Introducción al *Data Warehouse*

Alberto Abelló Gamazo
Josep Curto Díaz
Àngels Rius Gavídia
Montse Serra Vizern
José Samos Jiménez
Juan Vidal Gil

PID_00236069

UOC
Universitat Oberta de Catalunya

Índice

Introducción

Generalmente, el estudio de las bases de datos se inicia con las bases de datos relacionales, que son las que, de manera mayoritaria, están implantadas desde hace unas décadas en la industria. Este tipo de bases de datos permite almacenar los datos y procesar la información generada con la operativa diaria de la organización. Por ello se dice que las bases de datos ofrecen apoyo a la actividad de negocio dentro de las organizaciones. Así pues, están diseñadas para realizar 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 emitir una factura, llenar un historial médico, gestionar un seguro de vida, etc.

Esta asignatura presenta otro tipo de bases de datos distinto a los tradicionales, los que están orientadas a ofrecer apoyo a la toma de decisiones en la organización. Se trata de los denominados almacenes de datos, conocidos también como *Data Warehouse* (en este módulo y en esta asignatura hablaremos de ambos indistintamente).

El objetivo principal del almacén de datos o *Data Warehouse* es extraer rendimiento de la información almacenada, y esto quiere decir extraer los datos para un análisis posterior que ayude a tomar decisiones. Por tanto, vemos 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 ha surgido con mucha fuerza lo que se denomina bases de datos NoSQL (Not Only SQL): es una amplia clase de sistemas de gestión de bases de datos que difiere del modelo clásico del sistema de gestión de bases de datos relacionales en aspectos importantes, el más destacado es que no usan SQL como el principal lenguaje de consultas. En este tipo de bases de datos hay una variada tipología, como las bases de datos del tipo clave-valor, orientadas a columnas, orientadas a documentos, grafos, etc., ahora bien, este tema no lo trataremos en esta asignatura.

Objetivos

Los contenidos 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 sus características.

3. Saber distinguir entre bases de datos operacionales y almacenes de datos en diferentes niveles.

4. Comprender la importancia de los datos y los procesos en la toma de decisiones, así como el papel que desempeña el almacén de datos en la toma de decisiones de una organización dentro de un contexto más amplio y como parte del sistema de información de la misma.

5. Conocer los elementos principales que integran el contexto del almacén de datos y su finalidad.

6. Conocer las actuales tendencias acerca del almacén de datos.

1. Qué es un *Data Warehouse*. Características

El término *Data Warehouse* o almacén de datos ha sido concebido por Bill Inmon y R. D. Hackathorn.

La definición proporcionada por Bill Inmon es la siguiente: 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 se trata de un tipo de bases de datos, cuya importancia reside en el apoyo que puede ofrecer a las organizaciones desde un punto de vista estratégico y que, a primera vista, no parece muy difícil de construir. Sin embargo, la dificultad principal en el momento de crear un almacén de datos está en saber, *a priori*, qué datos se necesitan y de qué manera deben organizarse.

¿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 puede haber desconocimiento en lo relativo a los datos empresariales realmente necesarios. Algunas de estas empresas no saben que no tienen datos suficientemente precisos para introducir en el almacén de manera que luego sea posible extraer resultados que sirvan para la toma de decisiones.

1.1. Evolución histórica

Antes de tratar las características de un almacén de datos, puede ser interesante ver cómo han evolucionado los sistemas de información en lo que respecta al almacenamiento y explotación de información para su análisis.

Brevemente, podemos resumir la evolución de la siguiente manera:

- **Década de 1960**. *Reporting manual*: la información era difícil de encontrar y analizar. Por otro lado, los informes generados no presentaban ninguna flexibilidad al usuario, ante cada nuevo requerimiento era necesario reprogramar los informes.

- **Década de 1970**. Aparición de los sistemas de soporte a la decisión y los sistemas de información ejecutiva. Se trataba de información muy orientada a la dirección que trataba de ofrecer apoyo a la toma de decisiones.

La información aunque se consolida para los informes se encontraba muy dispersa y cada nuevo requerimiento implicaba reprogramación.

- **Década de 1980**. Herramientas *desktop* de análisis de datos. Se trata de aplicaciones de escritorio que utilizaban una interfaz de usuario más amigable que las anteriores, pero que debido al crecimiento de los sistemas operacionales tenían información de origen muy disperso, difícil de encontrar, creando silos de información.

- **Década de 1990**. Creación de los primeros *Data Warehouse* para centralizar la información proveniente de los sistemas operacionales y para facilitar los procesos de análisis. Inmon publicó el libro *Building the Data Warehouse* en 1992, año en el que se empezó a hacer extensivo el uso del término.

- **Año 2000**. Emergencia y desarrollo de las plataformas de Inteligencia de Negocio entorno al *Data Warehouse*. Ampliación de los sistemas de planificación empresarial (ERP) con un módulo de inteligencia de negocio.

- **Años 2003-2007**. Emergencia y desarrollo de las plataformas conocidas como *Data Warehouse Appliances* (DWA) que optimizan hardware y software para el trabajo analítico. Los fabricantes líderes lanzan al mercado diferentes tipos de DWA basados en arquitecturas de computación paralela de tipo MPP (*massively parallel processing*).

- **Años 2008-2015**. Emergencia y desarrollo de tecnologías *Big Data*. Aparición y desarrollo *framework Hadoop*. Convivencia entre estas tecnologías y los almacenes de datos tradicionales.

En la actualidad la inteligencia de negocio y la gestión de la información son actividades prioritarias en los departamentos de Tecnologías de la Información (TI) de las compañías.

1.2. Características de un *Data Warehouse*

Como se ha visto anteriormente, el almacén de datos representa un cambio 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.

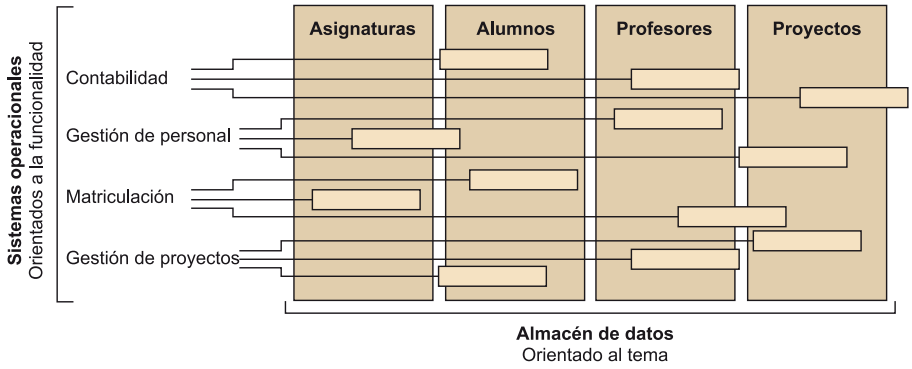
1.2.1. Orientado al tema

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 almacenados (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 (el uso que se le vaya a dar), sino dividida por temas de interés.

Figura 1. Orientación al tema de un almacén de datos



En la figura 1 se considera el caso de una universidad, en la cual cada sistema operacional accede exactamente a los datos que necesita y se supone que 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 de ellos no los requiere, como por ejemplo las notas de los estudiantes. En cambio, un almacén de datos guarda los datos según los posibles temas que se pueden analizar. Debemos tener en cuenta que *a priori* no sabemos qué utilidad concreta se dará a los datos almacenados. Simplemente se guardan para cuando sea necesario analizarlos. Además, tampoco se guardarán todos los datos de los sistemas operacionales debido a que algunos no pertenecen a ningún tema de análisis que resulte de interés, como por ejemplo podrían ser los números de teléfono de los estudiantes

1.2.2. 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 debe ser la integración de 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 usuarios 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, unificar niveles de granularidad, 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), garantizar la calidad de los datos integrados y realizar una gestión ágil de fuentes con altos volúmenes de datos.

Además, es necesario mencionar que los almacenes de datos disponen de un componente que ayuda a integrar: los metadatos, de los cuales hablaremos más adelante. Aunque hay que tener presente que los metadatos 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

1.2.3. 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 acontecimiento en el mundo real (la historicidad). El segundo cuándo tenemos constancia del 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 informe se hizo teniendo en cuenta unos datos u otros. La no volatilidad implica que no existan las operaciones de modificar y borrar propiamente dichas. Los datos no se borran o modifican, sino que se insertan correcciones y la fecha en la que se han registrado.

Ejemplo de historificación

Un caso de historificación puede ser la entidad que guarda la información relativa a las tarifas móvil de un operador de telecomunicaciones. Las características de estas tarifas varían en el tiempo. Por ejemplo, los minutos de llamadas o los megabytes de navegación de una tarifa pueden sufrir variaciones de acuerdo con la estrategia comercial de la compañía. En una base de datos operacional nos interesa tener la última foto de cada tarifa, mientras que en un almacén de datos guardaremos un histórico de cambios que nos permitirán realizar estudios a lo largo del tiempo.

Figura 2. Ejemplo de historicidad de almacén de datos

tarifa	minutos	navegación
100	200	1000

Base de datos operacional

id	tarifa	minutos	navegación	fecha inicio	fecha fin
10001	100	150	500	01/07/2014	31/01/2016
10002	100	200	1000	31/01/2016	–

Almacén de datos

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).

2. Objetivos de un *Data Warehouse*

En este apartado enumeraremos los principales objetivos que un almacén de datos debería lograr o cumplir, tanto en el ámbito empresarial como en el técnico: ser un repositorio central e integrado de información empresarial y ser un repositorio base para procesos de análisis y *reporting*.

2.1. Repositorio central e integrado de información empresarial

Como se ha comentado en apartados anteriores los sistemas operacionales contienen información valiosa para el negocio, pero dispersa en distintos sistemas y bases de datos. El *Data Warehouse* tiene como uno de sus principales objetivos el ser un repositorio central de información corporativa que puede provenir de diversos sistemas. Este repositorio tiene diversas funciones:

- Integrar información proveniente de los distintos sistemas de la compañía.
- Consolidar y homogeneizar esta información.
- Ser el punto central de información. Versión más fiel de la información, evitando tener diferentes versiones según la fuente que se consulte.
- Depurar y limpiar los datos, garantizando su calidad.
- Facilitar procesos de fusión empresarial, si se usa con este fin.

2.2. Repositorio base para procesos de análisis y *reporting*

Los procesos de *reporting* y análisis de la compañía necesitan nutrirse de una información de base. Esta información de negocio se recoge de las bases de datos operacionales, pero acceder a ellas para realizar un proceso de análisis o *reporting* que precisa de información de negocio diversa y recogida en diferentes sistemas, puede ser un proceso costoso y complejo debido a las diferentes ubicaciones de los datos y a su heterogeneidad. Por ello resulta más productivo acceder a un repositorio centralizado como es el *Data Warehouse*.

Partiendo de la información integrada, consolidada y depurada del Data Warehouse podemos realizar procesos de análisis y *reporting* de diferente naturaleza:

- Procesos *reporting* periódico recurrente.

- Cuadros de mando.

- Procesos de *reporting ad hoc* para necesidades de información concretas.

- Procesos de analítica avanzada (predicción de eventos de negocio, *forecasting* de evolución temporal).

Estos procesos nos informarán de la situación y evolución de la compañía desde múltiples perspectivas ayudándonos a comprender qué está pasando y por qué. Así mismo, nos ayudan a intentar predecir la evolución en un futuro. Estos análisis que apoyan la toma de decisiones serían menos ágiles sin la existencia de un *Data Warehouse*.

Ejemplo de proceso de análisis partiendo del *Data Warehouse*

Si tenemos que realizar un análisis de nuestros clientes en el que estimemos la probabilidad de abandono de la compañía en base a las distintas variables que caracterizan el cliente, será más sencillo y óptimo lanzar estos procesos de propensión de abandono partiendo de la base de datos de clientes del *Data Warehouse*, donde tenemos toda la información de nuestros clientes depurada e integrada, que tener que acceder a cada uno de los sistemas que recogen información de nuestros clientes: CRM, ERP, facturador, ventas, etc.

3. Comparativa entre *Data Warehouse* 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

Imaginemos 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, 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 la alta dirección de las instituciones o empresas tiene que manipular de una manera habitual y que ha propiciado la aparición en el mercado de herramientas de ayuda en la toma de decisiones.

3.1. Diferencias en el almacenamiento, el diseño y la estructura de los datos

1) 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.

2) 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.

3) 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.

4) 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 en intervalos de tiempo definidos. En la base de datos operacional las actualizaciones suelen ser atómicas (registro a registro) y en el almacén de datos por lotes (conjuntos de registros).

5) Estructura

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 deben de adaptarse 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 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, ya que favorece la rapidez de las consultas.

En cuanto a la integridad de la información vemos que las bases de datos operacionales usualmente garantizan la integridad definiendo restricciones en base de datos (claves primarias y foráneas), mientras que en los almacenes de datos, nos encontramos diseños en los que la integridad se garantiza en el proceso de carga (actualizaciones masivas) y no se definen restricciones en la base de datos de destino para mejorar el rendimiento de la actualización.

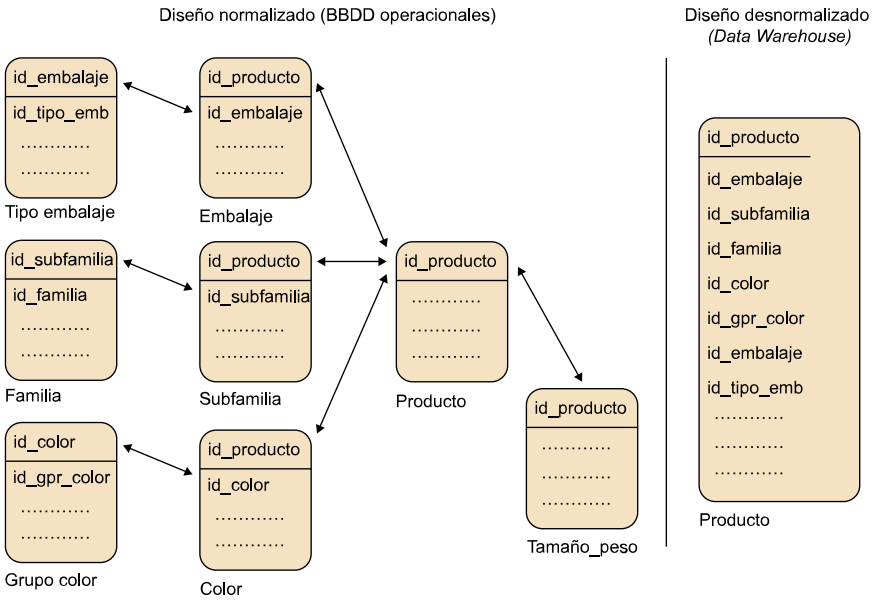
Ejemplo de diferencia de estructuración

Una empresa que comercializa un determinado grupo de productos tendrá una base de datos operacional con la información de sus productos. Entre la información que guardamos de los productos podemos tener características como el tamaño, el peso, la familia, la subfamilia, el embalaje, el tipo de embalaje, el color y el grupo de color. Son características de naturaleza diferente y existen agrupaciones entre las mismas (familia-subfamilia, embalaje-tipo embalaje, color-grupo color). Cabe la posibilidad de crear un diseño normalizado con una entidad de datos producto que tenga un identificador único que permita relacionarlo con otras entidades como tamaño-peso, subfamilia, color y embalaje. A su vez existirá una relación entre color y grupo de color, subfamilia y familia y embalaje y grupo de embalaje tal y como muestra la figura 3. Este diseño responde a las necesidades de un sistema operacional orientado a transacciones y pretende evitar redundancias de espacio.

En cambio, si trabajamos en un diseño de base de datos para un *Data Warehouse* nuestro diseño estará orientado a las consultas, en muchos casos masivas. Por ejemplo, si en una consulta analizamos un indicador según los productos y sus atributos, es posible que necesitemos realizar muchas combinaciones de tablas en la consulta si se ha diseñado la base de datos de acuerdo al modelo de datos operacional. Esto puede complicar y ralentizar la consulta. Por tanto, si realizamos un diseño orientado a consultas puede interesarnos crear una única entidad producto con todos los atributos que lo describen, ya que eso puede simplificar las consultas a una única entidad y mejorar los tiempos de respuesta de esta. Hay que considerar que este segundo diseño es más apropiado para

las consultas, pero tenemos redundancia en los datos, especialmente en los atributos de agrupación (familia, tipo de embalaje y grupo de color).

Figura 3. Diseño normalizado vs diseño desnormalizado



3.2. Diferencias en el tratamiento de la información

1) 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. Dentro de esta variedad de posibilidades sí es posible identificar entidades, agregaciones o cruces de uso frecuente de acuerdo con los cuales podemos definir vistas o tablas de bases de datos que contengan preagregados, índices en las tablas u otro tipo de estrategias de optimización.

2) Tiempo de respuesta

El tiempo de respuesta de las operaciones debe 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 necesariamente instantáneo. Las operaciones en los almacenes de datos suelen ser consultas masivas que es imposible obtener de forma instantánea, pero sí deben estar en un tiempo razonable acorde con el trabajo del analista. Hay informes que realizan un conjunto de consultas masivas y que pueden ser planificados en diferido para que se ejecuten

en *background* y puedan ser consultados posteriormente. El concepto de ejecución de consulta en diferido es mucho menos común en las bases de datos operacionales.

3.3. Diferencias en las funcionalidades

1) Actividades

La actividad de las bases de datos operacionales se produce día a día con las actividades del negocio, ya que se utilizan para la operativa o funcionamiento de la empresa. Por lo tanto, serán aplicaciones fáciles de manejar, en las que no se tendrá que pensar mucho en las opciones que ofrece y serán rápidas.

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

2) 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.

3) 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 hay menos usuarios y suelen definirse diferentes perfiles según la información que se va a consultar. Tradicionalmente ha existido un usuario de perfil directivo (dirección, marketing, planificación estratégica, control de gestión, etc.) que accede a datos agrupados y/o acumulados; también existen perfiles tácticos que acceden a datos agregados, pero con una visión más centrada en su departamento o línea de negocio y, por último, un perfil operativo que accede a información más relacionada con la operativa diaria.

4. La factoría de información corporativa

William Inmon presentó en 1998 lo que se denomina factoría de información corporativa. Se trata de un concepto para hacer referencia a un conjunto de componentes que interactúan para ayudar a gestionar todos los flujos de datos desde los sistemas operacionales de la empresa hasta los analistas. Su objetivo es transformar los datos de los sistemas operacionales (materias primas) en información útil para los analistas (producto elaborado) con el fin de utilizarla en los procesos de toma de decisiones en la organización. En este módulo veremos los diferentes componentes de esta factoría y cómo interactúan entre sí.

Estos componentes son los siguientes:

- Almacenes de datos departamental, corporativo y operacional.
- Componente de transformación e integración.
- Gestión de datos maestros.
- Metadatos.
- Estructuras multidimensionales.

4.1. Almacén de datos departamental

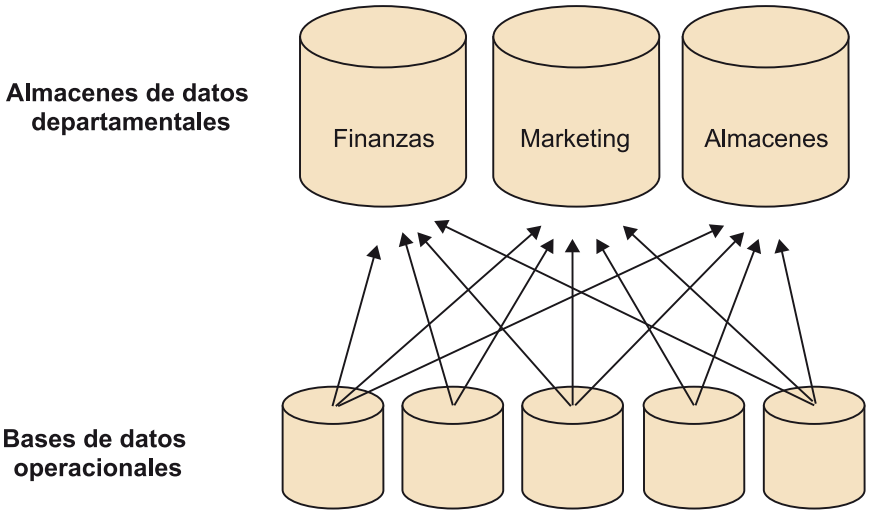
Construir un almacén de datos es muy costoso, además de tener unos requerimientos de rendimiento difíciles de conseguir. La solución para obtener un tiempo de respuesta bajo es disponer de diferentes almacenes solo con información parcial del negocio (únicamente la parte que interese a un departamento o conjunto de personas).

Estos almacenes de datos departamentales, también conocidos por su denominación en inglés *Data Mart*, normalmente estarán diseñados según el modelo de datos multidimensional, lo que facilita la mejora en el rendimiento mediante técnicas específicas de almacenamiento de los datos. Además, para no sobrecargar los sistemas con datos innecesarios, estos solo contienen datos históricos dentro del periodo de tiempo que sea estrictamente necesario.

Ejemplo de técnica de almacenamiento para mejorar el tiempo de respuesta

Una técnica para mejorar el tiempo de respuesta es la preagregación. Esta técnica consiste en guardar los resultados de las funciones de agregación (suma, media, mínimo, etc.) ya calculados para cuando el usuario los pida. Esto quiere decir que tenemos que conocer (o imaginar) qué consultas querrán hacerse para calcular previamente los resultados, de modo que el cálculo no se tenga que hacer en el momento concreto en que se solicita.

Figura 4. Almacenes de datos departamentales



Tal como puede verse en la figura 4 para cada departamento o grupo de usuarios se construye un almacén. Este solo contiene los datos necesarios para satisfacer las necesidades concretas del departamento o grupo de usuarios y los integra con independencia de la fuente de datos de procedencia. Estos datos se modelan siguiendo la visión de la realidad que tenga el departamento correspondiente y no hace falta que se consensúe con toda la empresa.

Un aspecto fundamental en el diseño de los almacenes de datos departamentales es la gestión de las entidades comunes entre almacenes. Por ejemplo, los departamentos de Marketing y Finanzas utilizarán entidades con datos de clientes o productos. Es importante que se utilice la misma entidad de cara a la integridad de datos entre almacenes de diferentes departamentos. Estas entidades comunes se denominan dimensiones conformadas en los modelos dimensionales y son entidades del tipo clientes, productos, proveedores, cuentas que por ser críticas para el negocio son utilizadas por muchos departamentos.

Otra ventaja de los almacenes de datos departamentales es que no necesitan tener los datos con el máximo nivel de detalle. Por ejemplo, si los analistas solo quieren ver los datos mensuales, no es necesario almacenar los datos diarios. De este modo, no habría que almacenar las ventas diarias de la empresa, sino solo el total que se ha vendido durante un mes, lo que representa un ahorro de espacio claro.

Tener muchos almacenes de datos pequeños permite abaratar costes, puesto que son más económicos que uno grande que satisfaga las necesidades de todo el mundo a la vez. Además, haciéndolo así, facilitamos la configurabilidad. Finalmente, también es más fácil controlar tanto los costes (que se imputarán al departamento correspondiente) como los accesos, procesos y configuración del sistema (que corresponderán a un conjunto de usuarios muy restringido).

Por otro lado, un almacén de datos departamental desde un punto de gestión de proyectos tiene un alcance más limitado y definido que un almacén de datos corporativo y el nivel de riesgo es menor. Facilita un planteamiento de proyecto por fases. Desde un punto de vista de negocio hay que señalar que

no todos los departamentos evolucionan al mismo ritmo en lo referente a necesidades analíticas, hay departamentos muy demandantes de analítica y datos como pueden ser Marketing o Finanzas y otros cuya demanda puede ser menor como Legal o RRHH.

Los almacenes de datos departamentales guardan una historia parcial de los datos que interesan a un departamento. Están diseñados para obtener un buen tiempo de respuesta ante las consultas de un conjunto de analistas.

4.2. Almacén de datos corporativo

Tener múltiples almacenes de datos departamentales independientes genera problemas a largo plazo, a pesar de que son más económicos y fáciles de construir a corto plazo. El primer problema es que, como podéis ver en la figura 4, tenemos procesos independientes de integración y transformación para cada almacén de datos departamental. Además, ¿dónde guardamos la información que actualmente no interesa a ningún departamento? No tenemos ningún lugar donde la podamos guardar y no la podemos despreciar. Hay que tener un almacén de datos corporativo que guarde toda la historia de todos los datos y siempre con el máximo nivel de detalle posible. Sin embargo, los almacenes de datos departamentales todavía son necesarios.

Conviene conocer la terminología anglosajona generalmente utilizada para referirse a cada uno de los almacenes.

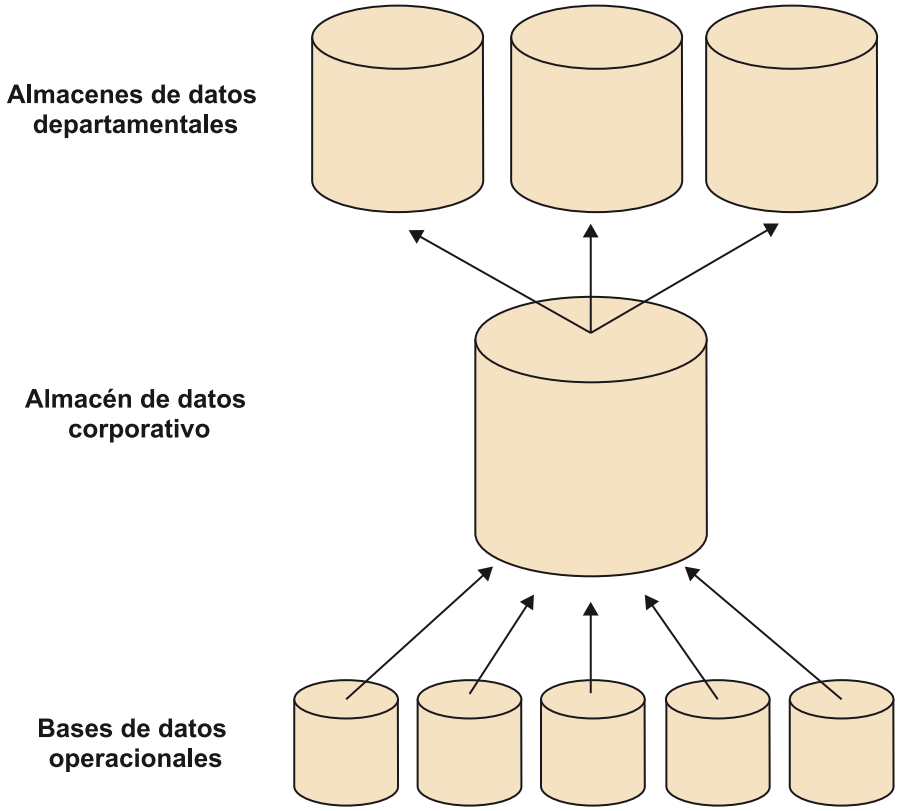
- Almacén de datos departamental: *Data Mart*.
- Almacén de datos corporativo: *Enterprise Data Warehouse*.

Generalmente cuando se habla de almacén de datos en general, sin especificar tipo, suele usarse el término anglosajón. Usualmente para referirnos a un almacén de datos corporativo utilizamos también *Data Warehouse*, aunque en ocasiones y para remarcar su carácter corporativo en el término empleamos *Enterprise Data Warehouse*. En ocasiones oímos hablar de *Data Warehouse* de marketing para referirse a un almacén de datos del departamento de Marketing, sería más apropiado denominarlo *Data Mart* de Marketing.

El almacén de datos corporativo no es apropiado para los usuarios finales, porque está diseñado para gestionar e integrar grandes cantidades de datos que, junto con el exceso de usuarios, degradan el tiempo de respuesta. No se puede diseñar para favorecer a un grupo de usuarios concreto, sino que tiene que servir a todos a la vez de la mejor manera posible.

De este modo, como se puede ver en la figura 5, el almacén de datos corporativo es el resultado de un proceso de integración y transformación de todas las fuentes de datos único y complejo, que estudiaremos en detalle en el apartado correspondiente de este módulo. Los almacenes de datos departamentales ahora se obtienen simplemente como resultado de un proceso de transformación a partir del almacén corporativo.

Figura 5. Almacén de datos corporativo



El almacén de datos corporativo guarda toda la historia de todos los datos de la empresa integrados. Está diseñado para almacenarlos de manera eficiente.

La tabla 1 resume las diferentes características de los dos tipos de almacén de datos que hemos visto hasta ahora:

Tabla 1

Característica	**Almacén de datos**	
	Departamental	Corporativo
Temática	Específica	Genérica
Fuentes de datos	Pocas	Muchas
Tamaño	Gigabytes	Terabytes
Tiempo de desarrollo	Meses	Años
Modelo de datos	Multidimensional	Relacional

En primer lugar, el almacén de datos corporativo tiene que ser genérico y debe guardar datos de toda la empresa siguiendo una visión consensuada del negocio. Por el contrario, los almacenes de datos departamentales son absolutamente específicos. Solo contienen los datos que pide un conjunto de usuarios, los guardan según la concepción que estos tienen del negocio y están optimizados para obtener un buen rendimiento ante las tareas de análisis que se desean realizar.

4.3. Almacén de datos operacional

Desgraciadamente, es posible que con los almacenes de datos departamentales y el corporativo todavía no tengamos cubiertas todas las necesidades de información de la empresa. Debido a su volumen de datos y a las técnicas de implementación que se utilizan, el almacén de datos corporativo (y, por lo tanto, los departamentales que se actualizan a partir de este) no se puede tener constantemente actualizado (solo se suele actualizar durante las noches o los fines de semana). Por otro lado, sus usuarios tampoco lo requieren, puesto que están más interesados en los datos históricos que en los actuales. Sin embargo, puede haber otros usuarios que también pidan datos integrados y que los quieran completamente actualizados. Aún necesitamos otro tipo de repositorio de información.

La aparición de este repositorio, también conocido por ODS (*Operational Data Store*), viene dada por la típica ponderación entre volumen de datos y velocidad del sistema. Hasta ahora, en los otros almacenes, lo que queríamos era tener absolutamente cualquier dato que pudiéramos llegar a necesitar para tomar una decisión. Como consecuencia de este requerimiento, el tiempo de respuesta puede llegar a degradarse y, en cualquier caso, nos vemos obligados a renunciar a tener los datos constantemente actualizados. En este caso, valoramos más el hecho de que los datos siempre estén actualizados, que no que los tengamos todos. Por lo tanto, renunciamos a tener datos históricos y disponemos de un repositorio volátil.

Este es el precio que se tiene que pagar para reducir el volumen de datos y poderlo mantener constantemente actualizado. De este modo, el almacén de datos operacional y el corporativo se complementan: el corporativo guarda todos los datos históricos, pero no está actualizado siempre, y el operacional siempre está actualizado, pero no contiene datos históricos.

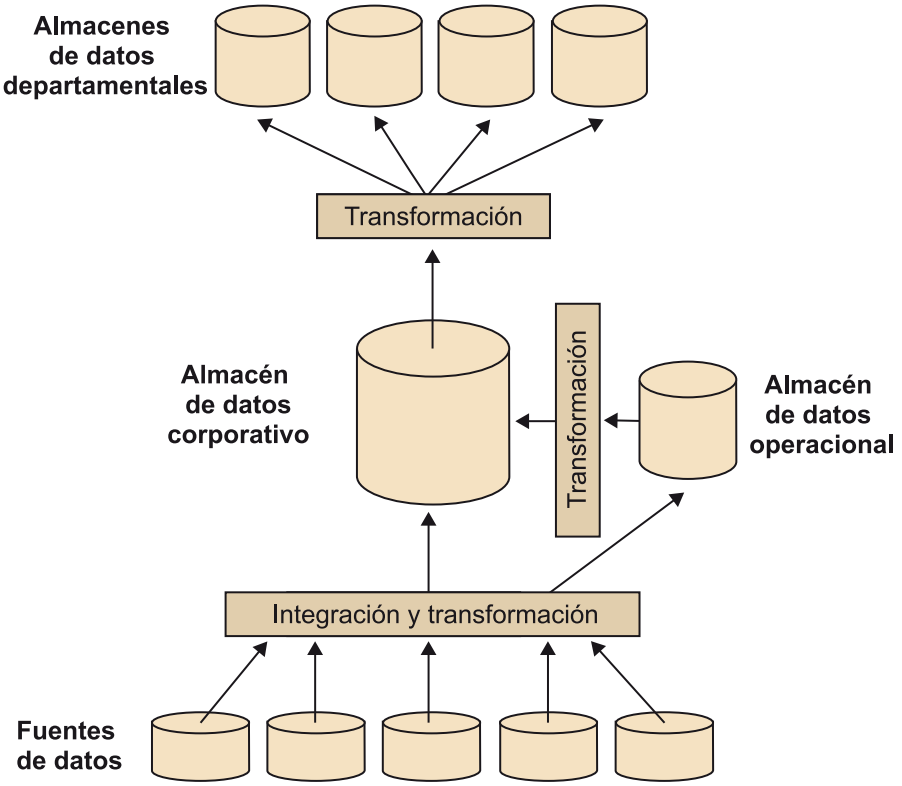
El almacén de datos operacional es una estructura a caballo entre el mundo operacional y el de la toma de decisiones. Está orientado al tema e integrado como cualquier almacén de datos, pero en este caso no contiene ningún tipo de información temporal.

4.4. El componente de integración y transformación

Como hemos visto en el apartado 3.2 «Diferencias en el tratamiento de la información», los sistemas operacionales de los que disponen las organizaciones generalmente no cumplen los requerimientos de los analistas. Como solución, se ha definido el concepto de almacén de datos, tanto en el ámbito departamental como en el corporativo, según las características de sus datos, que lo diferencian de los sistemas operacionales.

Aun así, los datos de los almacenes de datos se obtienen a partir de los sistemas operacionales de la empresa, así como de fuentes externas. Por sus características distintas en cuanto a estructura y organización, los datos obtenidos de las fuentes no se pueden utilizar directamente en el almacén de datos, sino que se tienen que adaptar a sus requerimientos en estos aspectos.

Figura 6. Componente de integración y transformación



La misión del componente de integración y transformación consiste en obtener los datos para los diferentes almacenes de datos de la organización. Este componente también se conoce por ETL, por sus siglas en inglés: *Extract, Transform and Load*.

Originalmente, los datos se obtienen a partir de los sistemas operacionales y otras fuentes de datos, y se deben transformar, depurar e integrar y, según la estructura de los esquemas de los almacenes de datos, también se deben transportar y cargar para que se puedan utilizar en los diferentes almacenes de datos de la organización.

A diferencia de los almacenes de datos, cuyo elemento principal es la base de datos, el elemento principal del componente de integración y transformación es el software encargado de llevar a cabo la misión descrita.

Tanto las fuentes de datos como los diferentes almacenes de datos se pueden encontrar en plataformas distintas, y, por lo tanto, el componente de integración y transformación tendrá elementos en las diferentes plataformas en las que esté el resto de los componentes de la FIC.

El componente de integración y transformación está formado por software que se ejecuta en las distintas plataformas en las que funciona el resto de los componentes de la FIC.

4.5. Gestión de datos maestros

Generalmente todos los almacenes de datos tienen una serie de entidades críticas en cuanto a la información que contienen como pueden ser: clientes, productos, proveedores o cuentas. Estas entidades intervienen en muchas consultas y su correcta actualización es fundamental para realizar un análisis preciso. Es común que dichas entidades sean compartidas por varios almacenes de datos departamentales y en ocasiones por sistemas no informacionales que acceden a estas entidades al ser el almacén de datos corporativo la imagen más fiel de las mismas. Así mismo estas entidades del almacén de datos terminan siendo entidades maestras que requieren una gestión especial de cara a la realización de actividades como pueden ser: consolidar toda la información relevante de la entidad que puede proceder de diferentes sistemas, asegurar la calidad de esta información, el refresco de la misma y la sincronización con otros sistemas, entre otras actividades.

Estas actividades se suelen denominar gestión de datos maestros o *Master Data Management* (MDM) y se encuadran dentro de las actividades de gobernanza de datos o *Data Governance* (DG).

La gestión de datos maestros tiene como principales objetivos:

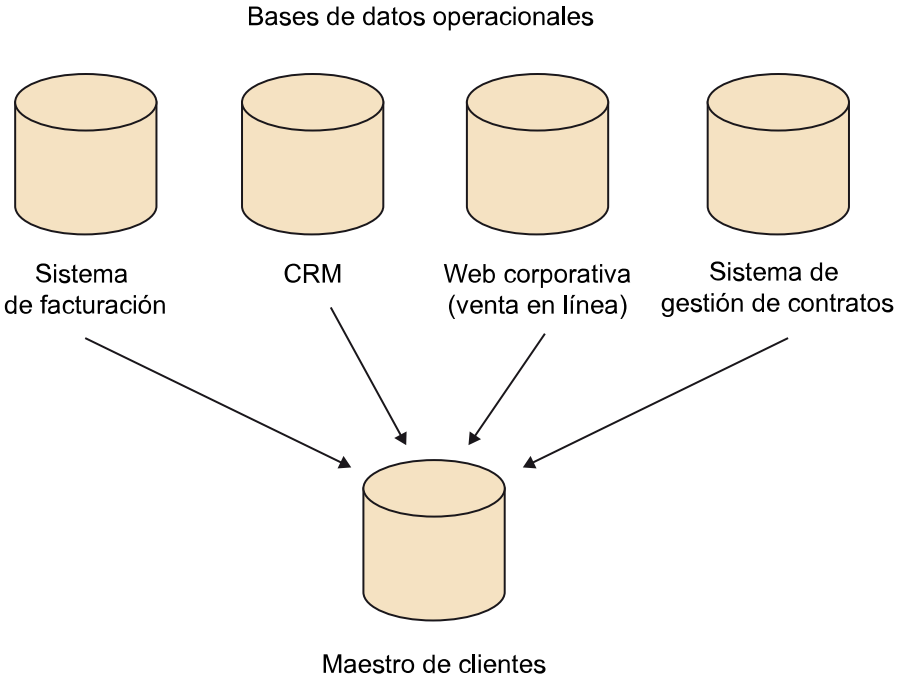
- Identificar las fuentes de origen de los datos maestros
- Identificar a los productores y consumidores de datos maestros.
- Recopilar y analizar metadatos sobre los datos maestros recopilados.

- Actualizar y mantener centralizadamente datos maestros. Procesos de consolidación y enriquecimiento. Generación de registros maestros.

- Determinar a los responsables (administradores) de los datos maestros.

- Asegurar la calidad del dato de estos maestros.

Figura 7. Maestro de clientes generado desde BBDD operacionales



Estas entidades maestras se integran dentro de los almacenes de datos.

Directamente relacionadas con las actividades de MDM están los procesos de seguimiento de la calidad del dato que se implementarán sobre los datos maestros y que permitirán monitorizar su calidad, revisando aspectos tales como la exactitud, integridad, consistencia y completitud.

4.6. Los metadatos

Los metadatos no son un elemento específico de la FIC: aparecen en muchos contextos del mundo del software. La definición más frecuente que hay del concepto de metadato está basada en su etimología: «Los metadatos son datos sobre datos». Los datos generalmente representan características de las entidades que modelan; en el caso de los metadatos, representan características de otros datos que facilitan su administración y uso. Es decir, lo que diferencia a un dato de un metadato, más que su estructura o contenido, es su propósito y uso.

Los metadatos describen sus características (por ejemplo, formato, origen, uso, etc.) sobre un conjunto de datos. Estos metadatos son datos y a su vez podemos tener otros metadatos que describan sus características (metadatos sobre metadatos), y así de manera sucesiva.

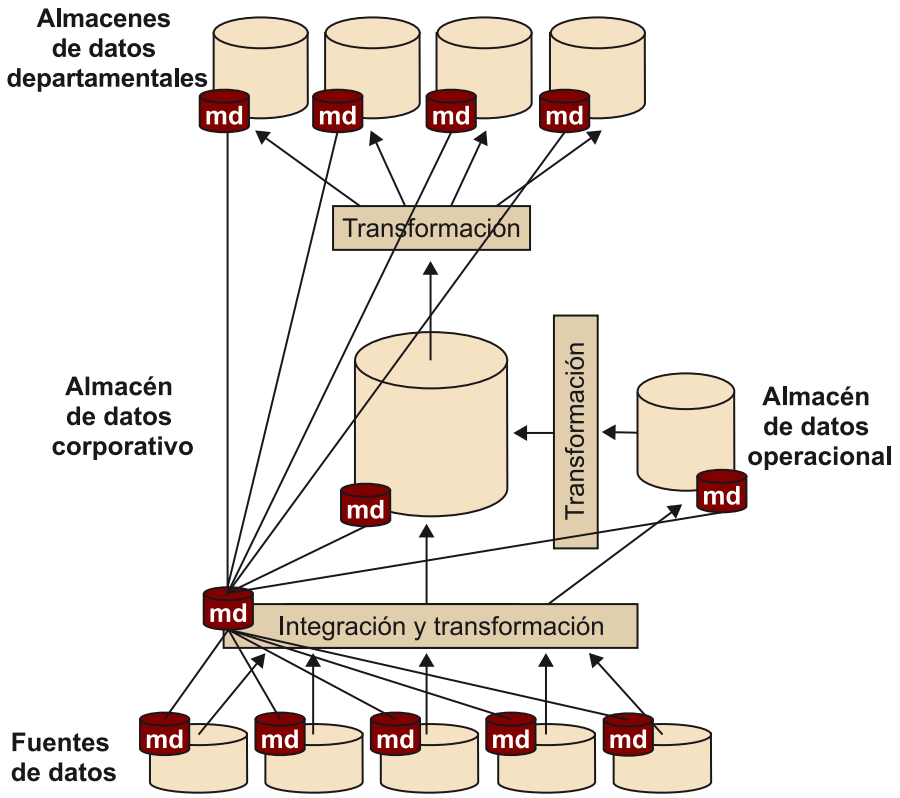
En este apartado, empezamos revisando el uso de los metadatos en la FIC, a continuación se presentan diferentes tipos de metadatos según el uso de los mismos. También se analiza la manera como se crean los metadatos, así como los estándares definidos para permitir compartirlos entre distintos componentes. Finalmente, se comenta la necesidad de utilizar diferentes versiones de metadatos en la FIC.

4.6.1. Metadatos y componentes de la FIC

En la FIC, se produce un flujo de datos desde las fuentes de estos hasta los analistas. Este flujo está compuesto por los datos propiamente dichos, que representan características de entidades del mundo real, y por los metadatos, datos que ofrecen información sobre los otros datos transferidos o almacenados.

Los metadatos están asociados a todos los componentes de la FIC (podéis ver la figura 8), pero son un componente por sí mismos. Inmon los define dentro de la FIC como el «pegamento» que mantiene unido el resto de los componentes, y por este motivo los considera como el componente más importante de la FIC.

Figura 8. Metadatos en la FIC



1) Metadatos de las fuentes de datos

Las bases de datos de los sistemas operacionales o las fuentes de datos en general, desde el punto de vista de la FIC, tienen como componente fundamental los datos. Sin embargo, además de estos, hay metadatos que están generados por herramientas CASE (*Computer Aided Software Engineering*) si estas se han utilizado en su construcción; en caso de estar construidos sobre un SGBD, tendremos aquellos que definen las bases de datos que intervienen y las relaciones entre sus elementos.

Generalmente, en las fuentes de datos los metadatos describirán, entre otras características, las estructuras según las que se almacenan los datos, la cantidad de registros almacenados, su forma de almacenamiento y las condiciones bajo las que se producen los datos.

2) Metadatos de los almacenes de datos

En los almacenes de datos, tendremos los metadatos asociados a los SGBD sobre los que están construidos, y encontramos algunos similares a los descritos para las fuentes de datos. Además, es posible encontrar información sobre el uso de los datos por parte de los usuarios: estadísticas de uso, información sobre seguridad (quién está autorizado a hacer qué operaciones), etc.

3) Metadatos en el componente de integración y transformación

El componente de integración y transformación utiliza los metadatos del resto de los componentes pero, además, puede definir como metadatos el origen de los datos, su destino, las transformaciones que se hacen en los datos de las fuentes para obtener los de los almacenes y la frecuencia o el resultado de estas transformaciones.

Una vez definidos todos los metadatos, a partir de estos se puede generar de manera automática el software que haga la función de este componente. Es más fácil y rápido mantener los metadatos que mantener un software desarrollado manualmente.

Los metadatos son el componente más importante de la FIC, puesto que cohesionan el resto de los componentes de los que también forman parte.

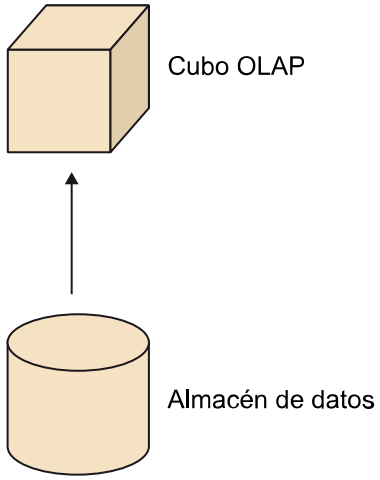
4.7. Estructuras multidimensionales

La información incorporada a los almacenes de datos de la FIC es explotada y visualizada desde la capa de presentación. En ocasiones, de cara a mejorar los tiempos de acceso a los almacenes se crea una estructural adicional con infor-

mación agregada. Teniendo en cuenta el alto número de métricas y dimensiones de análisis que podemos tener en un almacén, las agregaciones posibles son muy numerosas. Existen estructuras de almacenamiento que contemplan todas estas posibilidades de agregación como pueden ser los cubos OLAP. Se entiende por OLAP, o proceso analítico en línea, al método para organizar y consultar datos sobre una estructura multidimensional. A diferencia de las bases de datos relacionales, todas las potenciales consultas están calculadas de antemano, lo que proporciona una mayor agilidad y flexibilidad al usuario de negocio.

Un cubo OLAP es un conjunto de celdas de datos organizadas según diferentes dimensiones. Se trata de una forma de representación de una base de datos multidimensional, en la cual el almacenamiento físico de los datos se realiza mediante una estructura multidimensional. Los cubos se pueden considerar como una ampliación de las dos dimensiones de una tabla convencional. Esta disposición de datos permite un análisis rápido, al encontrarse gran cantidad de la información precalculada.

Figura 9. Cubo OLAP

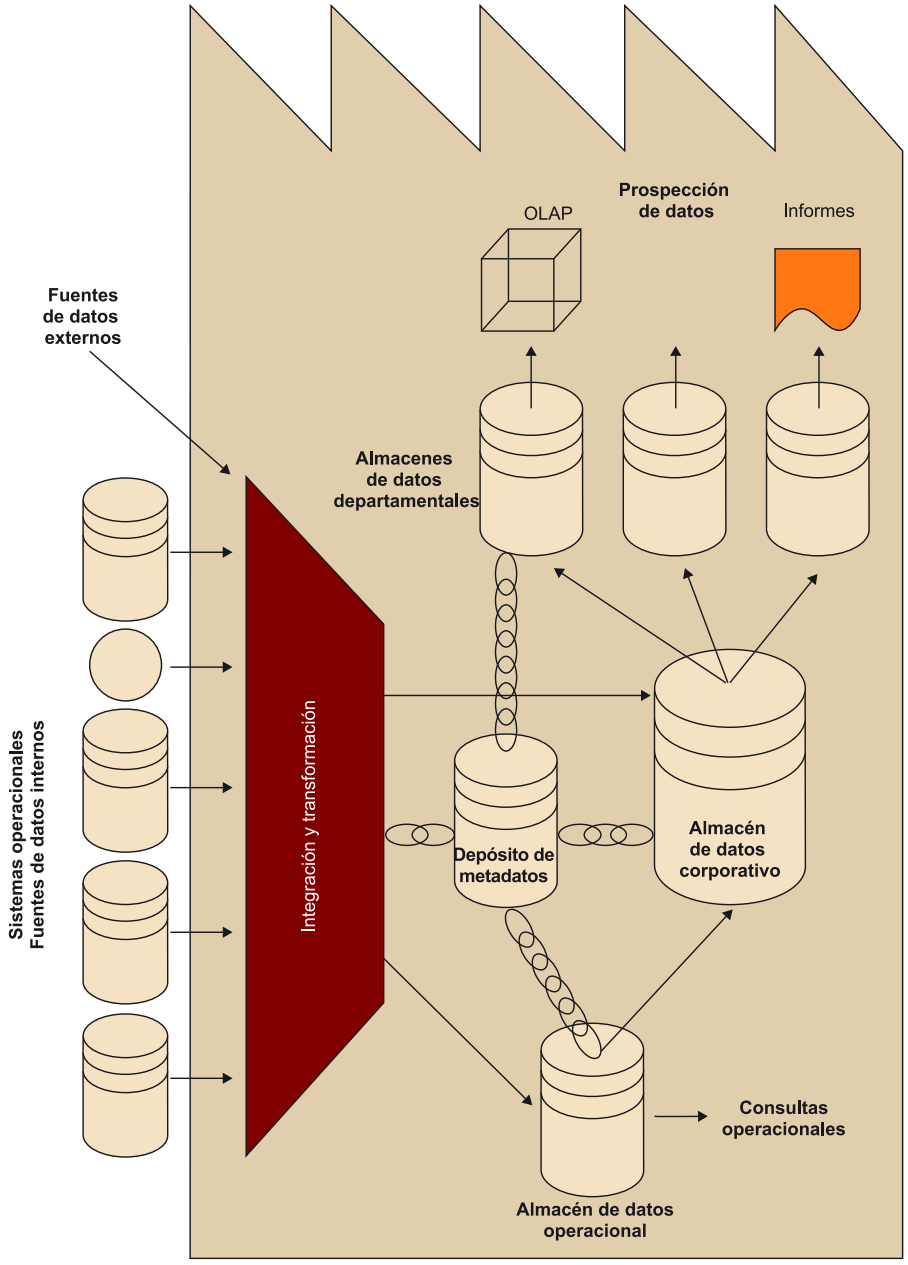


4.8. Integración componentes de la FIC

Llegados a este punto, y conociendo de manera global los componentes que forman la factoría de información corporativa, en este apartado veremos cómo todo converge en un solo bloque.

La figura 10 esquematiza todos los componentes de la factoría de información.

Figura 10. FIC completa



Los datos entran, provenientes de los sistemas operacionales de la misma empresa u otras fuentes de datos externos, directamente al componente de integración y transformación. Este componente de software los prepara para guardarlos en el almacén de datos operacional o directamente en el almacén de datos corporativo. También es este componente de transformación el que genera una parte de los metadatos que utilizarán el resto de los componentes en su funcionamiento. Los datos del almacén de datos operacional servirán tanto para ser consultados, como para alimentar el almacén de datos corporativo. Finalmente, según la utilidad que se dará a los datos, estos se depositan en pequeños almacenes de datos departamentales que están a punto para ser consultados o tratados.

Un error de terminología también bastante común es denominar el todo (la factoría de información corporativa) como si fuera solo una parte (el almacén de datos). Se habla de un componente en lugar de hablar del proceso que utiliza este componente. Stephen R. Gardner define el almacenamiento de datos como un proceso, no un producto, para reunir y gobernar datos de distintas procedencias con el fin de obtener una visión única y detallada, total o parcial, de un negocio. Esta idea no parece tan diferente de la factoría de información presentada por William Inmon. Más bien solo es otro punto de vista, que en cierto modo incluye el primero. El hecho de hablar de un proceso implica que haya elementos que lo hagan posible o, como mínimo, que ayuden a hacerlo posible.

Podemos considerar la factoría de información como el conjunto de elementos que hacen posible el proceso de almacenamiento de información. El almacén de datos simplemente sería un componente más, como también lo son el repositorio de metadatos, el componente de integración y transformación, etc.

En este punto, todavía nos podríamos plantear la necesidad de esta factoría de información. ¿Por qué hay que añadir toda esta complejidad a los sistemas de información de la empresa? Si ya tenemos los datos en los sistemas operacionales, ¿por qué los replicamos en la factoría de información? ¿Por qué los analistas no consultan los datos directamente en los sistemas operacionales? ¿No estamos derrochando recursos? Podéis encontrar las respuestas a estas preguntas más o menos implícitas en los apartados anteriores de este mismo módulo, pero ahora desmentiremos de manera explícita esta supuesta duplicidad de datos:

a) Los sistemas operacionales contienen los datos que la empresa utiliza en su día a día en la ejecución del negocio. En cambio, la factoría de información contiene datos de análisis, generalmente extraídos de estos sistemas operacionales, pero no necesariamente coincidentes. Puede haber datos operacionales (por ejemplo, el número de teléfono de los clientes) que no interesen para tomar decisiones y datos muy importantes para tomar decisiones (como el beneficio) que no se utilicen en el funcionamiento diario de la empresa.

b) Generalmente, los sistemas operacionales no contienen datos históricos para no retrasar de manera innecesaria su funcionamiento. En cambio, estos datos históricos son imprescindibles a la hora de tomar decisiones.

c) Los sistemas operacionales siempre guardan los datos detallados (por ejemplo, los artículos vendidos a cada cliente). En los sistemas decisionales, a veces, no interesa entrar en tanto detalle. Lo que solo se desea es el importe total de la venta, el gasto mensual del cliente o, simplemente, el total vendido durante el mes a todos los clientes.

d) Finalmente, otra diferencia entre las bases de datos de los sistemas operacionales y las de la factoría de información es que estas últimas contienen datos limpios. Durante la fase de entrada de datos a la factoría de información, estos se limpian, se sustituyen o se eliminan los valores nulos, se detectan inconsistencias, posibles contradicciones entre diferentes fuentes de datos, etc. En los sistemas operacionales, con una entrada continua de datos, no se puede garantizar esta pulcritud.

La factoría de información no contiene los mismos datos que los sistemas operacionales, a pesar de que la intersección no es vacía.

5. El almacén de datos dentro de un sistema de *Data Warehouse*

El almacén de datos que creamos gracias a los procesos y elementos de la factoría de información corporativa es un activo crítico para la toma de decisiones de la compañía y quedará integrado en los sistemas de tipo informacional o de inteligencia de negocio. Sistemas destinados a consultar y analizar información, que facilitan la toma de decisiones en la compañía dotando de capacidades de inteligencia a las actividades de gestión y dirección de la compañía.

Un sistema de inteligencia de negocio tiene diferentes subsistemas que se apoyan en el almacén de datos para generar información. Estos subsistemas consultan o analizan la información de diferentes formas:

- Informes estándares: informes predefinidos que se ejecutan periódicamente.
- Informes *ad hoc*: informes bajo petición para una consulta de negocio determinada.
- Análisis OLAP: análisis multidimensional que se apoya en los cubos OLAP.
- Cuadros de mando: cuadro resumen con los principales KPI e informes de control.
- Procesos de analítica avanzada: procesos de analítica predictiva, series temporales, etc.

Todos estos subsistemas se nutren del almacén de datos, y es crítica la correcta actualización del mismo en tiempo y forma.

6. 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, hay varios factores que condicionan la evolución de los almacenes de datos:

1) 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 en la gestión y manipulación de los datos. Nos referimos, pues, a un problema que tiene tres dimensiones: velocidad de generación de los datos, volumetría de los datos y variabilidad de los datos, que son las tres dimensiones que caracterizan las tecnologías denominadas *Big Data*.

Buena parte del crecimiento de datos proviene de información no estructurada. El almacenamiento y procesamiento de este tipo de información supone un reto y cambios en los almacenes de datos tradicionales, que se basaban en una estructura y esquema incompatible con la naturaleza de este tipo de información no estructurada.

2) Nuevas técnicas de modelización. Daniel Linstedt publicó en el año 2000 una nueva técnica denominada *data vault*. Su objetivo era la creación de almacenes de datos flexibles y auditables en tiempo real. Se trata de una técnica de modelización basada en tres tipos de entidades:

- *Hubs*: contiene los indicadores claves de negocio.
- *Links*: contiene las relaciones.
- *Satellites*: contiene las descripciones.

Plantea una situación intermedia entre la modelización en tercera forma normal y el esquema en estrella. En esta técnica prima la flexibilidad y la escalabilidad, y permite que el modelo pueda adaptarse de forma sencilla a los cambios en el negocio y en la organización.

Ved también

El esquema en estrella que se desarrolla en detalle en el módulo «Diseño e implementación multidimensional de un *Data Warehouse*» de esta asignatura.

El desarrollo de un *data vault* se realiza en una serie de etapas:

- Identificar los *hubs*.
- Establecer las relaciones (*links*).
- Establecer las descripciones (*satellites*).

- Añadir componentes independientes como calendarios o tablas de relación.

- Añadir tablas necesarias para mejorar rendimiento: tablas puentes, estructuras *point-in-time*, etc.

3) 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.

a) Análisis continuo de datos: mediante flujos continuos de datos, se 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.

b) Procesamiento de eventos complejos: 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.

c) BI de autoservicio: en la actualidad existen herramientas de *Business Intelligence* que acortan los ciclos de creación de cuadros de mando y facilitan una interfaz ágil y con importantes capacidades de conexión e integración. Estas soluciones pueden trabajar conectándose al almacén de datos o directamente a las fuentes de datos origen. El papel que tiene en este tipo de soluciones el almacén de datos no es tan crítico como en las soluciones de *Business Intelligence* tradicional.

d) Bases de datos en memoria: 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.

e) El uso de infraestructura de hardware que da soporte a los *Data Warehouse* ha cambiado y han aparecido nuevos tipos de almacenamiento físico como los discos de estado sólido que mejoran notablemente los procesos de lectura y escritura. Por otra parte, la opción de procesar y almacenar en entornos de computación en la nube cobra cada vez más relevancia en las organizaciones.

f) Los **cambios en las infraestructuras** y la necesidad creciente de **disponer de servidores dedicados para tareas analíticas** ha dado lugar a la aparición de los *Data Warehouse Appliances* que es una plataforma de hardware y software orientada a *datawarehousing* y procesos analíticos. Muchos fabricantes tienen este tipo de plataformas ofreciendo: hardware, sistema operativo y base de datos optimizados para *dawarehousing*.

g) *Data Warehouse* en la nube. Los principales fabricantes ofrecen soluciones para trabajar con almacenes de datos en la nube, bien centrando la solución en el *Data Warehouse*, o bien integrado dentro de una solución *Business Intelligence* completa. Existe la opción de la infraestructura híbrida que combina soluciones en la nube con soluciones en servidores e infraestructura propia.

h) *Virtual Data Marts*: aparición de *Data Marts* virtuales basados en federación de datos. Estos *Data Marts* virtuales no existen físicamente y se crean generando una capa virtual que parte del almacén de datos corporativo. Reducen el movimiento de datos, aunque presentan las limitaciones habituales en cuanto a tiempos de respuesta y soporte de reglas de negocio complejas que tiene la federación de datos.

i) Hadoop, MapReduce, Spark 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).

4) 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. Los almacenes de datos son empleados, con frecuencia, como origen de datos de estos procesos.

5) Convivencia entre los *Data Warehouse* y entornos *Big Data* como Hadoop. Tal y como se ha señalado las tecnologías *Big Data* han experimentado un importante crecimiento en los últimos años. Actualmente, en las compañías conviven tecnologías *Big Data*, como Hadoop, con los almacenes de datos. El ecosistema Hadoop ha ido adquiriendo estos últimos años una función cada vez más importante en la gestión de la información y los procesos de análisis de las compañías. Actualmente hay diferentes situaciones de convivencia entre los almacenes de datos y el ecosistema Hadoop de acuerdo al papel desempeñado por este último.

a) Hadoop como ODS (*Operational Data Store*): en una primera etapa de implantación, Hadoop recibe toda la información no estructurada o generada en tiempo real, dado que este tipo de información es muy costosa de almacenar y gestionar en las bases de datos relacionales. La infraestructura de almacenamiento de Hadoop es económica y muy escalable, lo que produce que Hadoop recopile también información estructurada y pueda almacenar fuentes de datos que alimentan los almacenes de datos, dentro de esta configuración, el crecimiento de Hadoop se dirige a convertirse en el ODS de los almacenes de datos, recopilando toda la información en bruto que posteriormente se consolida en los almacenes. Se trata de un ODS con gran capacidad de almacenamiento, de crecimiento y muy eficaz para recopilar datos generados en tiempo real. En esta configuración la interacción entre Hadoop y los almacenes de datos es la necesaria para comunicar los almacenes con el ODS.

b) Procesos de análisis y *Business Intelligence* (BI) sobre Hadoop. El papel que pueda desempeñar Hadoop dentro de los procesos de análisis y BI determina también la interacción entre Hadoop y los almacenes de datos. Cada vez hay más herramientas de análisis o visualización de datos que se integran con Hadoop y permiten realizar determinados procesos de BI en este entorno, de modo que será necesario traer información de los almacenes de datos a Hadoop para enriquecer estos procesos de análisis. Hoy en día hay muchas organizaciones que centralizan el BI sobre los almacenes de datos que a su vez reciben datos de entornos Hadoop, pero la tendencia de realizar BI en Hadoop es cada vez mayor y esto provocará, como se ha dicho, el trasiego de información desde los almacenes de datos hacia Hadoop.

Resumen

En este módulo hemos introducido el 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 y visto que en realidad no es un concepto nuevo, ya que de manera implícita se estaba utilizando con otras herramientas. Hemos visto que los centros de información han sido los precursores del almacén de datos. A continuación, hemos definido el almacén de datos según Inmon y hemos repasado sus características principales: orientación al tema, integración, no volatilidad y datos históricos.

Además, 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. Se han comparado las bases de datos operacionales con los almacenes de datos y se ha visto que las diferencias son realmente muy importantes.

También se ha introducido el concepto de la factoría de la información corporativa, detallando sus principales componentes: almacén de datos departamental, corporativo, operacional, el componente de integración y transformación, las estructuras multidimensionales y los metadatos. Así mismo, se ha analizado el papel del almacén de datos dentro de un sistema informacional.

Finalmente, se han repasado las tendencias actuales en los almacenes de datos, haciendo hincapié en la evolución de tecnologías tales como el *Big Data*, la analítica avanzada, el análisis continuo de datos, los cambios en la infraestructura (*cloud*, bases de datos en memoria...) o el crecimiento del *Business Intelligence* de autoservicio, entre otros.

Actividades

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

a) Explicad cuáles son los objetivos de este proyecto.

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

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

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?

3. ¿Cuál es la diferencia principal entre el almacén de datos corporativo y el departamental?

4. ¿Cómo justificaríais la necesidad del almacén de datos operacional?

5. ¿Por qué los almacenes de datos departamentales no se alimentan directamente de los sistemas operacionales, en lugar de hacerlo del almacén de datos corporativo?

6. ¿Qué operaciones lleva a cabo el componente de integración y transformación?

7. ¿Cuál es el elemento principal del componente de integración y transformación?

8. ¿Qué papel tienen los metadatos en la FIC?

9. ¿Hay redundancia entre los datos de las bases de datos operacionales y los de la factoría de información corporativa?

10. Rellenad la tabla siguiente indicando las principales diferencias que hay entre los sistemas operacionales y decisionales:

Característica	Sistemas operacionales	Sistemas decisionales
Usuarios típicos		
Número de usuarios		
Tuplas a las que se ha accedido		
Objetivo del sistema		
Funciones principales		
Diseño		
Características de los datos		
Uso		

Característica	Sistemas operacionales	Sistemas decisionales
Acceso		
Unidad de trabajo		
Requerimientos		
Tamaño		

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.

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.

3. La diferencia principal es el tamaño. Mientras el almacén de datos corporativo contiene todos los datos que interesan o pueden llegar a interesar a cualquiera de la empresa, un almacén departamental solo contiene aquellos que en un momento dado interesan a un conjunto de analistas.

4. El almacén de datos operacional sirve para satisfacer de manera eficiente y sin interferir en los sistemas operacionales las necesidades de acceso integrado a datos no históricos.

5. Si cargamos los datos de los almacenes de datos departamentales directamente de las bases de datos operacionales, multiplicamos los procesos necesarios de integración y transformación de los datos.

6. El componente de integración y transformación obtiene los datos de las fuentes de datos, los depura, transforma e integra, los transporta a los almacenes de datos y los carga allí. También obtiene datos del almacén de datos operacional y los transforma, transporta y carga en el almacén de datos corporativo. Además, hace la misma operación entre el almacén de datos corporativo y los almacenes de datos departamentales.

7. A diferencia de otros componentes de la FIC, cuyo elemento principal es la base de datos, el elemento principal del componente de integración y transformación es el software que implementa su misión.

8. Generalmente, los metadatos son datos que nos dan información sobre otros datos. En la FIC, son el componente que se encarga de cohesionar el resto de los componentes.

9. Sí, hay algunos datos que están en los dos sistemas. Sin embargo, esta redundancia es mínima y necesaria, puesto que los sistemas operacionales no guardan datos históricos, ni agregados, ni han pasado un proceso de limpieza e integración.

10. Las diferencias principales entre los sistemas operacionales y los decisionales son las siguientes

Característica	Sistemas operacionales	Sistemas decisionales
Usuarios típicos	Administrativos	Analistas(ejecutivos)
Número de usuarios	Miles	Centenares
Tuplas a las que se ha accedido	Centenares	Miles
Objetivo del sistema	Ejecución del negocio	Análisis del negocio

Característica	Sistemas operacionales	Sistemas decisionales
Funciones principales	Operaciones diarias (OLTP)	Toma de decisiones (OLAP)
Diseño	Orientado a la funcionalidad	Orientado al tema
Características de los datos	Actuales y actualizados, atómicos, normalizados, aislados	Históricos, resumidos (agregados), desnormalizados, integrados
Uso	Repetitivo y rutinario (consultas predeterminadas)	Esporádico e innovador (consultas *ad hoc*)
Acceso	R/W	Principalmente lectura
Unidad de trabajo	Transacciones simples	Consultas complejas
Requerimientos	Rendimiento de transacciones + consistencia de datos	Rendimiento de las consultas y precisión de los datos
Tamaño	MB/GB	GB/TB

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 apoyar en la toma de decisiones en *Data Warehouse*.

almacén de datos corporativo *m* Conjunto de datos que guarda integrados todos los datos históricos de la empresa.

almacén de datos departamental *m* Conjunto de datos que resuelve las necesidades de análisis de un departamento o conjunto de usuarios.

almacén de datos operacional *m* Conjunto de datos integrado y orientado al tema, pero sin datos históricos. Se suele utilizar como paso intermedio en la construcción del almacén de datos corporativo.

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

Big Data Conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos.

cloud Computación en la nube, conocida también como servicios en la nube, informática en la nube, nube de cómputo o nube de conceptos (del inglés *cloud computing*), es un paradigma que permite ofrecer servicios de computación a través de una red, que usualmente es Internet.

data governance Véase gobierno del dato.

data vault Conjunto de almacenes de datos flexibles y auditables en tiempo real.

Data Warehouse Véase almacén de datos.

Data Warehouse Appliances Plataforma de *hardware* y *software* orientada a *datawarehousing* y procesos analíticos.

dato (definición desde el punto de vista de los sistemas decisionales) *m* Medida, observación hecha y almacenada en algún sistema.

factoría de información corporativa *f* Conjunto de elementos de software y hardware que ayudan al análisis de datos para tomar decisiones. Sigla FIC

FIC *f* Véase factoría de información corporativa.

gestión de datos maestros *f* Metodología que identifica la información más crítica de una organización y crea una única fuente fiable

gobierno del dato *m* Metodología que tiene por objeto asegurar que los datos son siempre fiables y válidos en cada contexto empresarial, que la calidad se mantiene a lo largo del tiempo y que existen mecanismos de control sobre quién puede hacer qué con los datos en cada momento.

master data management Véase gestión de datos maestros.

metadato *m* Datos sobre datos.

OLAP Siglas que hacen referencia a las herramientas de análisis, normalmente multidimensional en *on-line analytical processing*.

OLTP Siglas de *on-line transaccional processing*.

SGBD Véase sistema de gestión de bases de datos.

sistema de gestión de bases de datos *m* *Software* que gestiona y controla bases de datos. Sus funciones principales son las de facilitar su uso simultáneo a muchos usuarios de distintos tipos, independizar al usuario del mundo físico y mantener la integridad de los datos. Sigla SGBD.

sistema de registro *m* Fuente de cada uno de los datos de los almacenes de datos, de entre todas las fuentes posibles.

sistema informacional *m* Sistema que apoya los procesos de toma de decisiones por parte de los analistas en la organización.

sistema operacional *m* Sistema que ayuda en las operaciones diarias del negocio de una organización.

sistema transaccional *m* Sistema basado en transacciones de lectura/escritura.

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*. Boston: 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.

Inmon, W. H.; Strauss, D.; Neushloss, G. (2010). *DW 2.0: The Architecture for the Next Generation of Data Warehousing*. Burlington, Mass.: Morgan Kaufman Series in Data Management Systems.

Kimball, R. (2002). *The Data warehouse toolkit: the complete guide to dimensional modeling*. Nueva York: Wiley.