

Cap 6

Fundamentos de la inteligencia de negocios: administración de bases de datos e información

6.1. Organización de datos en un entorno tradicional de archivos

Un sistema efectivo de información da a los usuarios información exacta, oportuna y relevante. La información exacta está libre de errores. La información es oportuna cuando está disponible en el momento que la requieren los encargados de la toma de decisiones. La información es relevante cuando es útil y adecuada para los tipos de trabajo y decisiones que la necesitan.

* Conceptos de organización de archivos

Un sistema de cómputos organiza los datos en una jerarquía que empieza con bits y bytes y avanza a campos, registros archivos y bases de datos. (Ver gráfico pág 236)

Un **Bit** representa la unidad más pequeña de datos que puede manejar una computadora.

Un grupo de bits, llamado un **Byte**, representa un carácter, el cual puede ser una letra, un número u otro símbolo.

Una agrupación de caracteres en una palabra, un grupo de palabras o un número completo se llama **Campo**.

Un grupo de campos relacionados, conforma un **Registro**.

Un grupo de registros del mismo tipo se llama **Archivo**.

Un grupo de archivos relacionados forma una **Base de datos**.

Un registro describe una entidad. Una **entidad** es una persona, lugar cosa o evento sobre el cual se almacena y conserva información.

Cada característica o cualidad que describe a una entidad en particular se llama **Atributo**.

* Problemas con el entorno tradicional de archivos

En la mayoría de las organizaciones, los archivos de datos y los sistemas marcaban una tendencia a crecer de manera independiente sin ajustarse a un plan a nivel de toda la empresa. Cada sector desarrollaba sus propios sistemas y archivos de datos.

Desde luego, para operar, cada aplicación requería sus propios archivos y su propio programa de cómputo. Al considerar a la compañía en su conjunto, este proceso conducía a múltiples archivos maestros creados, conservados y operados por divisiones o departamentos separados. Al cabo de 5 o 10 años, la organización se encuentra con una carga de cientos de programas y aplicaciones muy difíciles de mantener y manejar. Los problemas resultantes son redundancia e inconsistencia de datos, dependencia entre los programas y los datos, inflexibilidad, una escasa seguridad de datos y la incapacidad de compartir datos entre aplicaciones.

Redundancia de datos es la presencia de datos duplicados en múltiples archivos de datos, de tal manera que los mismos datos están almacenados en más de un lugar.

La redundancia de datos desperdicia recursos de almacenamiento y también conduce a la **Inconsistencia de datos**, en el cual el mismo atributo podría tener valores diferentes.

Al utilizar diferentes sistemas de codificación para representar los valores de un atributo se podría generar confusión adicional (ejemplo ropa extra grande vs XL). La confusión resultante dificultaría que las Compañías crearan sistemas de administración de las relaciones con el cliente, de administración de la cadena de suministro o empresariales, que integraran los datos desde fuentes diferentes.

Dependencia entre los programas y los datos: se refiere a la estrecha relación entre los datos almacenados en archivos y los programas específicos que se requieren para actualizar y mantener esos archivos, de tal manera que los cambios en los programas requieren cambios a los datos.

Carencia de flexibilidad: Un sistema tradicional de archivos puede enviar informes programados de rutina, después de extensos esfuerzos de programación, pero no pueden transmitir informes con fines específicos o responder de manera oportuna a requerimientos imprevistos de información.

Seguridad escasa: Dado que hay poco control o administración de datos, el acceso y difusión de la información podrían salirse de control. Es posible que la administración no tenga forma de saber quién esté teniendo acceso a los datos de la organización, o incluso modificándolos.

Carencia de compartición y disponibilidad de los datos: Debido a que la información está fragmentada en diferentes archivos y en distintas partes de la organización no se pueden relacionar entre sí, es prácticamente imposible que la información se comparta o se acceda de manera oportuna. La información no puede fluir libremente a través de las diferentes áreas funcionales o distintas partes de la organización.

6.2. Enfoque de las bases de datos para la administración de datos

Una **Base de Datos** es un conjunto de datos organizados para servir eficientemente a muchas aplicaciones al centralizar los datos y controlar su redundancia. Los datos se guardan físicamente para que se presenten a los usuarios como si estuvieran almacenados en un solo lugar. Una sala base de datos da servicios a múltiples aplicaciones.

Sistemas de administración de Base de Datos (DBMS)

Un sistema de administración de base de datos (DBMS) es el software que permite a una organización centralizar los datos, administrarlos eficientemente y proporcionar, mediante los programas de aplicación, el acceso a los datos almacenados.

El DBMS actúa como una interfaz entre los programas de aplicación y los archivos de datos físicos. Cuando el programa de aplicación solicita un elemento de datos, el DBMS encuentra este elemento en la base de datos y lo presenta al programa de aplicación.

Al separar las vistas lógicas y física de los datos, el DBMS libera al programador o al usuario final de la tarea de comprender dónde y cómo se almacenan realmente los datos.

La **vista lógica** presenta los datos como los deberían percibir los usuarios finales o los especialistas de la empresa.

La **vista física** muestra cómo están organizados y estructurados realmente los datos en un medio de almacenamiento físico.

El software de administración de bases de datos hace disponibles las diferentes vistas lógicas de la base de datos física requerida por los usuarios.

- **Cómo resuelve un DBMS los problemas del entorno de archivo tradicional**

Un DBMS reduce la redundancia y la inconsistencia de datos al minimizar la cantidad de archivos aislados en los cuales se repiten los mismos datos. El DBMS ayuda a la organización a controlar la redundancia de datos.

Un DBMS elimina la inconsistencia de los datos porque puede ayudar a la organización a asegurarse de que todas las ocurrencias de los datos redundantes tengan los mismos valores.

El DBMS elimina la dependencia entre los programas y los datos, permitiendo a estos últimos ser

autosuficientes. El acceso y la disponibilidad de la información se incrementarán y los costos del desarrollo y el mantenimiento de los programas se reducirán porque los usuarios y los programadores pueden realizar consultas con fines específicos de la información que contiene la base de datos.

El DBMS da a la organización la posibilidad de centralizar la administración de datos, su uso y seguridad.

- **DBMS relacional**

Las Bases de datos relacionales representan los datos como tablas bidimensionales (llamadas relaciones). Las tablas podrían considerarse como archivos. Cada tabla contiene datos acerca de una entidad y sus atributos.

Ej (Ver Figura 6.4 pág 231): La base de datos tiene una tabla independiente para la entidad *Proveedor* y otra para la entidad *Parte*. Ambas tablas constan de una cuadrícula de columnas y filas de datos. Los elementos de datos individuales para cada entidad se almacenan como campos independientes, y cada campo representa un atributo para esa entidad. Los campos de una base de datos relacional también se denominan **Columnas**.

Para la entidad *Proveedor*, el número, nombre, calle, del proveedor se almacenan como campos separados dentro de la tabla *Proveedor* y cada campo representa un atributo para la entidad *Proveedor*.

La información real sobre un solo proveedor que se encuentra en una tabla se denomina **Fila**. Por lo general se les llama **Registros**, o en términos muy técnicos **tuplas**. Los datos de la entidad *Parte* tienen su propia tabla separada.

El campo para el *Nº Proveedor* de la tabla *Proveedor* identifica de manera única cada registro con la finalidad de que se pueda recuperar, actualizar o clasificar, y se denomina **Campo clave**. Cada tabla de una base de datos relacional tiene un campo designado como **Clave principal**. Este campo clave es el identificador único para toda la información en cualquier fila de la tabla y su clave principal no se puede duplicar. Cuando la clave principal de una tabla aparece en otra tabla, se denomina **Clave externa** y en esencia es un campo para buscar datos sobre el proveedor de una parte específica.

- **Operaciones de un DBMS relacional**

Las tablas de una base de datos relacional se pueden combinar fácilmente para entregar los datos requeridos por los usuarios, siempre y cuando dos tablas compartan un elemento de datos común.

En una base de datos relacional se utilizan tres operaciones básicas para generar conjuntos de datos útiles: **seleccionar**, **proyectar** y **unir**.

La operación **seleccionar** crea un subconjunto que consta de todos los registros (filas) de la tabla que cumplen los criterios establecidos.

La operación **unir** combina tablas relacionales para dar al usuario más información de la que está disponible en las tablas individuales.

La operación **proyectar** crea un subconjunto que consiste en las columnas de una tabla, lo cual permite al usuario crear nuevas tablas que contengan solamente la información que se requiere.

- **DBMS jerárquico y de red**

Un DBMS **jerárquico** modela relaciones uno a muchos, en tanto que uno de **red** modela relaciones muchos a muchos. Ya no se utilizan porque son mucho menos flexibles que los DBMS relacionales y no soportan consultas con fines específicos, como las consultas de información en lenguaje natural. Estos sistemas se vuelven lentos si se requieren acceder una gran cantidad de veces a los datos almacenados en el disco para ejecutar los comandos seleccionar, unir y proyectar.

Sin embargo, es posible afinar la base de datos para acelerar las consultas predeterminadas.

- ***DBMS orientados a objetos***

Un DBMS orientado a objetos almacena los datos y los procedimientos que realizan operaciones sobre estos datos como objetos que se pueden recuperar y compartir de manera automática.

Los sistemas de administración de bases de datos orientadas a objetos (OODBMS) se pueden emplear para manejar los diversos componentes multimedia o subprogramas de Java que se utilizan en las aplicaciones para la Web, las cuales comúnmente integran la información fragmentada de diversas fuentes.

Aunque las bases de datos orientados a objetos pueden almacenar tipos de información más complejos que los DBMS relacionales, son relativamente lentas comparadas con los DBMS relacionales para procesar grandes cantidades de transacciones.

Existen sistemas híbridos ***DBMS relacional-orientado a objetos***, para proporcionar capacidades de DBMS orientados a objetos y relacional.

Capacidades de los sistemas de administración de bases de datos

Un DBMS incluye capacidades y herramientas para organizar, manejar y acceder a la información de una base de datos. Las más importantes son su capacidad de definición de datos, su diccionario y su lenguaje de manipulación de datos.

Un DBMS tiene una capacidad de ***definición de datos*** para especificar la estructura del contenido de la base de datos. Esta se podría utilizar para crear tablas de bases de datos y para definir las características de los campos de cada tabla. Esta información acerca de la base de datos podría documentarse en un ***diccionario de datos***, que es un archivo automatizado o manual que almacena definiciones de elementos de datos y sus características.

Los diccionarios de datos para grandes bases de datos corporativas podrían captar información adicional, como la propiedad (quien es el responsable de dar mantenimiento a los datos en la organización), autorización, seguridad, y los individuos, funciones de negocios, programas e informes que utiliza cada elemento de datos.

- ***Elaboración de consultas e informes***

Un DBMS incluye herramientas para acceder y manipular la información de una base de datos. La mayoría de los DBMS tiene un lenguaje especializado llamado ***lenguaje de manipulación de datos***, el cual se utiliza para agregar, cambiar, eliminar y recuperar la información de la base de datos. Este lenguaje contiene comandos que permiten que los usuarios finales y los especialistas en programación extraigan los datos de la base de datos para satisfacer requerimientos de información y desarrollar aplicaciones. El lenguaje actual de manipulación de datos más sobresaliente es el ***Lenguaje de Consulta Estructurado o SQL***.

Los usuarios de un DBMS para computadoras grandes o de rango medio, podrían utilizar SQL para recuperar la información que necesitaran de la base de datos.

Los DBMS incluyen características para generar informes que permitan desplegar los datos de interés en un formato más estructurado y refinado del que sería posible con la sola consulta.

Diseño de Bases de Datos

Para crear una base de datos se debe entender las relaciones entre los datos, el tipo de datos que se conservarán en la base de datos, cómo se utilizarán y cómo tendrá que cambiar la organización para

manejar los datos desde una perspectiva de la empresa como un todo. La base de datos requiere de un **diseño conceptual o lógico** tanto como de un **diseño físico**.

El **diseño conceptual o lógico**, de una base de datos es un modelo abstracto de la base de datos desde una perspectiva de negocios, en tanto que, **el diseño físico** muestra la manera real en que la base de datos está organizada en dispositivos de almacenamiento de acceso directo.

- **Normalización y diagramas entidad-relación**

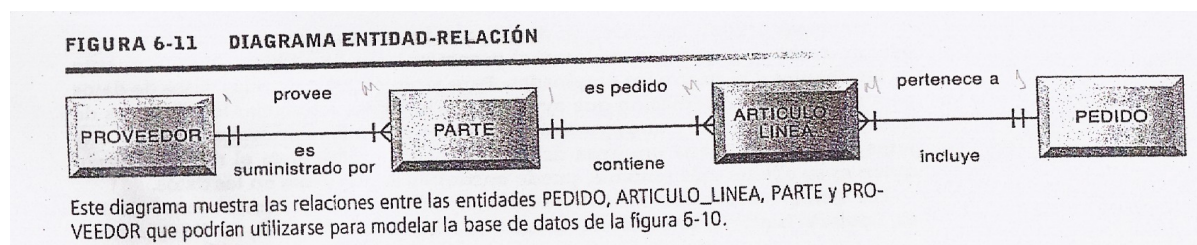
El diseño conceptual de la base de datos describe la manera en que se agruparán los elementos en la base de datos. El proceso de diseño identifica las relaciones entre los elementos de datos y la manera más eficiente de agruparlos para que cumplan los requerimientos de información de la empresa.

El proceso también identifica los elementos de datos redundantes y las agrupaciones de elementos de datos requeridos por programas de aplicación específicos.

Los grupos de datos se organizan, refinan y simplifican hasta que surge una vista lógica general de las relaciones entre todos los elementos de datos en la base de datos.

Para utilizar de manera efectiva un modelo de base de datos relacional, es necesario simplificar las agrupaciones complejas de datos para minimizar elementos de datos redundantes y complicadas relaciones de muchos a muchos. El proceso para crear estructuras de datos pequeñas, estables, aunque flexibles y adaptables a partir de grupos complejos de datos se denomina **normalización**.

Los diseñadores de bases de datos documentan sus modelos de datos con un **diagrama entidad-relación**. Los cuadros representan entidades. Las líneas que conectan los cuadros representan relaciones. Una línea que conecta dos entidades y termina con dos marcas cortas designa una relación uno a uno. Una línea que conecta dos entidades y termina con una pata de cuervo y una marca corta indica una relación uno a muchos.



Si la empresa no implementa su modelo de datos correcto, el sistema será incapaz de servir bien a la empresa.

- **Distribución de Bases de Datos**

El diseño de bases de datos también abarca la manera en que se distribuyen los datos. Se pueden diseñar sistemas de información con una base de datos **centralizada** que sea utilizada por un procesador central único o por múltiples procesadores en una red cliente/servidor.

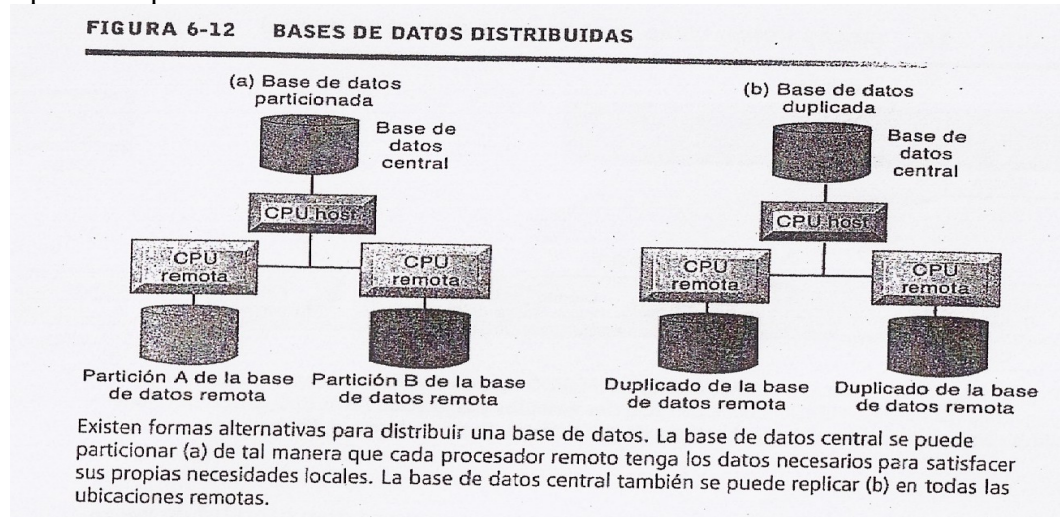
La base de datos también puede ser **distribuida**, que es aquella que se almacena en más de un lugar físico.

Hay dos formas principales de distribuir una base de datos:

- **Base de datos particionada**, unas partes se almacenan y mantienen en otros lugares de manera que cada procesador remoto tenga los datos necesarios para abastecer su área local. Los cambios en los archivos locales se pueden justificar con la base de datos central a partir

de lotes, con frecuencia por la noche. Otra estrategia es **replicar** la base de datos central en todas las ubicaciones remotas. Esta estrategia también requiere que la actualización de la base de datos central se haga en horas de menos carga.

Los sistemas distribuidos reducen la vulnerabilidad de un solo sitio central masivo. Incrementan el servicio y el nivel de respuesta para los usuarios locales y con frecuencia pueden correr en computadoras más pequeñas y menos caras. Sin embargo, las bases de datos locales en ocasiones pueden desviarse de los estándares y definiciones centrales de datos y plantear problemas de seguridad por la amplia distribución del acceso a datos delicados.



6.3. Uso de bases de datos para mejorar el desempeño empresarial y la toma de decisiones

Las empresas utilizan sus bases de datos para dar seguimiento a las transacciones básicas. Pero también, se las necesitan para suministrar información que ayudará a la empresa a manejar los negocios de manera más eficiente, y que ayudará a los gerentes y a los empleados a tomar mejores decisiones.

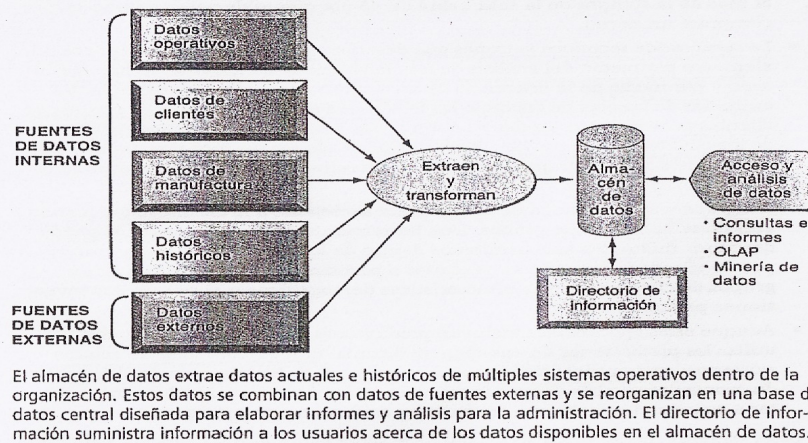
En una empresa grande, con robustas bases de datos o enormes sistemas para cada función (ventas, contabilidad, etc.) se requieren capacidades y herramientas especiales para analizar extensas cantidades de datos y para acceder datos desde múltiples sistemas. Estas incluyen: *el almacenamiento de datos, la minería de datos y herramientas para acceder a las bases de datos a través de la Web.*

- **Almacenes de datos**

Un almacén de datos es una base de datos que almacena datos actuales e históricos de potencial interés para los encargados de la toma de decisiones de toda la empresa. Los datos se originan en muchos sistemas de transacciones operativas esenciales, y podrían incluir datos derivados de transacciones en sitios Web. El almacén de datos consolida y estandariza información de diferentes bases de datos operativas con el propósito de que la información se pueda utilizar a través de la empresa para el análisis y la toma de decisiones por parte de la administración.

El almacén de datos pone los datos a disposición de quien los necesite, pero no se pueden modificar. También proporciona una variedad de herramientas de consulta con fines específicos y estandarizadas, herramientas analíticas y características gráficas para la elaboración de informes. Algunas empresas utilizan portales de intranets para poner la información de almacenes de datos a disposición de toda la empresa.

FIGURA 6-13 COMPONENTES DE UN ALMACÉN DE DATOS



- ***Mercado de Datos***

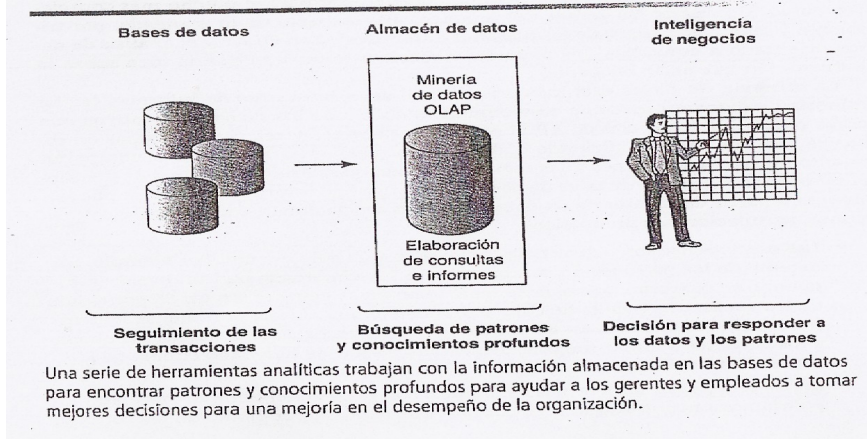
Un mercado de datos es un subconjunto de un almacén de datos en el que una parte resumida o altamente enfocada de los datos de la organización se coloca en una base de datos separada para una población específica de usuarios. Por lo común, un mercado de datos se enfoca en una sola área objetivo o línea de negocios, por lo que usualmente se puede construir más rápido y a un costo más bajo que un almacén de datos a nivel de toda la empresa.

Inteligencia de negocios, análisis de datos multidimensionales y minería de datos

Una serie de herramientas permite a los usuarios analizar estos datos para descubrir nuevos patrones, relaciones y conocimientos profundos para orientar la toma de decisiones. Estas herramientas para consolidar, analizar y dar acceso a extensas cantidades de datos para ayudar a los usuarios a tomar mejores decisiones de negocios con frecuencia se denominan ***Inteligencia de Negocios (BI)***. Las principales herramientas incluyen software para consultas e informes de bases de datos, herramientas para análisis de datos multidimensionales (procesamiento analítico en línea) y minería de datos.

La inteligencia de negocios da a las empresas la capacidad de acumular información; desarrollar conocimiento sobre clientes, competidores y operaciones internas, y cambiar la manera de tomar decisiones, para lograr una rentabilidad más alta y otras metas de negocios.

FIGURA 6-14 INTELIGENCIA DE NEGOCIOS



Las bases de datos operativas de la empresa dan seguimiento a las transacciones que se generan por

el manejo del negocio. Estas bases de datos alimentan información al almacén de datos. Los gerentes utilizan las herramientas de inteligencia de negocios para encontrar patrones y significados en los datos. A continuación, los gerentes aprovechan lo que han aprendido del análisis de los datos tomando decisiones de inteligencia de negocios más informadas.

- ***Procesamiento analítico en línea (OLAP)***

El procesamiento analítico en línea soporta el análisis de datos multidimensionales, el cual permite a los usuarios ver los mismos datos en diferentes formas utilizando múltiples dimensiones. Cada aspecto de la información -producto, precio, etc- representa una dimensión diferente. OLAP permite a los usuarios obtener respuestas en línea a preguntas específicas en un lapso de tiempo sumamente rápido, aun cuando los datos estén almacenados en bases de datos bastante grandes, como las cifras de ventas de varios años. (Ver Gálico pág 242)

- ***Minería de Datos***

Con OLAP y el análisis de datos orientado a consultas, los usuarios necesitan tener una buena idea sobre la información que están buscando.

La ***minería de datos*** está más orientada al descubrimiento de información. La minería de datos proporciona conocimientos acerca de los datos corporativos que no se podrían obtener con OLAP, al encontrar patrones y relaciones ocultas en robustas bases de datos en infiriendo reglas a partir de ellos para predecir el comportamiento futuro. Los patrones y las reglas se utilizan para orientar la toma de decisiones y predecir el efecto de tales decisiones. Los tipos de información que se pueden obtener a partir de la minería de datos incluyen ***asociaciones, secuencias, clasificaciones, agrupaciones y pronósticos***.

Las ***asociaciones*** son ocurrencias enlazadas a un solo evento. Cuando se compra...también se compra..., pero cuando hay una promoción de ...se compra este el ..% de las veces.

En las ***secuencias*** los eventos se enlazan con el paso del tiempo. Ej, si se compra una casa el ...% de las veces se comprará un artículo del hogar dentro de las siguientes 2 semanas.

La ***clasificación*** reconoce patrones que describen el grupo al cual pertenece un elemento por medio del análisis de los elementos existentes que se han clasificado y por medio de la inferencia de un conjunto de reglas. La clasificación ayuda a descubrir las características de los clientes que se podrían perder y así se pueden idear campañas especiales para retenerlos.

El ***agrupamiento*** funciona de una manera semejante a la clasificación cuando aún no se han definido grupos.

La ***generación de pronósticos*** emplea una serie de valores existentes para pronosticar qué otros valores habrá.

Uno de los usos populares de la minería de datos es ofrecer análisis detallado de patrones existentes en los datos de clientes para realizar campañas de MKT o para identificar clientes rentables.

El ***análisis predictivo*** utiliza técnicas de minería de datos, datos históricos y suposiciones sobre condiciones futuras para predecir resultados de eventos, como la probabilidad de que un cliente responderá a una oferta o comprará un producto específico.

La minería de datos plantea riesgos para la protección de la privacidad individual. La tecnología de

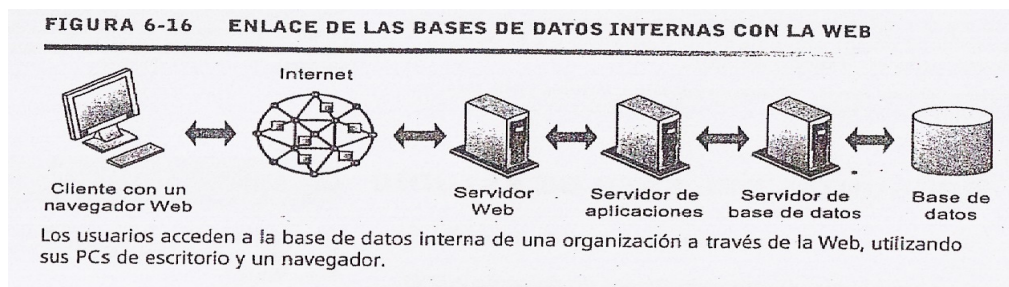
minería de datos puede combinar información de muchas fuentes diversas para crear una “imagen de datos” detallada sobre cada uno de nosotros.

Bases de datos y la Web

Muchas empresas utilizan ahora la Web para poner parte de la información de sus bases de datos internas a disposición de sus clientes y socios de negocios.

El usuario accede al sitio Web del detallista por medio de Internet, utilizando un navegador Web instalado en su Pc del cliente. El navegador Web del usuario solicita datos a la base de la organización, utilizando comandos HTML para comunicarse con el servidor Web.

El servidor Web pasa las solicitudes de datos a software que traduce los comandos HTML a SQL para que los pueda procesar el DBMS que trabaja con la base de datos. En un entorno cliente/servidor, el DBMS reside en una computadora dedicada denominada **servidor de base de datos**. El DBMS recibe las solicitudes de SQL y suministra los datos requeridos. El middleware transfiere información desde la base de datos interna de la organización de regreso al servidor Web para que se entregue en forma de página Web al usuario.



Las ventajas de utilizar la Web para acceder a las bases de datos internas de una organización:

1. Un navegador es más sencillo de utilizar que las herramientas de consulta propietarias.
2. La interfaz Web requiere pocos cambios o ninguno a la base de datos interna. Cuesta mucho menos agregar una interfaz Web a un sistema heredado que rediseñar y volver a construir el sistema para mejorar el acceso a los usuarios.

El acceso a bases de datos corporativas a través de la Web está creando nuevas eficiencias, oportunidades y modelos de negocios.

6.4. Administración de los recursos de datos

**** Establecimiento de una política de información***

Una **política de información** especifica las reglas de la organización para compartir, distribuir, adquirir, estandarizar, clasificar e inventariar la información. Esta establece procedimiento y responsabilidades específicos, que identifican cuáles usuarios y unidades de la organización pueden compartir información, dónde se puede distribuir la información y quién es responsable de actualizar y mantener la información.

La **administración de datos** es responsable de las políticas y procedimientos específicos por medio de los cuales se pueden manejar los datos como un recurso organizacional. Estas responsabilidades incluyen desarrollar la política de información, planificar los datos, supervisar el diseño lógico de la base de datos y el desarrollo del diccionario de datos, y vigilar la manera en que los especialistas en sistemas de información y los grupos de usuarios finales utilizan los datos.

El ***gobierno de información*** se encarga de las políticas y procesos para manejar la disponibilidad, usabilidad, integridad y seguridad de los datos empleados en una empresa, con un énfasis especial en promover la privacidad, la seguridad, la calidad de los datos, y el cumplimiento de las regulaciones gubernamentales.

Una organización grande tendrá un grupo de diseño y administración de bases de datos dentro de la división de sistemas de información corporativa, responsable de definir y organizar la estructura y contenido de la base de datos, así como su mantenimiento.

El grupo de diseño establece la base de datos física, las relaciones lógicas entre los elementos y las reglas de acceso y los procedimientos de seguridad. Las funciones que desempeñan se conocen como ***Administración de la base de datos***.

Aseguramiento de la calidad de datos

Una base de datos y una política de información bien diseñadas contribuirán a garantizar que la empresa tenga la información que necesita. Sin embargo, se deben emprender acciones adicionales para asegurar que la información de las bases de datos de la organización sea exacta y confiable.

Si una base de datos está debidamente diseñada y se han establecido estándares de datos para toda la empresa, los elementos de datos duplicados o inconsistentes deberán ser mínimos. Sin embargo, la mayoría de los problemas de calidad de datos, como nombres, se derivan de errores durante la captura de datos. La incidencia de tales errores se está incrementando a medida que las empresas desplazan sus negocios a la Web y permiten a los clientes y proveedores introducir datos en sus sitios Web que actualizan directamente sus sistemas internos.

Antes de que una nueva base entre en funcionamiento, las organizaciones tienen que identificar y corregir los datos erróneos y establecer mejores rutinas para editar datos una vez que sus bases de datos estén operando. El ***análisis de la calidad de los datos***, que es un sondeo estructurado de la exactitud y nivel de completitud de los datos que hay en un sistema de información. Las auditorías de la calidad de los datos se pueden realizar sondeando archivos de datos completos, sondeando muestras de los archivos de datos o sondeando a los usuarios finales acerca de su percepción de la calidad de los datos.

La ***depuración de datos*** conocida como ***limpieza de datos*** consiste en actividades para detectar y corregir datos de una base de datos que sean incorrectos o incompletos, tengan un formato inadecuado o sean redundantes. La depuración de datos no sólo corrige errores sino que también impone la consistencia entre diferentes conjuntos de datos originados en sistemas de información separados. Existe software de depuración de datos especializado para sondear automáticamente archivos de datos, corregir errores en los datos e integrar los datos en un formato consistente para toda la empresa.

