La factoría de información corporativa

Alberto Abelló Gamazo
José Samos Jiménez
Josep Curto Díaz

PID_00203541

Universitat Oberta de Catalunya

Índice

Introducción

Las empresas deben adaptarse a los cambios de su entorno (clientes, proveedores, tecnología, etc.). Por este motivo, cada día es más importante tomar decisiones. En este sentido los datos han ganado especial importancia, hasta llegar a ser un bien más de las empresas, como podrían serlo las materias primas, la energía, el capital o las personas. Hay que disponer de todos los datos posibles, tanto del negocio como de todo lo que lo rodea, y ser capaz de analizarlos de manera eficiente para decidir qué es lo mejor que se puede hacer en cada situación.

Generalmente, podemos considerar que ya tenemos la mayoría de la información necesaria para tomar decisiones en los sistemas operacionales de la empresa. En asignaturas relacionadas con las bases de datos y la ingeniería del software, ya hemos visto cuáles son las características de estos sistemas operacionales. Desgraciadamente, como hemos visto y continuaremos viendo más adelante en este módulo, este tipo de sistemas no son los más adecuados para tomar decisiones. Por lo tanto, se requiere la información y gestionarla de modo que se faciliten las tareas de análisis.

Sistemas operacionales

Entendemos por sistemas operacionales aquellos que nos ayudan en las operaciones diarias del negocio, en contraposición con los sistemas de análisis, que nos ayudan a tomar decisiones.

William Inmon presentó en 1998 lo que se denomina factoría de información corporativa. Se trata de 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 de apoyo a los analistas (producto elaborado), para 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.

Factoría de información corporativa

En inglés, *corporative information factory*. Lo abreviaremos como *FIC*.

Se trata de una factoría o fábrica de información en el ámbito de toda la organización o en el corporativo.

Objetivos

Este módulo presenta una arquitectura para gestionar información que ayuda a tomar decisiones. Se presentan las características de cada uno de los elementos de la arquitectura. Con este módulo, lograréis los objetivos siguientes:

1. Entender la necesidad de la factoría de información corporativa en la gestión del conocimiento para tomar decisiones.

2. Distinguir los diferentes elementos arquitectónicos de una factoría de información corporativa.

3. Ser capaces de razonar la necesidad de los diferentes sistemas de almacenamiento.

4. Saber qué son los metadatos y cuál es su papel como elemento integrador de los distintos subsistemas.

5. Reconocer la importancia de los metadatos en la toma de decisiones.

1. Usuarios y fuentes de información de los almacenes de datos

En primer lugar, en este apartado veremos los dos extremos de la cadena de obtención de información, es decir, quiénes son los usuarios de la FIC (qué queremos) y cuáles son las fuentes de información que han de satisfacer sus necesidades (qué tenemos). Esto nos hará pensar sobre qué debe haber en medio para hacer que las dos cosas sean compatibles.

1.1. Los usuarios

Los sistemas operacionales tienen muchos usuarios que acceden a muy pocos datos, mientras que, en lo que respecta a los sistemas de análisis, los utilizan muy pocos usuarios que quieren ver muchos datos. Recordemos que los sistemas operacionales se utilizan en el día a día de la empresa. Sirven para facilitar tareas rutinarias y repetitivas de los oficinistas. Cuando hablamos de tareas de análisis, las cosas se vuelven algo más complejas y podemos identificar **tres tipos diferentes de usuarios**, que podemos denominar: granjero, explorador y turista. Realmente, estos nombres no son muy importantes. Lo que sí importa son las características de cada uno y los requisitos que presentan.

Lectura recomendada

Podéis encontrar los tres tipos de usuarios extensamente explicados en la obra siguiente:

W. H. Inmon; C. Imhoff; R. Sousa (1998). *Corporate Information Factory*. EE. UU.: John Wiley & Sons, Inc.

Los analistas tienen requerimientos diferentes de los que presentan los oficinistas. Además, podemos diferenciar distintos tipos de analistas con características muy distitnas, que la FIC ha de tener en cuenta.

1.1.1. Granjero

Este primer tipo de usuario lleva a cabo accesos a la información absolutamente predecibles y repetitivos. De manera regular, encuentra cosas interesantes que ayudan a que la empresa funcione. En todo momento sabe qué quiere y cómo lo ha de obtener, porque, generalmente, repite las consultas de manera periódica. Podríamos decir que tiene su parcela de información y se dedica a cultivarla para extraer provecho de la misma regularmente. No accede a grandes cantidades de datos (puesto que nunca sale de su parcela) y los suele pedir resumidos, aunque le puede llegar a interesar ver diferentes niveles de detalle.

Este tipo de usuario suele utilizar **herramientas OLAP**[1]. Estas herramientas están pensadas para ser utilizadas por personal no informático. Son sencillas, comprensibles y ponen énfasis en la presentación de los resultados. Mediante

(1)Sigla de la expresión inglesa *online analytical processing*, 'procesamiento analítico en línea'.

el modelo multidimensional (muy cercano a la manera de entender el negocio de este tipo de usuarios), consiguen reflejar la complejidad que hay en las estructuras y relaciones de la vida real.

En este grupo tenemos a los empleados, los proveedores y los clientes a los que la organización proporciona servicios informacionales. Actualmente, la inteligencia de negocio operacional, que potencia el uso de estos sistemas en todas las capas de la organización, permite a los usuarios de negocio utilizar los datos y la información en los procesos de negocio de manera natural, sin tener que salir de sus aplicaciones. Esto se debe al hecho de que la información se encuentra integrada y en cualquier momento es accesible a los procesos de negocio, de modo que los usuarios mismos muchas veces no son conscientes ni del hecho de que usan el almacén de datos.

Ejemplo de análisis en línea

Como ejemplo de granjero, podemos pensar en la persona encargada de hacer previsiones de *stock* para los almacenes. Esta persona seguramente querrá disponer de los datos de *stock* de cada producto durante los últimos años, y también de los pedidos pendientes de servir. Basándose en estos datos, tendrá que decidir qué hay que comprar y cuándo. Si compráramos demasiado o a deshora, la empresa podría perder mucho dinero. No se tiene que confundir a este analista con la persona que simplemente registra las entradas y salidas del almacén, que no ha de tomar ninguna decisión.

1.1.2. Explorador

Hay otros usuarios analistas que, al contrario que los granjeros, tienen unos accesos totalmente imprevisibles e irregulares. Pasan una gran parte del tiempo sin consultar los datos, planificando o preparando su estudio y, cuando lo tienen todo a punto, empiezan a explorar de golpe una gran cantidad de datos tan detallados como sea posible. Realmente, no saben exactamente qué buscan hasta que lo encuentran, y los resultados en ningún caso están garantizados. Sin embargo, a veces encuentran algo realmente interesante que claramente mejora el negocio. Con frecuencia se conocen como usuarios exploradores (*power users*) de la organización.

Un usuario explorador suele ser informático y/o estadístico, experto en prospección de datos y por lo tanto, con dominio de herramientas de análisis estadístico. Estas herramientas tratan de extraer información oculta (no evidente) de un conjunto de datos. Generalmente, son semiautomáticas (como mínimo piden algunos parámetros o que los usuarios validen los resultados) y tienen que estar controladas por técnicos especializados.

En el contexto actual, como resultado de la problemática conocida como *big data*, la figura del explorador ha evolucionado hacia una nueva figura: el **científico de datos** (*data scientific*). Un científico de datos tiene que ser capaz de extraer información de grandes conjuntos de datos (en términos del problema de *big data*) de acuerdo con un objetivo claro de negocio, no aleatoriamente, y después presentarla de manera sencilla al resto de los usuarios no expertos de

Big data

Cuando hablamos de *big data* nos referimos al crecimiento de los datos en volumetría, en velocidad de generación y en variabilidad de origen y forma.

la organización. Por lo tanto, se trata de un perfil transversal con conocimientos de informática, matemáticas, estadística, minería de datos, diseño gráfico, visualización de datos y usabilidad.

Este perfil será clave para las organizaciones que quieren generar ventajas competitivas a partir de la información. En los próximos años, la demanda de este perfil se incrementará precisamente en aquellas organizaciones que ya tienen en consideración este tipo de necesidad y están desplegando iniciativas de analítica de negocio, es decir, en las organizaciones que ya han logrado un nivel de madurez alto en la explotación de datos y en la generación de información de valor.

Ejemplos de minería de datos

Podemos utilizar herramientas de minería de datos para reconocer patrones de comportamiento para detectar fraudes (facturas, hipotecas o llamadas telefónicas impagadas), generar reglas de manera automática para componer una cartera de valores invertidos en bolsa, encontrar factores de riesgo en un postoperatorio o descubrir relaciones entre las compras de ciertos productos en el supermercado (por ejemplo, pañales y cerveza).

1.1.3. Turista

Tendríamos que entender este último tipo de usuario como un equipo formado por dos o más personas. Por un lado, tendríamos a la persona que posee una visión global de la empresa a la que se le ocurre la posibilidad de hacer un estudio sobre un cierto tema. Por otro lado, habría un experto en informática, conocedor de los sistemas de análisis de la empresa, encargado de averiguar si el estudio es factible con los datos y las herramientas disponibles o no.

Este equipo mirará datos sin seguir ningún patrón de acceso y raramente observará dos veces los mismos datos. Por lo tanto, tampoco podemos conocer sus requerimientos *a priori*. Además de los datos, también estará especialmente interesado en consultar los metadatos. Las herramientas que utilizarán los turistas son **navegadores o buscadores** (tanto de datos como de metadatos) y el resultado de su trabajo serán proyectos que llevarán a cabo los granjeros o los exploradores.

Los metadatos

Los metadatos son datos sobre los datos. Los podéis ver con más detalle en el apartado "Los metadatos" de este mismo módulo.

Un usuario turista es, en definitiva, un usuario casual de la información.

1.2. Las fuentes de información y sus datos

La primera pregunta que nos tenemos que hacer es si ya tenemos la solución a los problemas que presentan estos usuarios y herramientas de análisis. Actualmente, lo que tienen las empresas es un conjunto de aplicaciones independientes, puestas en marcha en distintos momentos, que dan respuesta a diferentes requerimientos. La gran mayoría de estas aplicaciones solo están

concebidas para dar apoyo al proceso de negocio (por ejemplo, la gestión de personal, contabilidad, etc.). Ninguna de estas aplicaciones se diseñó para ser utilizada por los analistas.

El espaciamiento de los momentos de desarrollo de cada aplicación, las diferencias de presupuestos y requerimientos y la falta de planificación hacen que encontremos más heterogeneidades de las que querríamos entre los sistemas operacionales. Si los analistas quisieran acceder directamente a los datos de estas fuentes de información, lo primero que deberían hacer sería **superar estas heterogeneidades entre las aplicaciones**.

Tipos de heterogeneidades

Podemos encontrar tanto heterogeneidades semánticas (el mismo tipo de información representado de maneras diferentes), como de sistemas (por ejemplo, hardware diferente, sistema operativo distinto o simplemente sistema de gestión de base de datos –SGBD– diferente).

Aunque los analistas fueran capaces de acceder a los múltiples sistemas operacionales a la vez, hay que tener presente que cada uno de estos ha sido diseñado para resolver un cierto problema de manera eficiente. Una de las implicaciones de esto es que no guardan más datos de los necesarios para resolver el problema correspondiente, puesto que esto empeoraría su eficiencia. Concretamente, **no guardan datos históricos** si no hace falta. Este tipo de datos son imprescindibles para tener una referencia a la hora de tomar decisiones. Una implicación directa de esta necesidad de datos históricos es el gran volumen de datos que han de gestionar los sistemas de análisis.

Ejemplo de datos necesarios para un análisis de telefonía en Cataluña

Imaginemos que utilizamos 4 bytes para codificar el origen de una llamada y 4 más para codificar el destino. Además, también podemos codificar el momento en que se hace la llamada con 4 bytes y su duración con 2. En total, para cada llamada solo necesitaremos guardar 14 bytes. Pensemos que la media de llamadas por persona y día es aproximadamente de diez y que en Cataluña viven seis millones de personas. Si quisiéramos hacer un estudio de los últimos tres años, tendríamos que nos hace falta lo siguiente:

3 años × (365 días/año) × (6×10^6 personas) × (10 llamadas/persona y día) × (14 bytes/llamada) = 10^{12} bytes = 10^9 kB = 10^6 MB = 10^3 GB = 1 TB.

Esto quiere decir que necesitaríamos diez discos de 100 GB para guardar toda esta información y aproximadamente estaríamos dos horas y media leyendo todos los datos (si asumimos una velocidad de lectura de 100 MB/segundo). Este tiempo de respuesta resulta claramente inadmisible para cualquier persona que haga una consulta interactiva. Pensad que las técnicas de indexación habituales tampoco sirven de mucho cuando lo que queremos consultar es la suma, la media, el mínimo o el máximo de la duración de las llamadas.

En contraposición a los sistemas OLAP, podríamos identificar los sistemas operacionales con los procesamientos transaccionales en línea (*OLTP*, *on-line transactional processing*). Estos sistemas transaccionales están pensados para obtener datos. Por lo tanto, les resulta esencial evitar la introducción de datos erróneos y permitir accesos concurrentes de manera aislada, sin interferencias. Así pues, están diseñados para manipular una gran cantidad de pequeñas transacciones que implican modificaciones de los datos.

En cambio, para los analistas, la disponibilidad de los datos es mucho más importante que el aislamiento. Sus consultas son mucho más complejas e involucran muchos datos. Estos analistas no pueden esperar un cierto conjunto de datos bloqueado por alguien que los modifique, porque, con la gran canti-

dad de datos que consultan, la probabilidad de que alguien estuviera modificando alguno sería demasiado alta. Además, **los analistas solo quieren hacer consultas** (estamos en un entorno solo de lectura, *read only*), de modo que todas las precauciones del mundo transaccional (pensadas para entornos de lectura/escritura, *read/write*) son totalmente innecesarias.

Los sistemas operacionales piden un buen rendimiento en la ejecución de transacciones que siempre tienen que dejar la base de datos en un estado consistente, mientras que los sistemas de análisis requieren ejecutar consultas complejas que retornen datos precisos en un tiempo de respuesta bajo. Intentar compatibilizar requerimientos solo de lectura o de lectura/escritura sería malo para los dos entornos. Esto quiere decir que no podemos utilizar los sistemas operacionales, sino que debemos crear sistemas independientes para los analistas.

A pesar de que los datos de los sistemas operacionales que tiene la empresa nos resulten muy interesantes, estos sistemas no cumplen los requerimientos de los analistas. Hay que definir un sistema que aproveche estos datos y que satisfaga las necesidades de estos usuarios de manera adecuada.

1.3. El almacén de datos

La solución a las necesidades de los analistas es construir una base de datos de análisis, que denominaremos **almacén de datos**, partiendo de las bases de datos operacionales, pero que funcione de manera independiente de estas.

Ved también

Podéis ver la definición de almacén de datos que aparece en el módulo "Introducción al almacenamiento de datos".

2. Los almacenes de datos departamentales

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 (solo la parte que interese a un cierto departamento o conjunto de personas).

Estos **almacenes de datos departamentales**[2] normalmente estarán diseñados siguiendo el modelo 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, solo contienen datos históricos dentro del periodo de tiempo que sea estrictamente necesario.

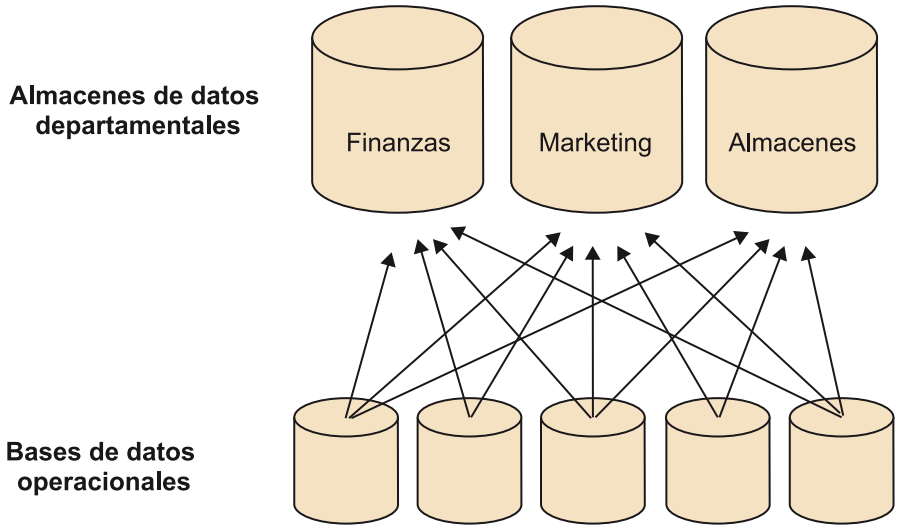
[(2)]En inglés, *data marts*.

Ved también

Veremos el modelo multidimensional en el módulo "Diseño multidimensional" de esta asignatura.

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 querrá hacer, para calcular previamente los resultados, de modo que el cálculo no se tenga que hacer en el momento concreto en que se solicitan.



Como podéis ver en la figura superior, para cada grupo de usuarios o departamento que lo requiera construimos uno de estos almacenes, que solo integra los datos de las fuentes de información que sean necesarios para satisfacer las necesidades concretas de su grupo de usuarios (lo que también facilita su funcionamiento). 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.

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).

Problemas de afinación

En cualquier base de datos, cuantos más usuarios tenemos más difícil se hace compatibilizar todos los requerimientos para conseguir el rendimiento óptimo.

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

3. El almacén de datos corporativo

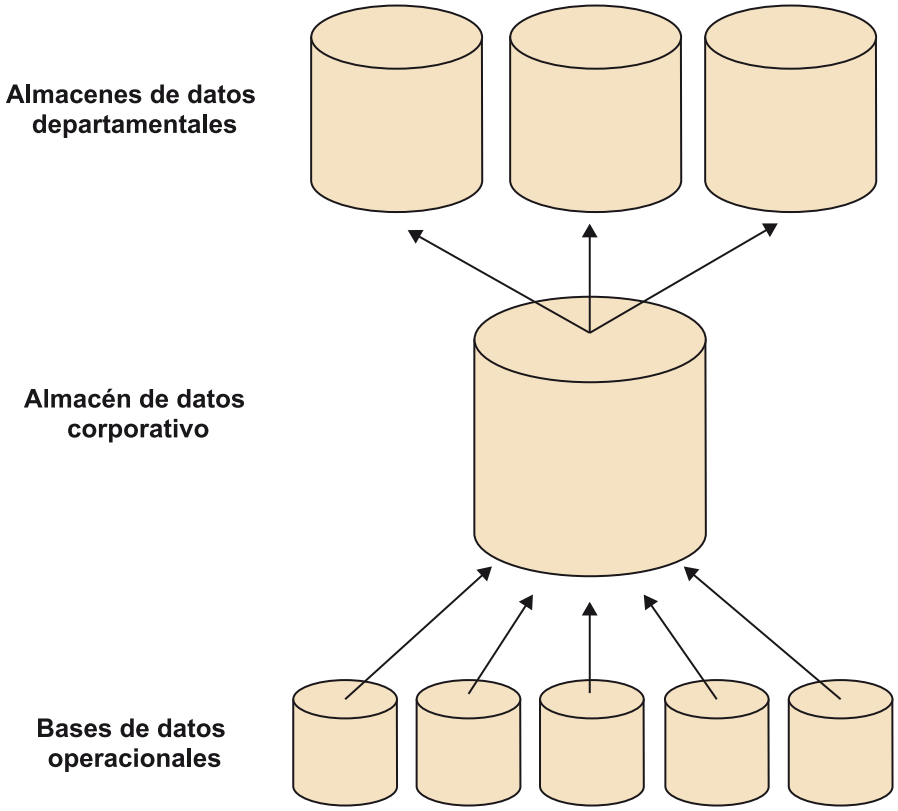
Almacén de datos corporativo

Acostumbramos a referirnos al almacén de datos corporativo con el término inglés *data warehouse*.

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 anterior, 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.

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 siguiente, el almacén de datos corporativo resulta 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.



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

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 cierto 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.

Realmente, los almacenes de datos departamentales no se alimentan directamente de las fuentes de datos, sino del almacén de datos corporativo. Sin embargo, por transitividad y sin olvidar esta puntualización, también podemos estudiar la diferencia que hay entre las fuentes de datos de los almacenes departamentales y de los corporativos. Como ya hemos visto, los almacenes de datos departamentales solo contienen los datos que interesan a un cierto grupo de analistas. Por lo tanto, solo guardarán datos que provengan de las fuentes de datos correspondientes. En cambio, el almacén de datos corporativo, dado que contiene todos los datos que interesan o pueden llegar a interesar a los analistas de cualquier departamento, debe alimentarse de la unión de todas las fuentes de información de todos los almacenes departamentales, además de aquellas fuentes de datos que hoy no interesan a ningún analista, pero que potencialmente pueden llegar a interesar a alguien.

Siguiendo el mismo razonamiento que acabamos de hacer para las fuentes de información, también podemos ver que los volúmenes de datos que contienen los dos tipos de almacenes de datos deben ser de órdenes de magnitud diferentes. Generalmente, podemos considerar que un almacén de datos departamental con datos solo de un cierto conjunto de temas y que no los contiene con el máximo nivel de detalle puede ocupar unos cuantos (posiblemente muchos) gigabytes (10^9 bytes) de datos. Por el contrario, el almacén de datos corporativo, que a la vez ha de contener los datos de todos los almacenes departamentales y no puede perder detalle (tiene que guardar la información tan detallada como sea posible, por si algún día alguien la necesita), ocupará un número de terabytes (10^{12} bytes) de datos.

Haciendo una simple regla de tres, podemos deducir que, si un tipo de almacén contiene muchos más datos que el otro, ha de integrar más fuentes de datos y tiene que ser mucho más genérico, entonces tardaremos mucho más tiempo en desarrollarlo y deberemos invertir en el mismo muchos más recursos (tanto económicos como humanos). Generalmente, podemos considerar que un proyecto para construir un almacén de datos departamental dura algunos meses, mientras que desarrollar el almacén de datos corporativo de la empresa es un proceso que dura años.

El almacén de datos corporativo acostumbra a estar implementado sobre un SGBD relacional simplemente por las prestaciones que ofrecen estos sistemas en la gestión de grandes volúmenes de datos, no porque estos sistemas estén especialmente concebidos para ello. Aún más, el rendimiento de los sistemas transaccionales suele ser especialmente malo en tareas de análisis, si no se extienden con mecanismos específicos de análisis (por ejemplo, nuevos tipos de índices, carga masiva de datos, etc.). En cambio, los almacenes de datos de-

Ved también

Veremos los diferentes sistemas multidimensionales (ROLAP, MOLAP, HOLAP, etc.) en el módulo "Diseño multidimensional".

partamentales, sin un requerimiento tan grande respecto al volumen de datos y valorando mucho más el tiempo de respuesta, se suelen implementar sobre sistemas que se basan en el modelo multidimensional.

Los sistemas operacionales están diseñados para responder bien ante pequeñas transacciones que modifican los datos. El modelo relacional ofrece toda la teoría de la normalización para conseguir que los SGBD tengan un buen rendimiento con este tipo de accesos, de modo que una modificación afecte a una sola tabla y se eviten las anomalías de inserción, actualización y borrado. Sin embargo, en el caso del almacén de datos corporativo, nos tenemos que plantear si, aunque utilicemos un SGBD relacional, realmente hay que normalizar nuestro esquema. La carga de datos se produce de manera masiva, toda de golpe y en el momento en que los usuarios no hacen consultas. ¿Qué tiene esto en común con las pequeñas transacciones de alta, baja, modificación y consulta de las bases de datos operacionales?

4. El 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.

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.

Lectura complementaria

Podéis encontrar mucha más información sobre el almacén de datos operacional en el libro siguiente:

W. Inmon; C. Imhoff; G. Batas (1996). *Building the Data Warehouse* (2.ª ed.). EE. UU.: John Wiley & Sons, Inc.

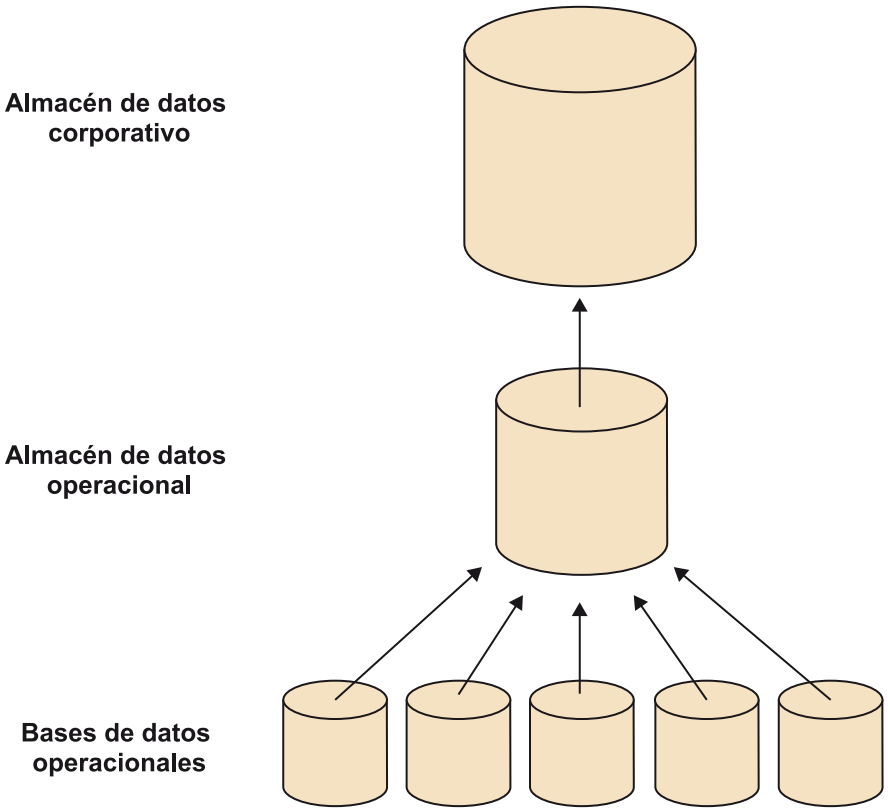
La aparición de este repositorio viene dada por la típica ponderación entre volumen de datos y velocidad del sistema. Hasta ahora, en los otros almacenes, lo que queríamos era tener absolutamente cualquier dato que pudiéramos llegar a necesitar para tomar una decisión. Como consecuencia de este requerimiento, el tiempo de respuesta puede llegar a degradarlo 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.

Además de permitir el acceso a datos operacionales integrados actualizados, como podemos ver en la figura siguiente, también facilita la construcción del almacén de datos corporativo. Se puede ver como una estrategia de dividir y vencer. En lugar de conseguir las cuatro características del almacén de datos corporativo a la vez, primero conseguimos dos de las mismas mediante el almacén de datos operacional (orientación al tema e integración) y en un se-

(3)En inglés, *back-up*.

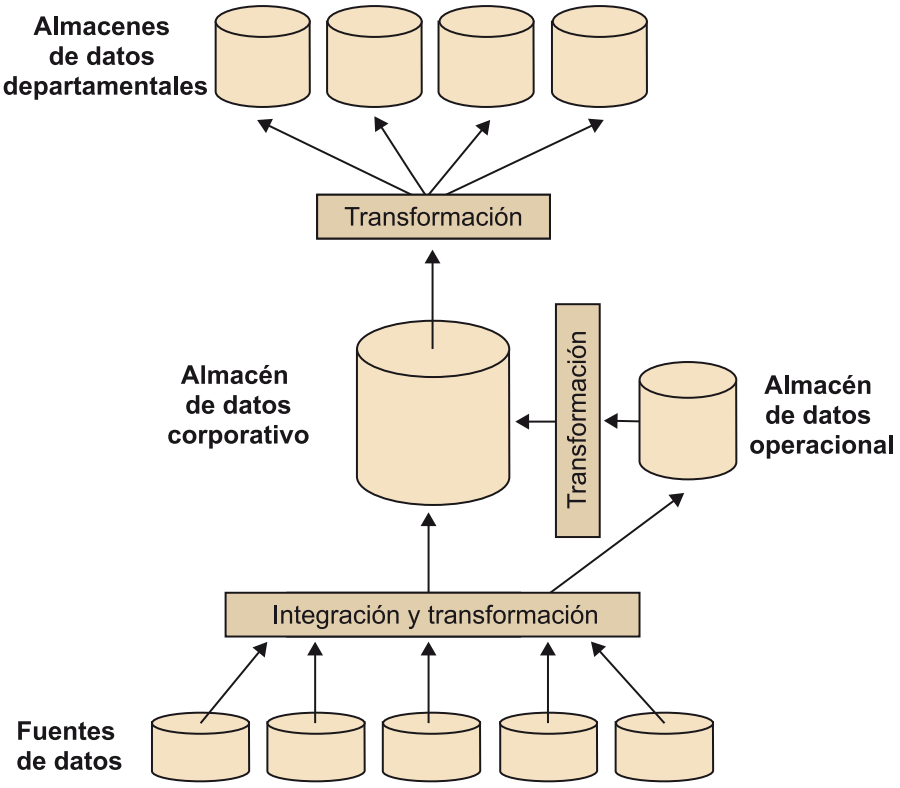
gundo paso añadimos la temporalidad (historicidad y no volatilidad). El paso del almacén de datos operacional al corporativo puede ser tan sencillo como hacer un volcado[3] de los datos.



5. El componente de integración y transformación

Como hemos visto en el apartado "Usuarios y fuentes de información de los almacenes de datos", 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.



Flujo de datos mediante el 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.

Los sistemas operacionales

Los sistemas operacionales son la puerta principal de entrada de datos a la organización, aunque esta también puede utilizar datos externos obtenidos de distintas fuentes (informes especializados, artículos de prensa, etc.). En los últimos tiempos, la fuente principal de datos externos para las organizaciones suele ser Internet.

Los términos integración y transformación

El componente de integración y transformación se suele referenciar mediante las siglas en inglés *I&T* (*integration and transformation*). También es frecuente encontrarlo referenciado como proceso *ETL* (*extraction, transformation and loading*). Para que su nombre fuera totalmente descriptivo, se tendría que denominar componente de obtención, transformación, depuración, integración, transporte y carga. Aquí utilizaremos la terminología definida en Inmon, Imhoff y Sousa (1998).

Integración y transferencia de los datos

Como podemos ver en la figura, los datos obtenidos de las fuentes de datos se cargan en el almacén de datos operacional y desde allí se transfieren al almacén de datos corporativo, aunque algunos también se pueden cargar directamente. Los datos ya integrados en el almacén de datos corporativo se transforman y cargan a los diferentes almacenes de datos departamentales.

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 estén 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 funcionan el resto de los componentes de la FIC.

En este apartado estudiaremos las actividades mencionadas del componente de integración y transformación.

5.1. Obtención de los datos

A partir de las fuentes de datos, tenemos que obtener los datos requeridos por los almacenes de datos de la FIC. Los almacenes de datos guardan la historia de los datos para permitir analizar su evolución; los sistemas operacionales generalmente tan solo guardan una imagen. Es decir, partiendo de sistemas que mantienen una "imagen" de los datos, tenemos que obtener aquellas que son necesarias para formar su historia, que podríamos representar como una "película" de los datos, entendiéndola como una secuencia de imágenes. Debemos obtener las actualizaciones que se producen sobre los datos para ir montando esta película[4].

(4) Fotografía o imagen; en inglés, se denomina *snapshot*.

Película

Aunque frecuentemente se hable de película para dar una visión de los datos en un almacén de datos, estos no estarán formados necesariamente por una sucesión de imágenes, sino que, entre otras posibilidades, pueden consistir en una imagen inicial y una serie de diferencias de las sucesivas imágenes obtenidas.

En la obtención de los datos de un almacén de datos, se distinguen dos fases:

- Obtención de los datos de la imagen inicial.
- Obtención de los datos para las actualizaciones.

Las fuentes de datos generalmente mantienen sus datos estructurados según la utilización que se haga de los mismos. Debido a la falta de integración entre las diferentes aplicaciones, resulta frecuente encontrar datos replicados entre sí. Es decir, para cada dato considerado podemos encontrar distintas fuentes disponibles.

El primer paso en la obtención de los datos consiste en determinar, de entre todas las fuentes posibles, cuál es la más adecuada para cada uno de los datos requeridos: debe determinarse lo que Inmon denomina el **sistema de registro**.

Obtendremos cada dato de la fuente o las fuentes que mejor se adapten a los requerimientos de calidad, precisión, estructura o disponibilidad de los almacenes de datos. Asimismo, para cada dato deberemos determinar qué fuente es la más adecuada para obtener la imagen inicial y de dónde obtendremos las sucesivas actualizaciones.

Actualmente, el patrón de crecimiento de los datos en el contexto empresarial ha cambiado por distintos motivos. Por un lado, por la aparición de múltiples dispositivos que generan datos de valor nuevos para las organizaciones (por ejemplo, sensores distribuidos en una ciudad para monitorizar la eficiencia del tráfico o los sistemas de distribución de agua, gas o electricidad). Por otro lado, el usuario cada vez presenta un comportamiento más activo por medio de las redes sociales, el comercio electrónico y los nuevos dispositivos inteligentes como teléfonos inteligentes (*smartphones*) y tabletas (*tablets*), y por lo tanto podemos decir que se convierte en el generador de datos principal. Como consecuencia de todo esto, el tamaño del fichero se ve reducido comparativamente y la cantidad de datos en tránsito genera una sombra digital de alto valor para las organizaciones (por ejemplo, en los sistemas de recomendación).

En este nuevo contexto, la información de valor para una organización no siempre se encuentra en los sistemas transaccionales y, a menudo, los datos no son estructurados. Según el grado de estructuración (y por extensión, la dificultad de extracción de la información) de los datos, los podemos clasificar en los tipos siguientes.

1) Datos estructurados: se caracterizan por tener una estructura conocida y se almacenan principalmente en bases de datos relacionales. La manipulación de los datos se hace por medio de gestores de bases de datos, y las consultas mediante SQL.

2) Datos semiestructurados: se encuentran encapsulados en ficheros semiestructurados como XML[5] o SGML[6]. En esta situación es posible trabajar con el contexto de negocio, lo que proporciona gran valor a las organizaciones. Ac-

(5)Del inglés, *extensible markup language*.

tualmente encontramos bases de datos especializadas en XML para manipular este tipo de datos, y también técnicas como *web-mining* (minería de datos aplicada a la web) que permiten recuperar información de páginas web.

(6)Del inglés, *standard generalized markup language*.

3) Datos no estructurados: encapsulados en objetos sin una estructura predefinida (audio, vídeo, PDF o Word) que requiere el uso de técnicas especiales como *text-mining* (minería de datos aplicada a ficheros de texto) o *information retrieval* (técnicas, con frecuencia estadísticas, aplicadas a encontrar información relacionada con un concepto en ficheros).

5.1.1. Obtención de los datos de la imagen inicial

Generalmente, los sistemas operacionales solo guardan una imagen de sus datos o bien una historia reducida de estos. Esta imagen es la que se ha de obtener para traspasarla a los almacenes de datos.

Si las diferentes imágenes almacenadas en los sistemas operacionales se han ido perdiendo a medida que se han hecho modificaciones, solo podremos disponer de la historia que almacenamos a partir del momento en que se construyan los almacenes de datos.

En algunos casos, por diferentes motivos (por ejemplo, por motivos legales), puede haber una historia más extensa de los datos, a veces fuera de los sistemas operacionales, aunque obtenida a partir de estos. En cada caso, tendremos que valorar si es útil para los analistas disponer en los almacenes de datos de la historia que había antes; en caso positivo, en lugar de partir de la imagen inicial partiríamos de una película inicial.

Datos históricos en un banco

En un banco, el sistema operacional de gestión de movimientos de las cuentas solo guarda los datos de los últimos doce meses. Los datos de los meses anteriores hasta un total de cinco años se tienen que almacenar por motivos legales. Aun así, estos movimientos históricos no se almacenan en el sistema operacional, sino que mensualmente se extraen de la base de datos del sistema y se almacenan en un medio de almacenamiento más económico. Aunque estos datos permanecen accesibles dentro de la organización, solo se accede a los mismos de manera puntual, por motivos operacionales, no para analizarlos.

Para obtener la imagen inicial, tendremos que desarrollar un conjunto de aplicaciones de obtención de los datos de las fuentes de datos, las cuales generalmente se ejecutarán una sola vez.

5.1.2. Obtención de los datos para las actualizaciones

Una vez tenemos la imagen inicial de los datos en los almacenes de datos, de manera repetitiva, según las necesidades de los analistas, deberemos obtener las modificaciones llevadas a cabo en las fuentes de datos. De este modo, iremos formando la película de los datos que ofreceremos a los usuarios para que la analicen.

Para obtener las actualizaciones, deberemos desarrollar un conjunto de aplicaciones de obtención de los datos de las fuentes de datos, las cuales se ejecutarán con frecuencia.

5.2. Transformación, depuración e integración de los datos

Cuando ya hemos obtenido los datos de las diferentes fuentes:

- Cada conjunto de datos puede tener una estructura distinta dependiendo de la fuente de la que proceda. Los tenemos que transformar para adaptarlos a la estructura del esquema del almacén de datos en el que se almacenarán.

- Tenemos que depurar los errores o conflictos que podamos encontrar dentro de los datos de cada una de las fuentes.

- Tenemos que integrar los datos depurando errores o conflictos entre datos de fuentes distintas.

Como resultado de este proceso, obtendremos un conjunto de datos directamente utilizable para actualizar el almacén de datos correspondiente.

A continuación, comentaremos algunos detalles de estas operaciones por separado. Esto no significa que se hagan de manera secuencial, sino que se pueden combinar o intercalar algunas de estas según las necesidades.

5.2.1. Transformación de los datos

Las transformaciones que hay que hacer sobre los datos pueden ser muy variadas. Entre las más frecuentes, encontramos las siguientes.

- Cambiar el formato o el tipo de los datos (por ejemplo, los campos de fecha).

- Cambiar la codificación (por ejemplo, EBCDIC a ASCII).

- Reestructurar los campos (por ejemplo, fusionar o dividir campos o cambiar su orden relativo).

- Cambiar las unidades o los códigos de representación (por ejemplo, cambios de moneda).

- Cambios en el grado de agregación (por ejemplo, calcular las ventas mensuales a partir de las diarias).

- Calcular campos derivados (por ejemplo, calcular la edad a partir de la fecha de nacimiento).

- Añadir información temporal (por ejemplo, periodo de validez de los datos).

Una de las transformaciones que generalmente siempre debe hacerse es la última mencionada, añadir a los datos información temporal. Se tendrá que añadir la información sobre el periodo de validez de los datos o el momento en el que se haya registrado la modificación (o en que se haya detectado), según sea requerido por el almacén de datos correspondiente. De este modo, secuenciamos las imágenes obtenidas para ir formando la película que almacena el almacén de datos.

5.2.2. Depuración de los datos

El objetivo de depurar los datos obtenidos de las diferentes fuentes es mejorar su calidad. Algunas de las incidencias más comunes que se producen son las siguientes:

- Detectar y corregir valores inconsistentes (por ejemplo, un atributo edad con un valor de trescientos cincuenta).

- Añadir valores por defecto a los campos con valores no definidos. Generalmente, se hace de acuerdo con criterios marcados por el almacén de datos al que se destinan estos según la fuente de datos. El valor suministrado puede ser constante, calculado o, en algunos casos, puede interesar dejarlo sin definir.

- Detectar y corregir información duplicada. A veces es difícil de detectar, puesto que se tienen distintas representaciones del mismo valor (por ejemplo, diferentes maneras de escribir el nombre de una calle en los datos del domicilio). Será más frecuente encontrar información duplicada entre distintas fuentes de datos, pero también la podemos encontrar dentro de una misma fuente.

5.2.3. Integración de los datos

Teniendo en cuenta los datos obtenidos de diferentes fuentes, los tenemos que integrar entre sí, y también con los datos del almacén de datos al que se destinan.

El proceso de integración será diferente dependiendo de si hacemos la carga inicial del almacén de datos, o bien una actualización de este.

Además del volumen de datos que hay que tratar, la diferencia principal reside en el hecho de que en las actualizaciones, para hacer la integración, podemos usar las correspondencias entre los datos de las fuentes y los del almacén de datos previamente establecidos en la carga inicial o en actualizaciones anteriores. Generalmente, en el proceso de carga inicial se hará una integración de todos los datos previa a la carga en el almacén de datos. Por otro lado, cuando se hace la actualización, es posible que no estén disponibles los datos de todas las fuentes al mismo tiempo e interese integrar los datos de las distintas fuentes por separado en el almacén de datos.

El problema principal con el que nos encontramos consiste en detectar qué datos representan el mismo concepto.

Si las diferentes fuentes de datos utilizan como clave el mismo campo de la entidad (por ejemplo, NIF), se pueden relacionar sin dificultad, excepto por errores en los datos. El problema surge cuando cada fuente de datos emplea su clave (por ejemplo, un código generado) y no hay campos comunes que puedan servir como clave alternativa para establecer relaciones entre sí o, si los hay, sus valores se representan de manera diferente entre las fuentes.

Durante el proceso de integración, se transformarán los datos para homogeneizar su representación y se eliminará la información duplicada.

Se tendrán que establecer los procedimientos adecuados para propagar las correcciones hechas hasta los sistemas operacionales de los que proceden los datos. Estas serán especialmente relevantes después de obtener los datos para la carga inicial del almacén de datos, pero también se deberán tener en cuenta las efectuadas en cada una de las actualizaciones de los datos.

Datos depurados

Si hemos dedicado un esfuerzo considerable para integrar los datos de clientes de distintos sistemas operacionales, depurándolos y eliminando duplicados, lo razonable es utilizar los datos depurados en los sistemas operacionales, en lugar de continuar utilizándolos con errores. Por este motivo, se tendrá que definir un sistema para propagar las correcciones hechas en los datos desde el componente de integración y transformación hasta los sistemas operacionales de los que proceden.

5.3. Transporte y carga de los datos

Cuando ya hemos obtenido los datos, se deben transportar desde las diferentes plataformas de las fuentes hasta las plataformas de los almacenes de datos a los que se incorporarán. También es necesario transportarlos entre distintos almacenes de datos (podéis ver la figura anterior).

> El componente de integración y transformación también se encarga de transportar los datos entre las diferentes plataformas y cargarlos en las bases de datos correspondientes.

Tanto en el transporte como en la carga de los datos hay que distinguir entre el proceso de carga inicial, que se ejecutará una sola vez, y el proceso de actualización, ejecutado con frecuencia. El transporte y la carga iniciales se pueden resolver de manera puntual, sin que haya la necesidad de dedicar recursos permanentes con esta finalidad. Solo será necesario tener disponibles de manera permanente los recursos para hacer las actualizaciones de los datos.

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[7]: "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.

(7)*Meta* en griego significa 'sobre'.

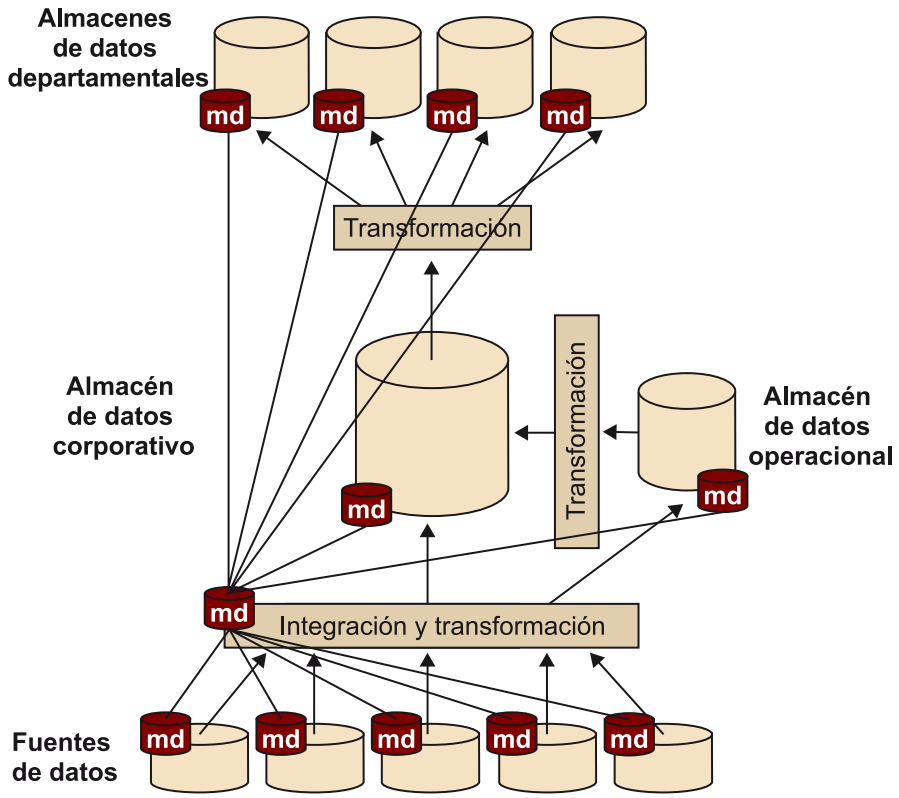
Teniendo en cuenta un conjunto de datos, los metadatos sobre estos describen sus características (por ejemplo, formato, origen, uso, etc.). Estos metadatos son datos y a su vez podemos tener otros metadatos que describan sus características (metadatos sobre metadatos), y así de manera sucesiva.

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. La sección acaba comentando la necesidad de utilizar diferentes versiones de metadatos en la FIC.

6.1. Metadatos y los 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 siguiente), pero son un componente por sí mismos. Inmon los define dentro de la FIC como la "goma de pegar" que mantiene unido el resto de los componentes, y por este motivo los considera como el componente más importante de la FIC.



Los metadatos como componente de la FIC.

6.1.1. Metadatos en 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 (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.

6.1.2. Metadatos en 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.

6.1.3. 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[8].

(8)Es decir, si se ha desarrollado manualmente, los metadatos mencionados formarían parte de su documentación.

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

6.2. Uso y tipos de metadatos

La información ofrecida por los metadatos nos permite entender mejor la estructura, el funcionamiento y los resultados de los sistemas que describen. Es decir, los metadatos resultan interesantes para el equipo de desarrollo del sistema, los técnicos que hacen que el sistema funcione y los usuarios finales que lo utilizan. De este modo, los podemos clasificar según el papel de las personas que los utilizan. Estos conjuntos de metadatos no son disjuntos, es decir, se utilizarán los mismos metadatos con objetivos diferentes.

6.2.1. Metadatos de construcción

Los equipos de desarrollo de los sistemas definen gran parte de los metadatos; posteriormente estos se usarán con otros objetivos diferentes en la construcción de los sistemas. En el caso de la FIC, definen la estructura de las distintas fuentes de datos, de los almacenes de datos, las transformaciones que hay que hacer, la planificación, etc.

Los metadatos de construcción tienen gran importancia, puesto que hacen que los sistemas sobre los que se definen sean más flexibles y fáciles de evolucionar.

Los metadatos son tan importantes en este aspecto de los sistemas que a veces este es el único uso que se les reconoce.

6.2.2. Metadatos de gestión

Durante el funcionamiento del sistema, para gestionarlo se utilizan algunos de los metadatos definidos durante la construcción y también se definen otros nuevos. Todos estos forman los metadatos de gestión. En la FIC se define a los usuarios que utilizarán los diferentes almacenes de datos, se almacena información sobre el uso que hacen de los mismos, sobre el resultado de las extracciones y las transformaciones de datos hechas, etc.

Los metadatos de gestión son utilizados por los técnicos que administran el sistema y hacen que este funcione.

6.2.3. Metadatos de uso

Los analistas generalmente no definirán metadatos (tampoco datos), al menos directamente, sino que se limitan a hacer consultas sobre estos. Además de consultar datos, también necesitan hacer consultas sobre los metadatos tanto de construcción como de gestión. No tendrán acceso a todos los metadatos definidos, sino solo a aquellos que los constructores del sistema hayan considerado de su interés según su perfil de usuario.

Ejemplo de metadatos de uso

Un analista necesita consultar los resultados de las ventas de una cadena de tiendas. Las ventas registradas en los sistemas operacionales de las tiendas se cargan cada día en el almacén de datos utilizado. El analista puede consultar los resultados propiamente dichos y el significado de un dato concreto (la fórmula utilizada para calcularlo: metadatos de construcción), y también puede hacer consultas sobre incidencias particulares que han tenido lugar para obtenerlos (si faltan los datos de alguna tienda: metadatos de gestión).

Generalmente, los usuarios de los sistemas operacionales solo necesitan trabajar con los datos de negocio almacenados en los sistemas. Los usuarios de los almacenes de datos, además de datos, necesitan metadatos.

Los metadatos, tanto de construcción como de gestión, tienen gran importancia para los usuarios de los almacenes de datos, ya que les suministran la información que necesitan sobre el significado o el estado de los datos que consultan.

6.3. El proceso de definición de los metadatos

Las ventajas de disponer de metadatos en cualquier sistema son indudables, puesto que proporcionan de manera explícita información que facilita la evolución, la gestión y el uso del sistema. Podemos definir los metadatos de ma-

nera manual o bien con el apoyo de alguna herramienta; asimismo, se pueden definir antes, durante o después de la construcción del sistema al que están asociados.

La situación ideal es que dispongamos de una herramienta para construir el sistema y que la definición de los metadatos asociados forme parte del proceso de construcción y mantenimiento, de modo que el sistema y los metadatos asociados evolucionen de manera conjunta.

Si no forman parte necesaria del proceso de desarrollo del sistema, se corre el riesgo de que, por limitaciones en el tiempo de desarrollo, en el presupuesto o por otros motivos, los metadatos no se actualicen con el sistema al que están asociados y se produzca, de este modo, una discordancia entre los metadatos y el sistema que describen.

Metadatos asociados al sistema

De este modo, como se presenta en Inmon, Imhoff y Sousa (1998), se consigue que los metadatos sean completos, no sean un elemento opcional, se actualicen de manera automática cada vez que se hagan modificaciones en el sistema y no requieran un esfuerzo adicional para mantenerlos.

6.4. Estándares de metadatos

En sistemas complejos (como el caso de la FIC), cada componente dispone de metadatos. Para definirlos, se han podido utilizar diferentes herramientas de apoyo: herramientas CASE, herramientas del SGBD, herramientas del componente de integración y transformación, etc. Por lo tanto, cada componente tiene sus metadatos, almacenados según su criterio y formato particular.

Para compartir los metadatos, los distintos componentes deben "hablar" el mismo idioma en este aspecto. Un estándar de definición de metadatos representa este idioma común.

A lo largo de la historia (es una historia relativamente corta, pues se inicia a principios de los noventa) se han definido diferentes estándares relacionados con metadatos. De manera particular, relacionados con la FIC, encontramos principalmente el *common warehouse metadata* (*CWM*). Su objetivo es definir un repositorio central que permita **integrar** los metadatos que hay definidos para las diferentes herramientas, de modo que mantenga una sola versión de todos estos. Para esto, el CWM define un modelo de datos para el almacenamiento formado por submodelos específicos para cada área, y un conjunto de capas de acceso al repositorio que ofrecen distintos grados de funcionalidad.

Lectura complementaria

Podéis encontrar más detalles sobre los estándares en un anexo dedicado a este tema en **W. A. Giovinazzo** (2000). *Object Oriented Data Warehouse Design*. Nueva Jersey: Prentice Hall PTR.

6.5. Metadatos históricos

Una de las características de los almacenes de datos es que almacenan datos históricos, como hemos visto en el apartado "Características de un almacén de datos" del módulo "Introducción al almacenamiento de datos". A lo largo del tiempo, las estructuras y otras características de los componentes de la FIC, los datos de las fuentes de datos y de los almacenes de datos, las correspon-

dencias entre estos y las transformaciones que se hacen han podido cambiar. Junto a estos componentes, también habrán cambiado los metadatos que los describen.

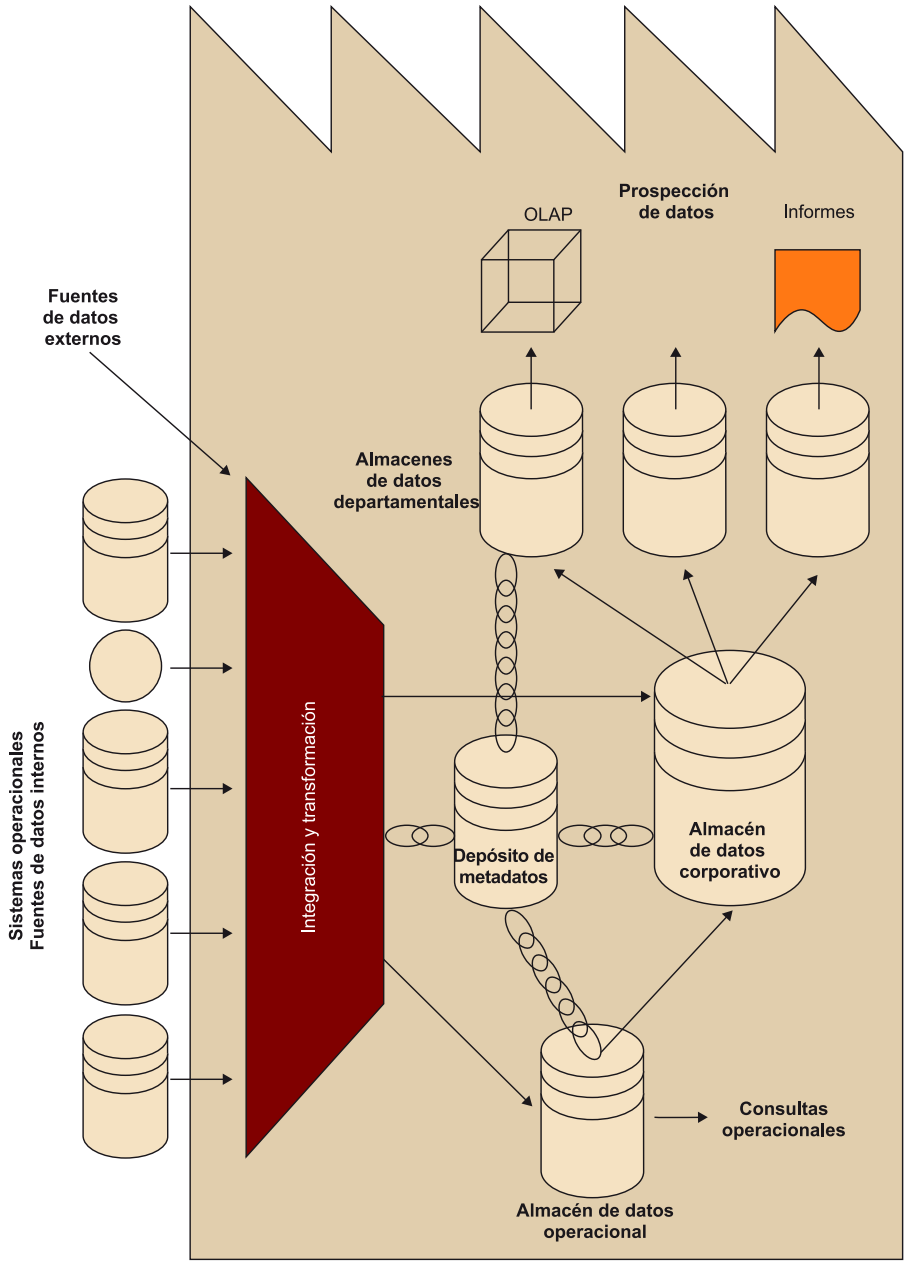
Por lo tanto, si en los almacenes de datos tenemos datos históricos con características diferentes, para cada conjunto de datos definidos bajo las mismas características tendremos que almacenar la versión de los metadatos que los definen. De este modo, los analistas podrán saber para cada dato cómo está almacenado o en qué condiciones se obtuvo.

Se necesita mantener un control de versiones de los metadatos de la FIC.

7. La factoría de información corporativa

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 siguiente esquematiza todos los componentes de la factoría de información.



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.

Probablemente, debido a la juventud del área, actualmente se produce una cierta confusión en los términos. Las empresas, por desconocimiento o simplemente por dar más importancia a su proyecto, suelen denominar "*data warehouse*" no al almacén de datos corporativo, sino a lo que realmente solo es un almacén de datos departamental (es decir, un *data mart*).

Otro abuso 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.

Lectura recomendada

Podéis ver esta definición del almacenamiento de datos en **R. S. Gardner** (1998, septiembre). "*Building the Data Warehouse*". *Communication of the ACM* (núm. 41, vol. 9, págs. 52-60).

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:

- 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.

- 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.

- 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.

- 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.

Resumen

En este módulo, hemos estudiado los diferentes componentes que constituyen la factoría de información corporativa. Hemos empezado por los dos extremos de la cadena (usuarios y fuentes de información) y hemos seguido con el resto de los componentes intermedios:

- Almacén de datos departamental
- Almacén de datos corporativo
- Almacén de datos operacional
- El componente de integración
- Los metadatos

Todos estos componentes se han estudiado haciendo referencia a los sistemas operacionales. Por lo tanto, podemos decir que en este módulo hemos completado el estudio comparativo entre almacenes de datos y bases de datos operacionales llevado a cabo en el módulo "Introducción al almacenamiento de datos".

Finalmente, hemos engranado todos estos componentes para constituir la arquitectura de la FIC. Hemos tenido en cuenta que los componentes que configuran la FIC son complementarios e interactúan entre sí para satisfacer las necesidades de los analistas.

Ejercicios de autoevaluación

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

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

3. ¿En qué se diferencian las herramientas OLAP de las de prospección de datos, en relación con sus necesidades de datos?

4. ¿Qué quiere decir que el almacén de datos corporativo no se puede diseñar teniendo en cuenta su funcionalidad?

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 hace el componente de integración y transformación?

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

8. ¿Qué es el sistema de registro?

9. ¿Qué dos fases podemos distinguir en la operativa del componente de integración y transformación?

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

11. ¿Qué tipos de metadatos encontramos en la FIC?

12. ¿Por qué son importantes los estándares de metadatos?

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

14. 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		
Datos, características de		
Uso		
Acceso		
Unidad de trabajo		
Requerimientos		
Tamaño		

Solucionario

1. 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 cierto conjunto de analistas.

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

3. Las herramientas OLAP corresponden a lo que hemos denominado granjeros, es decir, accesos regulares a pequeñas cantidades de datos normalmente resumidos para mostrarlos a los usuarios. Por el contrario, las herramientas de prospección de datos corresponden a lo que hemos denominado exploradores, accesos esporádicos a grandes cantidades de datos tan detallados como sea posible para hacer estudios estadísticos.

4. El ciclo de desarrollo de los sistemas operacionales empieza con la definición de los requerimientos o la funcionalidad que tienen que dar. En cambio, el almacén de datos corporativo se construye sin saber del todo cuál será la necesidad concreta que satisfará. Por lo tanto, se diseña según los temas interesantes que haya definidos.

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 aquí. 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. Es la fuente más adecuada de entre todas las fuentes posibles para los datos que se almacenan en el almacén de datos corporativo.

9. En una primera fase, se obtienen los datos que forman la imagen inicial de los datos. En una segunda fase, de manera iterativa se obtienen las actualizaciones que se han hecho sobre los datos para ir formando la "película" (secuencia de imágenes) que nos muestra la evolución de los datos.

10. 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.

11. Encontramos tres tipos de metadatos, según los usuarios que los generan o utilizan: metadatos de construcción (utilizados por desarrolladores), de gestión (utilizados por los técnicos que administran y gestionan los sistemas) y de uso (utilizados por los analistas). Un metadato puede ser de los tres tipos si es usado por los tres tipos de usuario.

12. Porque cada herramienta que utilizamos en la construcción de la FIC definirá sus metadatos utilizando un formato determinado. Los estándares permitirán a las distintas herramientas intercambiar y compartir metadatos.

13. 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.

14. 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

Característica	Sistemas operacionales	Sistemas decisionales
Objetivo del sistema	Ejecución del negocio	Análisis del negocio
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 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 cierto 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.

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**.

información (definición desde el punto de vista de los sistemas decisionales) *f* Datos relevantes para alguien que decide y que afectan a alguna de sus decisiones.

metadato *m* Datos sobre datos.

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

OLTP *On-line transaccional processing*.

SGBD *m* 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**
en database management system

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

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

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

Bibliografía

Asociación de Técnicos en Informática (1999, marzo-abril). *Novática* (núm. 138).

Giovinazzo, W. A. (2000). *Object Oriented Data Warehouse Design*. Nueva Jersey: Prentice Hall PTR.

Inmon, W. H.; Imhoff, C.; Sousa, R. (1998). *Corporate Information Factory*. EE. UU.: John Wiley & Sons, Inc.

Jarque, M.; Lenzerini, M.; Vassiliou, Y.; Vassiliadis, P. (2000). *Fundamentals of Data Warehouses*. Berlín: Springer-Verlag.