Administración de los datos

Jordi Serra Ruiz
Miquel Colobran Huguet
Josep Maria Arqués Soldevila
Eduard Marco Galindo

P08/81029/02269

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

Actualmente, uno de los grandes valores de todas las organizaciones es la información (también podríamos decir datos, aunque ya veremos qué diferencia hay). Ésta es la materia prima del sistema informático. Por lo tanto, es lo que se tiene que conocer mejor en cuanto a dónde está, porque como todos los ordenadores están conectados por las redes informáticas, corremos el peligro de que los datos estén dispersos por toda la organización en las estaciones de trabajo. Si eso pasa, no sabremos ni qué datos tenemos ni dónde los tenemos que ir a buscar.

Necesitaremos saber dónde está la información para hacer copias de seguridad, ya que en caso de que haya un desastre el software se puede reinstalar, pero los datos los ha creado la organización, no se pueden "comprar" en ningún sitio. Al ser uno de los activos más importantes actualmente, hay que protegerla.

También es importante saber dónde están estos datos porque los podemos combinar para obtener datos nuevos sobre nuestra organización.

Para la dirección (y en general para toda la organización), disponer de la información adecuada a tiempo permite tomar decisiones correctas en cada situación en el momento oportuno.

Objetivos

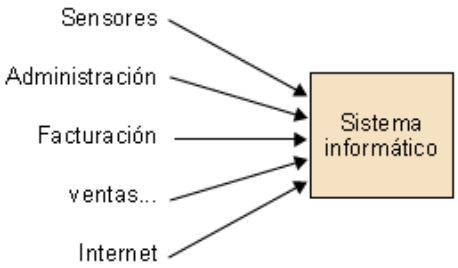
Los materiales de este módulo contienen las herramientas necesarias para que el estudiante alcance los objetivos siguientes:

1. Saber que la información de la organización es muy valiosa.
2. Comprender que es necesario saber dónde está toda la información de la organización.
3. Saber diferentes maneras de mantener la integridad de la información.
4. Saber por qué se han hecho populares los servidores de bases de dados y otras arquitecturas de datos.
5. Saber en qué lugares puede haber datos de la organización y cuáles son los mejores sitios donde podrían estar.
6. Conocer los fundamentos de seguridad de las bases de datos.

1. Los datos y la organización

Una organización crea datos continuamente, por lo que una manera de considerar el sistema informático es como si sólo se tratara de un almacén de datos.

Sistema informático como almacén de datos



Este "gran depósito" de datos en bruto no es útil de esta manera, porque si los datos no tienen una organización y una coherencia, con la gran cantidad que hay, no tendría ningún sentido.

Antes de continuar, cabe decir que hay una diferencia entre información y datos. La definición formal de cada uno de estos conceptos es la siguiente.

Los **datos** son los registros de los sucesos. La **información** es el procesamiento de los datos para que tengan sentido.

Información y datos

A pesar de la diferencia formal entre datos e información, a menudo los informáticos utilizamos estas dos palabras como sinónimas. Como el procesamiento para dar sentido a los datos se puede hacer en muchos lugares, nosotros también utilizaremos indiferentemente los términos *datos* e *información*.

Ejemplo sobre la diferencia entre dato e información

70293 puede querer decir 7 de febrero de 1993 o embalaje 7029 para el camión número 3. Otro ejemplo: 1813 puede ser un número de matrícula de coche o puede ser una hora, las 18 horas y 13 minutos.

Generalmente, se guardan datos y se presenta al usuario información, es decir, que el procesamiento de esta información para que tenga sentido se hace en el software en el momento de presentarla en el dispositivo de salida. En principio eso es lo que da más flexibilidad al sistema, ya que permite convertir los datos a cualquier formato de salida. Sin embargo, a menudo eso queda lejos de la realidad, porque los diferentes tipos de software guardan los datos con diferentes formatos. Además, la mayoría de software es incompatible entre sí, por lo que la recuperación cruzada de información no es nada sencilla.

Así pues, las organizaciones crean constantemente grandes cantidades de datos, y lo que hacen los sistemas informáticos es procesarlos y distribuirlos entre todos los elementos de la organización para aumentar la eficacia del conjunto.

El sistema informático pretende guardar la información de la organización para que sea fácil recuperarla posteriormente y de la misma manera en que se había guardado.

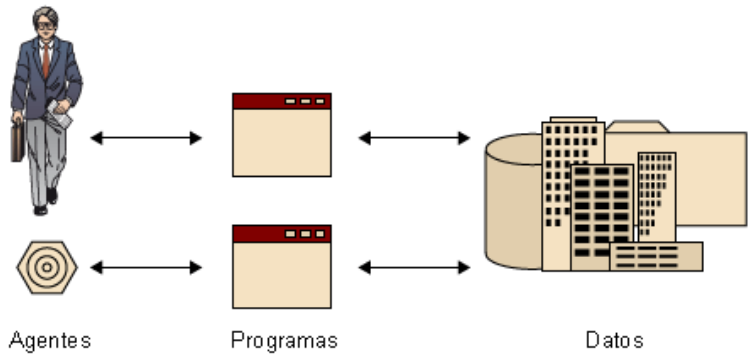
El primer objetivo del sistema informático es guardar la información, pero haciendo posible la recuperación igual que se había guardado. Ahora bien, con eso desperdiciamos el gran potencial de los ordenadores: la capacidad de manipular información.

En este contexto, "manipular información" tiene un sentido bastante amplio. Estos son algunos de los significados posibles:

- Poder añadir información nueva a la que ya teníamos.
- Poder borrar información obsoleta de la que ya teníamos.
- Poder modificar información incorrecta de la que ya teníamos.
- Poder relacionar diferentes informaciones para construir información nueva, contrastar la coherencia o la veracidad de la información que se quiere entrar, reducir la redundancia de la información o buscar nuevas relaciones.

La interacción de los datos con la organización la podemos ver de esta manera:

Esquema de interacción de datos con la organización



Los **agentes** pueden ser, por ejemplo, personas o sensores que a partir de programas acceden a los datos de la organización. También podríamos pensar en estos agentes como entidades que están fuera de la organización, en Internet, pero que interaccionan con el sistema informático y modifican de alguna manera los datos del sistema.

Los **programas** se utilizan para acceder a los datos de la organización de una manera ordenada y correcta. Parece lógico que el acceso a los datos, dado que éstos son un bien valioso de la organización, se tiene que proteger. Los programas se encargan (junto con el sistema operativo) de hacer todos los controles necesarios para evitar los accesos no deseados a los datos de la organización.

Con esta visión de los datos de la organización parece sencillo, viable y quizás incluso fácil resolver el problema de un usuario que nos diga que necesita extraer determinada información del sistema informático. Por lo que hemos visto aquí hasta ahora este problema se reduciría a crear un nuevo programa que accediera a los datos. De momento, por ejemplo, no hemos tenido en cuenta cómo se manejan los permisos y los grupos que incorporan los diferentes ficheros que integran la información de la organización para proteger el contenido, ya que la información se tiene que proteger de accesos indiscriminados.

El sistema informático recopila información, la guarda, la manipula y la presenta al usuario.

2. Dónde está la información

Desgraciadamente, la realidad no es tan simple. En el apartado anterior todo era ideal: guardábamos los datos y, cuando los necesitábamos, los cogíamos y los presentábamos (imprimíamos). Si era lo que nos hacía falta, ya habíamos acabado. La realidad incluye muchos más procesos que hacen que el sistema sea más complejo. Estos son algunos de los problemas:

- Los datos están guardados en ficheros.
- Los datos están guardados en un formato determinado.
- Los datos no siempre están guardados donde se necesitan.
- Las fuentes de datos, es decir, los sensores, las personas, los formularios web, etc. que generan datos, los generan en un formato determinado.
- Cada fuente de datos puede utilizar un formato diferente, y se puede guardar en ficheros potencialmente diferentes y en ubicaciones diferentes.

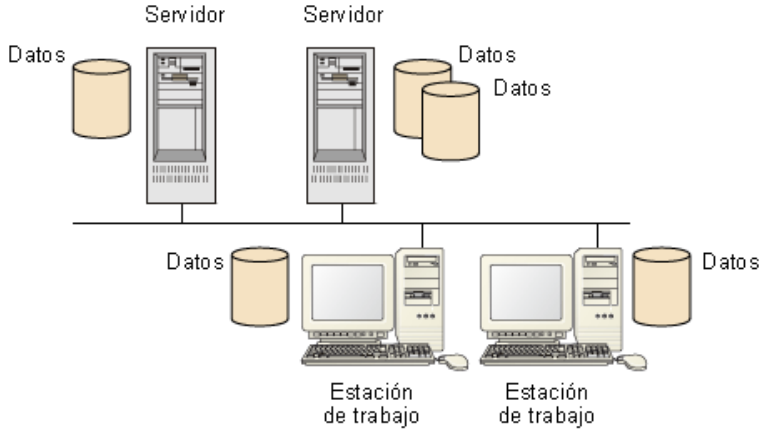
¿Información relacionada?

Formularios web diferentes pueden generar ficheros de texto diferentes, y un solo usuario, según la aplicación que utilice, creará información diferente, en formatos diferentes, que irá a parar a ficheros diferentes. Si, por ejemplo, el usuario hace un documento con procesadores de textos del balance anual contable para el banco, y lo hace en diferentes ficheros según las cuentas, creará información relacionada (del mismo tema) que será difícil de relacionar una con otra.

Lo que nos pasa en la práctica es que los datos de la organización estarán dispersos desde el punto de vista del software. No todos los programas podrán acceder a todos los datos, independientemente de los permisos del usuario.

El problema es todavía más grave si nos fijamos en la estructura de la red. Veamos los lugares donde puede haber potencialmente información. Los datos pueden estar en estos lugares porque se ha decidido así, o porque los usuarios, por desconocimiento, generan información y la guardan en estos lugares, sin consultarlo.

Esquema de una red



Podemos tener información en cualquier directorio de cualquier partición de cualquier disco de los servidores o de las estaciones de trabajo. Incluso en ordenadores portátiles.

Es decir, yendo al extremo, podemos tener información duplicada en máquinas diferentes, con sistemas operativos diferentes, que tengan sistemas de ficheros diferentes.

Lo peor es que parte de esta información puede ser crítica y mucha puede ser desconocida por el departamento de informática, o quizás sabe que existe y no la tiene controlada y, por lo tanto, no dispone de ningún mecanismo de recuperación ante un desastre.

Tenemos, pues, un problema que puede llegar a ser muy grave y que se ha repetido históricamente en todas las organizaciones. Veamos posibles soluciones:

- A partir de la tabla de aplicaciones y del diseño de los usuarios en general, sabemos qué software utilizan los usuarios. También sabemos, a partir de la tabla de aplicaciones, dónde está la información. El mantenimiento de la que esté en los servidores es responsabilidad del administrador de usuarios o del administrador de servidores. En cualquier caso, sabemos dónde está y disponemos de los mecanismos de seguridad, protección y recuperación adecuados en caso de emergencia.

Ved también

Recordad que hemos tratado cómo se hace el mantenimiento de la información de los usuarios en el módulo "Administración de usuarios".

- A partir de la tabla de aplicaciones y del diseño de los usuarios en general, sabemos cuál es el software que genera información en local. De todo este software se tiene que implementar el mecanismo de copias de seguridad que haga falta, y habilitar los permisos y la seguridad que sean necesarios para garantizar la confidencialidad y poder incorporarlas en el mecanismo general de copia para estar prevenidos en caso de fallo.

- Dentro de la categoría de información "no controlable", suele estar el software de ofimática (hojas de cálculo, procesadores de textos, pequeñas bases de datos, agendas, etc.). Se tiene que poner especial atención en el hecho de que los datos estén en la unidad de red privada del usuario (de la cual se hace una copia de seguridad cada día).

- Igualmente, se tiene que formar a los usuarios en el uso de las herramientas informáticas con las que trabajan para que aprendan a poner la información en su unidad personal de red en lugar de hacerlo en la unidad local (o grabarlo donde propone la aplicación por defecto, ya que entonces muchas veces ni el mismo usuario sabe dónde están realmente guardados los datos). Esta "cultura informática" puede evitar la dispersión de la información por todos los discos de la organización, que es uno de los grandes problemas de administración.

Informática portátil

Actualmente, con el crecimiento en las organizaciones de la informática portátil, hay un gran problema añadido al de la dispersión, ya que estos equipos necesitan autonomía y funcionan sin estar conectados a la red de la organización. La primera cuestión que se nos plantea es: ¿qué pasa con los datos que necesita para trabajar? Pueden ser sensibles si salen fuera de la organización sin ningún tipo de protección. La segunda cuestión es que estos equipos modifican o añaden datos al sistema informático (están desconectados de la red). ¿Cómo sincronizan esta información con el sistema de la organización? Una última cuestión sería que a veces necesitan funcionar conectados a la red de la organización como si fueran una estación de trabajo más, y a veces tienen que conectarse a la red de la organización desde fuera. ¿Cómo se puede hacer eso?

Se tienen que buscar maneras de saber dónde está toda la información de la organización. Nos servirá, por ejemplo, para garantizar la seguridad.

3. La consulta de la información

De la misma manera que la organización genera una gran cantidad de datos constantemente, los usuarios también necesitan nueva información de la organización muy a menudo. Eso puede querer decir nuevas consultas a los datos o generar nuevas fuentes de datos en el sistema. Por lo tanto, el administrador tiene que conocer todas las bases de datos de la organización.

El término *base de datos*

En este caso, utilizamos el término *base de datos* en su sentido más amplio, ya que engloba desde una hoja de cálculo hasta una base de datos completa, pasando incluso por un documento de un procesador de textos. Todo es información de la organización.

Ya que la seguridad de toda la información de la organización es responsabilidad del administrador, ha de:

- Conocer la localización.
- Disponer de un método de restauración en caso de problemas. Eso implica copias de seguridad, etc.
- Tener una idea general del contenido. Tiene que tener información de la información (podríamos decir metainformación).

La localización es importante porque tiene una gran relevancia en las políticas de copias de seguridad y en los tiempos de acceso para los usuarios, como también la facilidad de acceso, de ampliaciones futuras, de ampliación del grupo de usuarios que acceden a los datos, etc.

La idea general del contenido tiene utilidad para la optimización del sistema (no se tienen que repetir bases de datos ni datos que ya forman parte de él) y para las peticiones de la dirección, que generalmente van encaminadas a extraer información del sistema.

Minería de datos

Los métodos de minería de datos pretenden extraer información coherente a partir de cantidades de información no coherente y probablemente dispersa. Hay una asignatura que trata este "problema".

Cuando conseguimos tener todos los datos en los servidores (algunas particiones, de algunos discos, de algunos servidores) el problema no está resuelto, pero al menos hemos avanzado mucho. Ya podemos hacer algunas cosas:

- Que diferentes grupos/usuarios accedan a los mismos datos (los compartan). Eso les permite añadir elementos y consultar, modificar y borrar estos datos.

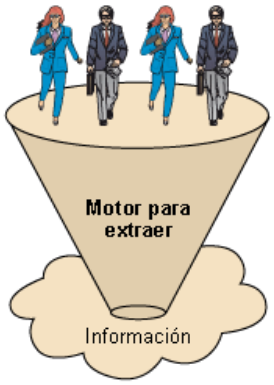
- Como los datos están en los servidores, la seguridad es mucho más alta.
  - Copias de la información.
  - Pérdida por fallo del equipo.
  - Robo.
  - Permisos y, por lo tanto, restricciones de acceso.

- Más facilidad de acceso y de difusión, porque en estas condiciones la información es potencialmente accesible para toda la organización (está controlada por la seguridad).

3.1. Las consultas de la dirección

También parece que nos podemos empezar a plantear un problema del cual hasta ahora no habíamos hablado: el problema de la dirección.

> La dirección ve el sistema informático como un almacén de información y hace consultas a esta información para tener una ayuda en la toma de decisiones.



La dirección necesita extraer información

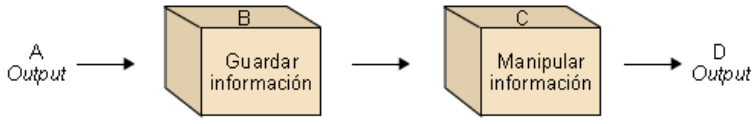
Por lo tanto, para tomar decisiones hacen peticiones del tipo: ¿cuál es el producto que ha dado menos horas de producción y más beneficio de ventas?, o ¿qué producto ha costado menos de vender a los comerciales?, o ¿qué productos tienen menos coste de transporte?, o ¿de qué componentes hemos tenido más pérdidas?, o bien, por ejemplo, ¿qué trabajadores han rendido menos?

Todas son preguntas a menudo bastante difíciles de contestar, aunque el administrador puede llegar a saber que la información está dentro del sistema.

Antes, con los datos distribuidos por toda la organización era imposible contestar a ninguna de estas consultas. Ahora, como los datos están en los servidores son todos accesibles. ¿Cuál es el problema?

Para contestar a esta pregunta, hay que volver un poco atrás y ver cómo se llega al punto actual sobre la información de la organización. El proceso de guardar/recuperar información lo podemos esquematizar de la manera siguiente:

Esquema del flujo de información en una organización



Input: Sensores, personas, aplicaciones, etc.
Output: Listados.

De momento sólo hemos conseguido solucionar el paso B, ya que tenemos toda la información accesible dentro de los servidores. Colateralmente, hemos mejorado el problema de seguridad y de compartimentación de la información, pero lo que no hemos mejorado nada es el punto A: la información continúa entrando con formatos completamente heterogéneos, está en una gran cantidad de ficheros diferentes y, por lo tanto, los programas para manipu-

lar esta información pueden ser extremadamente complejos de hacer (por no decir imposibles). Si manipular información en formatos heterogéneos para obtener información coherente es complejo, y la dirección pide consultas a menudo, el problema se convierte en intratable.

Si finalmente obtenemos algún resultado de algunos de estos ficheros con formatos heterogéneos, este resultado también estará en un formato heterogéneo (la misma información guardada de dos maneras). ¿Qué se tiene que hacer para fabricar la lista final (paso D)?

No basta con tener toda la información en los servidores para poder decir que ya la podemos manipular como queramos.

3.2. Servidores de bases de datos

Ante este conflicto aparece en el mercado un primer intento para solucionarlo: los servidores de bases de datos. La idea es muy simple. Integrar toda la información en un solo lugar. Este lugar tiene que ser rápido, seguro, fácil de utilizar, fácil de hacer crecer, fácil de hacer copias de seguridad y tiene que incorporar permisos.

Los servidores de bases son sistemas gestores de bases de datos (SGBD[1]), aunque actualmente también se conocen como *servidores para bases de datos*, porque funcionan con tecnología cliente/servidor, es decir, que las estaciones de trabajo obtienen la información de las bases de datos haciendo peticiones de información a un servidor a través de la red.

(1)SGBD es la sigla que se corresponde con la expresión inglesa *relational database management system* (RDBMS).

Sistemas gestores de bases de datos

En el mercado hay muchos sistemas gestores de bases de datos (SGBD) comerciales, y algunos gratuitos. En general, todos piden grandes cantidades de disco y de memoria RAM. Si las bases de datos tienen que ser grandes, entonces hace falta una unidad de control de proceso (CPU) rápida para poder procesar las peticiones SQL y las transacciones rápidamente. Si se prevé que el servidor irá cargado con bastante información, seguir las indicaciones del fabricante de la base de datos sobre los requisitos del ordenador en relación con el disco, la RAM y el CPU es muy importante para obtener un buen rendimiento.

Este software incorpora interfaces para poder interrogar o programar las bases de datos desde cualquier lenguaje, y una vez se ha creado la base de datos (base de datos, tablas, estructura, campos, etc.) se puede utilizar desde cualquier aplicación que haya sido preparada para hacerlo. Incluso desde un servidor web con extensiones.

Este conjunto de características ha hecho que muchas organizaciones hayan migrado sus aplicaciones hacia otras orientadas a servidores de bases de datos. De esta manera, obtenemos las ventajas siguientes:

- Toda la información está concentrada en los servidores.

- Toda la información, aunque pueda entrar en formatos heterogéneos, está guardada en formato homogéneo.
- Es posible relacionar información de fuentes diferentes, porque está guardada en un solo sitio y en un mismo formato.
- Permite hacer salidas mucho más fácilmente.
- Permite hacer pequeñas modificaciones de informes mucho más fácilmente.
- Posibilita tener todo el conjunto de datos de la organización mucho más integrado.
- Simplifica mucho las copias de seguridad.
- Facilita las consultas de la dirección.

Hay una solución para tener toda la información en un solo lugar, un servidor de bases de datos.

Ofimática e informática móvil

La ofimática y la informática móvil continúan estando presentes, por lo que no toda la información está dentro del servidor de bases de datos. Por lo tanto, se tiene que educar a los usuarios en el uso de las unidades personales situadas en los servidores como medida para proteger y poder compartir, si hace falta, la información.

3.3. ERP

Actualmente se ha llevado mucho más allá la idea de los servidores de bases de datos. Sobre un servidor de bases de datos se instalan integradas las aplicaciones que sabemos que utilizan todas las organizaciones. Después, se comercializa el producto.

El resultado se llama sistema de planificación de recursos de la empresa (ERP[2]). Es un conjunto de módulos o paquetes (las aplicaciones). Por ejemplo, contabilidad, facturación y nómina, todo perfectamente integrado y funcionando sobre un servidor de bases de datos.

[(2)]ERP es la sigla de *enterprise resource planning*.

Estos sistemas de gestión de la información, al integrar y automatizar muchos aspectos operacionales de la organización, homogeneiza todo el proceso de introducir información e informes. Al integrar toda la información en un mismo lugar, las consultas que se pueden crear son muchas y muy potentes.

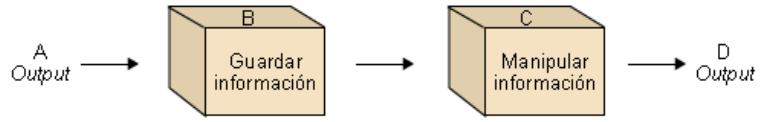
3.4. Almacén de datos

Con los SGBD[3] y los ERP[4] solucionamos el paso B y una parte del paso C del flujo de información en una organización, pero no hemos solucionado el problema de las consultas de la dirección. En especial porque, a menudo, las consultas hacen referencia a intervalos de tiempo que nuestras bases de datos no contemplan. Por ejemplo, las preferencias de nuestros clientes durante los últimos tres veranos ¿hacia qué productos se han decantado?

[(3)]SGBD es la sigla de *sistemas gestores de bases de datos*.

[(4)]ERP es la sigla de *enterprise resource planning*, en castellano, sistema de planificación de recursos de la empresa.

Esquema del flujo de información en una organización



Input: Sensores, personas, aplicaciones, etc.
Output: Listados.

Los almacenes de datos[5] son un paso más en el área del almacenamiento de datos. En lugar de estar orientados a la manipulación de datos (altas, modificaciones, bajas...), conocidos como datos operacionales, están orientados a guardar los datos para hacer búsquedas y consultas.

(5)Almacén de datos en inglés se expresa como *data warehouse*.

Un almacén de datos es una colección de datos orientada a un determinado ámbito, integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la organización en que se utiliza.

Podemos analizar la definición de almacén de datos:

- **Orientado a temas**. Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo acontecimiento u objeto del mundo real queden unidos entre sí.

- **Variable en el tiempo**. Los cambios producidos en los datos, a lo largo del tiempo, quedan registrados para que los informes que se puedan generar reflejen estas variaciones.

- **No volátil**. La información no se modifica ni se elimina. Cuando se ha almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.

- **Integrado**. La base de datos contiene los datos de todos los sistemas operacionales de la organización, y éstos tienen que ser consistentes.

Un almacén de datos, por lo tanto, es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes diferentes, para después procesarla permitiendo su análisis desde muchas perspectivas. Puede estar diseñado sobre un SGBD, pero no es lo mismo.

SGBD	Almacén de datos
Datos operacionales	Datos de la organización para obtener información
Orientado a la aplicación	Orientado al tema

SGBD	Almacén de datos
Actual	Actual e histórico
Detallado	Detallado y resumido
Cambia constantemente	Estable

Algunas de las ventajas de implantar un almacén de datos son:

- Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global de la organización.

- Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén.

- Permite aprender de los datos del pasado y predecir situaciones futuras en diversos escenarios.

- Simplifica, dentro de la organización, la implantación de sistemas de gestión integral.

3.4.1. Arquitectura de un almacén de datos

La arquitectura básica de un almacén de datos y sus elementos es la siguiente:

- **Bases de datos operacionales**. Bases de datos que registran las transacciones necesarias para un correcto funcionamiento de la organización. Estos datos son la fuente principal para el almacén de datos.

- **ETL**. Periódicamente se tienen que importar datos al almacén de datos. Es necesario normalizar los datos antes de introducirlos en el almacén de datos mediante herramientas de extracción, transformación y carga (ETL[6]). Estas herramientas leen los datos primarios (bases de datos OLTP[7] de la organización), realizan el proceso de transformación en el almacén de datos (filtraje, adaptación, cambios de formato...) y escriben en el almacén.

- **Almacén de datos**. El depósito donde se almacenan todos los datos de la organización.

- **Metadata**. Uno de los componentes más importantes de la arquitectura de un almacén de datos son los metadatos. Se trata de datos que describen cuál es la estructura de los datos que se almacenarán y cómo se relacionan. Los metadatos documentan, entre otras cosas, qué tablas hay en una base de datos, qué columnas tiene cada tabla y qué tipo de datos se pueden

(6)ETL son las siglas en inglés de extraer, transformar y cargar (*extract, transform and load*).

(7)OLTP es la sigla en inglés de procesamiento de transacciones en línea (*online transaction processing*). Son los SGBD.

almacenar. Los datos son de interés para el usuario final y los metadata es de interés para los programas que tienen que manejar estos datos.

- **Data Mart**. Un subgrupo lógico del almacén de datos completo. Los Data Mart son subconjuntos de datos con el fin de ayudar a que un área específica de la organización pueda tomar mejores decisiones. Los datos existentes en este contexto pueden ser agrupados y explorados de muchas maneras, por diversos grupos de usuarios según sus necesidades. Se podría decir que los Data Mart son pequeños almacenes de datos centrados en un tema o un área específica dentro de la organización.

Observación

Sobre los Data Mart se pueden construir EIS (*executive information systems*; en castellano, sistemas de información para directivos) y DSS (*decision support systems*, sistemas de ayuda a la toma de decisiones).

- **Aplicaciones de usuario**. Son un conjunto de herramientas que hacen las consultas, analizan y presentan la información. A menudo, se habla de un cliente de almacén de datos. Se pueden encontrar diversos tipos de aplicaciones de usuario.
  - Herramientas de consulta y creación de informes.
  - Herramientas de desarrollo de aplicaciones.
  - Sistemas de información ejecutiva (EIS[8]).
  - Herramientas de minería de datos.
  - Sistemas de apoyo a la toma de decisiones (DSS[9]).

Modelización multidimensional

Para la consulta y el modelado de datos, en el almacén de datos, se utiliza el modelado multidimensional, que es una alternativa a los modelos entidad-relación.

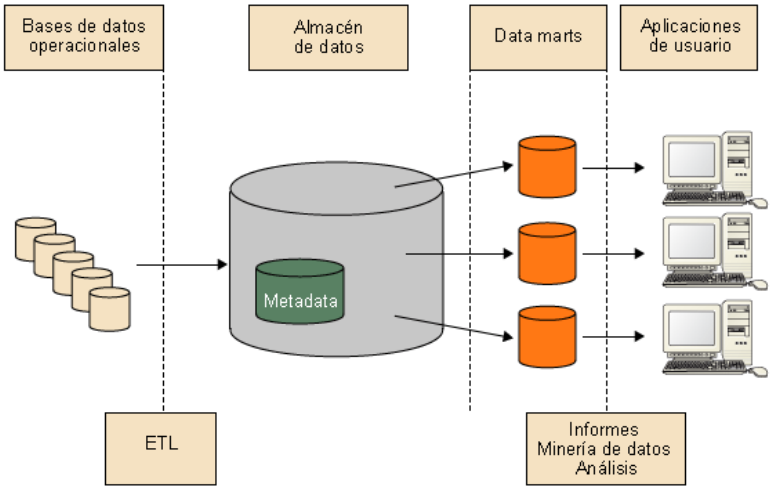
[8]EIS son las siglas de la expresión inglesa *executive information system*.

[9]DSS son las siglas de la expresión inglesa *decision support system*.

Minería de datos

Se conoce como *minería de datos* (en inglés, *data mining*) al proceso de análisis de grandes cantidades de datos con el objetivo de extraer información útil. Por ejemplo, para realizar clasificaciones o predicciones.

Esquema de la arquitectura básica de un almacén de datos



4. Protección de la información

La protección de la información se puede mirar desde dos vertientes. En la primera, sencillamente se trata de protegerla contra accesos no deseados. En la segunda, se trata de protegerla para no perderla.

La información es uno de los bienes más valiosos de la organización. Se tiene que invertir una parte importante de los esfuerzos para protegerla.

4.1. Seguridad de la red

La información está dentro de la red de la organización, en servidores, en estaciones de trabajo, en portátiles, etc. El personal accede a ella, y es con los mecanismos de acceso que se limita y controla quién puede acceder a esta información, de qué manera y qué puede hacer. Un diseño del entorno de usuarios correcto es el primer paso para evitar accesos no deseados a la información.

Ved también

Ved cómo se hace el diseño del entorno de los usuarios en el módulo "Administración de usuarios".

El otro paso para tener sistemas seguros es evitar intrusiones. Es muy difícil tener sistemas completamente seguros; por lo tanto, sólo podemos tratar de hacer el sistema tan seguro como sea posible.

Ved también

Ved maneras de optimizar la seguridad del sistema en el módulo "Administración de la seguridad".

4.2. Copias de seguridad

Las copias de seguridad de los datos de una organización es uno de los asuntos más importantes dentro de la administración de los datos, ya que sin una buena política de recuperación ante desastres la información se puede perder. Básicamente, se tiene que tener presente que los datos se tendrían que copiar a diario de una manera íntegra, y tener un historial de copias de seguridad, que según la organización puede ir desde pocos meses hasta años.

Si se ha seguido una política que refuerza que la información esté en los servidores, las copias son mucho más sencillas, ya que pueden estar perfectamente integradas dentro del proceso normal de copia de los servidores.

Ved también

Ved las políticas de copias de seguridad en el apartado 6 del módulo "Administración de servidores".

La política para reforzar que se guarde la información en el servidor pasa por llevar a cabo acciones como, por ejemplo, las siguientes:

- Que las aplicaciones guarden por defecto en los espacios personales del usuario del servidor. De esta manera, si el usuario no dice explícitamente dónde quiere guardar la información, irá al servidor.

- Formar al usuario en el entorno informático, explicarle las ventajas de guardar la información en la unidad de red como, por ejemplo, el cambio de ordenador por traslado de puesto de trabajo o la pérdida de datos en caso de fallo del ordenador.

- Formar al usuario en las aplicaciones que tiene que utilizar para enseñarle que lo tiene que guardar todo en unidades de red.

- Hacer que todas las aplicaciones corporativas guarden la información en las unidades de red.

4.3. Seguridad en bases de datos

Podemos ver una base de datos como un "conjunto de datos integrados, adecuado a diversos usuarios y a diferentes usos". De manera que los problemas de seguridad vendrán por el uso simultáneo de estos datos.

La protección de los datos se tiene que hacer contra fallos físicos, fallos lógicos y errores humanos (sean o no intencionados), que pueden alterar o corromper los datos, ocasionando que la base de datos se haga inútil para lo que se creó.

Cualquier SGBD tiene que proporcionar técnicas que permitan que los usuarios tengan acceso únicamente a una parte de la base de datos y no al resto. De manera que los SGBD tienen un subsistema de seguridad de autorización encargado de garantizar la seguridad de partes de la base de datos contra el acceso no autorizado.

Las bases de datos llevan mecanismos para prevenir fallos (subsistema de control), para detectarlos cuando se han producido (subsistema de detección) y para corregirlos después de que se han detectado (subsistema de recuperación).

Los aspectos fundamentales de la seguridad en las bases de datos son:

- Confidencialidad. No se tiene que proporcionar datos a usuarios no autorizados. Incluye aspectos de privacidad (protección de datos personales).

- Accesibilidad. La información tiene que estar disponible.

- Integridad. Se tiene que asegurar que los datos no han sido falseados.

4.3.1. Confidencialidad de la información

Para facilitar la administración, los SGBD incorporan el concepto de perfil, rol o grupo de usuarios que reúne un conjunto de privilegios. De manera que el usuario asignado a un grupo hereda todos los privilegios del grupo.

El subsistema de control de acceso se encarga de denegar o conceder el acceso a los usuarios. En un SGBD, puede haber diversos tipos de autorización:

- **Autorización explícita**. Usada en los sistemas tradicionales. Se trata de almacenar quién puede acceder a qué objetos de la base datos con qué privilegios. Se acostumbra a utilizar una matriz de control de accesos.

- **Autorización implícita**. La autorización sobre un objeto se puede deducir a partir de otros. Por ejemplo, si se puede acceder a una clase en un SGBD, se puede acceder a todas las instancias de la clase.

- **Autorización fuerte**. Cuando las autorizaciones deducidas no pueden ser invalidadas.

- **Autorización débil**. Se permiten excepciones sobre las autorizaciones implícitas.

- **Autorización positiva**. Si está, indica la existencia de la autorización.

- **Autorización negativa**. Es la negación explícita de una autorización.

4.3.2. Disponibilidad de la información

Los sistemas de bases de datos tienen que asegurar la disponibilidad de los datos a aquellos usuarios que necesitan el acceso. Así pues, hay mecanismos que permiten recuperar la base datos contra fallos lógicos o físicos que destruyen total o parcialmente los datos.

Utilidades de seguridad

Los SGBD contienen utilidades propias, pero desde un punto de vista de seguridad se tendría que pensar en alternativas, como servidores tolerantes a fallos.

El principio básico que sustenta la recuperación de la base de datos ante cualquier fallo es la redundancia física. Los fallos más habituales son los provocados por fallos eléctricos, fallos del hardware y fallos en los dispositivos de almacenamiento (discos).

Esta situación motivó la aparición del concepto de **transacción**. Ante cualquier fallo se tiene que poder asegurar que, después de una actualización, la base de datos quede en un estado consistente. Para conseguirlo se crean unas unidades de ejecución llamadas transacciones, que se pueden definir como secuencias de operaciones que se tienen que ejecutar de forma atómica. O se realizan todas las operaciones de la transacción globalmente o no se hace ninguna.

4.3.3. Integridad de la información

En este contexto se entiende por integridad la corrección, validez o precisión de los datos de la base de datos. El objetivo es proteger la base de datos contra operaciones que puedan introducir inconsistencias en los datos. El subsistema de integridad de un SGBD tiene que detectar y corregir, en la medida de lo posible, las operaciones incorrectas.

Hay dos tipos de operaciones que pueden violar la integridad de los datos: las operaciones semánticamente inconsistentes y las interferencias debidas a accesos concurrentes.

- **Operaciones contra la integridad semántica**. Existen operaciones que pueden vulnerar restricciones definidas en el diseño de la base de datos (por ejemplo, restricciones sobre los dominios o sobre los atributos). Estas restricciones pueden ser estáticas (también llamadas de estado o situación) o dinámicas (llamadas de transición).

- **Operaciones contra la integridad operacional**. En sistemas multiusuario es imprescindible un mecanismo de control de concurrencia para conservar la integridad de la base de datos. De lo contrario se podrían producir importantes inconsistencias derivadas del acceso concurrente.

5. Tareas/responsabilidades del administrador

Por lo tanto, con todo lo que hemos visto, una relación aproximada de las tareas/responsabilidades del administrador de datos es la siguiente:

- Velar para que los datos estén en los servidores.
- Cuidar de la copia de seguridad de los datos y diseñar, si hace falta, la política de copias de seguridad.
- Velar para que los permisos de todos los datos sean correctos y nadie pueda acceder a más información de la que necesita.
- Asegurar la disponibilidad de los datos a los usuarios.
- Velar por un tiempo de respuesta de las bases de datos correcto.
- Saber dónde están todos los datos/bases de datos de la organización.
- Evitar al máximo la duplicidad de información dentro de la organización.
- Resolver las consultas de la dirección al sistema informático.
- Configurar bases de datos corporativas.
- Establecer los permisos y los accesos a estas bases de datos por parte de los usuarios.
- Velar por el funcionamiento correcto de las bases de datos.
- Velar por una gestión y un almacenamiento correctos de los datos según los protocolos de actuación definidos en la Ley Orgánica de Protección de Datos (LOPD).

Ved también

Ved el módulo "Administración de la seguridad".

Resumen

Los datos son la base del sistema informático y también son de gran valor para la organización. El gran peligro es que pueden estar en muchos sitios. Tenemos que intentar educar a los usuarios para que los concentren en los servidores de una manera natural. Eso da uniformidad y seguridad a los datos.

A pesar de todo, esto no es suficiente para satisfacer una de las grandes demandas de la organización: extraer nuevas conclusiones a partir de la información. La única manera es concentrando la información en una única aplicación, un servidor de bases de datos.

Si optamos por esta solución, tenemos que instalar una interfaz homogénea sobre este servidor de bases de datos, una arquitectura de datos, un sistema de planificación de recursos de la empresa (ERP) o un almacén de datos.

Al haberse convertido en un activo muy valioso, la información tiene que ser protegida; así, son necesarias tanto políticas de copias de seguridad como protecciones de acceso a los datos.

Actividades

1. Observad organizaciones de vuestro entorno (tiendas, empresas, bancos, etc.) e intentad ver las fuentes de datos que puede haber.

2. Fijaos en el entorno. Veréis la misma información con muchos formatos diferentes, por ejemplo, la hora (digital, analógico, etc.). ¿Cómo la guardaríais en una base de datos? Buscad otros ejemplos de datos con multitud de formatos diferentes.

Ejercicios de autoevaluación

1. Enumerad algunas de las fuentes de datos de los lugares siguientes:
a) Un hospital.
b) Una discoteca.

2. Uno de los comerciales de la organización os dice que para hacer los cálculos de venta de los productos, para guardar los productos y los incrementos, no quiere utilizar una base de datos porque nunca ha trabajado así: él negocia con el cliente directamente y, como mucho, puede introducir los datos en una hoja de cálculo en su portátil (y esto como gran favor si lo hace). ¿Qué le diríais?

3. ¿Concentraríais los datos en los servidores?
a) Sí, porque así se pueden compartir más fácilmente.
b) No, es más práctico distribuirlos en los lugares en que se necesitan.
c) No, porque podríamos colapsar los servidores y la red.
d) Sí, porque es más sencillo hacer consultas en la información de la organización.
e) Sí, porque eliminamos completamente la redundancia.
f) a y d.

4. ¿Cuál de estas frases sobre los servidores de bases de datos es falsa?
a) Los servidores de bases de datos concentran la información dispersa en un solo sitio.
b) Los ERP tienen como una de sus bases un servidor de bases de datos.
c) Un servidor de bases de datos con la información de la organización facilita las consultas de la dirección.
d) Con un servidor de bases de datos, el tráfico de la red disminuye porque los formatos de salida son homogéneos.
e) Un servidor de bases de datos homogeneiza la manera de guardar la información.

5. Una de estas tareas no es responsabilidad del administrador de datos.
a) Evitar al máximo la duplicidad de información dentro de la organización.
b) Asegurar la disponibilidad de los datos a los usuarios.
c) Configurar bases de datos corporativas.
d) Conectar las bases de datos con el servidor web.
e) Velar por el funcionamiento correcto de las bases de datos.

Solucionario

Ejercicios de autoevaluación

1. a) Se genera información en muchos sitios. Algunos son:
- Recepción de pacientes.
- Extender recetas.
- Visitas a todas las consultas.
- Todo el departamento de contabilidad.
- Todo el departamento de facturación a pacientes.
- Todo el departamento de pedidos a proveedores.
- Departamento de nóminas.

b) También se genera información en muchos lugares. Algunos son:
- Marketing.
- Contratar espectáculos.
- Contabilidad.
- Facturación.
- Nóminas.
- Pedidos a proveedores.
- Seguridad.

2. Si una persona os expone este problema, hay que convencerlo con argumentos como, por ejemplo, los siguientes:
- La organización trabaja como una unidad y, por lo tanto, todos los precios están en un solo lugar, dentro de una base de datos en un servidor.
- Puede tomar los datos y manipularlos, pero si negocia cambios de precio, los tiene que reflejar en la base de datos para saber:
  - Cuál es el vendedor más competitivo.
  - Cuál es el vendedor que trabaja mejor.
  - Cuál es el vendedor que factura más.

De esta manera, la organización puede asignar incentivos (económicos) o de otro tipo a los comerciales con mejores índices de venta.
Si sólo tiene los datos en local y pierde el ordenador, se lo roban, etc., ni él ni nadie de la organización los podrá recuperar, por lo cual la organización tendrá una pérdida de información y él puede perder los incentivos que hemos dicho.
La información de la organización es propiedad de la organización. Por lo tanto, es la organización quien establece las directrices de funcionamiento, y no las personas que la manipulan.

3. f

4. d

5. d

Glosario

base de datos *f* Término genérico que indica un lugar para guardar datos.

dato *f* Registro de los sucesos.

data warehouse *m* Base de datos que almacena una gran cantidad de datos transaccionales integrados para ser usada por el análisis.

Data Mart *m* Conjunto de hechos y datos organizados para apoyo decisional basados en la necesidad de un área o departamento específico. Los datos están orientados a satisfacer las necesidades particulares de un departamento, teniendo sentido sólo para el personal de este departamento.

data mining *m* Análisis de los datos para descubrir relaciones, patrones o asociaciones desconocidas.

diccionario de datos *m* Un compendio de definiciones y especificaciones para las categorías de datos y sus relaciones.

dimensión *f* Entidad independiente, dentro del modelo multidimensional de una organización, que sirve como clave de búsqueda (actuando como índice) o como mecanismo de selección de datos.

DSS *m* Ved **sistema de apoyo de decisiones**.

enterprise resource planning *m* Ved **planeamiento de recursos de la empresa**.

ERP *m* Ved **planeamiento de recursos de la empresa**.

fuente de datos *f* Cualquier elemento (perteneciente o no a la organización) que crea datos que utiliza el sistema informático.

información *f* Procesamiento de los datos para que tengan sentido.

lenguaje de consultas estructurado *m* Lenguaje que se utiliza para interrogar a los gestores de bases de datos.
en structured query language.
sigla: **SQL**.

planeamiento de recursos de la empresa *m* Sistema de gestión de la información integrado que pretende dar una solución completa al problema de la información dentro de una organización.
en enterprise resource planning.
sigla: **ERP**.

OLAP *m* Sigla de *online analytical processing*. Conjunto de principios que provee un entorno de trabajo dimensional para apoyo decisional.

OLTP *m* Sigla de *online transaction processing*. Sistema transaccional que mantiene los datos operacionales de la organización.

servidor de bases de datos *m* Sistema que guarda datos y al que, por algún mecanismo, generalmente mediante peticiones a través de la red, se le pide información o se le introduce.

SGBD *m* Ved **sistema de gestión de bases de datos**.

sistema de gestión de bases de datos *m* Software que gestiona datos de una manera ordenada para guardarlos y recuperarlos.
sigla: **SGBD**.

sistema de apoyo a la toma de decisiones *m* Sistema de aplicaciones automatizadas de la organización que asiste en la toma de decisiones mediante un análisis estratégico de la información histórica.

SQL *m* Ved **lenguaje de consultas estructurado**.

structured query language *m* Ved **lenguaje de consultas estructurado**.

Bibliografía

Barcelo García, M.; Pastor i Collado, J. (1999). *Gestió d'una organització informàtica*. Barcelona: Universitat Oberta de Catalunya.

Date, C. J. (2000). *An Introduction to Database Systems*. Estados Unidos: Addison Wesley.

Elmasri, R.; Navothe, S. (2000). *Fundamentals of Database Systems*. Estados Unidos: Addison Wesley.

Inmon, W. (2005). *Building de Data Warehouse* (4.ª ed.). Estados Unidos: Wiley.

Kimball, R.; Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2.ª ed.). Estados Unidos: Wiley.

Pfleeger, C. (1997). *Security in Computing*. Estados Unidos: Prentice Hall.

Prague, C.; Irwin, M. (1996). *El libro de Access para Windows 95*. Madrid: Anaya multimedia.