

Capítulo 6

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

OBJETIVOS DE APRENDIZAJE

Después de leer este capítulo, usted podrá responder a las siguientes preguntas:

1. ¿Cuáles son los problemas de administrar los recursos de datos en un entorno de archivos tradicional y cómo se resuelven mediante un sistema de administración de bases de datos?
2. ¿Cuáles son las principales capacidades de los sistemas de administración de bases de datos (DBMS) y por qué es tan poderoso un DBMS?
3. ¿Cuáles son algunos principios importantes del diseño de bases de datos?
4. ¿Cuáles son las principales herramientas y tecnologías para acceder a la información de las bases de datos y mejorar tanto el desempeño de negocios como la toma de decisiones?
5. ¿Por qué son la política de información, la administración de datos y el aseguramiento de la calidad de los datos esenciales para administrar los recursos de datos de la empresa?

RESUMEN DEL CAPÍTULO

- | | |
|-----|--|
| 6.1 | ORGANIZACIÓN DE LOS DATOS EN UN ENTORNO DE ARCHIVOS TRADICIONAL
Términos y conceptos de organización de archivos
Problemas con el entorno de archivos tradicional |
| 6.2 | LA METODOLOGÍA DE LAS BASES DE DATOS PARA LA ADMINISTRACIÓN DE DATOS
Sistemas de administración de bases de datos
Capacidades de los sistemas de administración de bases de datos
Diseño de bases de datos |
| 6.3 | USO DE BASES DE DATOS PARA MEJORAR EL DESEMPEÑO DE NEGOCIOS Y LA TOMA DE DECISIONES
Almacenes de datos
Herramientas para la inteligencia de negocios: análisis de datos multidimensional y minería de datos
Las bases de datos y Web |
| 6.4 | ADMINISTRACIÓN DE LOS RECURSOS DE DATOS
Establecimiento de una política de información
Aseguramiento de la calidad de los datos |
| 6.5 | PROYECTOS PRÁCTICOS SOBRE MIS
Problemas de decisión gerencial
Obtención de la excelencia operacional: creación de una base de datos relacional para la administración del inventario
Mejora de la toma de decisiones: uso de las bases de datos en línea para buscar recursos de negocios en el extranjero |

MÓDULO DE TRAYECTORIAS DE APRENDIZAJE

Diseño de bases de datos, normalización y diagramas de entidad-relación
Introducción a SQL
Modelos de datos jerárquico y de red

Sesiones interactivas:

¿Qué pueden aprender las empresas de la minería de texto?

Errores del buró de crédito: grandes problemas de la gente

RR DONNELLEY TRATA DE DOMINAR SUS DATOS

Es probable que en estos momentos usted esté utilizando un producto de RR Donnelley. Esta empresa con base en Chicago es una gigantesca compañía de impresión y servicio comercial que provee servicios de impresión, formularios, etiquetas, correo directo y otros servicios. Quizás este libro de texto provenga de esas imprentas. La reciente expansión de la empresa se ha visto impulsada por una serie de adquisiciones, entre las que están, la imprenta comercial Moore Wallace en 2005 y la compañía de administración de la cadena de suministro e impresión llamada Banta, en enero de 2007. Los ingresos de RR Donnelley dieron un salto considerable, de \$2.4 mil millones en 2003 a más de \$9.8 mil millones en la actualidad.

Sin embargo, todo ese crecimiento generó desafíos en cuanto a la administración de la información. Cada compañía adquirida tenía sus propios sistemas, conjunto de datos de clientes, distribuidores y productos. Al provenir de tantas fuentes distintas, con frecuencia los datos eran inconsistentes, duplicados o incompletos. Por ejemplo, cada una de las diferentes unidades de la empresa podría tener un significado distinto para la entidad "cliente". Una podría definir "cliente" como una ubicación específica de facturación, mientras que otra lo podría definir como la entidad matriz legal de una compañía. Donnelley tuvo que utilizar procesos manuales que consumían mucho tiempo para reconciliar los datos almacenados en varios sistemas, para poder obtener una imagen clara a nivel empresarial de cada uno de sus clientes, ya que podrían estar haciendo negocios con varias unidades de la compañía. Estas condiciones aumentaron las ineficiencias y los costos.

RR Donnelley había crecido tanto que ya no era práctico almacenar la información de todas sus unidades en un solo sistema. No obstante, Donnelley aún necesitaba un solo conjunto claro de datos que fuera preciso y consistente para toda la empresa. Para resolver este problema, RR Donnelley recurrió a la administración de datos maestros (MDM). El objetivo de la MDM es asegurar que una organización no utilice varias versiones de la misma pieza de datos en distintas partes de sus operaciones, para lo cual fusiona los registros dispares en un solo archivo maestro autenticado. Una vez implementado el archivo maestro, los empleados y las aplicaciones acceden a una sola vista consolidada de los datos de la compañía. En especial, es útil para las compañías como Donnelley que tienen problemas de integración de sus datos como resultado de las fusiones y adquisiciones.

La implementación de la MDM es un proceso de varios pasos que incluye el análisis de los procesos de negocios, la limpieza de los datos, la consolidación, reconciliación de los datos y la migración de datos hacia un archivo maestro de toda la información de la compañía. Las compañías deben identificar qué grupo es "propietario" de cada pieza de datos y responsable de resolver las definiciones inconsistentes de datos, además de otras discrepancias. Donnelley lanzó su programa MDM a finales de 2005 y empezó a crear un solo conjunto de identificadores para los datos de sus clientes y distribuidores. La compañía optó por un modelo de registro mediante el concentrador de datos (Data Hub) de Purisma, en el cual los datos del cliente aún residen en el sistema en donde se originan, pero se registran en un "concentrador" o "hub" maestro y se hace referencia cruzada a ellos, de modo que las aplicaciones puedan encontrarlos, y los que están en el sistema de origen no se tocan.



Casi un año después, Donnelley lanzó su almacén de datos maestro de clientes (Customer Master Data Store), el cual integra los datos provenientes de numerosos sistemas debido a las adquisiciones de Donnelley. Los datos obsoletos, incompletos o que tienen un formato incorrecto se corrigen o eliminan. Al tener un solo conjunto consistente a nivel empresarial de datos con definiciones y estándares comunes, la gerencia puede averiguar con facilidad qué tipo de negocios y qué tanta actividad comercial tiene con un cliente específico para identificar los mejores clientes y las oportunidades de ventas. Y cuando Donnelley adquiera una compañía, podrá ver con rapidez una lista de los clientes que se traslapen.

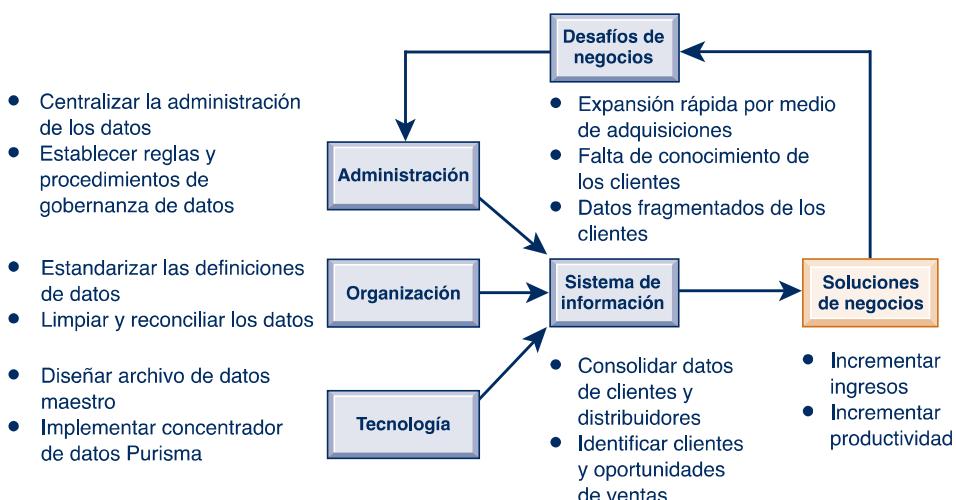
Fuentes: John McCormick, "Mastering Data at R.R. Donnelley", *Information Management Magazine*, marzo de 2009; www.rdonnelley.com, visitado el 10 de junio de 2010, y www.purisma.com, visitado el 10 de junio de 2010.

La experiencia de RR Donnelley ilustra la importancia de la administración de datos para las empresas. Donnelley ha experimentado un crecimiento fenomenal, en su mayor parte debido a las adquisiciones, aunque su desempeño de negocios depende de lo que pueda o no hacer con sus datos. La forma en que las empresas almacenan, organizan y administran sus datos tiene un tremendo impacto sobre la efectividad organizacional.

El diagrama de apertura del capítulo dirige la atención a los puntos importantes generados por este caso y por este capítulo. La gerencia decidió que la compañía necesitaba centralizar la administración de sus datos. La información sobre los clientes, distribuidores, productos y demás entidades importantes se había almacenado en varios sistemas y archivos distintos, de donde no era fácil recuperarlos y analizarlos. A menudo eran redundantes e inconsistentes, lo cual limitaba su utilidad. La gerencia no podía obtener una vista a nivel empresarial de todos sus clientes en todas sus adquisiciones para comercializar sus productos y servicios, además de proveer un mejor servicio y soporte.

En el pasado, RR Donnelley había utilizado mucho los procesos manuales en papel para reconciliar sus datos inconsistentes y redundantes, y para administrar su información desde una perspectiva a nivel empresarial. Esta solución ya no era viable a medida que la organización crecía cada vez más. Una solución más apropiada era identificar, solidificar, limpiar y estandarizar los datos de los clientes junto con los demás tipos de datos en un solo registro de administración de datos maestro. Además de usar la tecnología apropiada, Donnelley tuvo que corregir y reorganizar los datos en un formato estándar, así como establecer reglas, responsabilidades y procedimientos para actualizar y utilizar los datos.

Un sistema de administración de datos maestro ayuda a RR Donnelley a impulsar la rentabilidad, al facilitar la identificación de clientes y las oportunidades de ventas. También mejora la eficiencia organizacional y la toma de decisiones, al tener disponibles datos más precisos y completos sobre los clientes, y al reducir el tiempo requerido para reconciliar los datos redundantes e inconsistentes.



6.1

ORGANIZACIÓN DE LOS DATOS EN UN ENTORNO DE ARCHIVOS TRADICIONAL

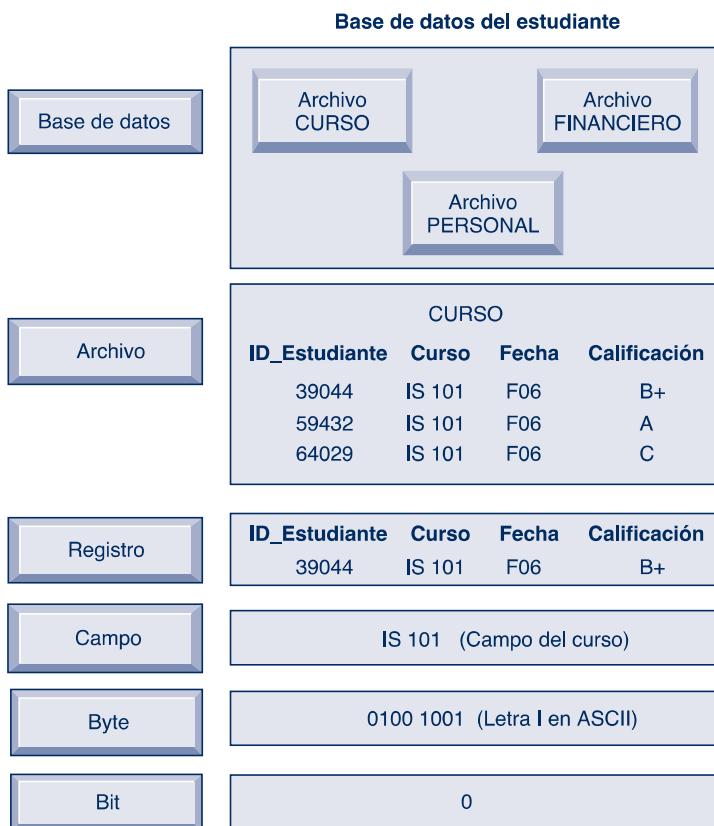
Un sistema de información efectivo provee a los usuarios información precisa, oportuna y relevante. La información precisa está libre de errores. La información es oportuna cuando está disponible para los encargados de tomar decisiones en el momento en que la necesitan. Así mismo, es relevante cuando es útil y apropiada tanto para los tipos de trabajos como para las decisiones que la requieren.

Tal vez le sorprenda saber que muchas empresas no tienen información oportuna, precisa o relevante debido a que los datos en sus sistemas de información han estado mal organizados y se les ha dado un mantenimiento inapropiado. Ésta es la razón por la que la administración de los datos es tan esencial. Para comprender el problema, veamos cómo es que los sistemas de información organizan los datos en archivos de computadora, junto con los métodos tradicionales de administración de archivos.

TÉRMINOS Y CONCEPTOS DE ORGANIZACIÓN DE ARCHIVOS

Un sistema computacional organiza los datos en una jerarquía que empieza con bits y bytes, y progresá hasta llegar a los campos, registros, archivos y bases de datos (vea la figura 6-1). Un bit representa la unidad más pequeña de datos que una computadora pue-

FIGURA 6-1 LA JERARQUÍA DE DATOS



Un sistema computacional organiza los datos en una jerarquía, la cual empieza con el bit, que representa 0 o 1. Los bits se pueden agrupar para formar un byte que representa un carácter, número o símbolo. Los bytes se pueden agrupar para formar un campo, y los campos relacionados para constituir un registro. Los registros relacionados se pueden reunir para crear un archivo, y los archivos relacionados se pueden organizar en una base de datos.

de manejar. Un grupo de bits, denominado byte, representa a un solo carácter, que puede ser una letra, un número u otro símbolo. Un agrupamiento de caracteres en una palabra, un conjunto de palabras o un número completo (como el nombre o la edad de una persona) se denomina **campo**. Un grupo de campos relacionados, como el nombre del estudiante, el curso que va a tomar, la fecha y la calificación, representan un **registro**; un grupo de registros del mismo tipo se denomina **archivo**.

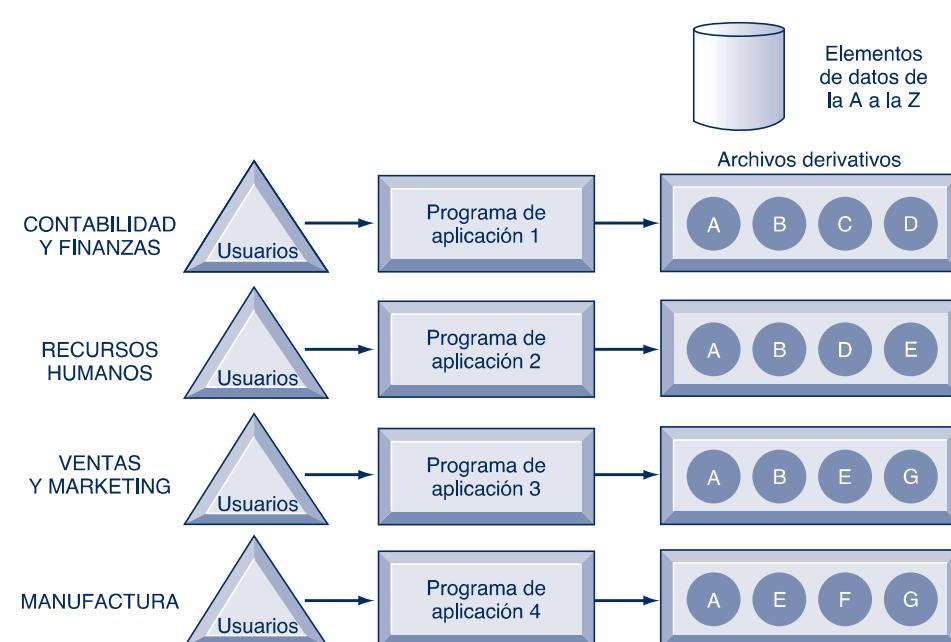
Por ejemplo, los registros en la figura 6-1 podrían constituir un archivo de cursos de estudiantes. Un grupo de archivos relacionados constituye una **base de datos**. El archivo de cursos de estudiantes que se ilustra en la figura 6-1 se podría agrupar con los archivos en los historiales personales de los estudiantes y sus antecedentes financieros, para crear una base de datos de estudiantes.

Un registro describe a una entidad. Una **entidad** es una persona, lugar, cosa o evento sobre el cual almacenamos y mantenemos información. Cada característica o cualidad que describe a una entidad específica se denomina **atributo**. Por ejemplo, ID_Estudiante, Curso, Fecha y Calificaciones son atributos de la entidad CURSO. Los valores específicos que pueden tener estos atributos se encuentran en los campos del registro que describe a la entidad CURSO.

PROBLEMAS CON EL ENTORNO DE ARCHIVOS TRADICIONAL

En la mayoría de las organizaciones, los sistemas tendían a crecer de manera independiente sin un plan a nivel de toda la compañía. Contabilidad, finanzas, manufactura, recursos humanos, ventas y marketing han desarrollado sus propios sistemas y archivos de datos. La figura 6-2 ilustra la metodología normal para el procesamiento de la información.

FIGURA 6-2 PROCESAMIENTO DE ARCHIVOS TRADICIONAL



El uso de una metodología tradicional para el procesamiento de archivos impulsa a cada área funcional en una corporación a desarrollar aplicaciones especializadas. Cada aplicación requiere un archivo de datos único que probablemente sea un subconjunto del archivo maestro. Estos subconjuntos producen redundancia e inconsistencia en los datos, inflexibilidad en el procesamiento y desperdicio de los recursos de almacenamiento.

Desde luego que cada aplicación requería sus propios archivos y programa para operar. Por ejemplo, el área funcional de recursos humanos podría tener un archivo maestro de personal, uno de nómina, uno de seguros médicos, uno de pensiones, uno de listas de correos y así en lo sucesivo, hasta que existieran decenas, tal vez cientos, de archivos y programas. En la compañía en general, este proceso condujo a varios archivos maestros creados, mantenidos y operados por divisiones o departamentos separados. A medida que continúa este proceso durante cinco o 10 años, la organización se ve atestada de cientos de programas y aplicaciones que son muy difíciles de mantener y administrar. Los problemas resultantes son la redundancia e inconsistencia de los datos, la dependencia programa-datos, la inflexibilidad, la seguridad defectuosa de los datos y la incapacidad de compartir datos entre aplicaciones.

Redundancia e inconsistencia de los datos

La **redundancia de los datos** es la presencia de datos duplicados en varios archivos, de modo que se almacenen los mismos datos en más de un lugar o ubicación. La redundancia ocurre cuando distintos grupos en una organización recolectan por separado la misma pieza de datos y la almacenan de manera independiente unos de otros. Desperdicia recursos de almacenamiento y también conduce a la **inconsistencia de los datos**, en donde el mismo atributo puede tener distintos valores. Por ejemplo, en las instancias de la entidad CURSO que se ilustran en la figura 6-1, la Fecha puede estar actualizada en algunos sistemas pero no en otros. El mismo atributo, ID_Estudiante, también puede tener distintos nombres en los distintos sistemas en toda la organización. Por ejemplo, algunos sistemas podrían usar ID_Estudiante y otros ID.

Asimismo se podría generar una confusión adicional al utilizar distintos sistemas de codificación para representar los valores de un atributo. Por ejemplo, los sistemas de ventas, inventario y manufactura de un vendedor minorista de ropa podrían usar distintos códigos para representar el tamaño de las prendas. Un sistema podría representar el tamaño como "extra grande", mientras que otro podría usar el código "XL" para el mismo fin. La confusión resultante dificultaría a las compañías el proceso de crear sistemas de administración de relaciones con el cliente, de administración de la cadena de suministro o sistemas empresariales que integren datos provenientes de distintas fuentes.

Dependencia programa-datos

La **dependencia programa-datos** se refiere al acoplamiento de los datos almacenados en archivos y los programas específicos requeridos para actualizar y dar mantenimiento a esos archivos, de tal forma que los cambios en los programas requieran cambios en los datos. Todo programa de computadora tradicional tiene que describir la ubicación y naturaleza de los datos con los que trabaja. En un entorno de archivos tradicional, cualquier cambio en un programa de software podría requerir un cambio en los datos a los que accede ese programa. Tal vez un programa se modifique de un código postal de cinco dígitos a nueve. Si el archivo de datos original se cambiara para usar códigos postales de nueve dígitos en vez de cinco, entonces otros programas que requirieran el código postal de cinco dígitos ya no funcionarían en forma apropiada. La implementación apropiada de dichos cambios podría costar millones de dólares.

Falta de flexibilidad

Un sistema de archivos tradicional puede entregar informes programados de rutina después de cierto esfuerzo extenso de programación, pero no puede entregar informes ad hoc ni responder de manera oportuna a los requerimientos de información no anticipados. La información requerida por las solicitudes ad hoc está en alguna parte del sistema, pero tal vez sea demasiado costoso recuperarla. Tal vez varios programadores tengan que trabajar durante semanas para reunir los elementos de datos requeridos en nuevo archivo.

Seguridad defectuosa

Como hay poco control o poca administración de los datos, el acceso a la información, así como su diseminación, pueden estar fuera de control. La gerencia tal vez no tenga forma de saber quién está accediendo a los datos de la organización, o incluso modificándolos.

Falta de compartición y disponibilidad de los datos

Como las piezas de información en los distintos archivos y las diferentes partes de la organización no se pueden relacionar entre sí, es casi imposible compartir o acceder a la información de una manera oportuna. La información no puede fluir con libertad entre áreas funcionales o partes de la organización distintas. Si los usuarios encuentran valores desiguales de la misma pieza de información en dos sistemas diferentes, tal vez no quieran usar estos sistemas debido a que no pueden confiar en la precisión de sus datos.

6.2

LA METODOLOGÍA DE LAS BASES DE DATOS PARA LA ADMINISTRACIÓN DE DATOS

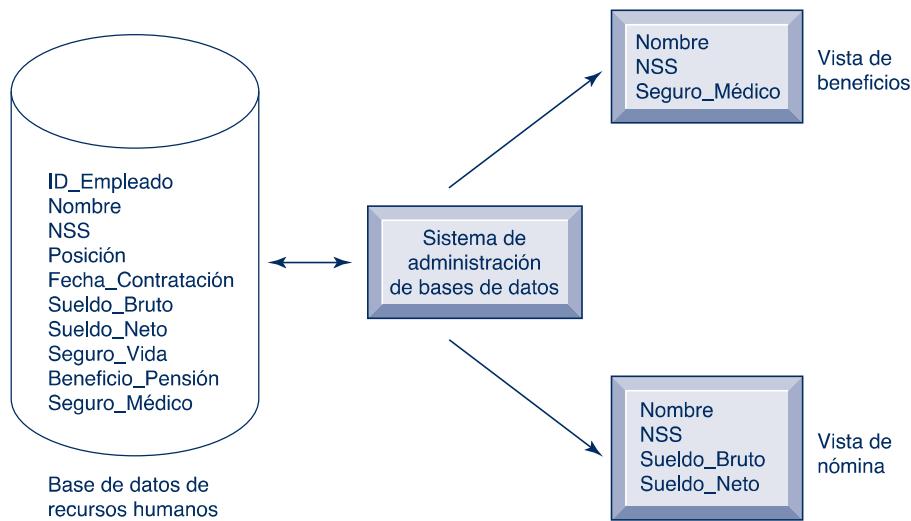
La tecnología de las bases de datos resuelve muchos de los problemas de la organización de los archivos tradicionales. Una definición más rigurosa de una **base de datos** es la de una colección de datos organizados para dar servicio a muchas aplicaciones de manera eficiente, al centralizar los datos y controlar los que son redundantes. En vez de guardar los datos en archivos separados para cada aplicación, se almacenan de modo que los usuarios crean que están en una sola ubicación. Una sola base de datos da servicio a varias aplicaciones. Por ejemplo, en vez de que una corporación almacene los datos de los empleados en sistemas de información y archivos separados para personal, nómina y beneficios, podría crear una sola base de datos común de recursos humanos.

SISTEMAS DE ADMINISTRACIÓN DE BASES DE DATOS

Un **Sistema de Administración de Bases de Datos (DBMS)** es software que permite a una organización centralizar los datos, administrarlos en forma eficiente y proveer acceso a los datos almacenados mediante programas de aplicación. 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, como el sueldo bruto, el DBMS lo busca en la base de datos y lo presenta al programa de aplicación. Si utilizará archivos de datos tradicionales, el programador tendrá que especificar el tamaño y formato de cada elemento de datos utilizado en el programa y después decir a la computadora en dónde están ubicados.

El DBMS libera al programador o al usuario final de la tarea de comprender en dónde y cómo están almacenados los datos en realidad, al separar las vistas lógica y física de los datos. La *vista lógica* presenta los datos según la manera en que los perciben los usuarios finales o los especialistas de negocios, mientras que la *vista física* muestra la verdadera forma en que están organizados y estructurados los datos en los medios de almacenamiento físicos.

El software de administración de bases de datos se encarga de que la base de datos física esté disponible para las diferentes vistas lógicas requeridas por los usuarios. Por ejemplo, para la base de datos de recursos humanos que se ilustra en la figura 6-3, un especialista de negocios podría requerir una vista que consista en el nombre del empleado, número de seguro social y cobertura del seguro médico. El miembro de un departamento de nómina podría necesitar datos tales como el nombre del empleado, el número de seguro social, el sueldo bruto y neto. Los datos para todas estas vistas se

FIGURA 6-3 BASE DE DATOS DE RECURSOS HUMANOS CON VARIAS VISTAS

Una sola base de datos de recursos humanos provee muchas vistas distintas de los datos, dependiendo de los requerimientos de información del usuario. Aquí se ilustran dos posibles vistas, una de interés para un especialista de beneficios y otra de interés para un miembro del departamento de nómina de la compañía.

almacenar en una sola base de datos, en donde la organización puede administrarlos con más facilidad.

Cómo resuelve un DBMS los problemas del entorno de archivos tradicionales

Un DBMS reduce la redundancia e inconsistencia de los datos al minimizar los archivos aislados en los que se repiten los mismos datos. Tal vez el DBMS no logre que la organización elimine la redundancia de datos en su totalidad, pero puede ayudar a controlarla. Incluso si la organización mantiene ciertos datos redundantes, el uso de un DBMS elimina la inconsistencia de los datos debido a que puede ayudar a la organización a asegurar que cada ocurrencia de datos redundantes tenga los mismos valores. El DBMS desacopla los programas y los datos, con lo cual estos últimos se pueden independizar. El acceso y la disponibilidad de la información serán mayores, a la vez que se reducirán los costos de desarrollo y mantenimiento de los programas debido a que los usuarios y programadores pueden realizar consultas ad hoc de la información en la base de datos. El DBMS permite a la organización administrar los datos, su uso y su seguridad en forma central.

DBMS relacional

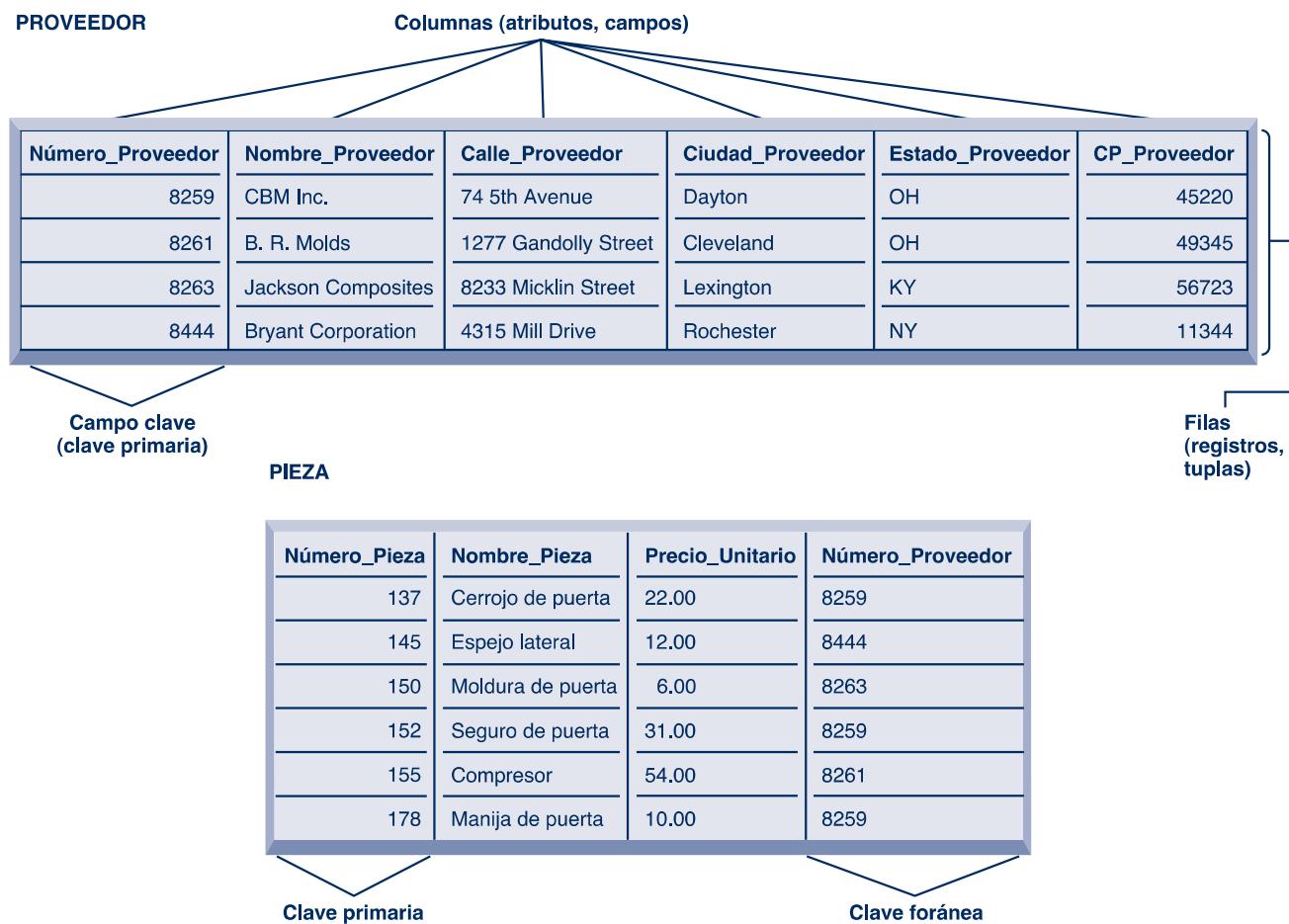
Los DBMS contemporáneos utilizan distintos modelos de bases de datos para llevar el registro de las entidades, atributos y relaciones. El tipo más popular de sistemas DBMS en la actualidad para las PCs, así como para computadoras más grandes y mainframes, es el **DBMS relacional**. Las bases de datos relacionales representan los datos como tablas bidimensionales (llamadas relaciones), a las cuales se puede hacer referencia como si fueran archivos. Cada tabla contiene datos sobre una entidad y sus atributos. Microsoft Access es un DBMS relacional para sistemas de escritorio, mientras que DB2, Oracle Database y Microsoft SQL Server son DBMS relacionales para las grandes mainframes y las computadoras de rango medio. MySQL es un popular DBMS de código fuente abierto; Oracle Database Lite es un DBMS para pequeños dispositivos de cómputo portátiles.

Veamos ahora cómo organiza una base de datos relacional la información sobre proveedores y piezas (vea la figura 6-4). La base de datos tiene una tabla separada para la entidad PROVEEDOR y una para la entidad PIEZA. Cada tabla consiste en una cuadrícula de columnas y filas de datos. Cada elemento individual de datos para cada entidad se almacena como un campo separado, y cada campo representa un atributo para esa entidad. Los campos en una base de datos relacionales también se llaman columnas. Para la entidad PROVEEDOR, el número de identificación de proveedor, nombre, calle, ciudad, estado y código postal 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 reside en una tabla se denomina fila. Por lo general las filas se conocen como registros, o en términos muy técnicos, como **tuplas**. Los datos para la entidad PIEZA tienen su propia tabla separada.

El campo para Nombre_Proveedor en la tabla PROVEEDOR identifica a cada registro en forma única, de modo que ese registro se pueda recuperar, actualizar u ordenar, y se denomina **campo clave**. Cada tabla en una base de datos relacional tiene un campo que se designa como su **clave primaria**. Este campo clave es el identificador único para toda la información en cualquier fila de la tabla y su clave primaria no puede estar duplicada. Numero_Proveedor es la clave primaria para la tabla PROVEEDOR y

FIGURA 6-4 TABLAS DE BASES DE DATOS RELACIONALES



Una base de datos relacional organiza los datos en forma de tablas bidimensionales. Aquí se ilustran las tablas para las entidades PROVEEDOR y PIEZA, las cuales muestran cómo representan a cada entidad y sus atributos. Numero_Proveedor es una clave primaria para la tabla PROVEEDOR y una clave foránea para la tabla PIEZA.

Numero_Pieza es la clave primaria para la tabla PIEZA. Observe que Numero_Proveedor aparece tanto en la tabla PROVEEDOR como en PIEZA. En la tabla PROVEEDOR, Numero_Proveedor es la clave primaria. Cuando el campo Numero_Proveedor aparece en la tabla PIEZA se denomina **clave foránea**, la cual es en esencia un campo de búsqueda para averiguar datos sobre el proveedor de una pieza específica.

Operaciones de un DBMS relacional

Las tablas de bases de datos relacionales se pueden combinar con facilidad para ofrecer los datos requeridos por los usuarios, siempre y cuando dos tablas cualesquiera comparten un elemento de datos común. Suponga que queremos encontrar en esta base de datos los nombres de los proveedores que nos puedan suministrar el número de pieza 137 o el 150. Necesitaríamos información de dos tablas: la tabla PROVEEDOR y la tabla PIEZA. Observe que estos dos archivos tienen un elemento de datos compartido: Numero_Proveedor.

En una base de datos relacional se utilizan tres operaciones básicas, como se muestra en la figura 6-5, para desarrollar conjuntos útiles de datos: seleccionar, unir y proyectar. La operación *seleccionar* crea un subconjunto que consiste en todos los registros del archivo que cumplen con criterios establecidos. En otras palabras, la selección crea un subconjunto de filas que cumplen con ciertos criterios. En nuestro ejemplo, queremos seleccionar registros (filas) de la tabla PIEZA en donde el Numero_Pieza sea igual a 137 o 150. La operación *unir* combina tablas relacionales para proveer al usuario más información de la que está disponible en las tablas individuales. En nuestro ejemplo, queremos unir la tabla PIEZA, que ya está recortada (sólo se presentarán las piezas 137 o 150), con la tabla PROVEEDOR en una sola tabla nueva.

La operación *proyectar* crea un subconjunto que consiste de columnas en una tabla, con lo cual el usuario puede crear nuevas tablas que contengan sólo la información requerida. En nuestro ejemplo queremos extraer de la nueva tabla sólo las siguientes columnas: Numero_Pieza, Nombre_Pieza, Numero_Proveedor y Nombre_Proveedor.

DBMS orientado a objetos

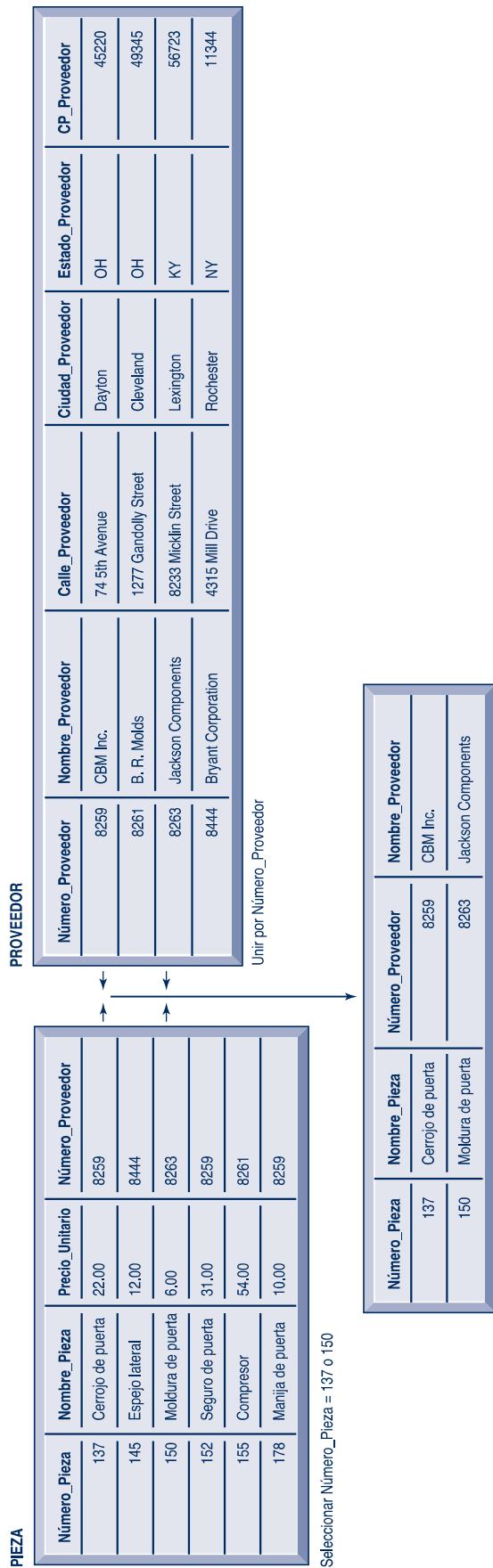
En la actualidad y en el futuro, muchas aplicaciones requerirán bases de datos que puedan almacenar y recuperar no sólo números y caracteres estructurados, sino también dibujos, imágenes, fotografías, voz y video en movimiento completo. Los DBMS diseñados para organizar datos estructurados en filas y columnas no se adaptan bien al manejo de aplicaciones basadas en gráficos o multimedia. Las bases de datos orientadas a objetos son más adecuadas para este propósito.

Un **DBMS orientado a objetos** almacena los datos y los procedimientos que actúan sobre esos datos como objetos que se pueden recuperar y compartir de manera automática. Los Sistemas de Administración de Bases de Datos Orientados a Objetos (OODBMS) están ganando popularidad debido a que se pueden utilizar para manejar los diversos componentes multimedia o los applets de Java que se utilizan en las aplicaciones Web, que por lo general integran piezas de información provenientes de una variedad de orígenes.

Aunque las bases de datos orientadas a objetos pueden almacenar tipos más complejos de información que los DBMS relacionales, son lentos en comparación con los DBMS relacionales para procesar grandes números de transacciones. Ahora hay sistemas **DBMS objeto-relacional** híbridos, que ofrecen las capacidades de los sistemas DBMS tanto orientados a objetos como relacionales.

Bases de datos en la nube

Suponga que su compañía desea utilizar los servicios de computación en la nube. ¿Hay alguna forma de administrar los datos en la nube? La respuesta es un "sí" condicional. Los proveedores de computación en la nube ofrecen servicios de administración de bases de datos, pero por lo general estos servicios tienen menos funcionalidad que sus contrapartes dentro de las premisas de la empresa. Por el momento, la base de clientes primordial para la administración de datos basados en la nube consiste en

FIGURA 6-5 LAS TRES OPERACIONES BÁSICAS DE UN DBMS RELACIONAL

Las operaciones seleccionar, unir y proyectar permiten combinar datos de dos tablas distintas y mostrar sólo los atributos seleccionados.

Proyectar columnas seleccionadas

empresas iniciales enfocadas en Web o negocios desde pequeños hasta medianos que buscan capacidades de bases de datos a un menor precio que el de un DBMS relacional estándar.

Amazon Web Services cuenta con una base de datos no relacional simple llamada SimpleDB y también con un servicio de bases de datos relacionales, el cual se basa en una implementación en línea de MySQL, el DBMS de código fuente abierto. Amazon Relational Database Service (Amazon RDS) ofrece el rango completo de capacidades de MySQL. El precio se basa en el uso (los costos varían desde 11 centavos por hora para una pequeña base de datos que utilice 1.7 GB de memoria del servidor, hasta \$3.10 por hora para una base de datos extensa que utilice 68 GB de memoria del servidor). También hay cargos por el volumen de datos almacenado, el número de solicitudes de entrada-salida, la cantidad de datos que se escriben en la base de datos y la cantidad que se leen de ella.

Además, Amazon Web Services ofrece a los clientes de Oracle la opción de obtener una licencia de Oracle Database 11g, Oracle Enterprise Manager y Oracle Fusion Middleware para ejecutarlos en la plataforma Amazon EC2 (nube de cómputo elástica).

Microsoft SQL Azure Database es un servicio de bases de datos relacionales basado en la nube y en el DBMS SQL Server de Microsoft. Ofrece un servicio de bases de datos con alta disponibilidad y escalable, hospedado por Microsoft en la nube. SQL Azure Database ayuda a reducir los costos al integrarse con las herramientas de software existentes y proveer simetría con las bases de datos tanto en las premisas de la empresa como en la nube.

TicketDirect, que vende boletos para conciertos, eventos deportivos, obras de teatro y películas en Australia y Nueva Zelanda, adoptó la plataforma de nube SQL Azure Database para poder mejorar la administración de las cargas pico del sistema durante los períodos con muchas ventas de boletos. Migró sus datos a la base de datos SQL Azure. Al cambiar a una solución en la nube, TicketDirect pudo escalar sus recursos de cómputo en respuesta a la demanda en tiempo real, al tiempo que mantuvo sus costos bajos.

CAPACIDADES DE LOS SISTEMAS DE ADMINISTRACIÓN DE BASES DE DATOS

Un DBMS incluye capacidades y herramientas para organizar, administrar y acceder a los datos en la base de datos. Las más importantes son: su lenguaje de definición de datos, el diccionario de datos y el lenguaje de manipulación de datos.

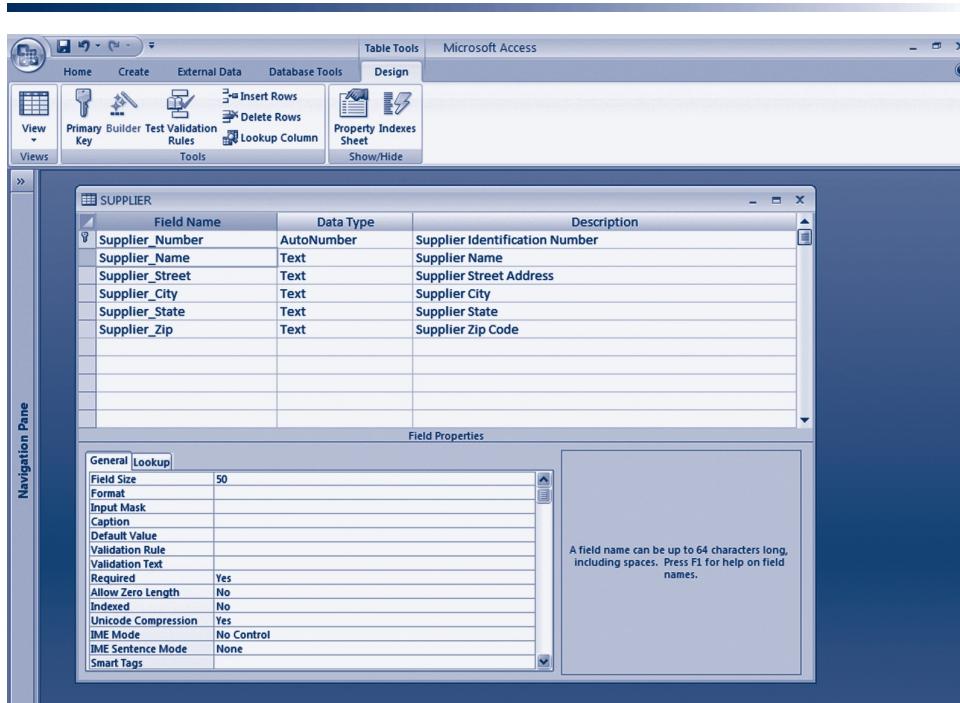
Los DBMS tienen una capacidad de **definición de datos** para especificar la estructura del contenido de la base de datos. Podría usarse para crear tablas de bases de datos y definir las características de los campos en cada tabla. Esta información sobre la base de datos se puede documentar en un **diccionario de datos**, el cual es un archivo automatizado o manual que almacena las definiciones de los elementos de datos y sus características.

Microsoft Access cuenta con una herramienta rudimentaria de diccionario de datos, la cual muestra información sobre el nombre, la descripción, el tamaño, tipo, formato y otras propiedades de cada campo en una tabla (vea la figura 6-6). Los diccionarios de datos para las grandes bases de datos corporativas pueden capturar información adicional, como el uso, la propiedad (quién en la organización es responsable de dar mantenimiento a la información), autorización, seguridad y los individuos, funciones de negocios, programas e informes que utilizan cada elemento de datos.

Consultas e informes

Un DBMS contiene herramientas para acceder a la información en las bases de datos y manipularla. La mayoría de los DBMS tienen un lenguaje especializado conocido como **lenguaje de manipulación de datos** el cual se utiliza para agregar, modificar, eliminar y recuperar los datos en la base. Este lenguaje contiene comandos que permiten a los usuarios finales y a los especialistas de programación extraer los datos de la base para satisfacer las solicitudes de información y desarrollar aplicaciones. El lenguaje de manipulación de datos más prominente en la actualidad es el **lenguaje de consulta estructurado**, o **SQL**. La figura 6-7 ilustra la consulta de SQL que produciría la nueva tabla

FIGURA 6-6 CARACTERÍSTICAS DEL DICCIONARIO DE DATOS DE MICROSOFT ACCESS



Microsoft Access cuenta con una herramienta rudimentaria de diccionario de datos, la cual muestra información sobre el tamaño, formato y otras características de cada campo en una base de datos. Aquí se muestra la información que se mantiene en la tabla PROVEEDOR. El pequeño ícono a la izquierda de Numero_Proveedor indica que es un campo clave.

resultante en la figura 6-5. En las Trayectorias de aprendizaje de este capítulo podrá averiguar más acerca de cómo realizar consultas de SQL.

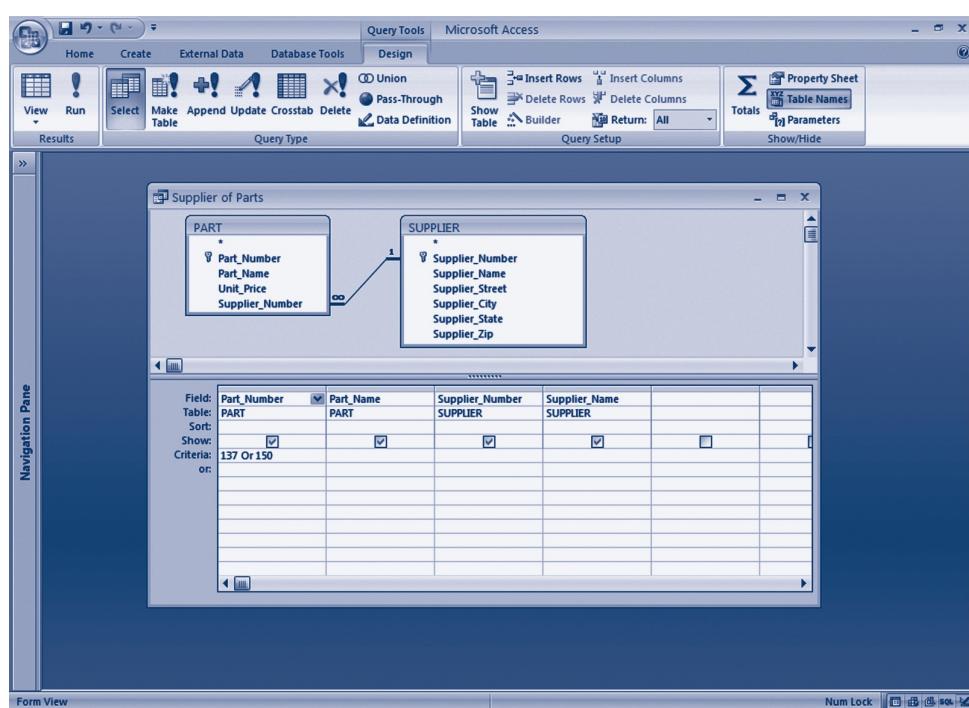
Los usuarios de DBMS para computadoras grandes y de rango medio, como DB2, Oracle o SQL Server, pueden emplear SQL para recuperar la información que necesitan de la base de datos. Microsoft Access también utiliza SQL, sólo que provee su propio conjunto de herramientas amigables para que el usuario realice consultas en las bases de datos, y para organizar la información de las bases de datos en reportes con una mejor presentación.

En Microsoft Access encontrará herramientas que permiten a los usuarios crear consultas al identificar las tablas y campos que desean junto con los resultados, para después seleccionar las filas de la base de datos que cumplan con ciertos criterios específicos. A su vez, estas acciones se traducen en comandos de SQL. La figura 6-8 ilustra cómo

FIGURA 6-7 EJEMPLO DE UNA CONSULTA SQL

```
SELECT PIEZA.Número_Pieza, PIEZA.Nombre_Pieza, PROVEEDOR.Número_Proveedor,
PROVEEDOR.Nombre_Proveedor
FROM PIEZA, PROVEEDOR
WHERE PIEZA.Número_Proveedor = PROVEEDOR.Número_Proveedor AND
Número_Pieza = 137 OR Número_Pieza = 150;
```

Aquí se ilustran las instrucciones de SQL para una consulta que selecciona los proveedores de las piezas 137 o 150. Se produce una lista con los mismos resultados que en la figura 6-5.

FIGURA 6-8 UNA CONSULTA EN ACCESS

Aquí se ilustra cómo se construiría la consulta de la figura 6-7 mediante las herramientas para crear consultas de Microsoft Access. Muestra las tablas, los campos y los criterios de selección utilizados para la consulta.

se construiría la misma consulta que la SQL para seleccionar piezas y proveedores, pero ahora mediante las herramientas para crear consultas de Microsoft.

Microsoft Access y otros sistemas DBMS tienen herramientas para generación de informes, de modo que se puedan mostrar los datos de interés en un formato más estructurado y elegante que el de las consultas. Crystal Reports es un popular generador de informes para los DBMS corporativos extensos, aunque también se puede utilizar con Access. Este último también cuenta con herramientas para desarrollar aplicaciones de sistemas de escritorio. Se incluyen herramientas para crear pantallas de captura de datos, para generar informes y desarrollar la lógica de procesamiento de transacciones.

DISEÑO DE BASES DE DATOS

Para crear una base de datos hay que comprender las relaciones entre la información, el tipo de datos que se mantendrán en la base, cómo se utilizarán y la forma en que tendrá que cambiar la organización para administrarlos desde una perspectiva a nivel de toda la compañía. La base de datos requiere tanto un diseño conceptual como uno físico. El diseño conceptual o lógico de la base de datos es un modelo abstracto de ésta desde una perspectiva de negocios, mientras que el diseño físico muestra la verdadera disposición de la base de datos en los dispositivos de almacenamiento de acceso directo.

Diagramas de normalización y de entidad-relación

El diseño de bases de datos conceptual describe la forma en que se deben agrupar los elementos de datos en la base. El proceso de diseño identifica las relaciones entre los elementos de datos y la manera más eficiente de agruparlos en conjunto para satisfacer los requerimientos de información de la empresa. Este proceso también identifica a los elementos de datos redundantes y las agrupaciones de elementos de datos requeridas

FIGURA 6-9 UNA RELACIÓN SIN NORMALIZAR PARA PEDIDO

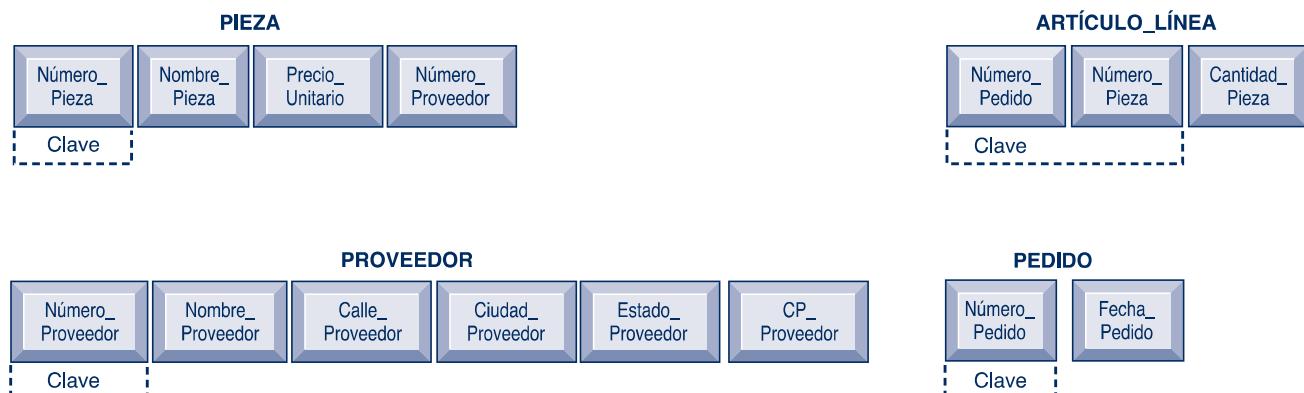
Una relación sin normalizar contiene grupos repetitivos. Por ejemplo, puede haber muchas piezas y proveedores para cada pedido. Sólo hay una correspondencia de uno a uno entre Número_Pedido y Fecha_Pedido.

para ciertos programas de aplicaciones específicos. Los grupos de datos se organizan, refinan y optimizan hasta que emerge una vista lógica general de las relaciones entre todos los datos en la base.

Para usar un modelo de base de datos relacional en forma efectiva, hay que optimizar los agrupamientos complejos de datos para minimizar los elementos de datos redundantes y las incómodas relaciones de varios a varios. El proceso de crear estructuras de datos pequeñas y estables pero a la vez flexibles y adaptativas a partir de grupos complejos de datos se denomina **normalización**. Las figuras 6-9 y 6-10 ilustran este proceso.

En la empresa específica que se modela aquí, un pedido puede tener más de una pieza, pero cada una sólo es proporcionada por un proveedor. Si creamos una relación llamada PEDIDO con todos los campos que se incluyen aquí, tendríamos que repetir el nombre y la dirección del proveedor para cada pieza del pedido, aun y cuando éste sea de piezas de un solo proveedor. Esta relación contiene lo que se denomina grupos de datos repetitivos, ya que puede haber muchas piezas en un solo pedido para un proveedor dado. Una manera más eficiente de ordenar los datos es dividir PEDIDO en relaciones más pequeñas, cada una de las cuales describe a una sola entidad. Si avanzamos paso a paso y normalizamos la relación PEDIDO, obtendremos las relaciones que se ilustran en la figura 6-10. Para averiguar más sobre la normalización, los diagramas entidad-relación y el diseño de bases de datos, consulte las Trayectorias de aprendizaje de este capítulo.

Los sistemas de bases de datos relacionales tratan de cumplir reglas de **integridad referencial** para asegurar que las relaciones entre las tablas acopladas permanezcan consistentes. Cuando una tabla tiene una clave foránea que apunta a otra, no es posible agregar un registro a la tabla con la clave foránea a menos que haya uno correspondiente en la tabla vinculada. En la base de datos que examinamos antes en el

FIGURA 6-10 TABLAS NORMALIZADAS CREADAS A PARTIR DE PEDIDO

Después de la normalización, la relación original PEDIDO se divide en cuatro relaciones más pequeñas. La relación PEDIDO se queda con sólo dos atributos y la relación ARTICULO_LINEA tiene una clave combinada, o concatenada, que consiste en Número_Pedido y Número_Pieza.

capítulo, la clave foránea Numero_Proveedor vincula la tabla PIEZA con la tabla PROVEEDOR. No podemos agregar un nuevo registro a la tabla PIEZA para una pieza con el Numero_Proveedor 8266 a menos que haya un registro correspondiente en la tabla PROVEEDOR para el Numero_Proveedor 8266. También debemos eliminar el registro correspondiente en la tabla PIEZA si quitamos el registro en la tabla PROVEEDOR para el Numero_Proveedor 8266. En otras palabras, ¡no debemos tener piezas de proveedores que no existen!

Los diseñadores de bases de datos documentan su modelo de datos con un **diagrama entidad-relación**, el cual se ilustra en la figura 6-11. Este diagrama muestra la relación entre las entidades PROVEEDOR, PIEZA, ARTICULO_LINEA y PEDIDO. Los cuadros representan las entidades, y las líneas que conectan los cuadros, las relaciones. Una línea que conecta dos entidades que termina en dos marcas cortas designa una relación de uno a uno. Una línea que conecta dos entidades y termina con una pata de cuervo y una marca corta encima de ella indica una relación de uno a varios. La figura 6-11 muestra que un PEDIDO puede contener varios ARTICULO_LINEA (es posible ordenar una PIEZA muchas veces y aparecer otras tantas como artículo de línea en un solo pedido). Cada PIEZA sólo puede tener un PROVEEDOR, pero muchos elementos PIEZA pueden ser proporcionados por el mismo PROVEEDOR.

No podemos enfatizarlo lo suficiente: si el modelo de datos de la empresa no es el correcto, el sistema no podrá dar buen servicio a la empresa. Los sistemas de la compañía no serán tan efectivos como podrían serlo debido a que tendrán que trabajar con datos que tal vez sean imprecisos, incompletos o difíciles de recuperar. Comprender los datos de la organización y la forma en que se deben representar en una base de datos es tal vez la lección más importante que puede usted aprender de este curso.

Por ejemplo, Famous Footwear, una cadena de zapaterías con más de 800 sucursales en 49 estados, no pudo lograr su objetivo de tener “el estilo correcto de zapato en la tienda apropiada para venderse al precio adecuado”, ya que su base de datos no estaba diseñada en forma correcta para ajustar con rapidez el inventario de las tiendas. La compañía tenía una base de datos relacional Oracle operando en una computadora IBM AS/400 de medio rango, pero el objetivo primordial para el que se diseñó la base de datos era producir informes estándar para la gerencia, en vez de reaccionar a los cambios en el mercado. La gerencia no pudo obtener datos precisos sobre artículos específicos en el inventario en cada una de sus tiendas. Para solucionar este problema, la compañía tuvo que crear una nueva base de datos en donde se pudieran organizar mejor los datos de las ventas y del inventario para realizar análisis y administrar el inventario.

6.3

USO DE BASES DE DATOS PARA MEJORAR EL DESEMPEÑO DE NEGOCIOS Y LA TOMA DE DECISIONES

Las empresas utilizan sus bases de datos para llevar el registro de las transacciones básicas, como pagar a los proveedores, procesar pedidos, llevar el registro de los clientes y pagar a los empleados. Pero también se necesitan bases de datos para proveer

FIGURA 6-11 UN DIAGRAMA ENTIDAD-RELACIÓN



El diagrama muestra las relaciones entre las entidades PROVEEDOR, PIEZA, ARTICULO_LINEA y PEDIDO que se podrían usar para modelar la base de datos de la figura 6-10.

información que ayude a la compañía a operar sus negocios con más eficiencia, y ayudar a los gerentes y empleados a tomar mejores decisiones. Si una compañía desea saber cuál producto es el más popular o quién es su cliente más rentable, la respuesta radica en los datos.

Por ejemplo, al analizar los datos de las compras de los clientes con tarjeta de crédito, la cadena de restaurantes Lousie's Trattoria de Los Ángeles descubrió que la calidad era más importante que el precio para la mayoría de sus clientes, que tenían educación universitaria y les gustaba el vino fino. Con base en esta información, la cadena introdujo platillos vegetarianos, más selecciones de mariscos y vinos más costosos, con lo cual se elevaron las ventas en más de un 10 por ciento.

En una compañía grande, con bases de datos o sistemas extensos para funciones separadas, como manufactura, ventas y contabilidad, se requieren capacidades y herramientas especiales para analizar enormes cantidades de datos y acceder a los datos de múltiples sistemas. Estas capacidades incluyen almacenes de datos, minería de datos y herramientas para acceder a las bases de datos internas a través de Web.

ALMACENES DE DATOS

Suponga que desea información concisa y confiable sobre las operaciones, tendencias y cambios actuales a través de toda la compañía. Si trabajara en una empresa grande, podría ser difícil obtener esta información debido a que, con frecuencia, los datos se mantienen en sistemas separados, como en ventas, manufactura o contabilidad. Tal vez algunos de los datos que llegara a necesitar estuvieran en el sistema de ventas, mientras que otros podrían encontrarse en el sistema de manufactura. Muchos son sistemas antiguos heredados que usan tecnologías de administración de datos o sistemas de archivos obsoletos, en donde es difícil para los usuarios acceder a la información.

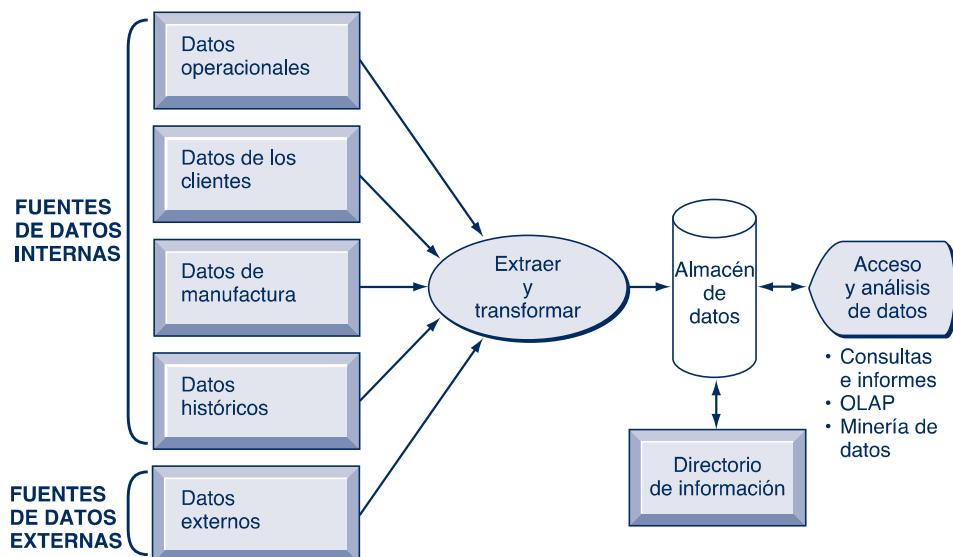
Tal vez tendría que invertir una cantidad exorbitante de tiempo para localizar y recopilar los datos que necesitara, o se vería obligado a tomar su decisión con base en un conocimiento incompleto. Y si deseara información sobre las tendencias, tal vez también tendría problemas para encontrar datos sobre los eventos anteriores, ya que en la mayoría de las empresas sólo sus datos actuales están disponibles de inmediato. Los almacenes de datos se encargan de estos problemas.

¿Qué es un almacén de datos?

Un **almacén de datos** es una base de datos que almacena la información actual e histórica de interés potencial para los encargados de tomar decisiones en la compañía. Los datos se originan en muchos sistemas de transacciones operacionales básicos, como los sistemas de ventas, las cuentas de clientes, la manufactura, y pueden incluir datos de transacciones de sitios Web. El almacén de datos consolida y estandariza la información de distintas bases de datos operacionales, de modo que se pueda utilizar en toda la empresa para el análisis gerencial y la toma de decisiones.

La figura 6-12 ilustra cómo funciona un almacén de datos. Éste pone los datos a disposición de cualquiera que los necesite, pero no se pueden alterar. Un sistema de almacén de datos también provee un rango de herramientas de consulta ad hoc y estandarizadas, herramientas analíticas y facilidades de informes gráficos. Muchas empresas usan portales de intranets para que la información del almacén de datos esté disponible en toda la empresa.

Catalina Marketing, una empresa de marketing global para importantes compañías y minoristas de bienes empaquetados para el consumidor, opera un almacén de datos gigante que incluye tres años de historial de compras para 195 millones de miembros del programa de lealtad de clientes en Estados Unidos en supermercados, farmacias y otros minoristas. Es la base de datos de lealtad más grande del mundo. Los clientes de la tienda minorista de Catalina analizan esta base de datos de históricos de compras de los clientes para determinar las preferencias de compras de los clientes individuales. Cuando un comprador paga en la caja registradora de uno de los clientes minoristas de Catalina, la compra se analiza al instante junto con el historial de compra de ese cliente

FIGURA 6-12 COMPONENTES DE UN ALMACÉN DE DATOS

El almacén de datos extrae los datos actuales e históricos de varios sistemas operacionales dentro de la organización. Estos datos se combinan con los provenientes de fuentes externas y se reorganizan en una base de datos central, diseñada para realizar informes y análisis gerenciales. El directorio de información da a conocer a los usuarios los datos disponibles en el almacén.

en el almacén de datos, para determinar qué cupones recibirá el cliente al momento de pagar, junto con su recibo.

El Servicio de Recaudación de Impuestos (IRS) de Estados Unidos mantiene un almacén de datos de conformidad que consolida la información de los contribuyentes que se ha fragmentado entre varios sistemas heredados distintos, contiene los datos personales sobre contribuyentes y devoluciones fiscales archivadas. Estos sistemas se habían diseñado para procesar formularios de devolución de impuestos de manera eficiente, pero sus datos eran muy difíciles de consultar y analizar. El almacén de datos de conformidad integra los datos de los contribuyentes de muchas fuentes dispares en una estructura relacional, lo cual facilita en gran medida las consultas y el análisis. Con una imagen completa y exhaustiva de los contribuyentes, el almacén ayuda a los analistas y al personal del IRS a identificar las personas que tienen más probabilidades de hacer trampa en sus pagos de impuestos y a responder con rapidez a las consultas de los contribuyentes.

Mercados de datos

A menudo las compañías crean almacenes de datos a nivel empresarial, en donde un almacén de datos central da servicio a toda la organización, o crean almacenes más pequeños y descentralizados, conocidos como mercados de datos. Un **mercado de datos** es un subconjunto de un almacén de datos, en el cual se coloca una porción con alto grado de enfoque en los datos de la organización en una base de datos separada para una población específica de usuarios. Por ejemplo, una compañía podría desarrollar mercados de datos sobre marketing y ventas para lidar con la información de los clientes. Antes de implementar un almacén de datos a nivel empresarial, la librería Barnes & Noble mantenía una serie de mercados de datos: uno para los datos sobre los puntos de venta en las tiendas minoristas, otro para las ventas de las librerías universitarias y un tercero para las ventas en línea. Por lo general, un mercado de datos se enfoca en un solo tema o línea de negocios, por lo que es común que se construya con más rapidez y a un menor costo que un almacén de datos a nivel empresarial.

HERRAMIENTAS PARA LA INTELIGENCIA DE NEGOCIOS: ANÁLISIS DE DATOS MULTIDIMENSIONAL Y MINERÍA DE DATOS

Una vez que los datos en línea se capturan y organizan en almacenes y mercados de datos, están disponibles para su posterior análisis mediante el uso de las herramientas para inteligencia de negocios, de las que hablamos brevemente en el capítulo 2. Las herramientas de inteligencia de negocios permiten a los usuarios analizar datos para ver nuevos patrones, relaciones y perspectivas que son útiles para guiar la toma de decisiones.

Las principales herramientas para la inteligencia de negocios incluyen el software para consultas e informes de bases de datos, herramientas para el análisis de datos multidimensional (procesamiento analítico en línea), y herramientas para la minería de datos. En esta sección le presentaremos estas herramientas; veremos más detalles sobre la ciencia del análisis de inteligencia de negocios y las aplicaciones al examinar la toma de decisiones en el capítulo 12.

Procesamiento analítico en línea (OLAP)

Suponga que su compañía vende cuatro productos distintos: tuercas, pernos, arandelas y tornillos en las regiones Este, Oeste y Central. Si desea hacer una pregunta bastante directa, como cuántas arandelas se vendieron durante el trimestre pasado, podría encontrar la respuesta con facilidad al consultar su base de datos de ventas. Pero, ¿qué pasaría si quisiera saber cuántas arandelas se vendieron en cada una de sus regiones de ventas, para comparar los resultados actuales con las ventas proyectadas?

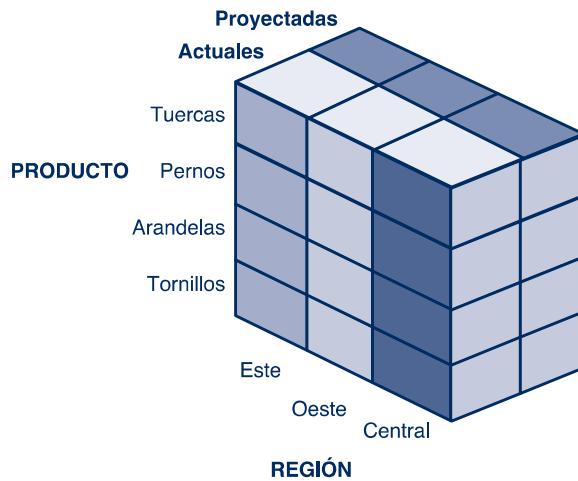
Para obtener la respuesta, necesitaría el **Procesamiento Analítico en Línea (OLAP)**. OLAP soporta el análisis de datos multidimensional, el cual permite a los usuarios ver los mismos datos de distintas formas mediante el uso de varias dimensiones. Cada aspecto de información —producto, precios, costo, región o periodo de tiempo— representa una dimensión distinta. Así, un gerente de productos podría usar una herramienta de análisis de datos multidimensional para saber cuántas arandelas se vendieron en el Este en junio, cómo se compara esa cifra con la del mes anterior y con la de junio del año anterior, y cómo se compara con el pronóstico de ventas. OLAP permite a los usuarios obtener respuestas en línea a las preguntas ad hoc tales como éstas en un periodo de tiempo bastante corto, incluso cuando los datos se almacenan en bases de datos muy grandes, como las cifras de ventas de varios años.

La figura 6-13 muestra un modelo multidimensional que podría crearse para representar productos, regiones, ventas reales y ventas proyectadas. Una matriz de ventas actuales se puede apilar encima de una matriz de ventas proyectadas para formar un cubo con seis caras. Si gira el cubo 90 grados en un sentido, la cara que se muestre será la de producto contra ventas actuales y proyectadas, si lo gira 90 grados de nuevo, verá la cara de región contra ventas actuales y proyectadas, si lo gira 180 grados a partir de la vista original, verá las ventas proyectadas y producto contra región. Se pueden anidar cubos dentro de otros cubos para crear vistas complejas de datos. Una compañía podría utilizar una base de datos multidimensional especializada o una herramienta que cree vistas multidimensionales de datos en las bases de datos relacionales.

Minería de datos

Las consultas en las bases de datos tradicionales responden a preguntas como: “¿Cuántas unidades del producto número 403 se enviaron en febrero de 2010?” El OLAP (análisis multidimensional) soporta solicitudes mucho más complejas de información, como: “Comparar las ventas del producto 403 relativas con el plan por trimestre y la región de ventas durante los últimos dos años”. Con OLAP y el análisis de datos orientados a consultas, los usuarios necesitan tener una buena idea sobre la información que están buscando.

La **minería de datos** está más orientada al descubrimiento, ya que provee perspectivas hacia los datos corporativos que no se pueden obtener mediante OLAP, al encontrar patrones y relaciones ocultas en las bases de datos grandes e inferir reglas a partir

FIGURA 6-13 MODELO DE DATOS MULTIDIMENSIONAL

La vista que se muestra es la de producto contra región. Si gira el cubo 90 grados, la cara mostrará la vista de producto contra las ventas actuales y proyectadas, si lo gira 90 grados otra vez, verá la vista de región contra ventas actuales y proyectadas. Es posible obtener otras vistas.

de estos patrones y relaciones, para predecir el comportamiento a futuro. Los patrones y reglas se utilizan para guiar la toma de decisiones y pronosticar el efecto de esas decisiones. Los tipos de información que se pueden obtener de la minería de datos son: asociaciones, secuencias, clasificaciones, agrupamientos y pronósticos.

- Las *asociaciones* son ocurrencias vinculadas a un solo evento. Por ejemplo, un estudio de los patrones de compra en supermercados podría revelar que, cuando se compran frituras de maíz, el 65 por ciento del tiempo se compra un refresco de cola, pero cuando hay una promoción, el 85 por ciento se compra un refresco de cola. Esta información ayuda a los gerentes a tomar mejores decisiones debido a que descubren la rentabilidad de una promoción.
- En las *secuencias*, los eventos se vinculan en el transcurso del tiempo. Por ejemplo, podríamos descubrir que si se compra una casa, el 65 por ciento del tiempo se compra un nuevo refrigerador dentro de las siguientes dos semanas, y el 45 por ciento se compra un horno dentro del mes posterior a la compra de la casa.
- La *clasificación* reconoce los patrones que describen el grupo al que pertenece un elemento, para lo cual se examinan los elementos existentes que hayan sido clasificados y se infiere un conjunto de reglas. Por ejemplo, las empresas como las compañías de tarjetas de crédito o las telefónicas se preocupan por la pérdida de clientes estables. La clasificación ayuda a descubrir las características de los clientes con probabilidades de dejar de serlo y puede proveer un modelo para ayudar a los gerentes a predecir quiénes son esos clientes, de modo que puedan idear campañas especiales para retenerlos.
- El *agrupamiento* funciona de una manera similar a la clasificación cuando aún no se han definido grupos. Una herramienta de minería de datos puede descubrir distintas agrupaciones dentro de los datos, como el hecho de encontrar grupos de afinidad para tarjetas bancarias o particionar una base de datos en grupos de clientes con base en la demografía y los tipos de inversiones personales.
- Aunque estas aplicaciones implican predicciones, el *pronóstico* utiliza las predicciones de una manera distinta. Se basa en una serie de valores existentes para pronosticar cuáles serán los otros valores. Por ejemplo, el pronóstico podría encontrar patrones en los datos para ayudar a los gerentes a estimar el futuro valor de variables continuas, como las cifras de ventas.

Estos sistemas realizan análisis de alto nivel de los patrones o tendencias, pero también pueden profundizar para proveer más detalles cuando sean necesarios. Existen

aplicaciones de minería de datos para todas las áreas funcionales de negocios, y también para el trabajo gubernamental y científico. Un uso popular de la minería de datos es el de proveer análisis detallados de los patrones en los datos de los consumidores para las campañas de marketing de uno a uno, o para identificar los clientes rentables.

Por ejemplo, Harrah's Entertainment, la segunda compañía de apuestas más grande en su industria, utiliza la minería de datos para identificar a sus clientes más rentables y generar más ingresos gracias a ellos. La compañía analiza en forma continua los datos sobre sus clientes que se recopilan cuando las personas juegan en las máquinas tragamonedas o utilizan los casinos y hoteles de Harrah's. El departamento de marketing de Harrah's utiliza esta información para crear un perfil de apuestas detallado, con base en el valor continuo de un cliente específico para la compañía. Por ejemplo, la minería de datos permite a Harrah's conocer la experiencia de juego favorita de un cliente regular en uno de sus casinos en los barcos de la región del Medio Oeste, junto con las preferencias de esa persona en cuanto al alojamiento, los restaurantes y el entretenimiento. Esta información guía las decisiones gerenciales sobre cómo cultivar los clientes más rentables y animarlos a que gasten más, y también sobre cómo atraer más clientes con un alto potencial de generación de ingresos. La inteligencia de negocios mejoró tanto las ganancias de Harrah's que se convirtió en la pieza central de la estrategia de negocios de la empresa.

El **análisis predictivo** utiliza las técnicas de minería de datos, los datos históricos y las suposiciones sobre las condiciones futuras para predecir los resultados de los eventos, como la probabilidad de que un cliente responda a una oferta o que compre un producto específico. Por ejemplo, la división estadounidense de The Body Shop International plc utilizó el análisis predictivo con su base de datos de clientes de catálogo, Web y de las tiendas minoristas para identificar a los clientes que tenían mayores probabilidades de realizar compras por catálogo. Esa información ayudó a la compañía a crear una lista de correo más precisa y dirigida para sus catálogos, con lo cual se pudo mejorar la tasa de respuesta en cuanto al envío de catálogos por correo y los ingresos por las ventas a través de este medio.

Minería de datos y minería Web

La principal función de las herramientas de inteligencia de negocios es lidiar con los datos que se han estructurado en bases de datos y archivos. Sin embargo, se cree que los datos no estructurados, que en su mayoría están organizados en forma de archivos de texto, representan más del 80 por ciento de la información útil de una organización. El correo electrónico, los memorándums, las transcripciones de los call centers, las respuestas a las encuestas, los casos legales, las descripciones de patentes y los informes de servicio son todos elementos valiosos para encontrar patrones y tendencias que ayuden a los empleados a tomar mejores decisiones de negocios. En la actualidad hay herramientas de **minería de texto** disponibles para ayudar a las empresas a analizar estos datos. Estas herramientas pueden extraer elementos clave de los conjuntos de datos extensos no estructurados, descubrir patrones y relaciones, así como sintetizar la información. Las empresas podrían recurrir a la minería de texto para analizar las transcripciones de los call centers de servicio al cliente para identificar las principales cuestiones de servicio y reparación.

La minería de texto es una tecnología relativamente nueva, pero la verdadera novedad es la cantidad de formas en que los consumidores generan datos no estructurados y los usos que dan las empresas a esos datos. La Sesión interactiva sobre tecnología explora algunas de las aplicaciones de los negocios de la minería de texto.

La Web es otra fuente extensa de información valiosa, y parte de ésta se puede explotar en busca de patrones, tendencias y perspectivas en relación con el comportamiento de los clientes. El descubrimiento y análisis de los patrones útiles y la información proveniente de World Wide Web se denominan **minería Web**. Las empresas podrían recurrir a la minería Web para que les ayude a comprender el comportamiento de los clientes, evaluar la efectividad de un sitio Web específico o cuantificar el éxito de una campaña de marketing. Por ejemplo, los comerciantes utilizan los servicios Google Trends y Google Insights for Search, que rastrean la popularidad de varias palabras y frases utilizadas en las consultas de búsqueda de Google para saber en qué están interesadas las personas y qué les gusta comprar.

SESIÓN INTERACTIVA: TECNOLOGÍA

¿QUÉ PUEDEN APRENDER LAS EMPRESAS DE LA MINERÍA DE TEXTO?

La minería de texto es el descubrimiento de patrones y relaciones a partir de grandes conjuntos de datos no estructurados; el tipo de datos que generamos en los correos electrónicos, las conversaciones telefónicas, lo que publicamos en los blogs, las encuestas en línea para los clientes y los tweets. La plataforma digital móvil ha amplificado la explosión en la información digital, en donde cientos de millones de personas llaman, envían mensajes de texto, buscan, usan "apps" (aplicaciones), compran bienes y escriben miles de millones de correos electrónicos mientras van de un lado a otro.

En la actualidad los consumidores son más que simples compradores: tienen más formas de colaborar, compartir información e influir en las opiniones de sus amigos y colegas; además, los datos que crean al hacerlo tienen un valor considerable para las empresas. A diferencia de los datos estructurados, que se generan a partir de eventos como completar una transacción de compra, los datos sin estructura no tienen una forma definida. Sin embargo, los gerentes creen que dichos datos pueden ofrecer perspectivas únicas en cuanto al comportamiento y las actitudes de los clientes que eran mucho más difíciles de terminar hace unos cuantos años.

Por ejemplo, en 2007 JetBlue experimentó niveles sin precedente de quejas de los clientes a raíz de una tormenta de hielo en febrero que provocó muchas cancelaciones de vuelos por todos lados y aviones varados en las pistas del Aeropuerto Kennedy. La aerolínea recibió 15 000 correos electrónicos al día de los clientes durante la tormenta y justo después de ella, mucho más de su volumen diario habitual de 400. Tan grande fue el volumen a comparación de lo usual, que JetBlue simplemente no podía leer todo lo que sus clientes decían.

Por fortuna, la compañía recién había contratado a Attensity, un distribuidor líder de software de análisis de texto, y pudo usar el software para analizar todo el correo que recibió durante dos días. De acuerdo con el analista de investigación de JetBlue llamado Bryan Jeppesen, el software Attensity Analyze for Voice of the Customer (VoC) permitió a JetBlue extraer con rapidez los sentimientos de los clientes, sus preferencias y las solicitudes que no pudo averiguar de ninguna otra forma. Esta herramienta utiliza una tecnología propietaria para identificar de manera automática hechos, opiniones, solicitudes, tendencias y puntos problemáticos a partir del texto no estructurado de las respuestas a las encuestas, notas de servicio, mensajes de correo electrónico, foros Web, mensajes publicados en blogs, artículos de noticias y otras comunicaciones de los clientes. La tecnología es capaz de identificar con precisión y de manera automática las muchas "voices" distintas que utilizan los clientes para expresar su opinión (como una voz negativa, positiva o condicional), la cual ayuda a las organizaciones a señalar los eventos y relaciones clave, como la intención de comprar, de salirse,

o los "deseos" de los clientes. Puede revelar cuestiones sobre productos y servicios específicos, reacciones a los esfuerzos de marketing y de relaciones públicas, e incluso señales de compra.

El software de Attensity integrado con las demás herramientas de análisis de clientes de JetBlue, como la métrica de Net Promoter de Satmetrix, que clasifica a los clientes en grupos que generan retroalimentación positiva, negativa o ningún tipo de retroalimentación sobre la compañía. Al usar el análisis de texto de Attensity en conjunto con estas herramientas, JetBlue desarrolló una declaración de derechos de los clientes que lidiaba con los principales problemas que tenían los clientes con la compañía.

Las cadenas hoteleras como Gaylord Hotels y Choice Hotels utilizan software de minería de texto para cosechar opiniones de las miles de encuestas de satisfacción del cliente que proporcionan sus huéspedes. Gaylord Hotels utiliza la solución de análisis de texto de Clarabridge que se ofrece a través de Internet como un servicio de software hospedado para recopilar y analizar la retroalimentación de los clientes proveniente de las encuestas, el correo electrónico, la mensajería instantánea, los call centers dotados de personal, y los foros en línea asociados con las experiencias de los huéspedes y los planificadores de reuniones en los centros de convenciones de la compañía. El software Clarabridge examina las encuestas de los clientes de la cadena hotelera y recopila los comentarios tanto positivos como negativos, para después organizarlos en una variedad de categorías y revelar opiniones menos obvias. Por ejemplo, los huéspedes se quejaron más sobre otros asuntos que sobre los cuartos ruidosos, pero las quejas de los cuartos ruidosos se correlacionaron con más frecuencia con las encuestas que indicaban una indisposición a regresar al hotel.

El análisis de las encuestas de los clientes solía tomar semanas, pero ahora sólo es cuestión de días gracias al software Clarabridge. Los gerentes de ubicaciones y los ejecutivos corporativos también han utilizado los hallazgos de la minería de texto para influir en las decisiones sobre mejoras en las empresas.

Wendy's International adoptó el software Clarabridge para analizar los cerca de 500 000 mensajes que recopila cada año de su foro de retroalimentaciones basado en Web, las notas del call center, los mensajes de correo electrónico, las encuestas en los recibos y los medios sociales. El equipo de satisfacción al cliente de la cadena había utilizado antes hojas de cálculo y búsquedas de palabras clave para repasar los comentarios de los clientes, una metodología manual muy lenta. La gerencia de Wendy's estaba en busca de una mejor herramienta para agilizar el análisis, detectar los problemas emergentes y señalar las áreas problemáticas de la empresa a nivel de tienda, regional o corporativo.

La tecnología de Clarabridge permite a Wendy's rastrear las experiencias de los clientes hasta el nivel de tienda en cuestión de minutos. Esta información oportuna ayuda a los gerentes de tiendas, regionales y corporativos a detectar y lidiar con los problemas relacionados con la calidad de los alimentos, la limpieza y la velocidad del servicio.

El software de análisis de texto tuvo auge primero con las agencias gubernamentales y las compañías de mayor tamaño con departamentos de sistemas de información que tenían los medios para usar de manera apropiada el complicado software, pero ahora Clarabridge ofrece una versión de su producto orientada hacia las pequeñas empresas. La tecnología ya ha tenido auge con las agencias policiales, las interfaces de las herramientas de búsqueda y las "plataformas de escucha" como Nielsen Online. Las plataformas de escucha son herramientas de minería de texto que se enfocan en la administración

de marcas para permitir a las empresas determinar cómo se sienten los clientes en cuanto a su marca y tomar acciones para responder al sentimiento negativo.

El análisis de datos estructurados no se hará obsoleto debido al análisis de texto, pero a las compañías que pueden usar ambos métodos para desarrollar una imagen más clara de las posturas de sus clientes les será más fácil establecer su marca y deducir las perspectivas que mejorarán la rentabilidad.

Fuentes: Doug Henschien, "Wendy's Taps Text Analytics to Mine Customer Feedback", *Information Week*, 23 de marzo de 2010; David Stodder, "How Text Analytics Drive Customer Insight", *Information Week*, 1 de febrero de 2010; Nancy David Kho, "Customer Experience and Sentiment Analysis", *KMWorld*, 1 de febrero de 2010; Siobhan Gorman, "Details of Einstein Cyber-Shield Disclosed by White House", *The Wall Street Journal*, 2 de marzo de 2010; www.attensity.com, visitado el 16 de junio de 2010, y www.clarabridge.com, visitado el 17 de junio de 2010.

PREGUNTAS DEL CASO DE ESTUDIO

MIS EN ACCIÓN

1. ¿Qué retos presenta para las empresas el aumento en los datos no estructurados?
2. ¿Cómo mejora la minería de texto el proceso de toma de decisiones?
3. ¿Qué tipos de compañías tienen más probabilidad de beneficiarse del software de minería de texto? Explique su respuesta.
4. ¿En qué formas podría la minería de texto conducir potencialmente a la erosión de la privacidad de la información personal? Explique.

Visite un sitio Web como QVC.com o TripAdvisor.com, en donde se detallen los productos o servicios que tienen reseñas de los clientes. Elija un producto, hotel u otro servicio que tenga al menos 15 reseñas de clientes y léalas, tanto las positivas como las negativas. ¿Cómo podría la minería de contenido Web ayudar a que la compañía mejore o comercialice de una mejor manera los productos o servicios que ofrece? ¿Qué piezas de información se deberían resaltar?

La minería Web busca patrones en los datos a través de la minería de contenido, la minería de estructura y la minería de uso. La minería de contenido Web es el proceso de extraer conocimiento del contenido de páginas Web, lo cual puede incluir datos de texto, imágenes, audio y video. La minería de estructura Web extrae información útil de los vínculos incrustados en documentos Web. Por ejemplo, los vínculos que apuntan a un documento indican su popularidad, mientras que los que salen de un documento indican la riqueza, o tal vez la variedad de temas cubiertos en él. La minería de uso Web examina los datos de interacción de los usuarios registrados por un servidor Web cada vez que se reciben solicitudes relacionadas con los recursos de un sitio Web. Los datos de uso registran el comportamiento del usuario cuando navega o realiza transacciones en el sitio Web y recopila los datos en un registro del servidor. Al analizar esos datos, las compañías pueden determinar el valor de ciertos clientes específicos, las estrategias de marketing cruzado entre los diversos productos y la efectividad de las campañas promocionales.

LAS BASES DE DATOS Y WEB

¿Alguna vez ha tratado de usar la Web para realizar un pedido o ver un catálogo de productos? Si su respuesta es positiva, es probable que haya usado un sitio Web vinculado a una base de datos corporativa interna. Ahora muchas compañías utilizan Web para poner parte de la información en sus bases de datos internas a disposición de los clientes y los socios de negocios.

Suponga por ejemplo que un cliente con un navegador Web desea buscar información de precios en la base de datos en línea de un vendedor minorista. La figura 6-14 ilustra la forma en que ese cliente podría acceder a la base de datos interna del vendedor a través de Web. El usuario accede al sitio Web del vendedor a través de Internet mediante el software de navegador Web en su PC cliente. El software de navegador Web del usuario solicita información a la base de datos de la organización, mediante comandos de HTML para comunicarse con el servidor Web.

Como muchas bases de datos de procesamiento en segundo plano (back-end) no pueden interpretar comandos escritos en HTML, el servidor Web pasa estas solicitudes de datos al software que traduce los comandos de HTML en SQL, de modo que el DBMS que trabaja con la base de datos pueda procesarlos. En un entorno cliente/servidor, el DBMS reside en una computadora dedicada llamada **servidor de bases de datos**. El DBMS recibe las solicitudes de SQL y provee los datos requeridos. El middleware transforma la información de la base de datos interna y la devuelve al servidor Web para que la ofrezca en forma de una página Web al usuario.

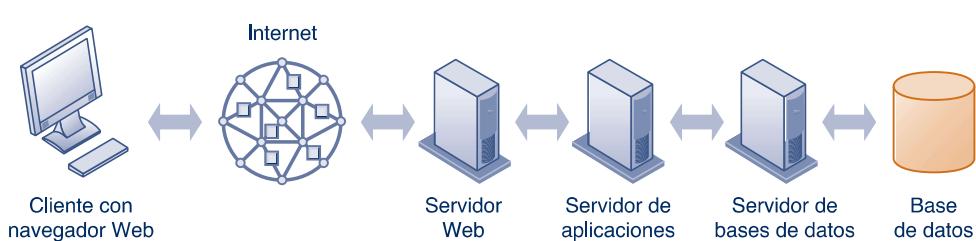
La figura 6-14 muestra que el middleware que trabaja entre el servidor Web y el DBMS es un servidor de aplicaciones que se ejecuta en su propia computadora dedicada (vea el capítulo 5). El software del servidor de aplicaciones maneja todas las operaciones de la aplicación, entre ellas el procesamiento de las transacciones y el acceso a los datos, entre las computadoras basadas en navegador y las aplicaciones o bases de datos de negocio de procesamiento en segundo plano (back-end) de una compañía. El servidor de aplicaciones recibe las solicitudes del servidor Web, ejecuta la lógica de negocio para procesar las transacciones con base en esas solicitudes y provee conectividad a los sistemas o bases de datos de procesamiento en segundo plano de la organización. De manera alternativa, el software para manejar estas operaciones podría ser un programa personalizado o una secuencia de comandos CGI: un programa compacto que utiliza la especificación *Interfaz de puerta de enlace común (CGI)* para procesar datos en un servidor Web.

Hay varias ventajas en cuanto al uso de Web para acceder a las bases de datos internas de una organización. En primer lugar, el software de navegador Web es mucho más fácil de usar que las herramientas de consulta propietarias. En segundo lugar, la interfaz Web requiere pocos o ningún cambio en la base de datos interna. Es mucho menos costoso agregar una interfaz Web frente a un sistema heredado que rediseñar y reconstruir el sistema para mejorar el acceso de los usuarios.

El acceso a las bases de datos corporativas por medio de Web está creando nuevas eficiencias, oportunidades y modelos de negocios. ThomasNet.com provee un directorio en línea actualizado de más de 600 000 proveedores de productos industriales, como químicos, metales, plásticos, goma y equipo automotriz. Antes conocida como Thomas Register, la compañía solía enviar enormes catálogos en papel con esta información y ahora la provee a los usuarios en línea a través de su sitio Web, gracias a lo cual se ha convertido en una compañía más pequeña y eficaz.

Otras compañías han creado empresas totalmente nuevas con base en el acceso a bases de datos extensas a través de Web. Un ejemplo de esto es el sitio de redes sociales MySpace, que ayuda a los usuarios a permanecer conectados entre sí o conocer nuevas

FIGURA 6-14 VINCULACIÓN DE BASES DE DATOS INTERNAS A WEB



Los usuarios acceden a la base de datos interna de una organización a través de Web, por medio de sus PCs de escritorio y el software de navegador Web.

personas. MySpace incluye música, comedia, videos y “perfiles” con información suministrada por 122 millones de usuarios sobre su edad, ciudad natal, preferencias en sus citas, estado civil e intereses. Mantiene una base de datos masiva para alojar y administrar todo su contenido. Facebook utiliza una base de datos similar.

6.4

ADMINISTRACIÓN DE LOS RECURSOS DE DATOS

El establecimiento de una base de datos es sólo el principio. Para poder asegurar que los datos para su empresa sigan siendo precisos, confiables y estén disponibles de inmediato para aquellos que los necesiten, necesitará políticas y procedimientos especiales para la administración de datos.

ESTABLECIMIENTO DE UNA POLÍTICA DE INFORMACIÓN

Toda empresa, ya sea grande o pequeña, necesita una política de información. Los datos de su empresa son un recurso importante, por lo que no es conveniente que las personas hagan lo que quieran con ellos. Necesita tener reglas sobre la forma en que se van a organizar y mantener los datos, y quién tiene permitido verlos o modificarlos.

Una **política de información** es la que especifica las reglas de la organización para compartir, diseminar, adquirir, estandarizar, clasificar e inventariar la información. La política de información establece procedimientos y rendiciones de cuentas específicos, identifica qué usuarios y unidades organizacionales pueden compartir información, en dónde distribuirla y quién es responsable de actualizarla y mantenerla. Por ejemplo, una política de información típica especificaría que sólo los miembros selectos del departamento de nómina y recursos humanos tendrían el derecho de modificar y ver los datos confidenciales de los empleados, como el salario o número de seguro social de un empleado, y que estos departamentos son responsables de asegurar que los datos de cada empleado sean precisos.

Si usted está en una empresa pequeña, los propietarios o gerentes son los que establecerían e implementarían la política de información. En una organización grande, administrar y planificar la información como un recurso corporativo requiere con frecuencia de una función de administración de datos formal. La **administración de datos** es responsable de las políticas y procedimientos específicos a través de los cuales se pueden administrar los datos como un recurso organizacional. Estas responsabilidades abarcan el desarrollo de la política de información, la planificación de los datos, la supervisión del diseño lógico de la base de datos, y el desarrollo del diccionario de datos, así como el proceso de monitorear la forma en que los especialistas de sistemas de información y los grupos de usuarios finales utilizan los datos.

Tal vez escuche que se utiliza el término **gobernanza de datos** para describir muchas de estas actividades. La gobernanza de datos es promovida por IBM y se encarga de las políticas y procedimientos para administrar la disponibilidad, utilidad, integridad y seguridad de los datos empleados en una empresa, con un énfasis especial en promover la privacidad, seguridad, calidad de los datos y el cumplimiento con las regulaciones gubernamentales.

Una organización grande también debe tener un grupo de diseño y administración de bases de datos dentro de la división de sistemas de información corporativos que sea responsable de definir y organizar la estructura y el contenido de la base de datos, y de darle mantenimiento. En una estrecha cooperación con los usuarios, el grupo de diseño establece la base de datos física, las relaciones lógicas entre los elementos, las reglas de acceso y los procedimientos de seguridad. Las funciones que desempeña se denominan **administración de la base de datos**.

ASEGURAMIENTO DE LA CALIDAD DE LOS DATOS

Una base de datos y una política de información bien diseñadas son un gran avance en cuanto a asegurar que la empresa tenga la información que necesita. Sin embargo, hay

que llevar a cabo ciertas acciones adicionales para asegurar que los datos en las bases de datos organizacionales sean precisos y permanezcan confiables.

¿Qué ocurriría si el número telefónico o el saldo de la cuenta de un cliente fueran incorrectos? ¿Cuál sería el impacto si la base de datos tuviera el precio incorrecto para el producto que usted vendió, o si su sistema de ventas y de inventario mostraran distintos precios para el mismo producto? Los datos imprecisos, inoportunos o inconsistentes con otras fuentes de información conducen a decisiones incorrectas, llamadas a revisión de los productos y pérdidas financieras. Los datos imprecisos en las bases de datos de justicia criminal y seguridad nacional podrían incluso someterlo a una vigilancia o detención innecesaria, como se describe en el caso de estudio al final del capítulo.

De acuerdo con Forrester Research, se devolvió el 20 por ciento de las entregas de paquetes de correo y comerciales en Estados Unidos debido a datos incorrectos en los nombres o las direcciones. Gartner Inc. informó que más del 25 por ciento de los datos críticos en las extensas bases de datos de las compañías Fortune 1000 son imprecisos o incompletos, incluyendo los códigos erróneos de productos y sus descripciones, las descripciones incorrectas en el inventario, los datos financieros erróneos, la información incorrecta de los proveedores y los datos erróneos de los empleados (Gartner, 2007).

Piense en todos los momentos que ha recibido varias piezas de la misma publicidad directa por correo el mismo día. Es muy probable que esto sea el resultado de que su nombre se repita varias veces en una base de datos. Tal vez lo hayan escrito mal o haya utilizado la inicial de su segundo nombre en una ocasión y en otra no, o quizás en un principio la información se capturó en un formulario en papel y no se digitalizó de manera apropiada para introducirlo al sistema. Debido a estas inconsistencias, ¡la base de datos lo consideraría como si fueran distintas personas! Nosotros a menudo recibimos correo redundante dirigido a Laudon, Lavdon, Lauden o Landon.

Si una base de datos está diseñada en forma apropiada y hay estándares de datos establecidos a nivel empresarial, los elementos de datos duplicados o inconsistentes deben reducirse al mínimo. Sin embargo, la mayoría de los problemas de calidad de los datos, como los nombres mal escritos, los números transpuestos y los códigos incorrectos o faltantes, se derivan de los errores durante la captura de los datos. La incidencia de dichos errores aumenta a medida que las compañías pasan sus negocios a Web y permiten que los clientes y proveedores introduzcan datos en sus sitios Web para actualizar de manera directa los sistemas internos.

Antes de implementar una nueva base de datos, las organizaciones necesitan identificar y corregir sus datos incorrectos y establecer mejores rutinas para editar los datos una vez que su base esté en operación. Con frecuencia, el análisis de la calidad de los datos empieza con una **auditoría de calidad de los datos**, la cual es una encuesta estructurada de la precisión y el nivel de su integridad en un sistema de información. Las auditorías de calidad de los datos se pueden realizar mediante la inspección de los archivos de datos completos, la inspección de muestras provenientes de los archivos de datos, o mediante encuestas a los usuarios finales sobre sus percepciones en cuanto a la calidad de los datos.

La **limpieza de datos**, conocida también en inglés como *data scrubbing*, consiste en actividades para detectar y corregir datos en una base que sean incorrectos, incompletos, que tengan un formato inapropiado o que sean redundantes. La limpieza de datos no sólo corrige los errores, sino que también impone la consistencia entre los distintos conjuntos de datos que se originan en sistemas de información separados. El software de limpieza de datos especializado está disponible para inspeccionar los archivos de datos de manera automática, corregir errores en los datos e integrarlos en un formato consistente a nivel de toda la compañía.

Los problemas de calidad de los datos no son sólo problemas de negocios, también representan serios problemas para los individuos, en cuanto a que afectan su condición financiera e incluso sus empleos. La Sesión interactiva sobre organización describe algunos de estos impactos, ya que detalla los problemas de calidad de los datos que se encuentran en las compañías que recolectan e informan sobre los datos de crédito de los consumidores. Cuando lea este caso analice los factores de administración, organización y tecnología detrás de este problema, y si las soluciones existentes son adecuadas o no.

SESIÓN INTERACTIVA: ORGANIZACIONES

ERRORES DEL BURÓ DE CRÉDITO: GRANDES PROBLEMAS DE LA GENTE

Acaba de encontrar el auto de sus sueños. Cuenta con un buen trabajo y suficiente dinero para el enganche. Todo lo que necesita es un préstamo por \$14 000. Tiene unas cuantas facturas de tarjetas de crédito, que paga con diligencia cada mes. Pero cuando solicita el préstamo, se lo rechazan. Cuando pregunta por qué, le dicen que tiene un préstamo vencido de un banco del cual nunca había escuchado antes. Acaba de convertirse en una de las millones de víctimas de los datos imprecisos u obsoletos en los sistemas de información de los burós de crédito.

La mayoría de los datos en los historiales de crédito de los consumidores en Estados Unidos se recolectan y mantienen a través de tres agencias de informes crediticios nacionales: Experian, Equifax y TransUnion. Estas organizaciones recolectan datos de varias fuentes para crear un expediente detallado de los hábitos de préstamos y pagos de un individuo. Esta información ayuda a los prestamistas a evaluar la capacidad crediticia de una persona, la capacidad de pagar un préstamo y puede afectar en la tasa de intereses y otros términos de un préstamo, como el hecho de si se puede otorgar o no un préstamo. Incluso puede afectar en la probabilidad de encontrar o mantener un empleo: por lo menos una tercera parte de los empleadores verifican los informes crediticios al tomar decisiones de contratación, despido o promoción.

Los burós de crédito en Estados Unidos recolectan información personal y datos financieros de una variedad de fuentes, entre ellos acreedores, prestamistas, empresas de servicios públicos, agencias de recolección de deudas y las cortes. Estos datos se agregan y almacenan en bases de datos masivas, cuyo mantenimiento está a cargo de los burós de crédito. A su vez, éstos venden la información a otras empresas para que evalúen los créditos.

Los burós de crédito afirman que saben qué tarjetas de crédito están en la cartera de cada cliente, cuánto deben de hipoteca y si la factura eléctrica se paga o no a tiempo. No obstante, si llega la información incorrecta a sus sistemas, ya sea por medio del robo de identidad o por los errores transmitidos por los acreedores, ¡tenga cuidado! Desenmarañar el enredo puede ser casi imposible.

Los burós comprenden la importancia de proporcionar información precisa tanto a los acreedores como a los consumidores. Pero también reconocen que sus propios sistemas son responsables de muchos errores en los informes crediticios. Algunos de estos errores ocurren debido a los procedimientos para asociar los préstamos con los informes crediticios individuales.

El volumen total de la información que se transmite de los acreedores a los burós de crédito incrementa la probabilidad de cometer errores. Por ejemplo,

Experian actualiza 30 millones de reportes de crédito a diario y alrededor de 2 mil millones de reportes de crédito al mes. Asocia la información de identificación personal en una solicitud o cuenta de crédito con la información de identificación personal en el archivo de crédito de un consumidor. La información de identificación personal contiene elementos como el nombre (primer nombre, apellido e inicial del segundo nombre), la dirección completa actual y el código postal, la dirección completa anterior y el código postal, y el número de seguro social. La nueva información de crédito pasa al archivo de crédito del consumidor que tenga la mejor coincidencia.

Los burós de crédito raras veces reciben información que coincida en todos los campos de los archivos de crédito, por lo que tienen que determinar cuánta variación permitir para poder seguirla considerando como coincidencia. Los datos imperfectos conducen a coincidencias no perfectas. Un consumidor podría proveer información incompleta o imprecisa en una solicitud de crédito. Un acreedor podría enviar información incompleta o imprecisa a los burós de crédito. Si la persona incorrecta coincide mejor que cualquier otra, los datos podrían por desgracia pasar a la cuenta incorrecta.

Tal vez el consumidor no escribió con claridad en la solicitud de la cuenta. Las variaciones en los nombres de las distintas cuentas de crédito también pueden producir coincidencias imperfectas. Considere como ejemplo el nombre Edward Jeffrey Johnson. Una cuenta podría decir Edward Jeffrey Johnson. Otra podría decir Ed Johnson. Otra más podría decir Edward J. Johnson. Suponga que los dos últimos dígitos del número de seguro social de Edward se transponen: hay más probabilidad de errores en la coincidencia.

Si el nombre o el número de seguro social en la cuenta de otra persona coinciden de manera parcial con los datos en su archivo, la computadora podría agregar los datos de esa persona a su registro. De igual forma, su registro podría corromperse si los trabajadores en las empresas que suministran datos fiscales y de bancarrota provenientes de los registros gubernamentales y de las cortes transponen de manera accidental un dígito o leen mal un documento.

Los burós de crédito afirman que es imposible para ellos monitorear la precisión de las 3.5 mil millones de piezas de información sobre las cuentas de crédito que reciben cada mes. Deben lidiar de manera continua con las reclamaciones falaces de los consumidores que falsifican información de las entidades crediticias o utilizan compañías sospechosas de reparación de crédito que desafian toda la información negativa en un informe de crédito, sin importar su validez. Para separar el bien del mal, los burós de crédito utilizan un sistema automatizado llamado e-OSCAR (solución electrónica en línea

para informes completos y precisos) para reenviar las disputas de los clientes a las entidades crediticias y que éstas las verifiquen.

Si su informe crediticio indica un error, los burós por lo general no se contactan de manera directa con la entidad crediticia para corregir la información. Para ahorrar dinero, los burós envían las protestas de los consumidores junto con la evidencia a un centro de procesamiento de datos operado por un contratista independiente. Estos contratistas sintetizan con rapidez cada queja con un breve comentario y un código de dos dígitos basado en un menú de 26 opciones. Por ejemplo, el código A3 indica que "pertenece a otro individuo con un nombre similar". Estos resúmenes son a menudo demasiado breves como para incluir los antecedentes que necesitan los bancos para comprender una queja.

Aunque este sistema corrige muchos errores (los datos se actualizan o corrigen en el 72 por ciento de las disputas), los consumidores tienen pocas opciones si el sistema falla. A los consumidores que presentan una segunda disputa sin proveer nueva información se les podría rechazar por ser "frívola". Si el consumidor trata de ponerse en contacto con la entidad crediticia que cometió el error por su cuenta, los bancos no tienen obligación de investigar la disputa: a menos que la envíe un buró de crédito.

Fuentes: Dennis McCafferty, "Bad Credit Could Cost You a Job", *Baseline*, 7 de junio de 2010; Kristen McNamara, "Bad Credit Derails Job Seekers", *The Wall Street Journal*, 16 de marzo de 2010; Anne Kadet, Lucy Lazarony, "Your Name Can Mess Up Your Credit Report", Bankrate.com, visitado el 1 de julio de 2009; "Credit Report Fix a Headache", *Atlanta Journal-Constitution*, 14 de junio de 2009, y "Why Credit Bureaus Can't Get It Right", *Smart Money*, marzo de 2009.

PREGUNTAS DEL CASO DE ESTUDIO

MIS EN ACCIÓN

1. Evalúe el impacto comercial de los problemas de la calidad de los datos de los burós de crédito para éstos, para los prestamistas y para los individuos.
2. ¿Se generan problemas éticos debido a los problemas en la calidad de los datos de los burós de crédito? Explique su respuesta.
3. Analice los factores de administración, organización y tecnología responsables de los problemas en la calidad de los datos de los burós de crédito.
4. ¿Qué se puede hacer para resolver estos problemas?

Vaya al sitio Web de Experian (www.experian.com) explórelo; ponga especial atención en sus servicios para empresas y negocios pequeños.

Después responda las siguientes preguntas:

1. Mencione y describa cinco servicios para negocios y explique cómo utiliza cada uno los datos de los consumidores. Describa los tipos de negocios que utilizarían estos servicios.
2. Explique cómo se ve afectado cada uno de estos servicios por los datos imprecisos de los consumidores.

6.5**PROYECTOS PRÁCTICOS SOBRE MIS**

Los proyectos en esta sección le proporcionan experiencia práctica para analizar los problemas de calidad de los datos, establecer estándares de datos a nivel de toda la compañía, crear una base de datos para administrar el inventario y utilizar Web para buscar recursos de negocios foráneos en las bases de datos en línea.

Problemas de decisión gerencial

1. Emerson Process Management, proveedor global de instrumentos y servicios de medición, análisis y monitoreo con base en Austin Texas, tenía un nuevo almacén de datos diseñado para analizar la actividad de los clientes y mejorar tanto el servicio como el proceso de marketing que estaba lleno de datos imprecisos y redundantes. Los datos en el almacén provenían de muchos sistemas de procesamiento de transacciones en Europa, Asia y otras ubicaciones alrededor del mundo. El equipo que diseñó el almacén supuso que los grupos de ventas en todas estas áreas introducirían los nombres y direcciones de los clientes de la misma forma, sin importar su ubicación. De hecho, las diferencias culturales combinadas con las complicaciones que se provocaban al absorber las compañías que Emerson había adquirido, condujeron a varias formas de introducir datos de cotizaciones, facturación, envíos y demás datos relacionados. Evalúe el impacto de negocios potencial de estos problemas de calidad de los datos. ¿Qué decisiones y acciones hay que tomar para llegar a una solución?
2. Su compañía proveedora industrial desea crear un almacén de datos en donde la gerencia pueda obtener una vista amplia a nivel corporativo de la información sobre las ventas críticas, para identificar los productos que se venden mejor en áreas geográficas específicas, los clientes clave y las tendencias de ventas. La información de sus ventas y productos se almacena en varios sistemas distintos: un sistema de ventas divisional que opera en un servidor Unix y uno corporativo de ventas que opera en una mainframe IBM. A usted le gustaría crear un solo formato estándar que consolide esos datos de ambos sistemas. Se ha propuesto el siguiente formato.

ID_PRODUCTO	DESCRIPCION_PRODUCTO	COSTO_POR_UNIDAD	UNIDADES_VENDIDAS	REGION_VENTAS	DIVISION	ID_CLIENTE

Los siguientes son archivos de ejemplo de los dos sistemas que proveerían la información para el almacén de datos:

SISTEMA CORPORATIVO DE VENTAS

ID_PRODUCTO	DESCRIPCION_PRODUCTO	COSTO_UNITARIO	UNIDADES_VENDIDAS	TERRITORIO_VENTAS	DIVISION
60231	Cojinete, 4"	5.28	900 245	Noreste	Piezas
85773	Unidad de montaje SS	12.45	992 111	Medio Oeste	Piezas

SISTEMA DE VENTAS DE LA DIVISIÓN DE PIEZAS MECÁNICAS

NUM_PROD	DESCRIPCION_PRODUCTO	COSTO_POR_UNIDAD	UNIDADES_VENDIDAS	REGION_VENTAS	ID_CLIENTE
60231	Cojinete de acero de 4"	5.28	900 245	N.E.	Anderson
85773	Unidad de montaje SS	12.45	992 111	M.O.	Kelly Industries

- ¿Qué problemas de negocios se crean al no tener estos datos en un solo formato estándar?
- ¿Qué tan fácil sería crear una base de datos con un solo formato estándar que pudiera almacenar los datos de ambos sistemas? Identifique los problemas con los que habría que lidiar.
- ¿Quiénes deben resolver los problemas, los especialistas de bases de datos o los gerentes generales de la empresa? Explique.
- ¿Quién debe tener la autoridad de finalizar un solo formato a nivel de toda la compañía para esta información en el almacén de datos?

Obtención de la excelencia operacional: creación de una base de datos relacional para la administración del inventario

Habilidades de software: diseño, consultas e informes de bases de datos

Habilidades de negocios: administración del inventario

Hoy en día las empresas dependen de las bases de datos para que les provean información confiable sobre los artículos en el inventario, sus costos y los que necesitan reabastecerse. En este ejercicio utilizará software de bases de datos para diseñar una base con la que se pueda administrar el inventario de una pequeña empresa.

La Tienda de Bicicletas de Sylvester, ubicada en San Francisco, California, vende bicicletas para camino regular, de montaña, híbridas, para paseo y para niños. En la actualidad, Sylvester compra bicicletas a tres proveedores pero planea agregar nuevos proveedores en un futuro cercano. Este negocio de rápido crecimiento necesita un sistema de bases de datos para administrar la información.

En un principio, la base de datos debe alojar información sobre proveedores y productos. Además, contener dos tablas: una de proveedores y una de productos. El nivel de reabastecimiento se refiere al número de artículos en el inventario que desencadena una decisión para pedir más artículos y evitar un desabastecimiento (en otras palabras, si el número de unidades de un artículo específico en el inventario disminuye a una cantidad inferior al nivel de reabastecimiento, hay que reabastecer el artículo). El usuario debe ser capaz de realizar varias consultas y producir diversos informes gerenciales con base en los datos que contienen las dos tablas.

Use la información que se incluye en las tablas en MyMISLab para crear una base de datos relacional simple que se pueda usar en la tienda de Sylvester. Una vez que cree la base de datos, realice las siguientes actividades:

- Prepare un informe que identifique las cinco bicicletas más costosas. El informe debe mostrar la lista de bicicletas en orden descendente, de la más costosa hasta la menos costosa, la cantidad en existencia de cada una y el porcentaje de ganancia en cada una de ellas.
- Prepare un informe que muestre una lista de cada proveedor, sus productos, las cantidades en existencia y los niveles de reabastecimiento asociados. El informe se debe ordenar en forma alfabética por proveedor. Dentro de cada categoría de proveedor hay que colocar los productos en orden alfabético.
- Prepare un informe que muestre una lista sólo de las bicicletas que tengan un nivel bajo de existencia y necesiten reabastecerse. El informe debe proveer información a los proveedores para los artículos identificados.
- Escriba una descripción breve de cómo se podría mejorar la base de datos para una administración más eficiente de la empresa. ¿Qué tablas o campos se deberían agregar? ¿Qué informes adicionales serían útiles?

Mejora de la toma de decisiones: uso de las bases de datos en línea para buscar recursos de negocios en el extranjero

Habilidades de software: bases de datos en línea

Habilidades de negocios: investigación de los servicios para operaciones en el extranjero

Los usuarios de Internet tienen acceso a muchos miles de bases de datos habilitadas para Web con información sobre servicios y productos en ubicaciones distantes. Este proyecto desarrolla habilidades en cuanto a cómo realizar búsquedas en estas bases de datos en línea.

Suponga que su compañía está ubicada en Greensboro, Carolina del Norte, y que fabrica muebles de oficina de diversos tipos. Hace poco adquirió varios nuevos clientes en Australia, y un estudio que comisionó indica que, si tuviera presencia ahí, podría incrementar de manera considerable sus ventas. Lo que es más, su estudio indica que obtendría más ganancias si empezara a fabricar muchos de sus productos en forma local (en Australia). En primer lugar, necesita abrir una oficina en Melbourne para establecer una presencia, y después empezar a importar de Estados Unidos. Después puede comenzar a producir en forma local.

Pronto viajará al área para planear el proceso de abrir una oficina, por lo que desea reunirse con organizaciones que le puedan ayudar con su operación. Tendrá que contactarse con personas u organizaciones que ofrezcan muchos de los servicios necesarios para que usted abra su oficina, como abogados, contadores, expertos en importación-exportación, equipo y soporte de telecomunicaciones, e incluso capacitadores que le puedan ayudar a preparar a sus futuros empleados a trabajar para usted. Empiece por buscar la recomendación del Departamento de Comercio de Estados Unidos acerca de cómo hacer negocios en Australia. Después pruebe las siguientes bases de datos en línea para localizar compañías con las que le gustaría reunirse durante su próximo viaje: el registro australiano de empresas (abr.business.gov.au/), Australia Trade Noe (australiatradenow.com/) y el directorio nacional de empresas de Australia (www.nationwide.com.au). Si es necesario, también podría probar los motores de búsqueda como Yahoo y Google. Después realice las siguientes actividades:

- Muestre una lista de compañías con las que quisiera ponerse en contacto para entrevistarlas en su viaje y determinar si le pueden ayudar con estas y otras funciones que piense que son vitales para establecer su oficina.
- Clasifique las bases de datos que utilizó en cuanto a la precisión en el nombre, integridad, facilidad de uso y utilidad en general.
- ¿Qué le indica este ejercicio acerca del diseño de las bases de datos?

MÓDULO DE TRAYECTORIAS DE APRENDIZAJE

Las siguientes Trayectorias de aprendizaje proporcionan contenido relevante a los temas que se cubrieron en este capítulo:

1. Diseño de bases de datos, normalización y diagramas entidad-relación
2. Introducción a SQL
3. Modelos de datos jerárquico y de red

Resumen de repaso

1. *¿Cuáles son los problemas de administrar los recursos de datos en un entorno de archivos tradicional y cómo se resuelven mediante un sistema de administración de bases de datos?*

Las técnicas tradicionales de administración de archivos dificultan a las organizaciones el proceso de llevar el registro de todas las piezas de datos que utilizan de una manera sistemática, y de organizarlos de modo que se pueda tener un fácil acceso a ellos. Se permitió a las distintas áreas y grupos funcionales desarrollar sus propios archivos en forma independiente. Con el tiempo, este entorno tradicional de administración de archivos crea problemas como la redundancia e inconsistencia de los datos, la dependencia programa-datos, inflexibilidad, mala seguridad, falta de compartición y disponibilidad de éstos. Un sistema de administración de bases de datos (DBMS) resuelve estos problemas mediante software que permite su centralización y administración, de modo que las empresas tengan una sola fuente consistente para todas sus necesidades de datos. El uso de un DBMS minimiza la cantidad de archivos redundantes e inconsistentes.

2. *¿Cuáles son las principales capacidades de los sistemas de administración de bases de datos (DBMS) y por qué es tan poderoso un DBMS?*

Las principales capacidades de un DBMS son: capacidad de definición de datos, capacidad de diccionario de datos y lenguaje de manipulación de datos. La capacidad de definición de datos especifica la estructura y el contenido de la base de datos. El diccionario de datos es un archivo automatizado o manual que almacena información sobre los datos en la base, entre estos, nombres, definiciones, formatos y descripciones de los elementos de datos. El lenguaje de manipulación de datos (como SQL) es un lenguaje especializado para acceder a los datos y manipularlos en la base.

La base de datos relacional es el método primario para organizar y dar mantenimiento a los datos en la actualidad en los sistemas de información, ya que es muy flexible y accesible. Organiza los datos en tablas bidimensionales conocidas como relaciones con filas y columnas. Cada tabla contiene información acerca de una entidad y sus atributos. Cada fila representa un registro y cada columna representa un atributo o campo. Cada tabla contiene también un campo clave para identificar de forma única cada registro para recuperarlo o manipularlo. Las tablas de las bases de datos relacionales se pueden combinar con facilidad para ofrecer los datos que requieren los usuarios, siempre y cuando dos tablas cualesquiera comparten un elemento de datos común.

3. *¿Cuáles son algunos principios importantes del diseño de bases de datos?*

Para diseñar una base de datos se requieren un diseño lógico y uno físico. El diseño lógico modela la base de datos desde una perspectiva de negocios. El modelo de datos de la organización debe reflejar sus procesos de negocios clave y los requerimientos para la toma de decisiones. El proceso de crear estructuras de datos pequeñas, estables, flexibles y adaptativas a partir de grupos complejos de datos al momento de diseñar una base de datos relacional se denomina normalización. Una base de datos relacional bien diseñada no debe tener relaciones de varios a varios, y todos los atributos para una entidad específica sólo se aplican a esa entidad. Esta base de datos trata de imponer las reglas de integridad referencial para asegurar que las relaciones entre tablas acopladas permanezcan consistentes. Un diagrama entidad-relación describe de forma gráfica la relación entre las entidades (tablas) en una base de datos relacional.

4. *¿Cuáles son las principales herramientas y tecnologías para acceder a la información de las bases de datos y mejorar tanto el desempeño de negocios como la toma de decisiones?*

Hay poderosas herramientas disponibles para analizar y acceder a la información en las bases de datos. Un almacén de datos consolida los datos actuales e históricos de muchos sistemas operacionales distintos en una base central diseñada para generar informes y realizar análisis. Los almacenes de datos soportan el análisis de datos multidimensional, también conocido como procesamiento analítico en línea (OLAP). El OLAP representa las relaciones entre los datos como una estructura multidimensional, que se puede visualizar en forma de cubos de datos y cubos dentro de cubos de datos, con lo cual se permite un análisis más sofisticado. La minería de datos analiza grandes reservas de datos, incluyendo el contenido de los almacenes de datos, para encontrar patrones y reglas que se puedan utilizar para predecir el comportamiento en un futuro y guiar la toma de decisiones. Las herramientas de minería de datos ayudan a las empresas a analizar extensos conjuntos de datos no estructurados que consisten en texto. Las herramientas de minería de datos se enfocan en el análisis de patrones e información útiles provenientes de World Wide Web; examinan la estructura de los sitios Web y las actividades de los usuarios de esos sitios Web, así como el contenido de las páginas Web. Las bases de datos convencionales se pueden vincular mediante middleware a Web o a una interfaz Web para facilitar el acceso de un usuario a los datos internos de la organización.

5. ¿Por qué son la política de información, la administración de datos y el aseguramiento de la calidad de los datos esenciales para administrar los recursos de datos de la empresa?

Para desarrollar un entorno de bases de datos se requieren políticas y procedimientos que ayuden a administrar los datos organizacionales, así como un buen modelo de datos y una tecnología de bases de datos eficiente. Una política de información formal gobierna el mantenimiento, la distribución y el uso de la información en la organización. En las grandes corporaciones, una función de administración de datos formal es responsable de la política de la información, así como de la planificación de los datos, el desarrollo del diccionario de datos y el monitoreo del uso de los datos en la empresa.

Los datos imprecisos, incompletos o inconsistentes crean serios problemas operacionales y financieros para las empresas, ya que pueden crear imprecisiones en los precios de los productos, las cuentas de los clientes y los datos del inventario, además de que conducen a decisiones imprecisas sobre las acciones que debe tomar la empresa. Las empresas deben realizar acciones especiales para asegurarse de tener un alto nivel de calidad en la información. Estas acciones incluyen el uso de estándares de datos a nivel empresarial, bases de datos diseñadas para minimizar los datos inconsistentes y redundantes, auditorías de calidad de los datos y software de limpieza de datos.

Términos clave

<i>Administración de la base de datos</i> , 230	<i>Entidad</i> , 210
<i>Administración de datos</i> , 230	<i>Gobernanza de datos</i> , 230
<i>Almacén de datos</i> , 222	<i>Inconsistencia de los datos</i> , 211
<i>Análisis predictivo</i> , 226	<i>Integridad referencial</i> , 220
<i>Archivo</i> , 210	<i>Lenguaje de consulta estructurado (SQL)</i> , 217
<i>Atributo</i> , 210	<i>Lenguaje de manipulación de datos</i> , 217
<i>Auditoría de calidad de los datos</i> , 231	<i>Limpieza de datos</i> , 231
<i>Base de datos</i> , 210	<i>Mercado de datos</i> , 223
<i>Base de datos (definición rigurosa)</i> , 212	<i>Minería de datos</i> , 224
<i>Campo</i> , 210	<i>Minería de texto</i> , 226
<i>Campo clave</i> , 214	<i>Minería Web</i> , 226
<i>Clave foránea</i> , 215	<i>Normalización</i> , 219
<i>Clave primaria</i> , 214	<i>Política de información</i> , 230
<i>DBMS objeto-relacional</i> , 215	<i>Procesamiento Analítico en Línea (OLAP)</i> , 224
<i>DBMS orientado a objetos</i> , 215	<i>Redundancia de los datos</i> , 211
<i>DBMS relacional</i> , 213	<i>Registro</i> , 214
<i>Definición de datos</i> , 217	<i>Servidor de bases de datos</i> , 229
<i>Dependencia programa-datos</i> , 211	<i>Sistema de Administración de Bases de Datos (DBMS)</i> , 212
<i>Diagrama entidad-relación</i> , 221	<i>Tupla</i> , 214
<i>Diccionario de datos</i> , 217	

Preguntas de repaso

- 1.** ¿Cuáles son los problemas de administrar los recursos de datos en un entorno de archivos tradicional y cómo se resuelven mediante un sistema de administración de bases de datos?
 - Mencione y describa cada uno de los componentes en la jerarquía de datos.
 - Defina y explique el significado de entidades, atributos y campos clave.
 - Mencione y describa los problemas del entorno tradicional de archivos.
 - Defina una base de datos y un sistema de administración de bases de datos; describa cómo resuelve los problemas de un entorno tradicional de archivos.

- 2.** ¿Cuáles son las principales capacidades de los sistemas de administración de bases de datos (DBMS) y por qué es tan poderoso un DBMS?
 - Nombre y describa con brevedad las capacidades de un DBMS.
 - Defina un DBMS relacional y explique cómo organiza los datos.
 - Mencione y describa las tres operaciones de un DBMS relacional.

- 3.** ¿Cuáles son algunos principios importantes del diseño de bases de datos?
 - Defina y describa la normalización y la integridad referencial; explique cómo contribuyen a una base de datos relacional bien diseñada.
 - Defina y describa un diagrama entidad-relación; explique su función en el diseño de bases de datos.

- 4.** ¿Cuáles son las principales herramientas y tecnologías para acceder a la información de las bases de datos y mejorar tanto el desempeño de negocios como la toma de decisiones?
 - Defina un almacén de datos; explique cómo funciona y cómo beneficia a las organizaciones.
 - Defina inteligencia de negocios y explique cómo se relaciona con la tecnología de bases de datos.
 - Describa las capacidades del procesamiento analítico en línea (OLAP).
 - Defina minería de datos; describa cómo difiere de OLAP y los tipos de información que proporciona.
 - Explique cómo difieren la minería de texto y la minería Web de la minería de datos convencional.
 - Describa cómo pueden los usuarios acceder a la información de las bases de datos internas de una compañía por medio de Web.

- 5.** ¿Por qué son la política de información, la administración de datos y el aseguramiento de la calidad de los datos esenciales para administrar los recursos de datos de la empresa?
 - Describa los roles de la política de la información y la administración de datos en cuanto a la administración de la información.
 - Explique por qué son esenciales las auditorías de calidad de los datos y su limpieza.

Preguntas de debate

- 1.** Se ha dicho que no es necesario el software de administración de bases de datos para crear un entorno de bases de datos. De su opinión al respecto.

- 2.** ¿En qué grado deben estar involucrados los usuarios finales en la selección de un sistema de administración de bases de datos y del diseño de la base de datos?

Colaboración y trabajo en equipo: identificación de las entidades y atributos en una base de datos en línea

Con su equipo de tres o cuatro estudiantes, seleccione una base de datos en línea para explorar, como AOL Music, iGo.com o Internet Movie Database (IMDb). Explore uno de estos sitios Web para ver qué información proporciona. Después haga una lista de las entidades y atributos que la compañía operadora del sitio Web debe registrar en sus bases de datos. Haga un diagrama

- 3.** ¿Cuáles son las consecuencias de que una organización no tenga una política de información?

de la relación entre las entidades que identifique. Si es posible, use Google Sites para publicar vínculos a páginas Web, anuncios de comunicación en equipo y asignaturas de trabajo; para lluvias de ideas, y para trabajar de manera colaborativa en los documentos del proyecto. Intentar usar Google Docs para desarrollar una presentación de sus hallazgos para la clase.

Los problemas de la base de datos de vigilancia de terroristas continúan

CASO DE ESTUDIO

Después de los ataques del 9-11, se estableció el Centro de Detección de Terroristas (TSC) del FBI para consolidar la información sobre los terroristas sospechosos de varias agencias gubernamentales en una sola lista para mejorar la comunicación entre las agencias. En ese entonces se creó una base de datos de terroristas sospechosos conocida como la lista de vigilancia de terroristas. Varias agencias gubernamentales de Estados Unidos habían estado manteniendo listas separadas y carecían de un proceso consistente para compartir información relevante.

Los registros en la base de datos TSC contienen información confidencial pero no clasificada sobre las identidades de los terroristas, como el nombre y la fecha de nacimiento, que se pueden compartir con otras agencias de detección. La información clasificada sobre las personas en la lista de vigilancia se mantiene en otras bases de datos de agencias policiales y de la agencia de inteligencia. Los registros de la base de datos de la lista de vigilancia se proveen a través de dos fuentes: el Centro Nacional Antiterrorista (NCTC) administrado por la oficina del director de inteligencia nacional provee información de identificación sobre individuos que tienen lazos con el terrorismo internacional. El FBI provee información de identificación sobre los individuos que tienen lazos con el terrorismo puramente nacional.

Estas agencias recolectan y mantienen información de los terroristas y nominan individuos para incluirlos en la lista de vigilancia consolidada del TSC. Tienen que seguir estrictos procedimientos establecidos por el jefe de la agencia correspondiente y deben ser aprobados por el ministro de justicia de Estados Unidos. El personal del TSC debe revisar cada registro enviado antes de agregarlo a la base de datos. Un individuo permanecerá en la lista de vigilancia hasta que el departamento o agencia correspondiente que nominó a esa persona para la lista determine que ésta se debe quitar de ella y eliminar de la base de datos.

La base de datos de la lista de vigilancia del TSC se actualiza a diario con nuevas nominaciones, modificaciones a los registros existentes y eliminaciones. Desde su creación, la lista creció de manera explosiva hasta llegar a 400 000 personas, registradas como 1.1 millones de nombres y alias, y sigue creciendo a una proporción de 200 000 registros por año. La información en la lista se distribuye a un amplio rango de sistemas de agencias gubernamentales para usarse en los esfuerzos por impedir o detectar los movimientos de los terroristas conocidos o presuntos.

Las agencias que reciben la lista son: FBI, CIA, Agencia de Seguridad Nacional (NSA), Administración de Seguridad en el Transporte (TSA), Departamento de Seguridad Nacional, Departamento de Estado, Aduanas y Protección Fronteriza, Servicio Secreto, Servicio de

Alguaciles Federales de Estados Unidos y la Casa Blanca. Las aerolíneas utilizan los datos suministrados por el sistema TSA en sus listas NoFly y Selectee para investigar previamente a los pasajeros, mientras que el Sistema de Aduanas y Protección Fronteriza de Estados Unidos utiliza los datos de la lista de vigilancia para ayudar a investigar a los viajeros que entran al país. El sistema del Departamento de Estado investiga a los que solicitan visas para entrar a Estados Unidos y a los residentes de que solicitan pasaportes, mientras que las agencias policiales estatales y locales usan el sistema del FBI para que les ayude con los arrestos, detenciones y otras actividades de justicia criminal. Cada una de estas agencias recibe el subconjunto de datos en la lista de vigilancia pertinente a su misión específica.

Cuando una persona hace una reservación en una aerolínea, llega a un puerto de entrada, solicita una visa para Estados Unidos o es detenido por la policía estatal o local dentro de este país, la agencia de investigación de primera línea o la aerolínea realizan una búsqueda basada en el nombre del individuo para compararlo con los registros de la base de datos de la lista de vigilancia de terroristas. Cuando el sistema computarizado para relacionar los nombres genera una "ocurrencia" (una coincidencia potencial de un nombre) con un registro de la lista de vigilancia, la aerolínea o agencia revisarán cada coincidencia potencial. Las coincidencias que sean claramente positivas o las exactas que no sean concluyentes (inciertas o difíciles de verificar) se envían al centro de inteligencia o de operaciones de la agencia de investigación aplicable, y también al TSC para un análisis más detallado. A su vez, el TSC revisa sus bases de datos y otras fuentes, como las bases de datos clasificadas que mantienen el NCTC y el FBI para confirmar si el individuo es una coincidencia positiva, negativa o inconclusa para el registro de la lista de vigilancia. El TSC crea un informe diario en el que sintetiza todas las coincidencias positivas con la lista de vigilancia y las distribuye a las diversas agencias federales.

El proceso de consolidar la información de distintas agencias ha sido lento y minucioso, ya que se requiere integrar por lo menos 12 bases de datos distintas. Dos años después de que se llevó a cabo el proceso de integración, se habían procesado 10 de las 12 bases de datos. Las dos restantes (el Sistema de Identificación Biométrica Automática del Servicio de Inmigración y Control de Aduanas de Estados Unidos, y el Sistema Automático Integrado de Identificación de Huellas Dactilares de Estados Unidos) son bases de datos de huellas dactilares. Aún queda más trabajo por realizar para optimizar la utilidad de la lista.

Los informes de la oficina de auditoría general y la oficina del inspector general aseguran que la lista con-

tiene imprecisiones y que las políticas departamentales del gobierno para nominar y quitar personas de la lista no son uniformes. También se ha generado una protesta pública debido al tamaño de la lista y los incidentes tan publicitados de personas que sin duda no son terroristas y descubren que se encuentran en la lista.

Para que la lista sea efectiva contra los terroristas, es necesario proteger con cuidado la información sobre el proceso de inclusión en ella. Los criterios específicos de inclusión no son del conocimiento público. Sin embargo, sabemos que para llenar sus listas de vigilancia, las agencias gubernamentales realizan rastreos amplios de información recopilada sobre los viajeros, en donde utilizan palabras mal escritas y variaciones alternativas de los nombres de los terroristas sospechosos. Esto a menudo provoca que se ingresen personas que no pertenecen a las listas de vigilancia, conocidas como "falsos positivos". También ocasiona que algunas personas aparezcan varias veces en la lista con sus nombres escritos de distintas formas.

Aunque estos criterios de selección pueden ser efectivos para rastrear tantos terroristas potenciales como sea posible, también provocan muchas más entradas erróneas en la lista de las que se generaría si el proceso requiriera información mucho más detallada para agregar nuevas entradas. Algunos ejemplos notables de 'falsos positivos' son; Michael Hicks, niño explorador de Nueva Jersey de ocho años, a quien detenían con frecuencia en el aeropuerto para una investigación adicional, y el fallecido senador Ted Kennedy, quien sufrió varios retrasos en el pasado debido a que su nombre se asemeja a un alias que alguna vez utilizó un presunto terrorista. Al igual que Kennedy, tal vez se haya agregado a Hicks debido a que su nombre es igual o similar al de otro presunto terrorista.

Estos incidentes cuestionan la calidad y la precisión de los datos en la lista de vigilancia de terroristas consolidada del TSC. En junio de 2005, un informe de la oficina del inspector general del Departamento de Justicia encontró cuentas de registros inconsistentes, registros duplicados y registros que carecían de campos de datos o tenían fuentes dudosas para sus datos. Aunque después el TSC mejoró sus esfuerzos por identificar y corregir los registros incompletos o imprecisos en la lista de vigilancia, el inspector general observó en septiembre de 2007 que la administración de la lista de vigilancia por parte del TSC aún mostraba ciertas debilidades.

Dada la opción entre una lista que detecte a todos los terroristas potenciales a costa de rastrear de manera innecesaria a algunos inocentes, y entre una lista que no pueda identificar a muchos terroristas por hacer el esfuerzo de evitar rastrear a los inocentes, muchos elegirían la lista que reconoce a todos los terroristas a pesar de las desventajas. Sin embargo, para empeorar las cosas para aquellos que ya sufren de la inconveniencia de haber sido incluidos de manera errónea en la lista, en la actualidad no hay un proceso simple y rápido de rectificación para los inocentes que esperan ser removidos de la lista.

El número de solicitudes para quitar personas de la lista de vigilancia sigue en aumento, con más de 24 000 solicitudes registradas (una cifra aproximada de 2 000 al mes), y sólo el 54 por ciento de ellas resueltas. El tiempo promedio para procesar una solicitud en 2008 era de 40 días, lo cual no era (ni lo es en la actualidad) lo bastante rápido como para estar a la par con el número de solicitudes entrantes para quitar personas de la lista. Como resultado, los viajeros que cumplían con la ley y se encontraban de manera inexplicable en la lista de vigilancia no son quitados con facilidad de ella.

En febrero de 2007, el Departamento de Seguridad Nacional instituyó su programa de solicitud de rectificación para viajeros (TRIP) para ayudar a que las personas que se agregaron por equivocación a las listas de vigilancia de terroristas se puedan quitar por sí solas, para así evitar los procesos adicionales de investigación y cuestionamiento. La madre de John Anderson afirmó que a pesar de sus mejores esfuerzos, no pudo quitar a su hijo de las listas de vigilancia. Según se informa, el senador Kennedy sólo pudo quitarse de la lista al llevar personalmente el asunto a Tom Ridge, que en ese entonces era director del Departamento de Seguridad Nacional.

Los oficiales de seguridad dicen que los errores como el que provocó que Anderson y Kennedy se incluyeran en las listas de vigilancia de pasajeros prohibidos (no-fly) y consolidada ocurren debido a la coincidencia de datos imperfectos en los sistemas de reservación de las aerolíneas con los datos imperfectos en las listas de vigilancia. Muchas aerolíneas no incluyen género, segundo nombre o fecha de nacimiento en sus registros de reservaciones, lo cual incrementa la probabilidad de falsas coincidencias.

Una manera de mejorar la detección y de ayudar a reducir el número de personas marcadas de manera errónea para una investigación adicional sería utilizar un sistema más sofisticado que involucre datos más personales sobre los individuos en la lista. La TSA está desarrollando un sistema así, conocido como "Vuelo Seguro" (Secure Flight), pero se ha retrasado en forma continua debido a las cuestiones de privacidad relacionadas con la sensibilidad y seguridad de los datos que recolectaría. Otros programas y listas de vigilancia similares, como los intentos de la NSA de recopilar información sobre presuntos terroristas, han provocado críticas por violaciones potenciales a la privacidad.

Además, la lista de vigilancia ha provocado críticas debido a su potencial de promover los perfiles y la discriminación racial. Algunos alegan que fueron incluidos por virtud de su raza y descendencia étnica, como David Fathi, un abogado para la ACLU de descendencia iraní, y Asif Iqbal, un ciudadano estadounidense de descendencia pakistaní con el mismo nombre que un detenido en Guantánamo. Los críticos abiertos de la política foránea de Estados Unidos, como algunos oficiales electos y profesores universitarios, también se han encontrado dentro de la lista.

Un informe liberado en mayo de 2009 por Glenn A. Fine, inspector general del Departamento de Justicia,

encontró que el FBI había mantenido de manera incorrecta casi 24 000 personas en su propia lista de vigilancia que suministra datos a la lista de vigilancia de terroristas, con base en información obsoleta o irrelevante. Después de examinar casi 69 000 casos de la lista del FBI, el informe descubrió que el 35 por ciento de esas personas permanecían en la lista a pesar de una justificación inadecuada. Lo más preocupante aún es que la lista no contenía los nombres de las personas que deberían de haber sido incluidas debido a sus lazos con el terrorismo.

Los oficiales del FBI afirman que la agencia ha realizado mejoras, incluyendo una mejor capacitación y un procesamiento más rápido de los casos, además de requerir que los supervisores de las oficinas regionales revisen la precisión e integridad de las nominaciones a la lista de vigilancia. No obstante, esta lista de vigilancia y las otras siguen siendo herramientas imperfectas. A principios de 2008, se reveló que 20 terroristas conocidos no estaban incluidos de manera correcta en la lista de vigilancia consolidada (lo que no queda claro es si estos individuos pudieron entrar a Estados Unidos como consecuencia).

Umar Farouk Abdulmutallab, el nigeriano que no pudo detonar explosivos plásticos en el vuelo de Northwest Airlines de Amsterdam a Detroit durante el día de Navidad de 2009, no aparecía dentro de la lista de pasajeros prohibidos. Aunque el padre de Abdulmutallaab había

reportado que estaba preocupado por la radicalización de su hijo al Departamento de Estado de Estados Unidos, éste no revocó la visa de Abdulmutallab debido a que su nombre estaba mal escrito en la base de datos de visas, por lo que se le permitió entrar a Estados Unidos. Faisal Shahzad, el bombardero del auto en Times Square, fue aprehendido el 3 de mayo de 2010, sólo unos momentos antes de que su vuelo de la aerolínea Emirates hacia Dubai y Pakistán despegara. La aerolínea no había verificado una actualización de último minuto a la lista de pasajeros prohibidos en donde se había agregado el nombre de Shahzad.

Fuentes: Scott Shane, "Lapses Allowed Suspect to Board Plane", *The New York Times*, 4 de mayo de 2010; Mike McIntire, "Ensnared by Error on Growing U.S. Watch List", *The New York Times*, 6 de abril de 2010; Eric Lipton, Eric Schmitt y Mark Mazzetti, "Review of Jet Bomb Pilot Shows More Missed Clues", *The New York Times*, 18 de enero de 2010; Lizette Alvarez, "Meet Mikey, 8: U.S. Has Him on Watch List", *The New York Times*, 14 de enero de 2010; Eric Lichtblau, "Justice Dept. Finds Flaws in F.B.I. Terror List", *The New York Times*, 7 de mayo de 2009; Bob Egelko, "Watch-list Name Confusion Causes Hardship", *San Francisco Chronicle*, 20 de marzo de 2008; "Reports Cite Lack of Uniform Policy for Terrorist Watch List", *The Washington Post*, 18 de marzo de 2008; Siobhan Gorman, "NSA's Domestic Spying Grows as Agency Sweeps Up Data", *The Wall Street Journal*, 10 de marzo de 2008; Ellen Nakashima y Scott McCartney, "When Your Name is Mud at the Airport", *The Wall Street Journal*, 29 de enero de 2008.

PREGUNTAS DEL CASO DE ESTUDIO

1. ¿Qué conceptos en este capítulo se ilustran en este caso?
2. ¿Por qué se creó la lista de vigilancia de terroristas consolidada? ¿Cuáles son los beneficios de la lista?
3. Describa algunas de las debilidades de la lista de vigilancia. ¿Qué factores de administración, organización y tecnología son responsables de estas debilidades?
4. ¿Qué tan efectivo es el sistema de las listas de vigilancia descrito en este caso de estudio? Explique su respuesta.
5. Si fuera responsable de la administración de la base de datos de la lista de vigilancia del TSC, ¿qué acciones tomaría para corregir algunas de estas debilidades?
6. ¿Cree usted que la lista de vigilancia de terroristas representa una amenaza considerable para la privacidad o los derechos constitucionales de los individuos? ¿Por qué sí o por qué no?

