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

CAPÍTULO 6

OBJETIVOS DE APRENDIZAJE

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

1. ¿Cuáles son los problemas de administrar los recursos de datos en 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 relacional?
3. ¿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?
4. ¿Por qué la política de información, la administración de datos y el aseguramiento de la calidad de los datos son esenciales para administrar los recursos de datos de la empresa?

CASOS DEL CAPÍTULO

Una mejor administración de los datos ayuda a que Toronto Globe and Mail llegue a sus clientes

Impulso de la gestión de flotillas de ARI mediante análisis en tiempo real

American Water mantiene el flujo de los datos

¿Acaso Big Data trae consigo grandes recompensas?

CASOS EN VIDEO

Dubuque usa la computación en la nube y sensores para crear una ciudad más inteligente

Almacenes de datos en REI: comprender al cliente

Inteligencia de negocios y bases de datos empresariales de Maruti Suzuki

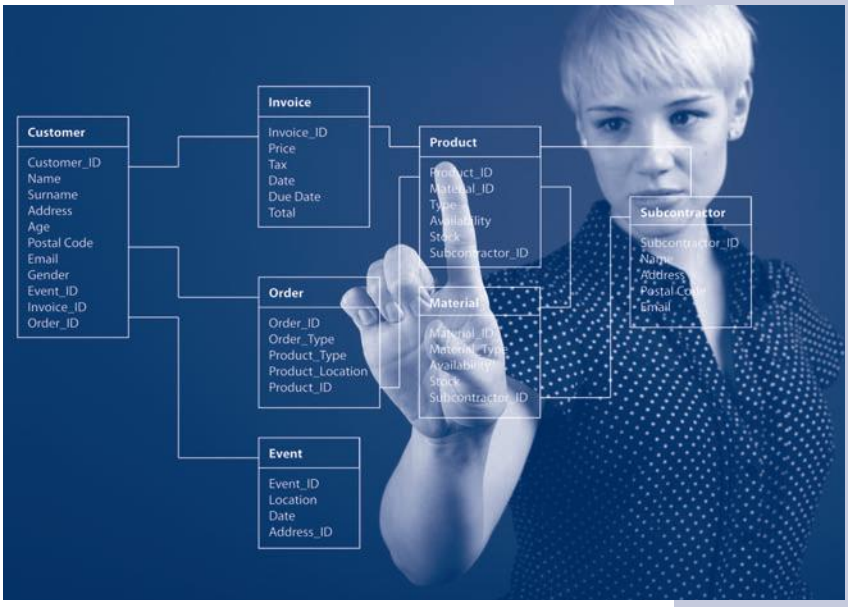
UNA MEJOR ADMINISTRACIÓN DE LOS DATOS AYUDA A QUE TORONTO GLOBE AND MAIL LLEGUE A SUS CLIENTES

¿Alguna vez ha recibido una nueva oferta de suscripción de un periódico o revista a la que ya está suscrito? Además de ser una molestia, enviar una oferta superflua a los clientes incrementa los costos de marketing. Entonces ¿por qué ocurre esto? La respuesta probable es debido a la mala administración de los datos. Es muy posible que el periódico no haya podido relacionar su lista de suscriptores existentes, a la cual mantiene en un lugar, con otro archivo que contenga su lista de prospectos de marketing.

The Globe and Mail, ubicado en Toronto, Canadá, era una de esas publicaciones que tenían estos problemas. Con un tiraje ininterrumpido durante 167 años, es el periódico más grande de Canadá, con una base de lectores acumulada de seis días de casi 3.3 millones. El periódico contaba con un programa de marketing bastante ambicioso, en el que veía como prospecto a cada uno de los hogares canadienses que no estaban ya inscritos. Pero había tenido problemas para alojar y gestionar los datos sobre estos prospectos.

Para operar un periódico importante se requiere administrar enormes cantidades de datos, incluyendo los datos de circulación, los de ingresos por publicidad, los datos de prospectos de marketing y los que "no deben contactarse", además de los datos de logística y entregas. Agregue a esto los datos requeridos para operar una empresa, como los datos financieros y de recursos humanos.

Durante muchos años The Globe and Mail alojó gran parte de sus datos en un sistema mainframe donde no era fácil usar y analizar los datos. Si los usuarios necesitaban información tenían que extraer los datos de la computadora mainframe y llevarlos a una de varias bases de datos locales para analizarlos, como las que se mantenían en Microsoft Access, Foxbase Pro y Microsoft Excel. Esta práctica generó numerosas concentraciones de datos que se mantenían en bases de datos aisladas para fines específicos, pero no había un repositorio central en el que se pudiera tener acceso a los datos más actualizados



© Semisatch/Shutterstock

desde un solo lugar. Con los datos esparcidos en tantos sistemas distintos en toda la empresa, era muy difícil contrastar los suscriptores con los prospectos a la hora de desarrollar la lista de correo para una campaña de marketing. También estaban las cuestiones de seguridad: The Globe and Mail recolecta y almacena la información de pago de los clientes; alojar estos datos confidenciales en varios lugares hace aún más difícil el poder asegurar que se implementen los controles de seguridad de datos correctos.

En 2002 el periódico comenzó a lidiar con estos problemas al implementar un sistema empresarial SAP con un almacén de datos SAP NetWeaver BW, el cual contendría todos los datos de la empresa provenientes de sus diversos orígenes de datos en una sola ubicación donde los usuarios de negocios pudieran acceder a ellos y analizarlos de una manera fácil.

Los primeros datos que ocuparon el almacén de datos fueron los de ventas por publicidad, que son una de las principales fuentes de ingresos. En 2007 The Globe and Mail agregó datos de circulación al almacén, incluyendo los detalles sobre los datos de entrega como el tiempo restante en la suscripción de un cliente y los datos sobre prospectos de marketing de fuentes independientes. También se agregaron al almacén los datos sobre los prospectos.

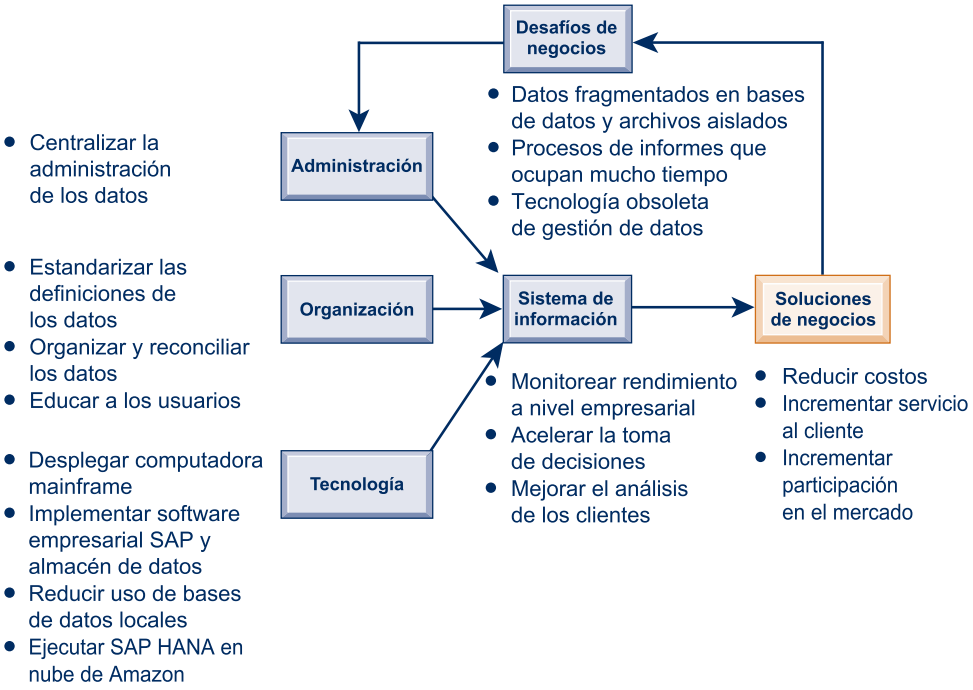
Con todos estos datos en un solo lugar, el periódico puede relacionar fácilmente los datos de los prospectos y de los clientes para no dirigirse a los clientes existentes con promociones de suscripción. También puede asociar los datos con los que "no deben contactarse" y los datos sobre el área de entregas para determinar si es posible entregar o no un periódico, o si hay que dirigirse a un cliente con una promoción de suscripción digital.

A pesar de los beneficios obvios del nuevo almacén de datos, no todos los usuarios de negocios de The Globe and Mail se incorporaron de inmediato. Las personas que solían extraer los datos del sistema mainframe y manipularlos en sus propias bases de datos o archivos siguieron haciendo lo mismo después de que el almacén de datos entró en operación. No entendían el concepto de un almacén de datos ni la necesidad de trabajar en torno a la gestión de datos a nivel empresarial. La gerencia de The Globe and Mail decidió atacar este nuevo problema educando a sus usuarios, en especial a los profesionales de marketing, con el valor de tener todos los datos de la organización en un almacén de datos y las herramientas disponibles para acceder a estos datos y analizarlos.

Las nuevas capacidades de análisis de datos de The Globe and Mail produjeron ahorros gracias a las eficiencias y los procesos modernizados que pagaron la inversión en un año. Las campañas de marketing que anteriormente tardaban dos semanas en completarse ahora sólo requieren un día. El periódico puede determinar sus tasas de saturación en cierta área para guiar sus planes de marketing. Y hay menos quejas de los suscriptores y suscriptores potenciales en cuanto a que se les contacte innecesariamente.

Para capitalizar aún más en cuanto a la gestión y el análisis de los datos, The Globe and Mail recurrió a la nube. Una meta de negocios clave para la empresa era reforzar el contenido en línea e incrementar la base de suscriptores digitales del periódico. The Globe and Mail dedicó más recursos al contenido en línea digital, con distintas tarifas de suscripción para los clientes que sólo accedían a través de Internet y para los clientes que recibían el periódico impreso. Para cortejar de manera agresiva a los suscriptores digitales, The Globe and Mail tenía que extraer sus datos sobre el flujo de clics que registraban las acciones del usuario en Web, para enfocarse en los potenciales suscriptores digitales con base no sólo en sus intereses específicos, sino también en sus intereses en un día en particular. El volumen de datos era demasiado grande como para que lo pudiera manejar la base de datos convencional Oracle de la empresa. La solución era usar el software de computación "en memoria" (in-memory) SAP HANA ONE y ejecutarlo en la plataforma de computación en la nube de Amazon Web Services, que acelera el análisis de datos y el procesamiento al almacenar los datos en la memoria principal de la computadora (RAM) en vez de hacerlo en dispositivos de almacenamiento externos. Esta solución en la nube permite a The Globe and Mail pagar sólo por las capacidades que usa cada hora.

Fuentes: www.theglobeandmail.com, visitado el 1 de marzo de 2014; "The Globe and Mail Uses SAP HANA in the Cloud to row Its Digital Audience", *SAP Insider Profiles*, 1 de abril de 2013, y David Hannon, "Spread the News", *SAP Insider Profiles*, octubre-diciembre de 2012.



La experiencia de The Globe and Mail ilustra la importancia de la administración de datos. El rendimiento de negocios depende de lo que una empresa puede o no hacer con sus datos. The Globe and Mail era una empresa grande y próspera, pero tanto la eficiencia operacional como la toma de decisiones de la gerencia se veían obstaculizadas por los datos fragmentados almacenados en varios sistemas a los que era difícil tener acceso. La forma en que las empresas almacenan, organizan y administran sus datos tiene un enorme impacto en la eficacia organizacional.

El diagrama de apertura del capítulo dirige la atención hacia los puntos importantes generados por este caso y por este capítulo. Los usuarios de negocios de The Globe and Mail mantenían sus propias bases de datos locales porque era muy difícil acceder a los datos de la empresa en el sistema mainframe tradicional del periódico. Las campañas de marketing tardaron más de lo necesario porque los datos requeridos tardaban mucho tiempo en ensamblarse. La solución fue consolidar los datos organizacionales en un almacén de datos de toda la empresa que proporcionara una sola fuente de datos para informes y análisis. El periódico tuvo que reorganizar sus datos en un formato estándar a nivel empresarial, establecer reglas, responsabilidades y procedimientos para acceder a los datos y usarlos, proporcionar herramientas para que los datos fueran accesibles y que los usuarios los utilizaran en consultas e informes, y educar a sus usuarios en cuanto a los beneficios del almacén.

El almacén de datos impulsó la eficiencia al facilitar la localización de los datos del Globe y el ensamble de los mismos para generar informes. El almacén de datos integró los datos de la empresa, de todas sus fuentes dispares hacia una sola base de datos exhaustiva que podía consultarse directamente. Se reconciliaron los datos para evitar errores, como contactar a los suscriptores existentes con ofertas de suscripción. La solución mejoró el servicio al cliente y al mismo tiempo redujo los costos. The Globe and Mail incrementó su capacidad de analizar con rapidez enormes cantidades de datos mediante el uso de SAP HANA que se ejecuta en el servicio en la nube de Amazon.

He aquí algunas preguntas a considerar: ¿cuál fue el impacto de negocios de los problemas de administración de datos de The Globe and Mail? ¿Qué trabajo tuvo que realizar tanto el personal de negocios como el técnico para asegurarse de que el almacén de datos produjera los resultados previstos por la gerencia?

6.1 ¿CUÁLES SON LOS PROBLEMAS DE ADMINISTRAR LOS RECURSOS DE DATOS EN UN ENTORNO TRADICIONAL DE ARCHIVOS?

Un sistema eficaz de información proporciona 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. Asimismo, es relevante cuando es útil y apropiada tanto para los tipos de trabajo 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. Esta es la razón por la que la administración de los datos es tan esencial. Para comprender el problema, veamos cómo 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 progresa 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 puede 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 de 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 sobre los historiales personales de los estudiantes y sus antecedentes financieros, para crear una base de datos de estudiantes.

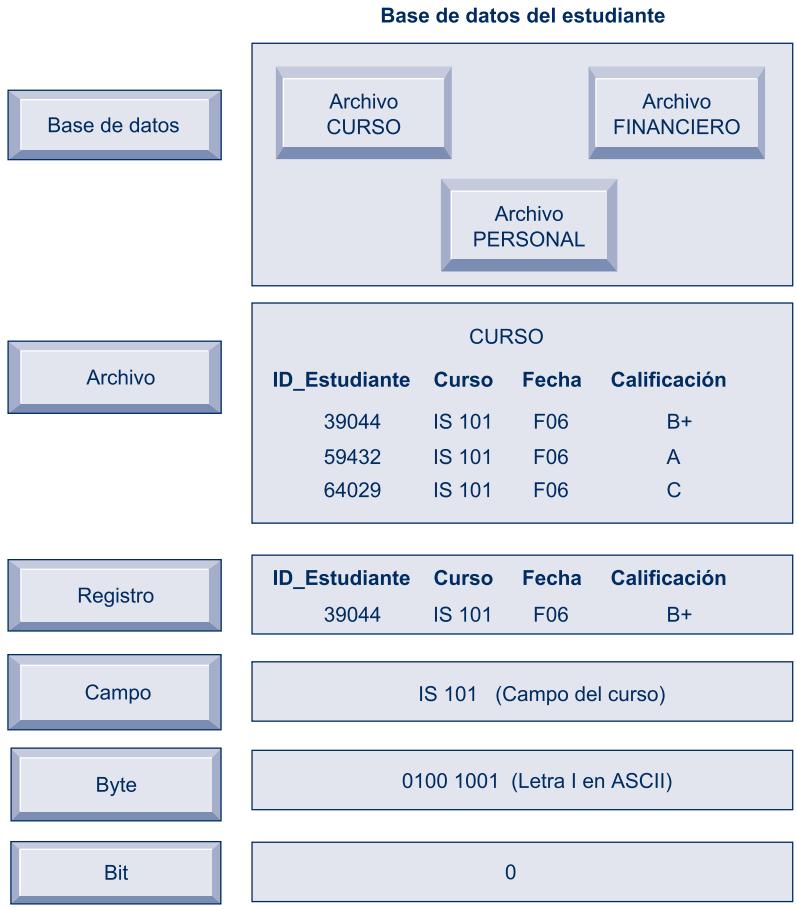
Un registro describe a una entidad. Una **entidad** es una persona, lugar, cosa o suceso sobre el cual almacenamos y mantenemos información. Cada característica o cualidad que describe a una entidad específica se denomina **atributo**. Por ejemplo, ID_Estudiante, Curso, Fecha y Calificaciones, son atributos 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 TRADICIONAL DE ARCHIVOS

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 desarrollaban sus propios sistemas y archivos de datos. La figura 6.2 ilustra la metodología normal para el procesamiento de la información.

Desde luego, cada aplicación requería sus propios archivos y su programa de cómputo para funcionar. 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, etc., hasta que hubiera decenas, tal vez cientos, de archivos y programas. En la compañía como un todo, este proceso condujo a varios archivos maestros creados, mantenidos y operados por divisiones o departamentos separados. A medida que este proceso avanza durante 5 o 10 años, la organización se satura con

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.

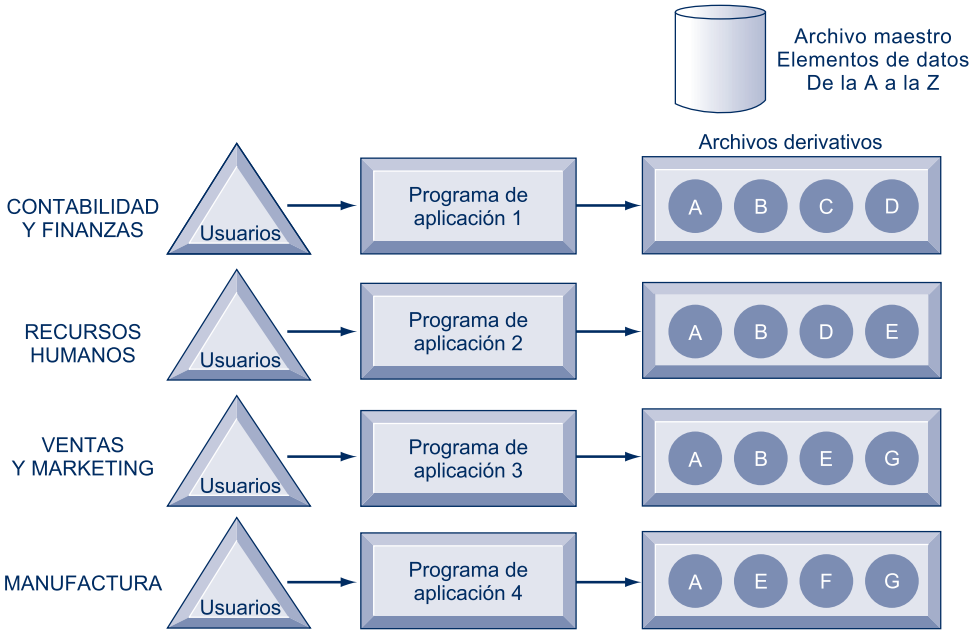
cientos de programas y aplicaciones que son muy difíciles de mantener y administrar. Los problemas resultantes son la redundancia e inconsistencia de los datos, la dependencia de programa-datos, la inflexibilidad, la seguridad defectuosa de los datos, y la incapacidad de compartir datos entre aplicaciones.

Redundancia e inconsistencia de los datos

La **redundancia de los datos** es la presencia de datos duplicados en varios archivos, de modo que los mismos datos se almacenan 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. La redundancia desperdicia recursos de almacenamiento y también conduce a la **inconsistencia de los datos**, 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 nombres diferentes 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

FIGURA 6.2 PROCESAMIENTO TRADICIONAL DE ARCHIVOS



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.

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 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 apropiadamente. La implementación apropiada de dichos cambios podría costar millones de dólares.

Falta de flexibilidad

Un sistema tradicional de archivos puede entregar informes programados de rutina después de extensos esfuerzos 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. Quizá varios programadores tengan que trabajar durante semanas para reunir los elementos de datos requeridos en un 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 archivos distintos y 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 distintas áreas funcionales o partes de la organización. 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 ¿Cuáles son las principales capacidades de los sistemas de administración de bases de datos (DBMS) y por qué es tan poderoso un DBMS relacional?

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 un conjunto de datos organizados para dar servicio de manera eficiente a muchas aplicaciones 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.

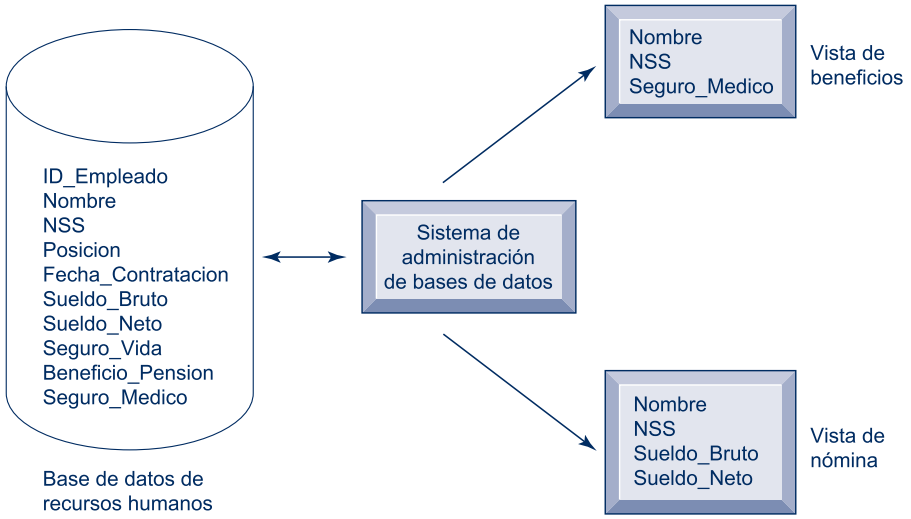
SISTEMAS DE ADMINISTRACIÓN DE BASES DE DATOS

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, por ejemplo el sueldo bruto, el DBMS lo busca en la base de datos y lo presenta al programa de aplicación. Si utilizara archivos de datos tradicionales, el programador tendría que especificar el tamaño y formato de cada elemento de datos utilizado en el programa y después decir a la computadora dónde están ubicados.

El DBMS libera al programador o al usuario final de la tarea de entender dónde y cómo están almacenados realmente los datos al separar las vistas lógica y física de los datos. La *vista lógica* presenta los datos según los perciben los usuarios finales o los especialistas de negocios, en tanto 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 conste del nombre del empleado,

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.

número de seguro social y cobertura del seguro médico. Un miembro de un departamento de nómina podría necesitar datos como el nombre del empleado, el número de seguro social, el sueldo bruto y neto. Los datos para todas estas vistas se almacenan en una sola base de datos, donde la organización puede administrarlos con más facilidad.

Cómo resuelve un DBMS los problemas del entorno de archivos tradicional

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 del todo la redundancia de datos, pero puede ayudar a controlarla. Aun cuando si la organización mantiene algunos 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 los datos se pueden independizar. La descripción de los datos que usa el programa no tiene que especificarse con detalle cada vez que se escribe un programa diferente. 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 para muchas aplicaciones simples sin tener que escribir programas complicados. El DBMS permite que la organización administre de manera central los datos, su uso y su seguridad. Es más fácil compartir datos en toda la organización debido a que los datos se presentan a los usuarios en una sola ubicación, en vez de fragmentarlos en muchos sistemas y archivos distintos.

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 PC, 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,

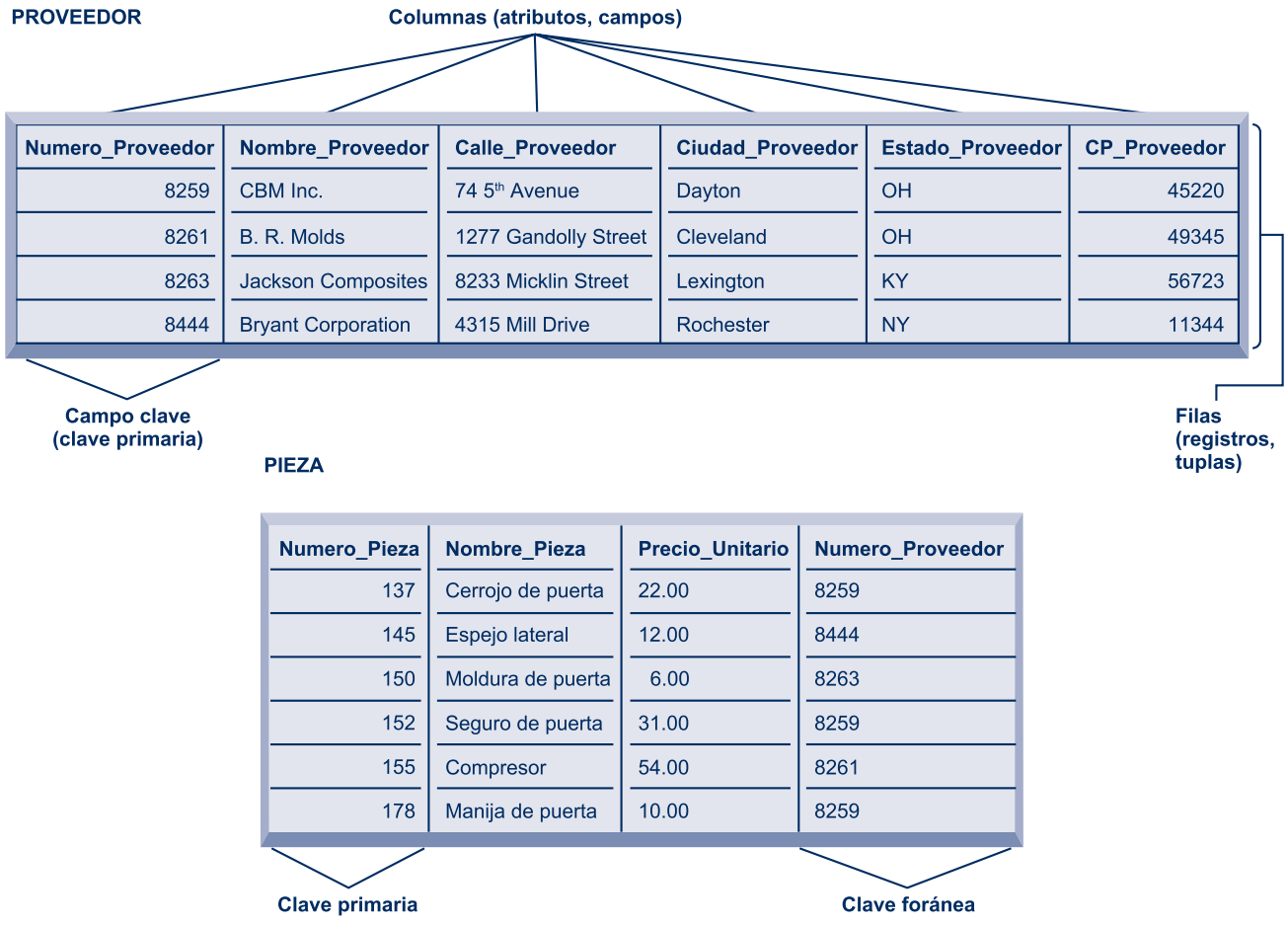
Oracle Database y Microsoft SQL Server son DBMS relacionales para las grandes mainframes y las computadoras de rango medio. MySQL es un popular DBMS de código fuente abierto.

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 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

FIGURA 6.4 TABLAS DE BASES DE DATOS RELACIONALES



PROVEEDOR

Numero_Proveedor	Nombre_Proveedor	Calle_Proveedor	Ciudad_Proveedor	Estado_Proveedor	CP_Proveedor
8259	CBM Inc.	74 5th Avenue	Dayton	OH	45220
8261	B. R. Molds	1277 Gandolly Street	Cleveland	OH	49345
8263	Jackson Composites	8233 Micklin Street	Lexington	KY	56723
8444	Bryant Corporation	4315 Mill Drive	Rochester	NY	11344

PIEZA

Numero_Pieza	Nombre_Pieza	Precio_Unitario	Numero_Proveedor
137	Cerrojo de puerta	22.00	8259
145	Espejo lateral	12.00	8444
150	Moldura de puerta	6.00	8263
152	Seguro de puerta	31.00	8259
155	Compresor	54.00	8261
178	Manija de puerta	10.00	8259

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.

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 Numero_Pieza es la clave primaria para la tabla PIEZA. Observe que Numero_Proveedor aparece tanto en la tabla PROVEEDOR como en la tabla 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 fácilmente para ofrecer los datos requeridos por los usuarios, siempre y cuando dos tablas compartan 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 cumplan 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 la que el Numero_Pieza sea igual a 137 o 150. La operación *unir* combina tablas relacionales para proveer al usuario con 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.

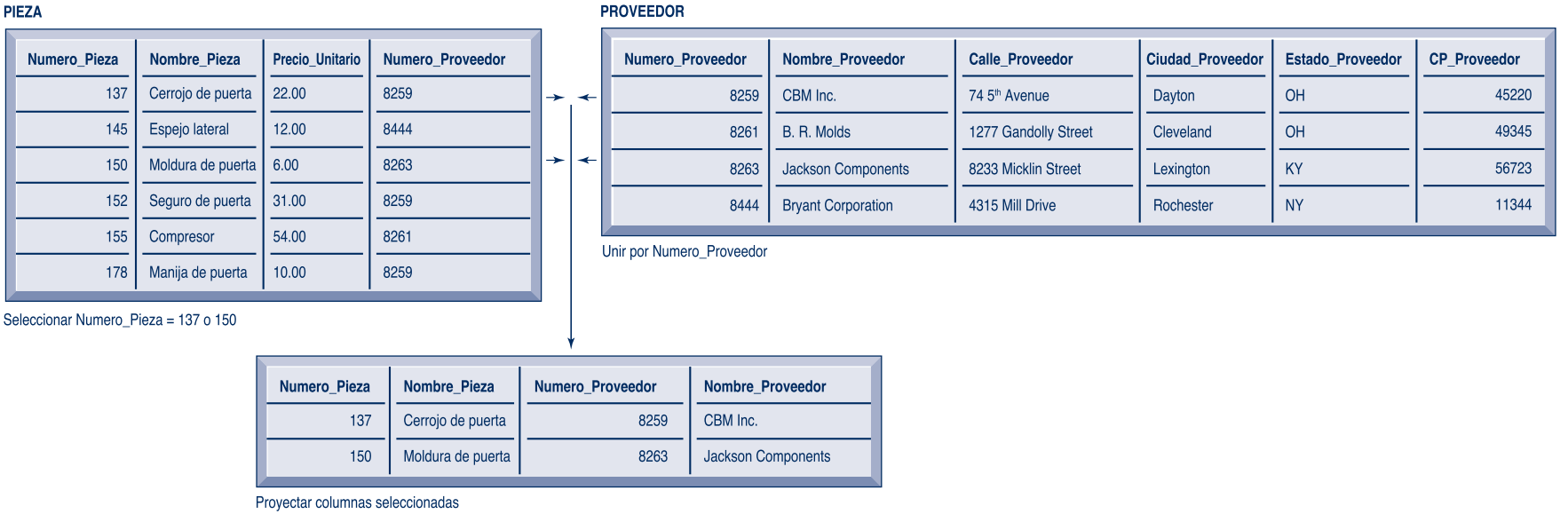
La operación *proyectar* crea un subconjunto que consiste en columnas en una tabla, con lo cual el usuario puede crear nuevas tablas que contengan solamente la información requerida. En nuestro ejemplo queremos extraer de la nueva tabla sólo las siguientes columnas: Numero_Pieza, Nombre_Pieza, Numero_Proveedor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

Bases de datos no relacionales y bases de datos en la nube

Durante más de 30 años, la tecnología de bases de datos relacionales ha sido el estándar de oro. La computación en la nube, los volúmenes de datos sin precedentes, las enormes cargas de trabajo para los servicios Web y la necesidad de almacenar nuevos tipos de datos requieren alternativas de bases de datos con respecto al modelo relacional tradicional de organizar los datos en forma de tablas, columnas y filas. Las empresas están recurriendo a las tecnologías de bases de datos no relacionales "*NoSQL*" para este fin. Los **sistemas de administración de bases de datos no relacionales** usan un modelo de datos más flexible y están diseñados para manejar grandes conjuntos de datos entre varias máquinas distribuidas, además de que pueden escalar fácilmente para aumentar o reducir su tamaño. Son útiles para acelerar las consultas simples contra grandes volúmenes de datos estructurados y no estructurados, ya sea en Web, social media, gráficos y demás formas de datos difíciles de analizar con herramientas tradicionales basadas en SQL.

Existen varios tipos distintos de bases de datos NoSQL, cada una con sus propias características técnicas y comportamiento. La base de datos NoSQL de Oracle es un ejemplo, al igual que SimpleDB de Amazon, uno de los servicios Web que se ejecutan en la nube. SimpleDB provee una interfaz de servicios Web sencilla para crear y almacenar varios conjuntos de datos, consultar datos con facilidad y devolverlos. No hay necesidad de predefinir una estructura de bases de datos formal o de cambiar esa definición si después se agregan nuevos datos. Por ejemplo, MetLife decidió emplear la base de datos NoSQL MongoDB de código abierto para integrar con rapidez los datos dispares y ofrecer una vista consolidada del cliente. La base de datos de MetLife reúne los datos de más de 70 sistemas administrativos separados, sistemas de reclamos y demás fuentes de datos, incluyendo los datos semiestructurados y no estructurados, como las imágenes de los registros de salud y certificados de defunción. La base de datos NoSQL es capaz de ingerir información

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 solamente los atributos seleccionados.

estructurada, semiestructurada y no estructurada sin requerir una asignación de bases de datos tediosa, costosa y que consuma mucho tiempo (Henschen, 2013).

Amazon y otros distribuidores de computación en la nube también proporcionan servicios de bases de datos relacionales. Amazon Relational Database Service (Amazon RDS) ofrece MySQL, SQL Server u Oracle Database como motores de bases de datos. Los precios se basan en el uso. Oracle tiene su propio servicio en la nube de bases de datos, al usar su base de datos Oracle relacional, y Microsoft AQL Azure es un servicio de bases de datos relacionales basado en la nube que utiliza el DBMS SQL Server de Microsoft. Los servicios de administración de datos basados en la nube tienen un atractivo especial para las empresas jóvenes enfocadas en Web o los negocios de tamaño pequeño a mediano que buscan capacidades de bases de datos a un precio más bajo que el de los productos de bases de datos de uso interno.

Además de los servicios de administración de datos basados en nubes públicas, las empresas tienen ahora la opción de usar bases de datos en nubes privadas. Por ejemplo, Sabre Holdings, el proveedor más grande del mundo de software como un servicio (SaaS) para la industria de la aviación, tiene una nube de bases de datos privada que soporta más de 100 proyectos y 700 usuarios. Una base de datos consolidada que abarca una reserva de servidores estandarizados que ejecutan Oracle Database provee servicios de bases de datos para varias aplicaciones.

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.

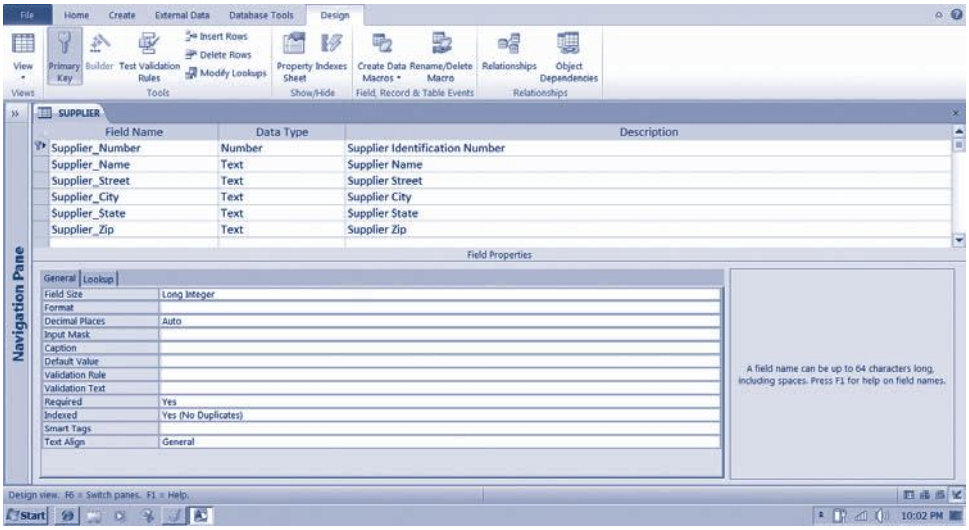
Microsoft Access cuenta con una herramienta rudimentaria de diccionario de datos, la cual muestra información sobre el nombre, la descripción, el tamaño, tipo, formato y otras propiedades de cada campo en una tabla (vea la figura 6.6). Los diccionarios de datos para las grandes bases de datos corporativas pueden capturar información adicional, como el uso, la propiedad (quién es el responsable en la organización 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 y manipular la información en las bases de datos. La mayoría de los DBMS tienen un lenguaje especializado llamado **lenguaje de manipulación de datos** el cual se utiliza para agregar, modificar, eliminar y recuperar los datos en la base de datos. 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 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.

FIGURA 6.6 CARACTERÍSTICAS DEL DICCIONARIO DE DATOS DE 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 icono a la izquierda de Numero_Proveedor indica que es un campo clave.

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 se construiría la misma consulta que la de 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, el cual, igualmente, cuenta con herramientas para desarrollar aplicaciones de sistemas de escritorio. Ambos incluyen herramientas para crear pantallas de captura de datos, generar informes y desarrollar la lógica de procesamiento de transacciones.

DISEÑO DE BASES DE DATOS

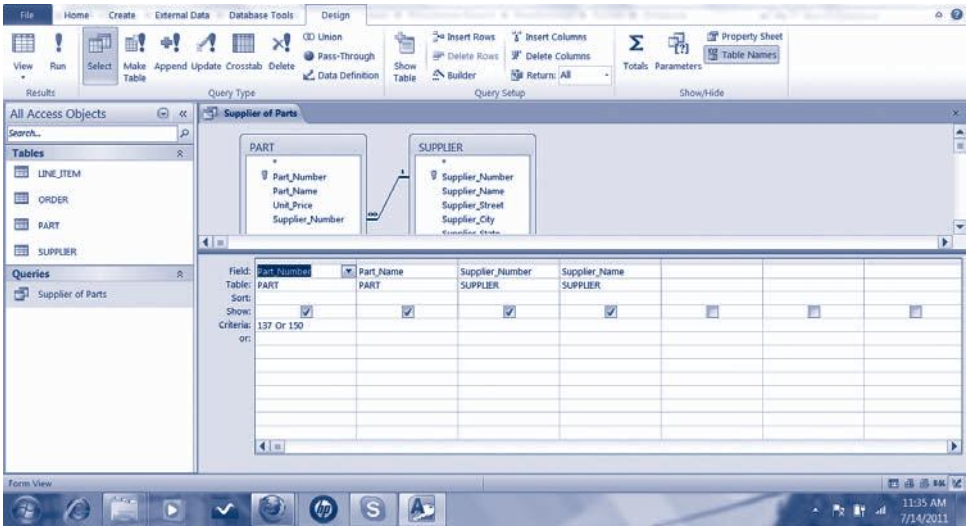
Para crear una base de datos hay que entender las relaciones entre la información, el tipo de datos que se mantendrán en la base, cómo se utilizarán y la forma en que la organización tendrá que cambiar para administrarlos desde una perspectiva a nivel

FIGURA 6.7 EJEMPLO DE UNA CONSULTA SQL

```
SELECT PIEZA.Numero_Pieza, PIEZA.Nombre_Pieza, PROVEEDOR.Numero_Proveedor,
PROVEEDOR.Nombre_Proveedor
FROM PIEZA, PROVEEDOR
WHERE PIEZA.Numero_Proveedor = PROVEEDOR.Numero_Proveedor AND
Numero_Pieza = 137 OR Numero_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 usando las herramientas de Microsoft Access para crear consultas. Muestra las tablas, los campos y los criterios de selección utilizados para la consulta.

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 la base de datos desde una perspectiva de negocios, en tanto que el diseño físico muestra la verdadera disposición de la base de datos en los dispositivos de almacenamiento de acceso directo.

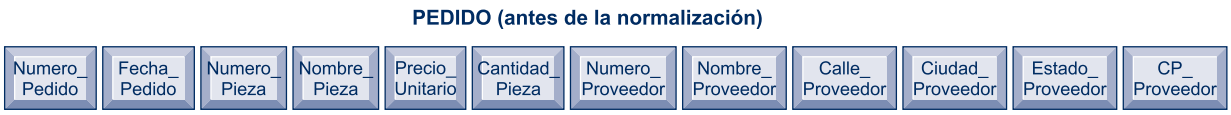
Diagramas de normalización y de entidad-relación

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 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 de datos.

Para usar un modelo de base de datos relacional en forma eficaz, hay que optimizar los agrupamientos complejos de datos para minimizar los elementos de datos redundantes y las incómodas relaciones de varios a varios. Al proceso de crear estructuras de datos pequeñas y estables pero a la vez flexibles y adaptivas a partir de grupos complejos de datos se le denomina **normalización**. Las figuras 6.9 y 6.10 ilustran este proceso.

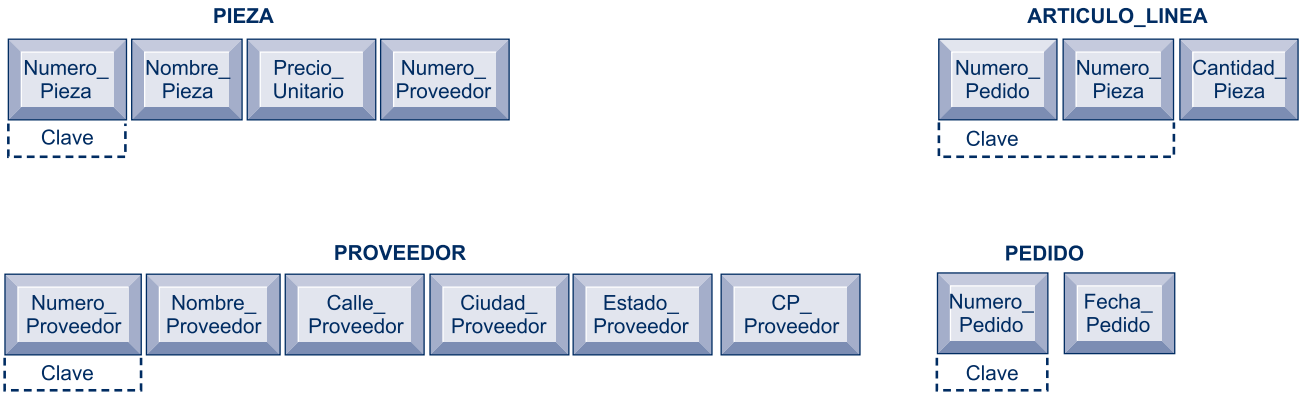
En la empresa específica que se modela aquí, un pedido puede tener más de una pieza, pero cada una sólo es proporcionada por un proveedor. Si creamos una relación

FIGURA 6.9 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 Numero_Pedido y Fecha_Pedido.

FIGURA 6.10 TABLAS NORMALIZADAS CREADAS A PARTIR DE PEDIDO



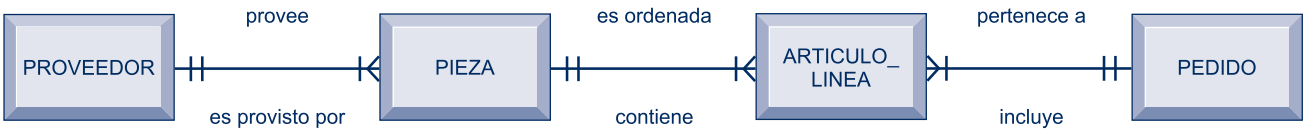
Después de la normalización, la relación original PEDIDO se ha dividido en cuatro relaciones más pequeñas. La relación PEDIDO se queda con sólo dos atributos y la relación ARTICULO_LINEA tiene una clave combinada, o concatenada, que consiste en Numero_pedido y Numero_Pieza.

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 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 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. Es decir, ¡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.

FIGURA 6.11 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.

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 que aparezca otras tantas como artículo de línea en un solo pedido.) Cada PIEZA solo 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 usted puede 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 correctamente diseñada para ajustar con rapidez el inventario de las tiendas. La compañía tenía una base de datos relacional Oracle operando en una computadora 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 la que se pudieran organizar mejor los datos de las ventas y del inventario para realizar análisis y administrar el inventario.

6.3 ¿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?

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 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.

EL DESAFÍO DE BIG DATA

La mayoría de los datos recolectados por las organizaciones solían ser los datos de transacciones que podían caber fácilmente en filas y columnas de sistemas de administración de bases de datos relacionales. Ahora, somos testigos de una explosión de datos provenientes del tráfico Web, mensajes de correo electrónico y contenido de medios sociales (tweets, mensajes de estado), así como los datos generados por máquinas de los sensores (utilizados en medidores inteligentes, sensores de fabricación y medidores eléctricos) o de sistemas de e-commerce. Estos datos pueden ser estructurados o no estructurados y, por ende, tal vez no sean adecuados para productos de bases de datos relacionales que organicen los datos en forma de columnas y filas. Ahora usamos el término **big data** para describir estos conjuntos de datos con volúmenes tan grandes que están más allá de la capacidad de un DBMS común para capturar, almacenar y analizar.

Big Data no se refiere a una cantidad específica, sino por lo general a los datos en el rango de los petabytes y exabytes; es decir, de miles de millones a billones de registros,

todos de orígenes distintos. Los Big Data se producen en cantidades mucho mayores y con mucha más rapidez que los datos tradicionales. Por ejemplo, un solo motor de jet es capaz de generar 10 terabytes de datos en sólo 30 minutos, y hay más de 25,000 vuelos de aerolíneas a diario. Aun cuando los "tweets" se limitan a 140 caracteres cada uno, Twitter genera más de 8 terabytes de datos por día. De acuerdo con la empresa de investigación de tecnología International Data Center (IDC), los datos se duplican con creces cada dos años, por lo que la cantidad de datos disponibles para las organizaciones está aumentando en forma indiscriminada.

A las empresas les interesan los Big Data debido a que pueden revelar más patrones y anomalías interesantes que los conjuntos de datos más pequeños, con el potencial de proveer nuevas perspectivas en cuanto al comportamiento de los clientes, los patrones de clima, la actividad del mercado financiero u otros fenómenos. Sin embargo, para derivar un valor de negocios de estos datos, las organizaciones necesitan nuevas tecnologías y herramientas capaces de administrar y analizar datos no tradicionales junto con sus datos empresariales tradicionales.

INFRAESTRUCTURA DE INTELIGENCIA DE NEGOCIOS

Suponga que desea información concisa y confiable sobre las operaciones, tendencias y cambios actuales en toda la empresa. Si trabajara en una empresa de gran tamaño, tendría que reunir los datos necesarios de sistemas separados, como ventas, manufactura y contabilidad, e incluso desde fuentes externas, como los datos demográficos o de las competencias. Es probable que cada vez fuera más necesario usar Big Data. Una infraestructura contemporánea para la inteligencia de negocios tiene una variedad de herramientas para obtener información útil de todos los tipos diferentes de datos que usan las empresas en la actualidad, incluyendo Big Data semiestructurados y no estructurados en grandes cantidades. Estas capacidades incluyen almacenes de datos y mercados de datos, Hadoop, computación en memoria y plataformas analíticas. Algunas de estas capacidades están disponibles como servicios en la nube.

Almacenes de datos y mercados de datos

La herramienta tradicional para analizar datos corporativos durante las últimas dos décadas ha sido el 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 básicos de transacciones operacionales, como los sistemas de ventas, las cuentas de clientes, la manufactura, y pueden incluir datos de transacciones de sitios Web. El almacén de datos extrae los datos actuales e históricos de varios sistemas operacionales dentro de la organización. Estos datos se combinan con los datos de fuentes externas y se transforman al corregir los datos imprecisos e incompletos y reestructurar los datos para generar informes gerenciales y realizar análisis antes de cargarlos en el almacén de datos.

El almacén de datos pone los datos a disposición de todos según sea necesario, pero no se puede 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.

A menudo las empresas crean almacenes de datos a nivel empresarial, donde un almacén de datos central da servicio a toda la organización, o crean almacenes de datos 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 sintetizada o 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 lidiar con la información de los clientes. El vendedor de libros Barnes & Noble solía mantener 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.

Hadoop

Los productos de DBMS relacionales y almacenes de datos no se adaptan bien para organizar y analizar Big Data o datos que no caben fácilmente en las columnas y filas utilizadas en sus modelos de datos. Para manejar datos no estructurados y semiestructurados en grandes cantidades, así como datos estructurados, las organizaciones usan **Hadoop**, que es un marco de trabajo de software de código abierto, administrado por la Fundación de Software Apache, lo que permite el procesamiento paralelo distribuido de enormes cantidades de datos a través de computadoras económicas. Descompone un problema de Big Data en varios subproblemas, los distribuye entre miles de nodos de procesamiento de computadoras económicas y luego combina el resultado en un conjunto de datos de menor tamaño que es más fácil de analizar. Tal vez usted ya haya usado Hadoop para encontrar la mejor tarifa aérea en Internet, obtener indicaciones para llegar a un restaurante, realizar una búsqueda en Google o conectarse con un amigo en Facebook.

Hadoop consta de varios servicios clave: el sistema de archivos distribuidos Hadoop (HDFS) para almacenamiento de datos y MapReduce para procesamiento de datos en paralelo de alto rendimiento. HDFS enlaza entre sí los sistemas de archivos en los numerosos nodos en un clúster Hadoop para convertirlos en un gran sistema de archivos. MapReduce de Hadoop se inspiró en el sistema MapReduce de Google para desglosar el procesamiento de enormes conjuntos de datos y asignar trabajo a los diversos nodos en un clúster. HBase, la base de datos no relacional de Hadoop, ofrece un acceso rápido a los datos almacenados en HDFS y una plataforma transaccional para ejecutar aplicaciones en tiempo real de alta escala.

Hadoop puede procesar grandes cantidades de cualquier tipo de datos, incluyendo datos transaccionales estructurados, datos poco estructurados como las fuentes de Facebook y Twitter, datos complejos como los archivos de registro de servidor Web y datos de audio y video no estructurados. Hadoop se ejecuta en un clúster de servidores económicos y pueden agregarse o eliminarse procesadores según sea necesario. Las empresas usan Hadoop para analizar volúmenes muy grandes de datos, así como para un área de concentración para datos no estructurados y semiestructurados antes de cargarlos en un almacén de datos. Facebook almacena gran parte de sus datos en un enorme clúster Hadoop, que contiene cerca de 100 petabytes, alrededor de 10,000 veces más información que la Biblioteca del Congreso estadounidense. Yahoo usa Hadoop para rastrear el comportamiento de los usuarios de modo que pueda modificar su página de inicio y adaptarla a sus intereses. La empresa de investigación de ciencias de la vida NextBio usa Hadoop y HBase para procesar datos para empresas farmacéuticas que realizan investigación genómica. Los principales distribuidores de bases de datos como IBM, Hewlett-Packard, Oracle y Microsoft tienen sus propias distribuciones de software de Hadoop. Otros distribuidores ofrecen herramientas para meter y sacar datos de Hadoop, o para analizarlos dentro de Hadoop.

Computación en memoria

Otra forma de facilitar el análisis de Big Data es utilizar la **computación en memoria**, que depende principalmente de la memoria principal (RAM) de la computadora para el almacenamiento de datos (los DBMS convencionales usan sistemas de almacenamiento de datos). Los usuarios acceden a los datos almacenados en la memoria principal del sistema, con lo cual se eliminan los cuellos de botella por los procesos de recuperación y lectura de datos en una base de datos tradicional basada en discos, y se reducen de manera drástica los tiempos de respuesta de las consultas. El procesamiento en memoria hace posible que conjuntos muy grandes de datos, del tamaño de un mercado de datos o de un almacén pequeño de datos, residan totalmente en la memoria. Los cálculos de negocios complejos que solían tardar horas o días pueden completarse en cuestión de segundos, y esto puede lograrse incluso en dispositivos portátiles (vea la Sesión interactiva: tecnología).

El capítulo anterior describe algunos de los avances en la tecnología de hardware de computadora contemporánea que hacen posible el procesamiento en memoria, como los poderosos procesadores de alta velocidad, el procesamiento multinúcleo y los precios cada vez menores de la memoria de computadora. Estas tecnologías ayudan a las empresas a optimizar el uso de la memoria y aceleran el rendimiento del procesamiento, a la vez que reducen los costos.

SESIÓN INTERACTIVA: TECNOLOGÍA

IMPULSO DE LA GESTIÓN DE FLOTILLAS DE ARI CON ANÁLISIS EN TIEMPO REAL

Automotive Resources International©, mejor conocida como ARI©, es la empresa privada más grande del mundo para servicios de administración de flotillas de vehículos. ARI tiene sus oficinas generales en Mt. Laurel, Nueva Jersey, con 2,500 empleados y oficinas en Norteamérica, Europa, el Reino Unido y Hong Kong. La empresa administra más de 1'000,000 de vehículos en Estados Unidos, Canadá, México, Puerto Rico y Europa.

Las empresas que necesitan vehículos para envíos (camiones, vans, automóviles, barcos y vagones de ferrocarril) pueden optar por gestionar su propia flotilla de vehículos o bien subcontratar la gestión de flotillas con empresas como ARI, que se especializan en estos servicios. ARI se encarga de todo el ciclo de vida y la operación de una flotilla de vehículos para sus clientes, desde la especificación inicial y la adquisición hasta la reventa, incluyendo servicios financieros, de mantenimiento, de gestión del combustible y administración del riesgo como la capacitación de seguridad de los conductores y la administración de accidentes. ARI también mantiene seis call centers en Norteamérica que operan 24/7, los 365 días del año para dar soporte a las operaciones de flotillas de los clientes, brindando asistencia relacionada con reparaciones, descomposturas, respuesta a los accidentes, mantenimiento preventivo y demás necesidades de los conductores. Estos call centers manejan cerca de 3.5 millones de llamadas por año de clientes, conductores y proveedores que esperan el acceso a la información práctica en tiempo real.

La acción de proporcionar esta información se ha convertido en un desafío cada vez mayor. Al operar una sola flotilla grande de vehículos comerciales se generan altos volúmenes de datos complejos, como la información sobre el consumo de combustible, mantenimiento, licencias y cumplimiento. Por ejemplo, una transacción de combustible requiere datos sobre los impuestos estatales que se pagan, el grado del combustible, la venta total, el monto vendido y tanto la hora como el lugar de la compra. Un trabajo simple de frenos y una revisión de mantenimiento preventivo generan docenas de registros para cada componente al que se da servicio. Cada pieza y servicio que se realiza sobre un vehículo se rastrea mediante códigos de la Asociación estadounidense del transporte de carga. ARI recolecta y analiza más de 14,000 piezas de datos por vehículo. Después multiplica los datos por cientos de flotillas, algunas con hasta 10,000 vehículos, todos operando al mismo tiempo a nivel mundial; así puede darse una idea del enorme volumen de datos que ARI necesita administrar, tanto para sus propias operaciones como para sus clientes.

ARI proporcionaba a sus clientes información detallada sobre las operaciones de sus flotillas, pero el tipo de información que podía ofrecer era muy limitado. Por ejemplo, podía generar informes detallados sobre los gastos por partidas, las compras de vehículos, los registros de mantenimiento y demás información operacional, los cuales se presentaban como simples hojas de cálculo, tablas o gráficos, pero no era posible analizar todos los datos para detectar tendencias y hacer recomendaciones. ARI podía analizar los datos cliente por cliente, pero no era capaz de agregar esos datos en toda su base de clientes. Por ejemplo, si ARI administraba la flotilla de vehículos de una compañía farmacéutica, sus sistemas de información no podían marcar como referencia el rendimiento de esa flotilla y compararlo con el resto de la industria. Ese tipo de problema requería demasiado trabajo manual y tiempo, y de todas formas no ofrecía el nivel de perspectiva que la gerencia consideraba posible.

Además, para crear los informes ARI tenía que recurrir a expertos internos en la materia, en varios aspectos de operaciones de flotilla, a quienes se les conocía como "usuarios avanzados de generación de informes". Cada solicitud de información se pasaba a estos usuarios avanzados. Una solicitud de un informe tardaría 5 días en completarse. Si el informe no era satisfactorio, regresaría a quien había escrito el informe para que realizara modificaciones. El proceso de ARI para analizar sus datos era demasiado prolongado.

A mediados de 2011 ARI implementó SAP BusinessObjects Explorer para dar a los clientes la capacidad mejorada de acceder a los datos y ejecutar sus propios informes. SAP BusinessObjects Explorer es una herramienta de inteligencia de negocios que permite a los usuarios de negocios ver, ordenar y analizar la información de inteligencia de negocios. Los usuarios realizan búsquedas a través de los datos y los resultados se muestran con una tabla que indica la mejor coincidencia de información. La representación gráfica de los resultados cambia a medida que el usuario hace más preguntas de los datos.

A principios de 2012 integró SAP BusinessObjects Explorer con HANA, la plataforma de computación en memoria de SAP que puede implementarse como aplicación dentro de las premisas (hardware y software) o en la nube. HANA está optimizada para realizar análisis en tiempo real y manejar volúmenes muy altos de datos operacionales y transaccionales en tiempo real. Los análisis en memoria de HANA consultan los datos almacenados en la memoria de acceso aleatorio (RAM) en vez de usar un disco duro o almacenamiento tipo flash.

Después de eso, las cosas comenzaron a ocurrir con rapidez. Cuando el controlador de ARI necesitaba un análisis de impacto de los mejores 10 clientes de la empresa, SAP HANA produjo el resultado en un lapso de 3 a 3.5 segundos. En el antiguo entorno de sistemas de ARI, esta tarea se habría asignado a un usuario avanzado especializado en el uso de herramientas de informes, habría que dibujar especificacio-

nes y diseñar un programa para esa consulta específica, un proceso que hubiera tomado 36 horas.

Ahora, usando HANA, ARI puede extraer rápidamente sus amplios recursos de datos y generar predicciones con base en los resultados. Por ejemplo, la empresa puede producir cifras precisas sobre los costos de operar una flotilla de cierto tamaño a través de determinada ruta en industrias específicas durante cierto tipo de clima y predecir el impacto de los cambios en alguna de esas variables. Y puede hacerlo casi con tanta facilidad como la de proveer a sus clientes un historial simple de sus gastos de combustible. Con esta información tan útil ARI provee más valor a sus clientes.

HANA también redujo el tiempo requerido para cada transacción manejada por los call centers de ARI (desde el momento en que un miembro del personal del call center toma una llamada hasta la recuperación y entrega de la información solicitada) en un 5%. Como el personal de los call centers representa el 40% de la sobrecarga directa de ARI, esa reducción en tiempo se traduce en grandes ahorros en costo.

ARI planea tener algunas de estas capacidades de generación de informes y análisis en tiempo real disponibles en dispositivos móviles, lo cual permitirá a los clientes aprobar al instante varios procedimientos operacionales, como la autorización de reparaciones de mantenimiento. Los clientes también podrán usar las herramientas móviles para una perspectiva instantánea de las operaciones de sus flotillas, con un nivel de detalle como el historial de los neumáticos de un vehículo específico.

Fuentes: "Driving 2 Million Vehicles with SAP Data", www.sap.com, visitado el 1 de febrero de 2014; www.arifleet.com, visitado el 1 de febrero de 2014, y "ARI Fleet Management Drives Real-Time Analytics to Customers", *SAP InsiderPROFILES*, 1 de abril de 2013.

PREGUNTAS DEL CASO DE ESTUDIO

1. ¿Por qué era tan problemática la administración de datos en ARI?
2. Describa las capacidades anteriores de ARI en cuanto a análisis de datos y generación de informes, y su impacto en el negocio.
3. ¿Fue SAP HANA una buena solución para ARI? ¿Por qué?
4. Describa los cambios en los negocios como resultado de adoptar HANA.

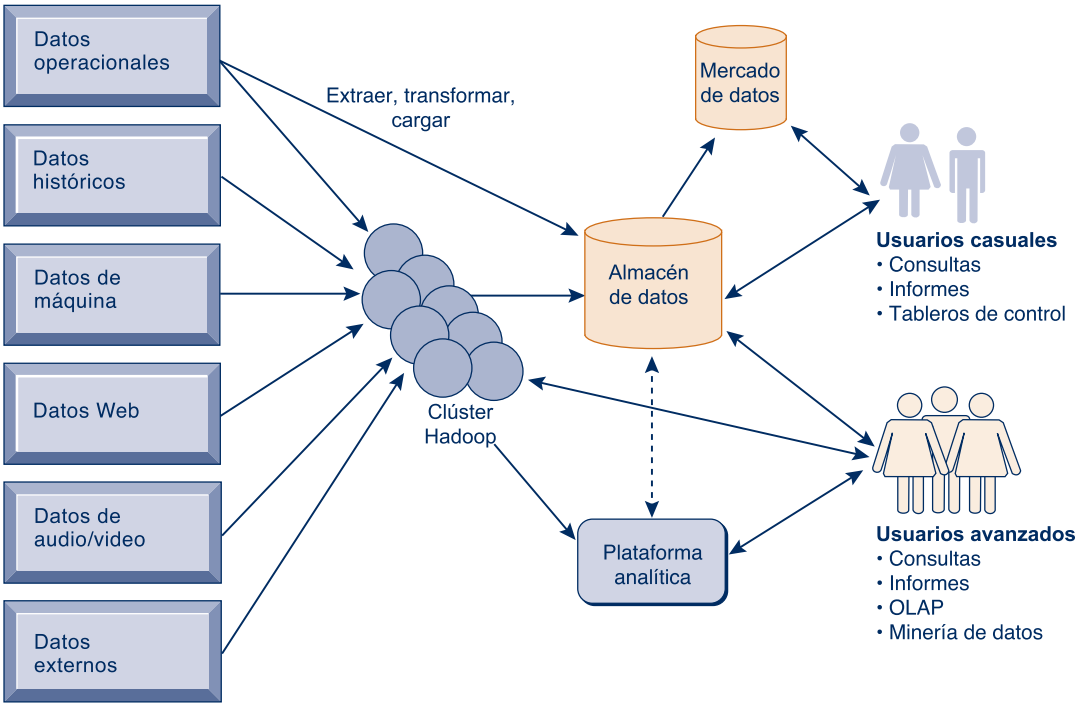
Los principales productos comerciales para la computación en memoria son: High Performance Analytics Appliance (HANA) de SAP, y Oracle Exalytics. Cada uno ofrece un conjunto de componentes de software integrados, incluyendo software de base de datos en memoria y software de análisis especializado, que se ejecutan en hardware optimizado para el trabajo de cómputo en memoria.

Plataformas analíticas

Los distribuidores de bases de datos comerciales han desarrollado **plataformas analíticas** especializadas de alta velocidad que utilizan tecnología tanto relacional como no relacional y están optimizadas para analizar conjuntos de datos de gran tamaño. Las plataformas analíticas como IBM Netezza y Oracle Exadata cuentan con sistemas de hardware-software preconfigurados que están diseñados de manera específica para el procesamiento de consulta y los análisis. Por ejemplo, IBM Netezza tiene componentes de base de datos, servidor y almacenamiento estrechamente integrados que manejan consultas analíticas complejas 10 a 100 veces más rápido que los sistemas tradicionales. Las plataformas analíticas también incluyen sistemas en memoria y sistemas de administración de bases de datos no relacionales. Ahora, las plataformas analíticas están disponibles como servicios en la nube.

La figura 6.12 ilustra una infraestructura de inteligencia de negocios contemporánea que usa las tecnologías que acabamos de describir. Los datos actuales e históricos se extraen de varios sistemas operacionales junto con datos Web, datos generados por máquinas, datos de audio/visuales no estructurados y datos provenientes de fuentes externas, que se han reestructurado y organizado para generación de informes y análisis. Los clústeres Hadoop preprocesan los Big Data para usarlos en el almacén de datos, mercados de datos o en una plataforma analítica, o para que los usuarios avanzados los

FIGURA 6.12 INFRAESTRUCTURA CONTEMPORÁNEA DE INTELIGENCIA DE NEGOCIOS



Una infraestructura contemporánea de inteligencia de negocios cuenta con capacidades y herramientas para administrar y analizar grandes cantidades y distintos tipos de datos provenientes de varias fuentes. Se incluyen herramientas de consulta y generación de informes fáciles de usar para los usuarios de negocios casuales y conjuntos de herramientas analíticas más sofisticadas para usuarios avanzados.

consulten de manera directa. Los resultados incluyen informes y tableros de control, así como resultados de las consultas. En el capítulo 12 veremos con mayor detalle los diversos tipos de usuarios BI y generación de informes BI.

HERRAMIENTAS ANALÍTICAS: RELACIONES, PATRONES, TENDENCIAS

Una vez que los datos se capturan y organizan mediante el uso de las herramientas para inteligencia de negocios que acabamos de describir, están disponibles para un posterior análisis utilizando el software para consultas e informes de bases de datos, el análisis de datos multidimensional (OLAP) y la minería de datos. En esta sección le presentaremos estas herramientas; en el capítulo 12 veremos más detalles sobre el análisis de inteligencia de negocios y aplicaciones.

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 deseara hacer una pregunta muy directa, por ejemplo, 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 preguntas ad hoc como éstas en un tiempo muy corto, incluso cuando los datos se almacenan en bases de datos muy grandes, como las cifras de ventas de varios años.

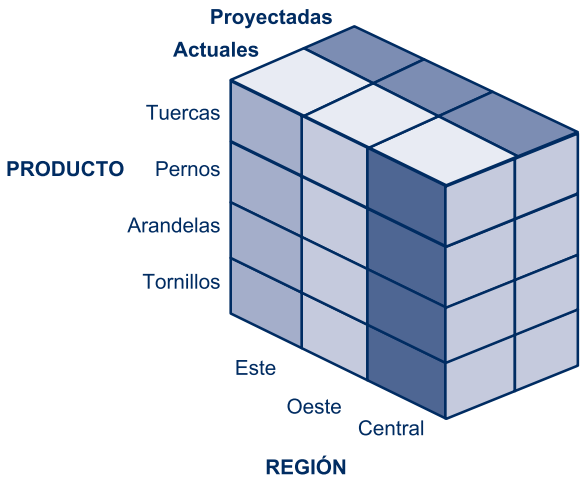
La figura 6.13 muestra un modelo multidimensional que podría crearse para representar productos, regiones, ventas reales y ventas proyectadas. Una matriz de ventas actuales se puede apilar encima de una matriz de ventas proyectadas para formar un cubo con seis caras. Si gira el cubo 90° en un sentido, la cara que se muestre será la del producto contra ventas actuales y proyectadas; si lo gira de nuevo 90°, verá la cara de la región contra ventas actuales y proyectadas, y si lo gira 180° 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 2013?" 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 de estos patrones y relaciones, para predecir el comportamiento a futuro. Los patrones y

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.

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% de veces se compra un refresco de cola, pero cuando hay una promoción, es el 85% de veces. 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% de veces se compra un refrigerador nuevo dentro de las siguientes dos semanas, y el 45% 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. Hay 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 a clientes rentables.

Entertainment, anteriormente conocida como Harrah's Entertainment, es la segunda compañía de apuestas más grande del mundo. Analiza continuamente los datos sobre sus clientes que se recopilan cuando las personas juegan en las máquinas tragamonedas o utilizan sus casinos y hoteles. El departamento de marketing corporativo 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 Caesars conocer la experiencia de juego favorita de un cliente regular en uno de sus casinos en los barcos, 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 Caesars que se convirtió en la pieza central de la estrategia de negocios de la empresa.

Minería de texto y minería Web

Se cree que los datos no estructurados, que en su mayoría están organizados en forma de archivos de texto, representan más del 80% de la información útil de una organización y son una de las principales fuentes de Big Data que las empresas desean analizar. 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 center de servicio al cliente para identificar las principales cuestiones de servicio y reparación, o para medir el sentimiento de los clientes con respecto a su empresa. El software de análisis de opiniones es capaz de extraer los comentarios de texto en un mensaje de correo electrónico, blog, conversación de social media o formulario de encuesta para detectar las opiniones favorables y desfavorables sobre temas específicos.

Por ejemplo, el corredor de saldos Charles Schwab usa el software Attensity Analyze para analizar cientos de miles de interacciones de sus clientes cada mes. El software analiza las notas de servicio de los clientes de Schwab, los correos electrónicos, las respuestas de las encuestas y las discusiones en línea para descubrir señales de descontento que puedan provocar que un cliente deje de usar los servicios de la empresa. Attensity puede identificar automáticamente las diversas "voces" que usan los clientes para expresar su retroalimentación (como una voz positiva, negativa o condicional) para señalar la intención de una persona de comprar, su intención de abandonar, o la reacción a un producto o mensaje de marketing específico. Schwab usa esta información para tomar acciones correctivas como establecer una comunicación directa del corredor con el cliente y tratar de resolver con rapidez los problemas que lo tienen descontento.

Web es otra fuente de datos extensos no estructurados para revelar 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 interesa comprar.

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 examina los datos relacionados con la estructura de un sitio Web específico. Por ejemplo, los vínculos que apuntan a un documento indican su popularidad, en tanto 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 recolecta 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.

El caso al final del capítulo describe las experiencias de las organizaciones al usar las herramientas analíticas y las tecnologías de inteligencia de negocios que hemos descrito para lidiar con los desafíos de los "big data".

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.

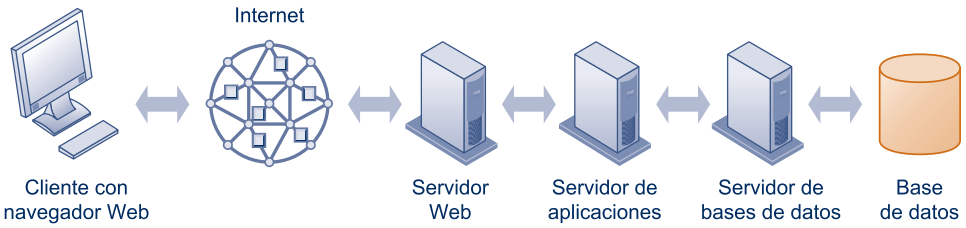
Dado que muchas bases de datos de procesamiento en segundo plano (back-end) no pueden interpretar comandos escritos en HTML, el servidor Web pasa estas solicitudes de datos al software que traduce los comandos de HTML en SQL, de modo que el DBMS que trabaja con la base de datos pueda procesarlos. En 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 negocios de procesamiento en segundo plano (*back-end*) de una compañía. El servidor de aplicaciones recibe las solicitudes del servidor Web, ejecuta la lógica de negocios 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 700,000 proveedores de productos industriales como químicos, metales, plásticos, goma y equipo automotriz. Antes conocida como Thomas

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 equipos PC de escritorio y el software de navegador Web.

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 Facebook, que ayuda a los usuarios a permanecer conectados entre sí o conocer nuevas personas. Facebook incluye "perfiles" con información suministrada por 1,300 millones de usuarios activos sobre sí mismos, incluyendo intereses, amigos, fotos y grupos a los que están afiliados. Mantiene una base de datos masiva para alojar y administrar todo su contenido. También hay muchas bases de datos habilitadas para Web en el sector público que ayudan a los consumidores y ciudadanos a acceder a información útil.

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

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 quienes 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 solamente miembros seleccionados 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 una función formal de administración de datos. La **administración de datos** es responsable de las políticas y procedimientos específicos a través de los cuales se pueden gestionar 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 haya escuchado que el término **gobernanza de datos** se emplea para describir muchas de estas actividades. La gobernanza de datos, promovida por IBM, 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, la seguridad, la calidad de los datos y el cumplimiento de 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 estuvieran 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. Gartner Inc. informó que más del 25% 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. Un estudio de Sirius Decisions sobre "El impacto de datos erróneos en la creación de demanda" descubrió que del 10 al 25% de los registros de clientes y prospectos contienen errores críticos de datos. Al corregir estos errores en su origen y seguir las prácticas recomendadas para promover la calidad de los datos, aumentó la productividad del proceso de ventas y se generó un incremento del 66% en los ingresos.

Algunos de estos problemas de calidad se deben a datos redundantes e inconsistentes producidos por varios sistemas que alimentan un almacén de datos. Por ejemplo, el sistema de pedidos de ventas y el sistema de administración de inventario podrían mantener datos sobre los productos de la organización. Sin embargo, el sistema de pedidos de ventas podría usar el término *Número de artículo* y el sistema de inventario podría llamar al mismo atributo *Número de producto*. Los sistemas de ventas, inventario o manufactura de un minorista de ropa podrían usar distintos códigos para representar valores para un atributo. Un sistema podría representar el tamaño de la ropa como "extra grande", mientras que el otro sistema podría usar el código "XL" para el mismo fin. Durante el proceso de diseño para la base de datos del almacén, las entidades de descripción de datos (como cliente, producto o pedido) se deben nombrar y definir de manera consistente para todas las áreas de negocios que usen la base de datos.

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 adecuadamente 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 traspuestos 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 la 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 funcionando. 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 por 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 estén incorrectos, incompletos, que tengan un formato inadecuado 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 especializado de limpieza de datos está disponible para inspeccionar automáticamente los archivos de datos, 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. Por ejemplo, la información imprecisa u obsoleta sobre los historiales crediticios de los consumidores que mantienen los burós de crédito pueden evitar que individuos solventes obtengan préstamos o se reduzca su probabilidad de encontrar o conservar un empleo.

La Sesión interactiva sobre administración ilustra la experiencia de American Water con la administración de datos como un recurso. Cuando lea este caso trate de identificar las políticas, procedimientos y tecnologías que se requirieron para mejorar la administración de datos en esta empresa.