. Estos cambios no están previstos en los sistemas de información tradicionales.
- Se requiere obtener información más detallada, dinámica y por periodos de tiempo.

Punto de vista informático

Las herramientas informáticas, a partir de 1994, han progresado de tal manera, tanto desde el punto de vista cuantitativo como desde el punto de vista cualitativo, que son posibles tantas operaciones analíticas como consultas de usuario final sobre volúmenes de datos enormes.

La llegada de las arquitecturas cliente/servidor ha sido lo que ha convertido los almacenes de datos en una auténtica disciplina para la profesión.

3. Características de un almacén de datos

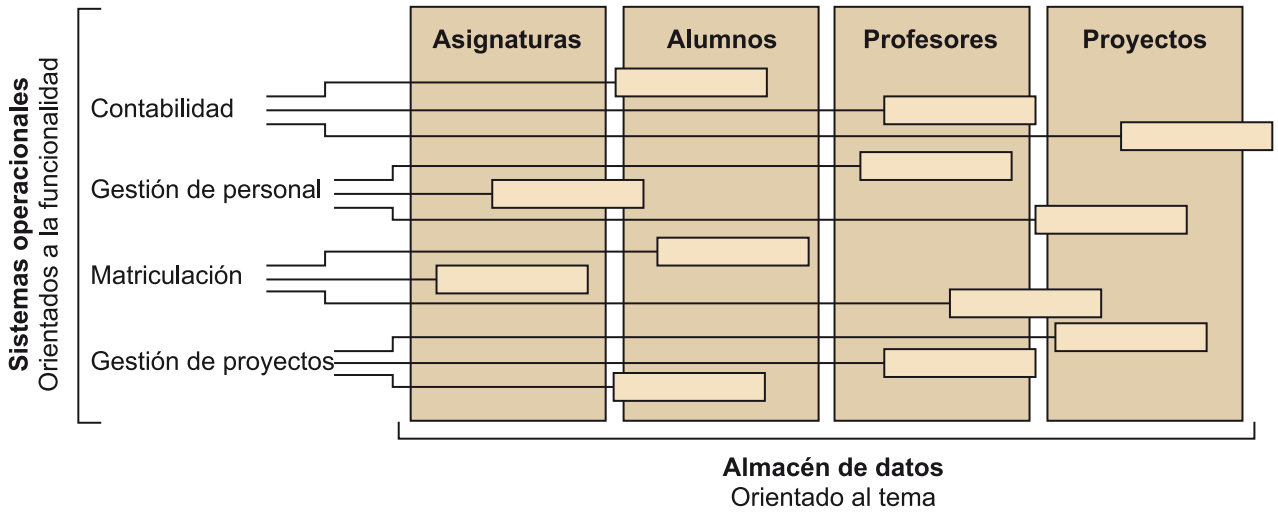
Como se ha visto anteriormente, el almacén de datos representa una novedad en el tratamiento de la información. Para llevar a cabo un tratamiento adecuado de la información, el almacén de datos debe cumplir un conjunto de características: que esté orientado al tema, que los datos estén integrados y que la información sea histórica y no volátil.

3.1. Orientado al tema

Esta primera característica hace referencia a las directrices de los diseñadores de los almacenes de datos. El diseño de los sistemas operacionales viene dado por un conjunto de requerimientos, puesto que se construyen para satisfacer una necesidad concreta y muy conocida. De este modo, hablamos de orientación a la funcionalidad.

Por el contrario, cuando diseñamos un almacén de datos, no sabemos cuáles serán las necesidades de los analistas. No podemos saber cuáles son los requerimientos concretos que tienen, ni el uso que se puede llegar a hacer de los datos que guardamos (esto se decidirá mucho después, cuando aparezca la necesidad de hacer un estudio concreto). Por consiguiente, lo único que el diseñador puede considerar en este caso son las áreas o los posibles temas de análisis.

Dado que no podemos conocer los requerimientos de los usuarios en el momento en que se construye el almacén de datos, la información no se estructura según su funcionalidad (por la cual se utilizarán), sino dividida por temas de interés.



Como se ve en la figura superior, cada sistema operacional (en este caso, de una universidad) accede exactamente a los datos que le hacen falta y de la manera más eficiente posible. Por ejemplo, la aplicación de contabilidad accederá a datos tanto de alumnos, como de profesores o de proyectos con empresas. Sin embargo, probablemente no accederá a todos porque algunos, como por ejemplo las notas de los estudiantes, no le hacen falta. En cambio, un almacén de datos guarda los datos según los posibles temas que se pueden analizar. Tened en cuenta que no sabemos qué utilidad concreta se dará a los datos almacenados. Simplemente se guardan para cuando haya que analizarlos (sin embargo, por ahora no sabemos ni cuándo ni cómo se deberá hacer). Además, no guardaremos todos los datos de los sistemas operacionales, porque algunos no pertenecen a ningún tema de análisis que nos interese (por ejemplo, los números de teléfono de los estudiantes).

3.2. Integración de datos

Sabemos que los sistemas operacionales de las empresas son heterogéneos: funcionan sobre hardware y software diferentes, utilizan modelos de datos distintos (unos orientados al objeto, otros relacionales, etc.) y presentan el negocio desde diferentes puntos de vista (finanzas, ventas, gestión de personal, etc.). Por lo tanto, el primer paso para ofrecer todos los datos a los analistas tiene que ser integrar todos estos sistemas, de modo que los analistas, a pesar de que los datos provengan de fuentes distintas, lo vean como si provinieran de una única fuente. El sistema debe facilitar la resolución de heterogeneidades tanto de semántica como de sistema.

Debemos tener presente que no se trata de informáticos, sino de usuarios no expertos a los que se tiene que facilitar el trabajo. Además, la integración también ayudará a encontrar contradicciones entre las fuentes de datos distintas.

La integración de los datos presenta múltiples problemas, que no siempre son fáciles de resolver. Por mencionar solo algunos de estos, podríamos hablar de unificar los tipos y las estructuras de datos, definir claves primarias comunes, encontrar una convención en la terminología y definiciones o definir un esquema de datos común (capaz de representar la información de todas las fuentes a la vez).

Integrar

Integrar no es simplemente poner los datos en un repositorio común. Estos datos tienen que pasar un proceso de integración y transformación que veremos con detalle más adelante, en el módulo "La factoría de información corporativa".

Además, es necesario mencionar que los almacenes de datos disponen de un componente que ayuda a integrar: los metadatos.

Los metadatos

Estudiaremos los metadatos con detalle en el módulo "La factoría de información corporativa", pero ya podemos avanzar que permiten simplificar y automatizar la obtención de la información desde los sistemas operacionales hasta los sistemas informacionales y, por lo tanto, son básicos para el proceso de integración.

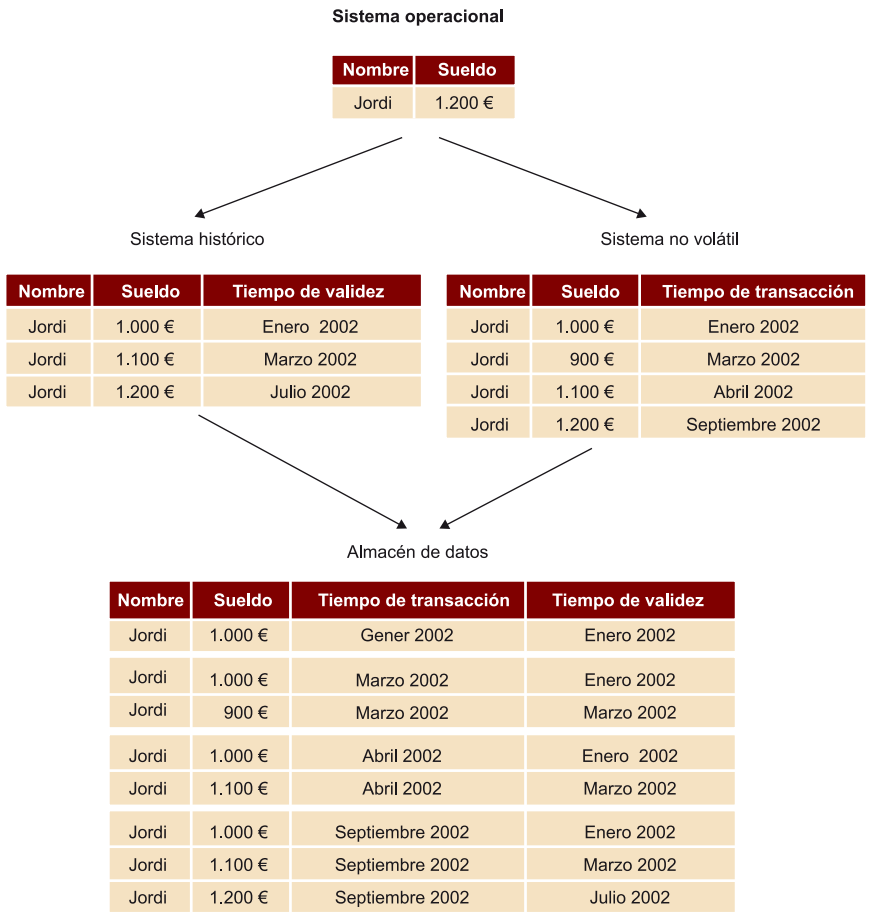
3.3. Información histórica y no volátil

Las dos últimas características de los almacenes de datos hacen referencia al tiempo. Como ya hemos comentado antes, los datos temporales son especialmente importantes en tareas de análisis.

> Hay que distinguir dos tipos de información temporal. El primer tipo nos indica cuándo se produce un cierto acontecimiento en el mundo real (la historicidad). El otro nos indica cuándo tenemos constancia de este hecho en nuestra base de datos (la no volatilidad).

La historicidad es importante para analizar cómo han evolucionado las cosas, para ver una película en lugar de una fotografía. Cualquier dato en el almacén de datos debe ir acompañado de su periodo de validez. En cambio, la no volatilidad nos muestra cuándo nos hemos enterado de los hechos y nos sirve para saber si un cierto informe se hizo teniendo en cuenta unos datos u otros. La no volatilidad implica que no haya las operaciones de modificar y borrar propiamente dichas. Los datos no se borran o modifican, sino que se insertan las correcciones y la fecha en la que se han hecho.

Ejemplo de historicidad y no volatilidad



Nombre	Sueldo	Tiempo de transacción	Tiempo de validez
Jordi	1.000 €	Gener 2002	Enero 2002
Jordi	1.000 €	Marzo 2002	Enero 2002
Jordi	900 €	Marzo 2002	Marzo 2002
Jordi	1.000 €	Abril 2002	Enero 2002
Jordi	1.100 €	Abril 2002	Marzo 2002
Jordi	1.000 €	Septiembre 2002	Enero 2002
Jordi	1.100 €	Septiembre 2002	Marzo 2002
Jordi	1.200 €	Septiembre 2002	Julio 2002

En esta figura, podemos ver la diferencia que hay si guardamos el sueldo de Jorge en un tipo de sistema temporal o en otro. Antes que nada, vemos que si guardamos el sueldo en un sistema relacional operacional (no temporal), solo necesitamos una tupla que contiene el sueldo actual de Jorge. Si disponemos de un sistema histórico, vemos cómo ha evolucionado el sueldo a lo largo del tiempo y sabemos en qué momento es válido cada valor. En cambio, si el sistema es simplemente no volátil, vemos cuáles son nuestros conocimientos/creencias según las transacciones que se hayan ejecutado en cada momento. Por ejemplo, nos permite saber que, en marzo, creíamos (de manera errónea) que Jorge cobraba 900 euros (esto no coincidió con la realidad en ningún momento; por lo tanto, un sistema histórico no lo refleja). Finalmente, vemos que en un almacén de datos tenemos los dos tipos de información temporal. Para ver el interés de esto, fijémonos en la última tupla. Nos indica que en septiembre supimos que Jorge cobraba 1.200 euros desde julio (cosa que no podemos reflejar en ninguno de los sistemas anteriores). Observad la diferencia entre el volumen de datos en un sistema operacional y en un almacén de datos (en este caso, una tupla frente a ocho tuplas).

La historicidad nos servirá para hacer estudios sobre la evolución del negocio, mientras que la no volatilidad garantiza que no perdemos ningún dato (ni siquiera los erróneos).

4. Objetivos de un almacén de datos

En este apartado, enumeraremos distintos objetivos que un almacén de datos debería lograr o cumplir, además de mencionar los objetivos tanto en el ámbito empresarial como en el técnico.

Los objetivos más destacados son los siguientes:

- Ayudar en la toma de decisiones.
- Segmentar los datos de negocio.
- Gestionar el conocimiento de la empresa.
- Depurar los datos.

Ayudar en la toma de decisiones

Este sería el objetivo principal del almacén de datos, puesto que debe ser un instrumento útil para ofrecer apoyo en la toma de decisiones. Este sistema se tiene que construir basándose en los requerimientos de aquellas personas que lo van a utilizar.

> Cuando se construye un almacén de datos, el punto más importante a tener en cuenta es obtener la información necesaria sobre el negocio con el que posteriormente se puedan desarrollar nuevas oportunidades que permitan mejorar los resultados de la empresa.

El hecho de que se segmenten los datos, que estos se encuentren disponibles con más rapidez, que se puedan hacer análisis financieros "fácilmente" sin depender de los informáticos y que estén depurados lleva a la toma de decisiones sea más fácil y que esta se haga con más fundamento.

Segmentar los datos de negocio

Toda organización necesita saber su posicionamiento dentro del mercado, procurar liderar en el sector y proyectarse adecuadamente hacia el futuro.

Segmentar los datos de negocio permite hacer particiones según uno o más criterios.

Ejemplo de segmentación

Un ejemplo clásico consiste en saber cuáles de los productos de la empresa son los que se venden más, en qué intervalos temporales y en qué situación geográfica.

La segmentación también sirve para avisar dónde puede haber futuros problemas que se nos escapan en una primera mirada.

Ejemplo empresa de venta de material eléctrico

Imaginemos que en una empresa de material eléctrico la mayoría de la gente que va a comprar son lampistas. Si segmentamos las ventas por tramos de edad, nos podemos encontrar con el hecho de que la mayoría de los lampistas tienen entre sesenta y sesenta y cinco años. Está claro que de aquí a cinco años el negocio habrá bajado mucho si no hacemos algo.

Por lo tanto, la segmentación nos permite hacer un estudio más profundo de los datos y, por consiguiente, establecer comparativas que nos ayudarán a valorarlos de una manera más correcta y a detectar tendencias.

Gestionar el conocimiento de la empresa

También es posible ver el almacén de datos como contenedor de datos de la empresa, los cuales serán sometidos a un análisis posterior.

Desde este punto de vista, el almacén de datos puede ser útil en los casos siguientes:

- Para gestionar empresas con sistemas distribuidos, puesto que ejerce funciones de integración de la información.
- Para gestionar empresas con sistemas heterogéneos, puesto que ejerce funciones de homogeneización.
- Para gestionar empresas con sistemas centralizados, porque pueden hacer la selección adecuada al tema/usuario.
- Para gestionar procesos de fusión empresarial, si se usa con el fin de consolidar.

Depurar los datos

Una de las carencias que hay cuando diseñamos un almacén de datos es la poca fiabilidad de algunos de los datos que tiene la empresa y su redundancia, con el consiguiente peligro de que se generen problemas de integridad en el almacén de datos.

La excusa que representa la creación de un almacén de datos pasa por mejorar, adecuar y racionalizar los datos que hay en el sistema operacional.

En la práctica, este hecho representa la mejora de algunos procesos operacionales que en condiciones normales no se habría podido hacer, pero con la excusa de la implantación de un almacén de datos es posible llevarla a cabo.

5. Comparativa: almacén de datos y bases de datos operacionales

Una manera de iniciar la comparativa entre los almacenes de datos y las bases de datos operacionales será a partir de los ejemplos siguientes:

Ejemplo 1

Imaginémonos la base de datos que puede utilizar un trabajador de banca de una sucursal cuando trabaja en la atención al público por ventanilla. Es cierto que el volumen de datos global de la base de datos puede ser muy alto, pero los datos que se manipulan en cada una de las transacciones son muy simples: la operación de un ingreso o de un reintegro en la base de datos probablemente solo involucre la inserción en una determinada tabla de una tupla que refleje esto.

Por lo tanto, en cada una de las operaciones (de manera general) se involucran muy pocos datos, pero es cierto que el volumen global resulta enorme y, dado que se acumulan a diario, tiende a crecer muy rápidamente. Además, la disponibilidad de la base de datos tiene que ser total: sería inaceptable que un cliente de esta sucursal se viera obligado a esperar quince minutos a que el sistema gestor hiciera la transacción que refleje un reintegro para disponer de dinero.

Ejemplo 2

Continuamos con la sucursal bancaria. Resulta evidente que, si el director de esta sucursal quiere decidir si potenciar un determinado producto financiero o no y para esto necesita analizar la evolución del índice de morosidad del último año de sus clientes, no debe tener en cuenta si un determinado cliente ha ido por la mañana a hacer movimientos en su cuenta y si este hecho ha variado la morosidad (exceptuando casos significativos). Las necesidades del director son más globales: necesita conocer la evolución ascendente o descendente de este índice sin entrar en detalle.

Como se puede comprobar, la función que lleva a cabo cada una de las bases de datos en los ejemplos anteriores es muy distinta. En el primer caso se trata de una base de datos operacional y en el segundo caso, de un almacén de datos.

Actualmente, las bases de datos relacionales son operativas en un entorno muy concreto que responde a las necesidades para las que se crearon. Estas necesidades suelen involucrar entornos de gestión puros en los que hay simplicidad de las estructuras y de los tipos de datos, utilización de transacciones cortas, etc.

Por otro lado, las necesidades actuales de información de las organizaciones han variado. La disponibilidad de gran cantidad de información es de vital importancia para los negocios, puesto que las decisiones de futuro se suelen tomar a partir de esta información.

Continuemos con los ejemplos

Está claro, por lo tanto, que los hechos que la base de datos operativa tiene no son los que el director necesita. De todas maneras, la globalización de los datos que busca el director se basa claramente en la información reflejada en esta base de datos, pero organizada de otro modo (en este caso, resumida).

Este tipo de necesidades para reflejar tendencias, evoluciones, hechos históricos en el negocio y posibilidades futuras son factores que el alta directiva de las instituciones o empresas tiene que manipular de una manera habitual y que ha causado que hayan aparecido en el mercado herramientas de ayuda en la toma de decisiones.

5.1. Diferencias en el almacenamiento, el diseño y la estructuración de los datos

Temporalidad

Los datos se tienen que guardar el tiempo que sea necesario. En las bases de datos operacionales este tiempo normalmente oscila entre uno y dos años, y en el almacén de datos se amplía de cinco a diez años. Más allá de estos intervalos de tiempo, los datos se dejan de considerar útiles.

Volumen

Evidentemente, la característica de la temporalidad nos condiciona el volumen. No es lo mismo guardar los datos un año que diez. Por lo tanto, en las bases de datos operacionales el volumen será relativamente pequeño y en el almacén de datos, será mucho mayor.

El volumen de los almacenes de datos

En Estados Unidos, cuando se refieren al volumen de los almacenes de datos siempre hablan de terabytes de datos.

Nivel de agregación

El nivel de agregación permite el cúmulo de los datos. En un nivel 0, tendríamos todos los datos de manera detallada. Este nivel de agregación en las bases de datos operacionales suele ser único y bastante bajo. En cambio, en el almacén de datos se suelen dar distintos niveles. Este hecho nos indica que algunas veces tenemos los datos duplicados de manera implícita.

Actualización

La actualización de los datos en una base de datos operacional se hace constantemente; por lo tanto, la información es muy cambiante. Por el contrario, en el almacén de datos se hace de una manera periódica y, dentro de este periodo, una sola vez.

Estructuración

El hecho de que las bases de datos operacionales y los almacenes de datos tengan objetivos distintos implica que necesitarán una estructuración diferente de los datos para lograr los objetivos que tienen asignados.

En el caso de las bases de datos operacionales, tendrán una estructura relacional, en la que se da mucha importancia a la estabilidad. Este hecho representa tener bases de datos estáticas, que no cambian con frecuencia su estructura.

En cambio, en los almacenes de datos habrá una visión multidimensional y a la vez serán muy dinámicos: estos se tienen que adaptar rápidamente a las necesidades del negocio para ser útiles en los procesos de toma de decisiones.

En el diseño del almacén de datos, hay que tener presente que estará el componente tiempo, mientras que en las bases de datos operacionales no es necesario.

En el diseño de las bases de datos operacionales, tiene que ser más importante que el acceso sea inmediato a un dato en concreto, mientras que en los almacenes de datos suelen predominar las consultas masivas de datos.

Otra diferencia importante es el hecho de que el diseño de las bases de datos convencionales tiene que ser normalizado, mientras que en los almacenes de datos es mejor la desnormalización.

5.2. Diferencias en el tratamiento de la información

Explotación de la información

En el entorno de las bases de datos operacionales, con frecuencia los usuarios finales acceden a los datos mediante aplicaciones predefinidas.

En los almacenes de datos, las consultas suelen ser imprevistas. Puede haberlas predefinidas, pero la variedad de posibilidades que encontramos hace imposible prever cuáles serán las necesidades de los usuarios finales. Además, estas consultas están orientadas a áreas de interés del negocio que con frecuencia son cambiantes.

Tiempo de respuesta

El tiempo de respuesta de las operaciones ha de ser instantáneo cuando hablamos de bases de datos operacionales, debido a la frecuencia con la que se actualizan los datos. Por el contrario, en el caso de los almacenes de datos, este tiempo debe ser rápido, pero no instantáneo, puesto que el tiempo no es crítico.

Tiempo de respuesta

No es nada fácil que las respuestas a las peticiones que se hacen a los almacenes de datos sean rápidas cuando hablamos de terabytes de datos.

5.3. Diferencias de funcionalidades

Actividades

La actividad de las bases de datos operacionales es del día a día; en definitiva, se trata de la operatividad para el funcionamiento de la empresa. Por lo tanto, serán aplicaciones fáciles de manejar –no se tendrá que pensar mucho en las opciones que hay– y rápidas.

Al contrario, la actividad de los almacenes de datos es de análisis y decisión estratégica. Las aplicaciones tendrán unas funcionalidades diferentes que en el entorno operacional, las cuales se complementarán con múltiples opciones y permitirán muchas opciones de libre aplicación.

Importancia de los datos

Como ya hemos dicho anteriormente, el dato es muy importante en los dos entornos. En el caso de la base de datos operacional, lo importante es el dato actual, mientras que en el caso del almacén de datos la importancia está en los datos históricos.

Usuarios

En las bases de datos operacionales, los usuarios suelen ser muchos. Este hecho se complementa con el nivel de usuario, puesto que no todo el mundo puede hacer de todo. Los usuarios suelen ser de la estructura media-baja de la empresa.

En el entorno del almacén de datos, los usuarios son muy pocos, suelen tener acceso a determinados datos agrupados y/o acumulados y acostumbran a estar en lo alto de la empresa: dirección, marketing, planificación estratégica, control de gestión, etc.

5.4. Tendencias actuales

Desde la concepción del almacén de datos, las tecnologías y técnicas de implementación han evolucionado para adaptarse a las necesidades de las organizaciones. En la actualidad, varios factores condicionan la evolución de los almacenes de datos:

a) Crecimiento exponencial del universo digital. Los usuarios y las redes de sensores duplican anualmente los datos de las organizaciones, y con frecuencia estos no están estructurados. Este crecimiento no solo plantea un reto en cuanto al almacenamiento, sino también a la gestión y manipulación de los datos. Nos referimos, pues, a un problema que tiene tres dimensiones: 1) velocidad de generación de los datos, 2) volumetría de los datos y 3) variabilidad de los datos.

b) Nuevas técnicas de modelización. Daniel Linstedt publicó en el 2000 una nueva técnica denominada *data vault*. El objetivo de esta tendencia era la creación de almacenes de datos flexibles y auditables en tiempo real.

c) Madurez de tecnologías de manipulación de datos. Las organizaciones actuales necesitan apoyo en la toma de decisiones y esta se fundamenta en datos de negocio que a menudo requieren tiempo. Este hecho ha motivado la aparición de tecnologías de complemento del almacén de datos tradicional. A continuación, se mencionan las siguientes.

- Análisis continuo de datos[1]: mediante flujos continuos de datos, permite analizar datos en tiempo real de manera continua. Un posible caso de uso podría contextualizarse en la monitorización del tráfico de una ciudad. Supongamos que hay que identificar los puntos donde se producen incidencias, habilitar en tiempo real una alerta basada en patrones y a continuación, automatizar algunas acciones que hay que tomar para reducir el número de incidencias. Estas acciones podrían consistir en avisar al personal de mantenimiento o cambiar el comportamiento de los elementos de la red.

(1)En inglés, *data streaming*.

- Procesamiento de acontecimientos complejos[2]: permite identificar patrones dentro de los procesos de negocio y automatizar algunas acciones que se repiten. Por ejemplo, si se identifican clientes que cumplen ciertas características, se pueden automatizar ofertas dirigidas a clientes que siguen un mismo patrón.

(2)En inglés, *complex event processing*.

- Bases de datos en memoria[3]: mediante la memoria de un servidor que utiliza técnicas OLAP, estas bases de datos permiten analizar datos de gran volumetría en tiempo real. Con frecuencia, esta tecnología da apoyo a las tecnologías anteriores.

(3)En inglés, *in-memory*.

- Hadoop, MapReduce y otras tecnologías equivalentes: empresas como Google, Amazon o Facebook gestionan a diario gran cantidad de datos que tienen que ser introducidos en el sistema y consultados en tiempo real. Con esta finalidad, con frecuencia se trabaja con redes de servidores que se consultan en paralelo y con bases de datos en columnas u otros SGBD no relacionales. Este enfoque se conoce como NOSQL (puesto que no solo utiliza el lenguaje SQL).

d) Analítica de negocio. Utiliza técnicas estadísticas y de minería de datos en procesos operativos de negocio. El objetivo es facilitar las decisiones relativas a la operativa y proponer tácticas de negocio basadas en predicciones. Algunos fabricantes especializados en almacenes de datos incluyen algoritmos para facilitar la creación de este tipo de ventajas competitivas.

Resumen

En este primer módulo hemos hecho una introducción al concepto de almacén de datos para disponer de los fundamentos suficientes para el resto de la asignatura.

Primero, hemos explicado qué es un almacén de datos. Como hemos visto, no es un concepto nuevo, puesto que de manera implícita se estaba utilizando aunque con otras herramientas: los centros de información son los precursores del almacén de datos.

Posteriormente, hemos definido el almacén de datos según Inmon y hemos repasado sus características principales: orientación al tema, integración, no volatilidad y datos históricos.

Hemos visto que los almacenes de datos no son otro tipo de organización de bases de datos, sino que otorgan un valor añadido muy importante a la organización por el hecho de aportar más conocimiento a la empresa y ayudarla en la toma de decisiones.

Finalmente, y para remarcar la idea anterior, se han comparado las bases de datos operacionales con los almacenes de datos y hemos visto que las diferencias son realmente muy importantes.

Actividades

1. Proponed en el foro qué proyecto de almacén de datos querríais desarrollar que corresponda, si es posible, con vuestra área de actividad profesional.

a) Explicad cuáles serían los objetivos de este proyecto.

b) ¿Qué datos creéis que serían relevantes para conseguirlo?

c) ¿Qué diferencia veis con el proyecto de base de datos operacional en el caso de que haya uno?

2. Buscad por la Red los cinco proyectos de almacén de datos que están desarrollados y que creáis que son más interesantes.

a) ¿Cuáles son los objetivos que tiene cada proyecto?

b) ¿Os sorprende alguno de estos? ¿Por qué?

c) Compartid estas experiencias en el foro.

Ejercicios de autoevaluación

1. ¿En qué características se basan los almacenes de datos?

2. Tenemos una base de datos operacional que está perfectamente normalizada y los procesos que trabajan sobre esta son muy rápidos.

a) ¿Nos serviría esta estructura para hacer procesos para tomar decisiones?

b) Si es que no, ¿qué diferencias habría que implementar para construir un almacén de datos?

Solucionario

Ejercicios de autoevaluación

1. Las características principales de un almacén de datos son la orientación a temas, la integración, la no volatilidad y los datos históricos. Estas características se basan en la filosofía que Inmon describió.

2.a) No sirve la misma estructura.

2.b) Desde el punto de vista de diseño, hay diferencias en la temporalización, el volumen de datos, el nivel de agregación, la actualización y la estructuración. Desde el punto de vista del tratamiento de la información, las diferencias son de explotación de la información y de tiempo de respuesta. Para acabar, desde el punto de vista de funcionalidades, hay diferencias en las actividades, en la importancia de los datos y en los usuarios finales.

Glosario

almacén de datos *m* Bases de datos orientadas a áreas de interés de la empresa que integran datos de distintas fuentes con información histórica y no volátil y que tienen como objetivo principal hacer de apoyo en la toma de decisiones.
en data warehouse

base de datos operacional *f* Base de datos destinada a gestionar el día a día de una organización, es decir, almacena la información en lo referente a la operativa diaria de una institución.

centro de información *m* Conjunto de ficheros y bases de datos precursor de los almacenes de datos que se basaba en datos del entorno operacional para extraer información y la almacenaba para usuarios y procesos. La información no estaba compartida.

cliente/servidor *m* Entorno mediante el cual se establecen relaciones entre agentes por medio de una red de transmisión de datos, de modo que los agentes clientes reclaman servicios ofrecidos por agentes servidores.

data warehouse Véase **almacén de datos**.

globalización *f* Proceso que incluye aspectos generales de la economía y que afecta mucho al desarrollo de los almacenes de datos.

transacción *f* Conjunto de operaciones de lectura y/o actualización de la base de datos que acaba confirmando o cancelando los cambios que se han llevado a cabo.

Bibliografía

Davenport, T.; Harris, J. (2008). *Competing on Analytics*. EE. UU.: Harvard Business School Press.

Franco, J. M.; EDS-Institut Prométhéus (1997). *El Data Warehouse - El Data Mining*. Barcelona: Gestión 2000.

Gill, H. S.; Rao, P. C. (1996). *Data Warehousing. La integración para la mejor toma de decisiones*. México: Prentice Hall.

Inmon, W. H.; Hackathorn, R. D. (1994). *Using the data warehouse*. Nueva York: Wiley.