Wirtschaftsinformatik

Eine Einführung

3., vollständig überarbeitete Auflage

Kenneth C. Laudon
Jane P. Laudon
Detlef Schoder

ALWAYS LEARNING

PEARSON

6.4.6 Online Analytical Processing – OLAP

Nachdem die Daten in Data Warehouses und Data Marts übertragen worden sind, können sie für weitere Analysen im Kontext der Business Intelligence herangezogen werden. Die Haupttools für Business Intelligence & Analytics schließen die Software für Datenbankabfragen und -berichte ebenso ein wie Tools für die mehrdimensionale Datenanalyse (Online Analytical Processing – OLAP) und Data-Mining.

Im herkömmlichen Sinne versteht man unter „Intelligenz" die Fähigkeit des Menschen, erworbenes Wissen mit neuen Informationen zu kombinieren und sein Verhalten so zu verändern, dass er sich neuen Aufgaben und Situationen optimal anpassen kann. Ebenso versteht man unter Business Intelligence die Fähigkeit von Unternehmen, Informationen anzuhäufen, sich Wissen über Kunden, Wettbewerber und interne Prozesse anzueignen und damit Entscheidungsfindungsprozesse in der Unternehmung im Hinblick auf eine höhere Rentabilität oder anderer Geschäftsziele positiv zu beeinflussen.

Zum Beispiel analysiert Harrah's Entertainment, die zweitgrößte Glücksspielfirma der Welt, fortlaufend eine Reihe Kundendaten, die anfallen, wenn Kunden an Spielautomaten spielen oder Kasinos und Hotels besuchen. Diese Daten werden von Harrah's Marketingabteilung dazu benutzt, detaillierte Spielerprofile zu erstellen und den Wert eines Kunden zu berechnen. Diese Informationen können dem Management bei Entscheidungen darüber helfen, wie die Kundenloyalität der rentabelsten Kunden gesteigert werden kann, wie Kunden dazu animiert werden können, mehr Geld auszugeben, und wie vermehrt Kunden gewonnen werden können, die potenziell große Umsätze generieren. Durch den Einsatz von Business Intelligence & Analytics konnte Harrah's seine Rentabilität so sehr steigern, dass Business Intelligence inzwischen das Kernstück von Harrah's Geschäftsstrategie geworden ist.

▶*Abbildung 6.16* zeigt eine typische Architektur von BI&A-Systemen. In den operativen Datenbanken eines Unternehmens werden Daten über die Transaktionen des Unternehmens gesammelt. Diese Daten werden aus den operativen Datenbanken extrahiert, transformiert und ins Data Warehouse geladen (ETL: Extract-Transform-Load). Manager nutzen dann Business-Intelligence-Lösungen, um Muster und Beziehungen in den Daten aufzudecken. Die Entscheidungsfindung der Manager basiert danach im Wesentlichen auf diesen Datenanalysen, da die Analyseergebnisse besser informierte und intelligentere Entscheidungen ermöglichen.

Dieser Abschnitt stellt die wichtigsten Komponenten einer BI&A-Architektur vor. Den Zusammenhang zwischen Business Intelligence & Analytics und Entscheidungsunterstützende Systeme finden Sie in *Kapitel 12* erläutert.

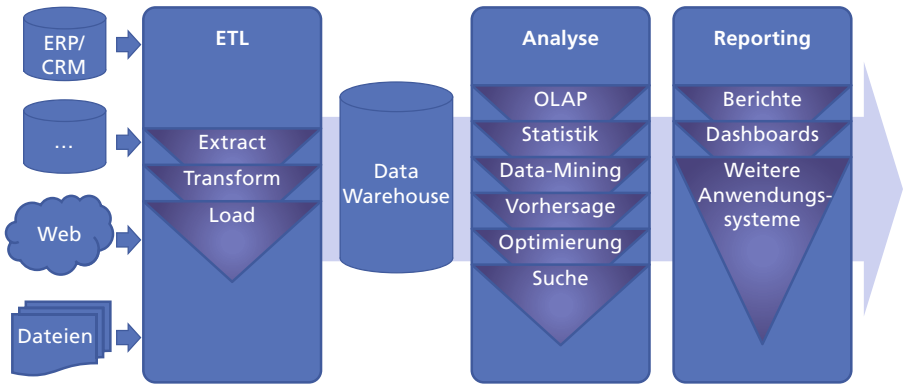

Abbildung 6.16: Business Intelligence & Analytics
Eine Reihe von Techniken wird auf die in verschiedenen Datenbanken gespeicherten Daten angewendet, um Muster herauszufinden und Managern und Angestellten so Hilfestellungen bei ihrer Entscheidungsfindung zu bieten und die Rentabilität eines Unternehmens zu steigern.

Quelle: S. Chaudhuri, U. Dayal und V. Narasayya, Communications of the ACM 54(8), S. 88-98 (2011).

ETL: Extract-Transform-Load

Extraktion Zur Datenextraktion muss zunächst ein technischer Zugriff auf die relevanten Datenquellen zumeist über bekannte Schnittstellen oder Konnektoren hergestellt werden.

Transformation Ziel hierbei ist es, die aus unterschiedlichen Quellen stammenden Daten, die höchstwahrscheinlich unterschiedlich strukturiert sind, in eine einheitliche Datenstruktur zu überführen. Dabei werden in mehrerlei Hinsicht die Daten integriert und bereinigt, etwa Dubletten entfernt, Datumsformate und Einheiten standardisiert, Aggregationsgrade angeglichen.

Laden Das Laden des transformierten Datensatzes umfasst die technische Anbindung der ETL-Komponente an ein Data Warehouse.

Angenommen, ein Unternehmen vertreibt vier verschiedene Produkte (Muttern, Bolzen, Unterlegscheiben und Schrauben) in den Regionen Ost, West und Mitte. Die Unternehmensführung könnte nun die tatsächlichen Produktumsätze pro Region ermitteln und diese mit Umsatzvorhersagen vergleichen wollen. Diese Analyse erfordert eine mehrdimensionale Sicht auf die gespeicherten Daten, d.h., dieselben Daten müssen unter Verwendung mehrerer Dimensionen auf verschiedene Weisen betrachten werden können. Jedes Kriterium – Produkt, Preis, Kosten, Region und Zeitraum – repräsentiert eine andere Dimension.

Um diese Art von Information zu aggregieren, kann das Unternehmen spezielle Analysewerkzeuge einsetzen, die mehrdimensionale Sichten von Daten aus relationalen Datenbanken erstellen. So könnte beispielsweise ein Produktmanager mithilfe sogenannter **Online-Analytical-Processing**-Werkzeuge **(OLAP)** in Erfahrung bringen, wie viele Unterlegscheiben im Monat Juni in der Region Ost verkauft wurden, wie sich dieser Umsatz zum Umsatz des Vormonats oder zum Umsatz im Monat Juni des letzten Jahres verhält und ob dieser Umsatz der Umsatzvorhersage entspricht.

▶*Abbildung 6.17* zeigt ein entsprechendes mehrdimensionales Modell zur Darstellung von Produkten, Regionen, tatsächlichen Umsätzen und Umsatzvorhersagen. OLAP bietet eine Reihe von Grundoperationen wie Slicing, Dicing, Pivoting, Roll-Up und Drill-Down. Beim Slicing wird nur ein Ausschnitt („eine Scheibe“) des Würfels betrachtet und nur diese Daten werden analysiert. Beim Dicing wird der

Online Analytical Processing (OLAP) | Technik, um Daten nach mehreren Dimensionen bzw. aus mehreren Perspektiven zu analysieren.

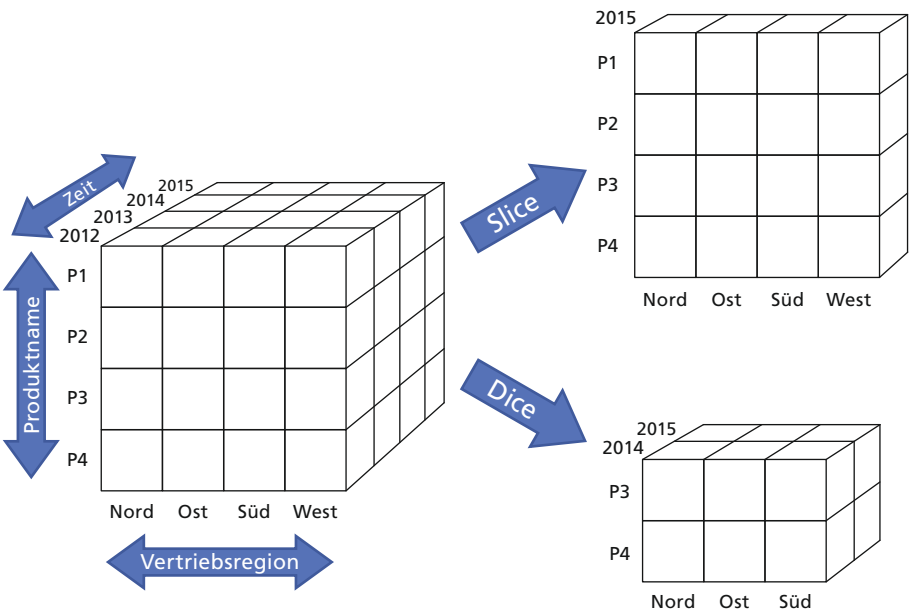

Abbildung 6.17: OLAP-Würfel sowie Slicing- und Dicing-Operationen (Chaudhuri und Dayal, 1997)

Würfel in mehreren Dimensionen gleichzeitig zugeschnitten, wodurch wieder ein (kleinerer) Würfel entsteht. Beim Pivoting wird der gesamte Würfel gedreht und die Daten werden aus einer anderen Dimension betrachtet. Unterschiedliche Aggregationsgrade werden durch Drill-Down (feiner auflösend) und Roll-Up (aggregierend) erreicht. Durch die Schachtelung von Würfeln können komplexe Datenansichten erzeugt werden. Zudem können Benutzer den Würfel entlang einer Dimension „zerschneiden", um so weitere Datenansichten zu erzeugen.

6.4.7 Data-Mining

Traditionelle Datenbankabfragen beantworten Fragen wie etwa „Wie viele Einheiten des Produkts mit der Nummer 403 wurden im Januar 2009 ausgeliefert?". OLAP, eine mehrdimensionale Analyse, unterstützt komplexere Informationsanforderungen, wie beispielsweise: „Vergleiche die Verkäufe für das Produkt 403 in Hinblick auf den Vierteljahresplan und die Verkaufsbereiche für die beiden vergangenen Jahre." Dennoch benötigt der Anwender für OLAP und eine abfrageorientierte Datenanalyse bereits eine klare Vorstellung darüber, welche Frage mit den Daten beantwortet werden soll.

Data-Mining hingegen ist im Kern ein explorativer Prozess. Unter Data-Mining versteht man den Einsatz verschiedener Techniken, um verborgene Muster und Beziehungen in großen Datenbeständen ausfindig zu machen und daraus auf Regeln zu schließen, die zur Vorhersage künftigen Verhaltens und als Orientierungshilfe für eine Entscheidungsfindung genutzt werden können (Fayyad et al., 2002; Hirji, 2001). Diese Muster und Regeln werden dazu benutzt, die Entscheidungsfindung zu unterstützen und die Auswirkungen dieser Entscheidungen vorherzusagen. Die Arten von Informationen, die durch Data-Mining abgeleitet werden können, sind u.a. Assoziationen, Sequenzen, Klassifizierungen, Cluster und Prognosen.

Data-Mining | Analyse großer Datenbestände, um Zusammenhänge, Muster und Regeln zu finden, die als Orientierungshilfe bei der Entscheidungsfindung und der Vorhersage künftiger Entwicklungen dienen können.

- **Assoziationen** sind Ausprägungen, die einem einzelnen Ereignis zugeordnet sind. Beispielsweise könnte eine Studie von Einkaufsmustern, die sogenannte Warenkorbanalyse, im Supermarkt zeigen, dass beim Kauf von Kartoffelchips in 65% aller Fälle auch Cola gekauft wird, während für den Fall einer speziellen Werbekampagne in 85% aller Fälle Cola gekauft wird. Mit dieser Information können Manager bessere Entscheidungen treffen, weil sie damit z.B. zielgerichteter werben oder die Produkte bedarfsgerechter anordnen können.
- In **Sequenzen** sind die Ereignisse über die Zeit verknüpft. Man könnte beispielsweise feststellen, dass beim Kauf eines Hauses in 65% aller Fälle innerhalb von zwei Wochen ein neuer Kühlschrank gekauft wird und in 45% aller Fälle innerhalb eines Monats ein Herd.
- Die **Klassifizierung** erkennt Muster, die die Gruppe beschreiben, zu denen ein Artikel gehört, indem sie vorhandene Artikel untersucht, die bereits klassifiziert wurden, und eine Regelmenge ableitet. Beispielsweise befürchten etwa Kreditkarten- oder Telekommunikationsanbieter die Abwanderung von Stammkunden. Die Klassifizierung kann helfen, die Eigenschaften von Kunden zu erkennen, die wahrscheinlich wechseln wollen, und ein Modell bereitstellen, das den Managern hilft vorherzusehen, um wen es sich dabei handelt, sodass man spezielle Kampagnen durchführen kann, um solche Kunden zu halten.
- **Clustering** arbeitet ähnlich wie die Klassifizierung, wenn noch keine Gruppen definiert wurden. Ein Data-Mining-Werkzeug erkennt verschiedene Gruppierungen innerhalb von Daten; beispielsweise kann es Ähnlichkeitsgruppen für Bankkarten ermitteln oder die Kundendatensätze einer Datenbank in Kundengruppen segmentieren, die auf Umfragen und den persönlichen Ausgabeverhalten beruhen.
- Die **Prognose-** oder **Forecasting-Technik** bedient sich statistischer Regressions- und Zeitreihenanalysen. Regressionsanalysen lassen sich heranziehen, um bei gegebenen (hypothetischen) Datenwerten, zukünftige Werte oder Ereignisse auf der Basis historischer Trends und Statistiken vorherzusagen (z.B. die Vorhersage des Absatzes von Fahrradzubehör auf Basis des Fahrradabsatzes im letzten Quartal). Im Gegensatz dazu sagen Zeitreihen nur zeitabhängige Datenwerte voraus (z.B.

die Stauwahrscheinlichkeit in der Ferienzeit auf Basis der tatsächlichen Staus in den Ferien des Vorjahres). Statistische Methoden und Prognosen basieren auf vom Anwender vorgegebenen Mustern (Modellen). Dabei werden konkrete Werte ermittelt. Hingegen sollen beim Data-Mining ohne diese Vorgabe Muster und Zusammenhänge erkannt werden.

Alle diese Data-Mining-Methoden sind dazu geeignet, Analysen über Muster und Trends auf einer sehr hohen Analyseebene aggregiert auszuführen, sie sind auch dazu geeignet, detaillierte Berichte zu erstellen, falls dies gewünscht wird. Data-Mining-Lösungen können in allen funktionellen Bereichen eines Unternehmens oder einer Verwaltung eingesetzt werden. Eine populäre Anwendung von Data-Mining ist die detaillierte Analyse von Mustern in Kundendaten zur Identifikation rentabler Kunden oder zur Planung von Eins-zu-eins-Marketingkampagnen.

Beispielsweise hat Virgin Mobile Australia Data Warehousing und Data-Mining dazu eingesetzt, die Kundenloyalität zu erhöhen und neue Services auf den Markt zu bringen. Im Data Warehouse von Virgin Mobile werden Daten aus dem ERP-System, dem CRM-System und dem Rechnungswesen in einer großen Datenbank konsolidiert. Dabei liefert das Data-Mining dem Management die Informationen, um die demografischen Profile neuer Kunden zu bestimmen und in Beziehung dazu zu setzen, welche Telefonmodelle diese Kunden gekauft haben. Außerdem helfen diese Datenanalysen dem Management dabei, die Rentabilität einer jeden Filiale sowie den Erfolg von Point-of-Sale-Kampagnen zu beurteilen. Auch die Reaktionen der Kunden auf neue Produkte und Dienstleistungen, die Kundenschwundquote und die durch jeden Kunden generierten Erlöse können durch Data-Mining besser eingeschätzt werden.

Predictive Analytics verwendet Techniken des Data-Mining. Ziel ist, historische Daten und Annahmen über zukünftige Umweltzustände in einem Modellansatz zu integrieren, um den Ausgang von Ereignissen in der Zukunft vorherzusagen, wie etwa die Wahrscheinlichkeit, dass ein Kunde auf ein Angebot reagiert oder ein spezifisches Produkt kauft. Beispielsweise wendete die US-Filiale von The Body Shop International plc. Predictive Analytics auf ihren Datenbanken der Katalog-, Web- und Einzelhandelskunden an, um Kunden zu identifizieren, die am ehesten aus einem Katalog bestellen würden. Diese Informationen halfen dem Unternehmen dabei, genauere und zielgerichtete Verteiler für ihre Kataloge zu erstellen, sodass die Beantwortungsquote der Kataloge und die Erlöse durch Katalogverkäufe gesteigert werden konnten.

6.4.8 Text-Mining und Web-Mining

Viele Werkzeuge für Business Intelligence & Analytics beschäftigen sich vorwiegend mit Daten, die in Datenbanken und Dateien strukturiert worden sind. Allerdings sollen Schätzungen zufolge unstrukturierte Daten, die meisten davon in Form von Textdateien, über 80% der nützlichen Informationen einer Organisation ausmachen. E-Mails, Memos, Callcenter-Transkripte, Umfrageantworten, Rechtsfälle, Patentbeschreibungen und Serviceberichte sind äußerst wertvoll für das Aufdecken von Mustern und Trends, die für Mitarbeiter zu einer besseren Entscheidungsgrundlage führen. Heute stehen den Unternehmen Text-Mining-Tools für die Analyse dieser Daten zur Verfügung. Mit diesen Tools lassen sich Schlüsselelemente aus großen unstrukturierten Datensets extrahieren, Muster und Beziehungen aufdecken und Informationen zusammenfassen. Die Unternehmen könnten Text-Mining so etwa für die Analyse von Aufzeichnungen der Anrufe bei Kundenservicezentren einsetzen, um größere Defizite bei Service und Reparaturen zu identifizieren.

Air Products and Chemicals in Allentown, Pennsylvania, verwendet Text-Mining zur Unterstützung bei der Identifizierung von Dokumenten, für die spezielle Aufbewahrungsverfahren gemäß dem Sarbanes-Oxley Act gelten. Das Unternehmen besitzt mehr als 9 Terabyte an unstrukturierten Daten (ohne E-Mails). Die SmartDiscovery-Software von Inxight Software klassifiziert und organisiert diese Daten so, dass das Unternehmen Geschäftsregeln auf eine Kategorie von Dokumenten statt auf Einzeldokumente anwenden kann. Wenn sich herausstellt, dass ein Dokument sich mit Vorgängen befasst, die unter das Sarbanes-Oxley-Gesetz fallen, kann das Unternehmen sicherstellen, dass das Dokument vorschriftsgemäß aufbewahrt wird.

Das Web ist eine weitere reiche Quelle von wertvollen Informationen, von denen einige nach Mustern, Trends und Einsichten in das Kundenverhalten durchsucht werden können. Das Aufdecken und die Analyse nützlicher Muster und Informationen aus

dem World Wide Web wird als Web-Mining bezeichnet. Unternehmen könnten Web-Mining für ein besseres Verständnis des Kundenverhaltens, die Bewertung der Effektivität einer bestimmten Website oder die Quantifizierung des Erfolgs einer Marketingkampagne nutzen. Zum Beispiel setzen Werbungstreibende Google Trends und Google Insights für Suchdienste ein, die die Beliebtheit verschiedener Wörter und Sätze verfolgen, die in Google-Suchanfragen verwendet werden, um herauszufinden, wofür sich die Leute interessieren und was sie gerne kaufen.

Web-Mining sucht nach Mustern in den Daten mittels Content-Mining, Structure-Mining und Usage-Mining. Als Web-Content-Mining wird der Prozess bezeichnet, in dem Wissen aus dem Content von Webseiten extrahiert wird. Diese Informationen können Text, Bilder, Audio- und Videodaten umfassen. Beim Web-Structure-Mining werden Daten überprüft, die mit der Struktur einer bestimmten Website in Beziehung stehen. Zum Beispiel geben Links, die auf ein Dokument verweisen, die Beliebtheit des Dokuments an, während Links, die von einem Dokument ausgehen, die Themenfülle oder auch die Bandbreite der Themen angeben, die in dem Dokument behandelt werden. Beim Web-Usage-Mining werden die Benutzerinteraktionsdaten geprüft, die von einem Webserver beim Erhalten von Anforderungen für die Ressourcen einer Website aufgezeichnet werden. In den Nutzungsdaten wird das Verhalten des Benutzers aufgezeichnet, wenn der Benutzer das Web durchsucht oder Transaktionen auf der Website vornimmt. Die Daten werden in einem Server-Protokoll gesammelt. Die Analyse solcher Daten kann Unternehmen dabei unterstützen, unter anderem den Wert bestimmter Kunden, produktübergreifende Cross-Marketingstrategien und die Effektivität von Werbekampagnen zu bestimmen.

Data-, Text- und Web-Mining sind mächtige und hilfreiche Werkzeuge, was gemäß datenschutzrechtlicher Überlegungen jedoch auch kritisch betrachtet werden kann. Mithilfe von Mining-Techniken können Daten aus verschiedenen Quellen zu einem detaillierten „Datenbild“ (Datenschatten) einzelner Personen kombiniert werden, das Auskunft über Attribute wie Einkommen, Fahrgewohnheiten, Freizeitaktivitäten, Familienmitglieder, politische Interessen etc. geben kann. Auf die Frage, ob und wie es Unternehmen erlaubt sein sollte, derart detaillierte Daten über einzelne Personen zu erfassen, wurde in *Kapitel 4* näher eingegangen.

Blickpunkt Technik

Big Data – großer Nutzen

Unternehmen heute haben mit einer wahren Flut an Daten aus Social Media, Suchabfragen und Sensoren sowie aus traditionellen Quellen zu kämpfen. 2012 soll die Menge der erzeugten digitalen Daten Schätzungen zufolge bei 988 Exabytes liegen, was einem Stapel Bücher von der Sonne zum Planet Pluto und zurück entspricht. Die Interpretation der „Big Data“ ist eine der größten Herausforderungen für Unternehmen aller Arten und Größen, aber sie bietet auch neue Möglichkeiten. Und es stellt sich die Frage, wie Unternehmen diese Möglichkeiten von Big Data zu ihrem Vorteil nutzen.

Die British Library musste sich erst an den Einsatz von Big Data gewöhnen. Jedes Jahr werden von den Besuchern der British-Library-Website über 6 Millionen Datenrecherchen durchgeführt und die Nationalbibliothek ist außerdem dafür zuständig, nicht mehr existierende britische Websites aus historischen Gründen zu bewahren, wie beispielsweise Websites zu ehemaligen Politikern. Die bisherigen Datenmanagementverfahren erwiesen sich für die Archivierung dieser Millionen von Websites als ungeeignet und die veralteten Analysetools konnten aus den riesigen Datenmengen keine nützlichen Informationen extrahieren. Angesichts dieser Herausforderungen suchte die British Library in Zusammenarbeit mit IBM nach einer Lösung für ihr Big-Data-Problem. IBM BigSheets ist eine Insight-Engine, die der Bibliothek dabei hilft, riesige Mengen unstrukturierter Webdaten zu sammeln, mit Anmerkungen zu versehen, zu analysieren und zu visualisieren, und am Ende die extrahierten Ergebnisse über einen Webbrowser auszugeben. So können Nutzer sich beispielsweise die Suchergebnisse als Tortendia-

▶ Forts.

gramm anzeigen lassen. IBM BigSheets setzt auf das Hadoop-Framework auf, das eine schnelle und effiziente Verarbeitung riesiger Datenmengen garantiert.

Strafverfolgungsbehörden auf allen Ebenen analysieren Big Data auf verborgene Muster bei Straftaten, wie Korrelationen zwischen Zeit, Gelegenheit und Organisationen, oder auf nicht direkt ersichtliche Beziehungen (*Kapitel 4*) zwischen Einzelpersonen und kriminellen Vereinigungen, die aus kleineren Datenmengen nicht abzulesen sind. Verbrecher und kriminelle Vereinigungen nutzen das Internet, um ihre Verbrechen zu koordinieren oder zu verüben. Neue Tools bieten Behörden die Möglichkeit, Daten aus den verschiedensten Quellen zu analysieren und anhand der Ergebnisse zukünftige Kriminalitätsmuster vorauszusagen. Das bedeutet, dass die Polizei Verbrechen proaktiver bekämpfen kann und im Idealfall durch rechtzeitige Präsenz verhindert, dass es überhaupt zu einer Straftat kommt.

Das Data Warehouse des Real Time Crime Center in New York City speichert Millionen von Datenpunkten zu Großstadtkriminalität und Straftätern. IBM und das New York Police Department (NYPD) haben zusammen ein Warehouse eingerichtet, das Daten zu mehr als 120 Millionen Strafanzeigen, 31 Millionen nationale Strafregistereintragungen und 33 Milliarden öffentlich verfügbare Registereinträge enthält. Mit den Suchfunktionen des Systems hat das NYPD schnellen Zugriff auf die Daten all dieser Datenquellen. Informationen zu Straftätern, wie Fotos der Verdächtigen, genaue Angaben zu ihren früheren Straftaten oder Adressen mit Wegbeschreibung, können in Sekunden auf einer Videowand angezeigt oder direkt einem Polizisten am Ort des Verbrechens übermittelt werden.

Andere Organisationen nutzen diese Daten, um sich umweltfreundlicher aufzustellen, oder, wie im Fall von Vestas, noch umweltfreundlicher aufzustellen. Vestas, mit Firmensitz in Dänemark, ist mit 43.000 Windkraftanlagen in 66 Ländern der größte Windenergieerzeuger der Welt. Standortdaten sind für Vestas sehr wichtig, damit es seine Anlagen genau dort aufstellen kann, wo die Windverhältnisse für die Energieerzeugung optimal sind. In Gebieten mit zu wenig Wind wird nicht genug Strom erzeugt und in Gebieten mit zu viel Wind können die Windräder Schaden nehmen. Deshalb verlässt sich Vestas bei der Wahl des besten Aufstellungsorts vornehmlich auf diese Standortdaten.

Vestas arbeitet mit einer Windbibliothek, die die Daten globaler Wettersysteme mit den Daten bereits im Betrieb befindlicher Windkraftanlagen abgleicht, um die Voraussetzungen für einen optimalen Standort zu ermitteln. Die vorherige Windbibliothek des Unternehmens legte den Informationen ein Messraster zugrunde, dessen Quadrate eine Größe von 27 × 27 Kilometern hatten. Den Ingenieuren bei Vestas ist es gelungen, die räumliche Auflösung auf 10 × 10 Meter zu verringern, um sich ein genaues Bild von den Windströmungsmustern an einem bestimmten Standort zu machen. Um die Genauigkeit seiner Standortbestimmungsmodelle jedoch noch weiter zu verbessern, musste Vestas die Rasterquadrate noch kleiner machen, was eine leistungsstärkere Datenmanagementplattform für das 10-Fache an Daten verglichen zu früherer forderte.

Das Unternehmen implementierte eine Lösung, die aus der IBM-Software InfoSphere BigInsights auf einem leistungsstarken Server vom Modell IBM System x iDataPlex bestand. (InfoSphere BigInsights ist ein Bündel von Softwaretools für die Analyse und Visualisierung von Big Data auf der Basis von Apache Hadoop.) Mit diesen Technologien konnte Vestas seine Windbibliothek stark erweitern und die Wetter- und Standortdaten mit viel genaueren und leistungsstärkeren Modellen verwalten und analysieren. Vestas' Windbibliothek umfasst zurzeit 2,8 Petabytes an Daten auf der Basis von ungefähr 178 Parametern wie Luftdruck, Luftfeuchtigkeit, Windrichtung, Temperatur, Windgeschwindigkeit und weitere historische Unternehmensdaten. Vestas plant für die Zukunft, globale Entwaldungsdaten, Satellitenbilder, Geodaten und Daten zu Mond- und Gezeitenphasen ebenfalls in die Datenbank mit aufzunehmen.

Nachdem das Unternehmen die Auflösung seines Winddatenrasters um fast 90 Prozent auf 3 × 3 Kilometer reduzierte, konnte es den optimalen Standort für eine Windkraftanlage in 15 Minuten ermitteln und nicht wie früher in drei Wochen, was den Vestas-Kunden eine wesentlich schnellere Rendite beschert.

▶ Forts.

Big-Data-Lösungen werden von Unternehmen aber auch zur Analyse des Konsumverhaltens genutzt. Der Autovermieter Hertz zum Beispiel sammelt Daten von Internetumfragen, E-Mails, Textnachrichten, Website-Verkehrsmustern sowie Daten, die an allen 8.300 Hertz-Standorten in 146 Ländern erhoben werden, und verwaltet diese Daten jetzt zentral anstatt in den einzelnen Niederlassungen. Das reduzierte den Zeitaufwand für die Datenverarbeitung und verbesserte die Reaktionszeit des Unternehmens auf Kundenfeedback und geändertes Konsumverhalten. Die Analyse der Daten aus mehreren Quellen ergab zum Beispiel, dass es in Philadelphia zu bestimmten Uhrzeiten zu Verzögerungen bei der Rückgabe kam. Nach Erkennen dieser Anomalie konnte das Unternehmen rechtzeitig gegensteuern und während dieser Spitzenzeiten die Anzahl der Mitarbeiter in der Philadelphia-Niederlassung erhöhen, sodass immer ein Ansprechpartner bei Problemen zur Verfügung stand. Das verbesserte nicht nur die Performance von Hertz, sondern auch die Kundenzufriedenheit.

Doch die Nutzung von Big Data hat auch Grenzen. In Zahlen zu schwimmen, heißt nicht unbedingt, dass die richtigen Daten gesammelt oder die besseren Entscheidungen getroffen werden. Letztes Jahr warnte ein Bericht des McKinsey Global Institutes vor einem Mangel an Spezialisten, die die richtigen Schlüsse aus all diesen Informationen ziehen können. Nichtsdestotrotz gibt es keine Anzeichen dafür, dass das Interesse an Big Data nachlässt; vielmehr ist zu erwarten, dass es in Zukunft noch viel mehr Big Data geben wird.

Quellen: Samuel Greengard, „Big Data Unlocks Business Value", Baseline, Januar 2012; Paul S. Barth, „Managing Big Data: What Every CIO Needs to Know", CIO Insight, 12. Januar 2012; IBM Corporation, „Vestas: Turning Climate into Capital with Big Data", 2011; IBM Corporation, „Extending and enhancing law enforcement capabilities", „How Big Data is Giving Hertz a Big Advantage" und „British Library and J Start Team Up to Archive the Web", 2010.

FRAGEN ZUR FALLSTUDIE

1. Beschreiben Sie die Art der Daten, die von den Organisationen in dieser Fallstudie gesammelt werden.
2. Nennen und beschreiben Sie die Business-Intelligence-Technologien, die in dieser Fallstudie angesprochen werden.
3. Aus welchen Gründen haben sich die Unternehmen in dieser Fallstudie dafür entschieden, Big Data zu sammeln und zu analysieren? Welche Vorteile hatten sie dadurch?
4. Nennen Sie drei Entscheidungen, die durch Heranziehen von Big Data verbessert wurden.
5. Welche Arten von Organisationen werden wahrscheinlich Big-Data-Management und Analysetools am nötigsten brauchen? Warum?

6.4.9 Data-Mining-Prozessmodelle

■ Daten-Wertschöpfungskette (Data Value Chain)

Der Begriff Wertschöpfungskette (*value chain*) wurde von Porter (1987) als Reihe von Aktivitäten, die Wertschöpfung erzeugen und aufbauen, definiert (siehe *Kapitel 1* und *3*). Miller und Mork (2013) haben dieses Konzept auf die speziellen Herausforderungen von Big Data angewendet und eine **Data Value Chain** (DVC) definiert, die Big-Data-Prozesse von der Datensammlung über die Datenanalyse bis zur Entscheidungsfindung unterstützen soll und dabei nicht an bestimmte Stakeholder und Technologien gebunden ist.

Die DVC ist dabei in drei Teilaufgaben Datenentdeckung (*Data Discovery*), Datenintegration und Datenerschließung (*Data Exploitation*) mit weiteren Unteraufgaben strukturiert (Miller und Mork, 2013, ▶*Abbildung 6.18*).

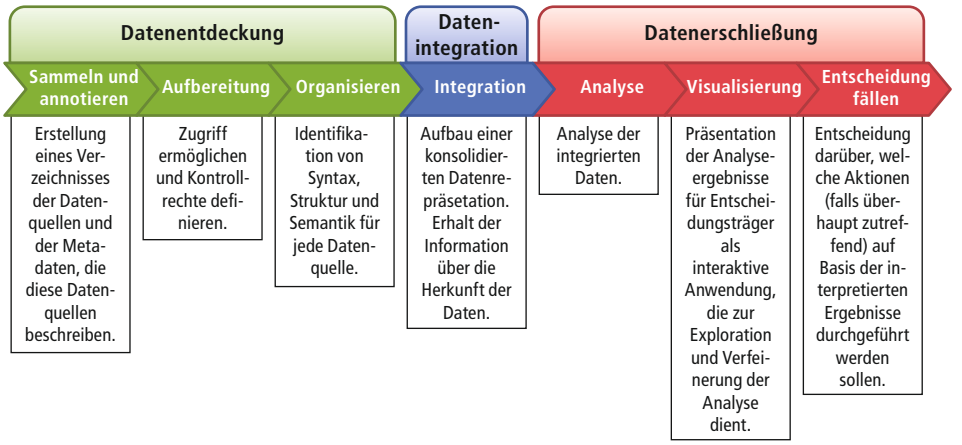

Abbildung 6.18: Die Data Value Chain (nach Miller und Mork, 2013, S. 58)

Die Teilaufgaben im Überblick:

- Datenentdeckung (Data Discovery): Bevor eine Analyse möglich ist, müssen die Daten aus verschiedensten Quellen gesammelt, mit Metadaten versehen und zugänglich gemacht werden.
- Datenintegration: Um eine (bestimmte) Datenanalyse durchführen zu können, müssen die Daten zu einer konsolidierten Repräsentation zusammengefasst werden. Dies geschieht etwa über Mappings, die eine Relation zwischen der Datenquelle und der konsolidierten Repräsentation definieren.
- Datenerschließung (Data Exploitation): Die Daten sind jetzt bereit, analysiert zu werden. Entscheidungsträger können sich dabei zumeist auf mehrere Analysen stützen, um fundierte Entscheidungen zu treffen. Es gelangen verschiedenste Verfahren zum Einsatz (siehe oben). Die Analysen sollten ausreichend dokumentiert sein, damit sie durch andere Analysten nachvollzogen werden können.

Prozessmodell Knowledge Discovery in Databases (KDD)

Die ungelenkte Anwendung von Data-Mining-Algorithmen führt meistens dazu, dass nur wenig bis nichts aussagende Muster (Patterns) in den Daten gefunden werden und die besonders wertvollen Informationen womöglich verborgen bleiben (Fayyad et al.,1996). Um dieses Problem zu entschärfen, begann im Jahr 1989 die Entwicklung am Prozessmodell **Knowledge Discovery in Databases, kurz: KDD** (Fayyad et al., 1996). Viele der darauf folgenden Data-Mining-Prozessmodelle bauen darauf auf und ähneln sich daher auch stark. KDD versteht sich als nichttrivialer Prozess der Identifizierung von validen, neuartigen, potenziell nützlichen und letztendlich verständlichen Patterns in Daten. Es handelt sich bei KDD um ein Prozessmodell, da es alle Schritte eines Data-Mining-Projekts definiert, aber es ist keine Methodik, da es nicht festlegt, wie die einzelnen Aufgaben ausgeführt werden sollen (Marbán et al, 2009). In einem komplexen Datensatz gibt es Hunderte Attribute und viele Datensätze, die für die gewünschte Data-Mining-Aufgabe irrelevant oder redundant sind (Relich und Muszynski, 2014). Am Anfang von KDD stehen daher eine Reihe von unterstützenden Phasen, die dazu dienen, die Daten vorzubereiten und die richtigen Data-Mining-Algorithmen auszuwählen. Kern bildet die Phase „Data Mining“, in der die ausgewählten Algorithmen auf die Daten angewendet werden. Danach folgt die Phase „Interpretation“, bevor mit der Phase „Using Discovered Knowledge“ der Abschluss gebildet wird. KDD ist iterativ: Von jeder Phase im Prozess kann in jede vorausgegangene Phasen zurückgesprungen werden, so entstehen gewollte Schleifen. KDD ist außerdem ein interaktives Prozessmodell, da der Benutzer im Laufe des Prozesses viele Entscheidungen treffen muss.

Nach diesem „Strickmuster“ sind viele weitere Modelle entstanden, z.B. SEMMA (Azevedo und Santos, 2008) und CRISP-DM. Letzteres wird im folgenden Abschnitt vorgestellt.

Cross-Industry Standard Process for Data Mining – CRISP-DM

Der **Cross-Industry Standard Process for Data Mining**, kurz: **CRISP-DM**, wurde 1996 von den damaligen Marktführern auf dem Gebiet des Data-Mining, Daimler-Benz, Integral Solutions Ltd., NCR und OHRA, ins Leben gerufen (Shearer, 2000). CRISP-DM ist ein frei verfügbares, hersteller-, werkzeug- und anwendungsunabhängiges Data-Mining-Prozessmodell. Das Modell ist sehr stark strukturiert: Die Inputs, Aufgaben und Outputs jeder Phase sind klar definiert und dokumentiert (zu finden in der CRISP-DM-Anleitung „CRISP-DM 1.0“, Chapman et al., 2000). Das Modell soll dadurch für Data-Mining-Einsteiger leicht zu erlernen, es soll aber trotzdem auf die speziellen Anforderungen jeder Anwendung anpassbar sein (Shearer, 2000). Die Version 1.0 wurde im Jahr 2000 durch die in der Zwischenzeit gebildete CRISP-DM Special Interest Group veröffentlicht (Shearer, 2000). In den Folgejahren hat sich CRISP-DM in der Praxis zum De-facto-Standard unter den DM-Prozessmodellen entwickelt (Marbán et al., 2009; Rennolls und Al-Shawabkeh, 2008).

CRISP-DM besteht aus 6 Phasen (Shearer, 2000): Geschäftsverständnis (*Business Understanding*), Datenverständnis (*Data Understanding*), Datenvorbereitung (*Data Preperation*), Modellierung (*Modeling*), Evaluierung (*Evaluation*) und Bereitstellung (*Deployment*). Diese Phasen sind ihrerseits in Unteraktivitäten weiter aufgeteilt. Es handelt sich um ein iteratives Modell, da die Erkenntnisse und Erfahrungen einer Iteration die Grundlage für einen neuen fundierteren Durchlauf bilden (▶*Abbildung 6.19*).

Die Phasen im Einzelnen (in Klammern jeweils die Bezeichnung der Subphasen):

Phase 1 – Geschäftsverständnis/Business Understanding: In der ersten Phase wird ein betriebswirtschaftlicher Blick auf die aktuelle Situation geworfen, um in den späteren Phasen zu verstehen, welche Daten mit welchen Methoden analysiert werden müssen. Es wird festgelegt, welche Fragen aus betriebswirtschaftlicher Sicht beantwortet, welche Ziele erreicht werden sollen und wie der Erfolg des DM-Projekts gemessen wird („Determine the Business Objective“). Dann soll der Data Scientist/Data Miner sich einen Überblick über die für das Data-Mining-Projekt zur Verfügung stehenden Ressourcen (Personal, Software und insbesondere Daten) verschaffen („Assess the Situation“). Außerdem werden die Projektrisiken identifiziert und Lösungspläne erstellt. Die Geschäfts-/Auswerteziele werden in die technische Perspektive mit der entsprechenden Terminologie übersetzt („Determine the Data Mining Goals“). Über einen Projektplan („Produce a Project Plan“) werden alle Planschritte aufgestellt, um die Auswerteziele zu erreichen.

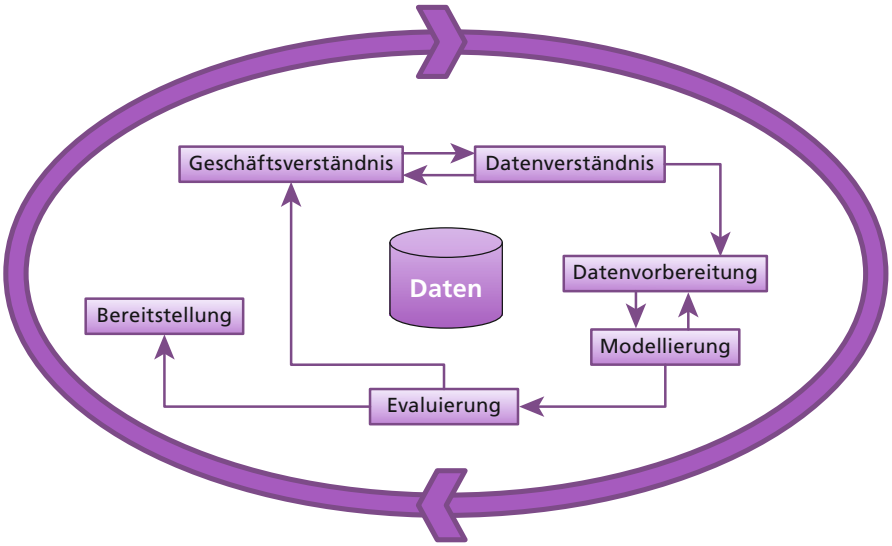

Abbildung 6.19: Der CRISP-DM-Lifecycle (Azevedo und Santos, 2008, Nachbildung)
Die Pfeile stellen die häufigsten Phasenübergänge dar (Shearer, 2000; Chapman et al., 2000).

Phase 2 – Datenverständnis/Data Unterstanding: Zunächst werden die zu betrachtenden Daten geladen bzw. zugänglich gemacht. Im Anschluss folgen weitere Aktivitäten, um mit den Daten vertraut zu werden, die Schlüsselattribute und Beziehungen zwischen den Daten zu identifizieren, interessante Subsets zu finden und Qualitätsprobleme in den Daten festzustellen.

Phase 3 – Datenvorbereitung/Data Preperation: Diese Phase enthält alle Aktivitäten, die nötig sind, um aus den Rohdaten den Datensatz zu formen, auf dem die Modellerstellungen und die Analysen ausgeführt werden können. Die Gesamtdatenmenge wird dabei hinsichtlich des Auswerteziels, der Qualität und technischen Beschränkungen reduziert („Select Data"). Qualitätsprobleme wie Datenlücken oder fehlerhafte Daten werden bereinigt, etwa durch Löschung ausgewählter Daten, Einsetzen von plausiblen Werten oder Schätzung („Clean Data"). Je nach Erfordernis werden abgeleitete Attribute oder gänzlich neue Datensätze erstellt, die das Modell vereinfachen („Construct Data"). Daten aus verschiedenen Quellen und Einträgen werden kombiniert, um neue Werte und Einträge zu erhalten („Integrate Data"). Gegebenenfalls müssen syntaktische Änderungen an den Datensätzen vorgenommen werden („Format Data").

Phase 4 – Modellierung/Modeling: Unterschiedliche Modellierungen („Select the Modeling Technique") und Verfahren werden ausgewählt und angewendet, ihre Parameter werden kalibriert. Wenn ein Verfahren nicht auf den vorhandenen Datensatz angewendet werden kann, muss zu Phase 3 zurückgesprungen werden. Ein oder mehrere konkrete Modeling-Verfahren werden ausgewählt. Bevor das Modeling ausgeführt wird, wird festgelegt, wie das resultierende Modell auf Qualität und Validität getestet werden kann („Generate Test Design"). Die Modellierung wird ausgeführt, ein oder mehrere (Analyse-) Modelle sind dabei der Output („Build the Model"). Der Data Mining Analyst interpretiert die erhaltenen Modelle in Kooperation mit Business-Analysten und Fachbereichsexperten und beurteilt den Erfolg hinsichtlich der Geschäftsziele („Assess the Model").

Phase 5 – Evaluierung: In dieser Phase wird das gesamte bisherige Data-Mining- Projekt evaluiert. Das erhaltene Modell und alle Schritte, die zu seiner Generierung geführt haben, müssen zu den Geschäfts-/Auswertezielen passen. Wenn Zeit und Budget es erlauben, können die Ergebnisse anhand einer praktischen Anwendung getestet werden. Zum Abschluss der Phase muss der Projektleiter entscheiden, ob die Ergebnisse bereit für das sogenannte Deployment (auf Deutsch etwa „Bereitstellung", „Einsatz") oder ob weitere Iterationen nötig sind.

Phase 6 – Bereitstellung/Deployment: Die gewonnen Kenntnisse müssen so strukturiert und präsentiert werden, dass der Anwender sie nutzen kann. Je nach Projekt kann die Deployment-Phase aus der Generierung eines Berichts bestehen oder die Implementierung eines wiederholbaren (automatisierten) Data-Mining-Prozesses umfassen. Im Rahmen einer Deployment-Strategie lässt sich dies festlegen („Plan Deployment"). Durch Überwachung und Wartung soll die unzweckmäßige Nutzung der DM-Ergebnisse vermieden werden („Plan Monitoring and Maintenance"). Ein finaler Bericht wird erstellt, der alle vorherigen Outputs und die DM-Ergebnisse enthalten soll („Produce Final Report"). Schließlich erfolgt eine Beurteilung des Projekts hinsichtlich Erfolge und Misserfolge, um für zukünftige DM-Projekte aus den Erfahrungen zu lernen („Review Project").

Datenbanken und das Web 6.5

Wenn ein Nutzer versucht, im Web einen Auftrag zu erteilen oder einen Produktkatalog anzuzeigen, greift er in diesem Moment vermutlich auf eine Webseite zu, die mit einer internen Unternehmensdatenbank verknüpft ist. Viele Unternehmen verwenden inzwischen das Web, um Kunden und Geschäftspartnern Informationen aus ihren internen Datenbanken zur Verfügung zu stellen. ▶*Abbildung 6.20* veranschaulicht, wie der Zugriff eines Kunden auf eine interne Datenbank des Händlers über das Web erfolgen könnte, wenn der Kunde mit einem Webbrowser in der Onlinedatenbank des Händlers nach Preisinformationen sucht.

Der Anwender greift über das Internet mit einem Webbrowser auf seinem Client-PC auf die Händler-Website zu. Der Browser des Anwenders fordert Daten aus der Unternehmensdatenbank an und verwendet HTTP zur Kommunikation mit dem Webserver. Weil viele Backend-Datenbanken über andere Protokolle als HTTP kommunizieren, übergibt der Webserver diese Datenanforderungen an Software, die eine Transformation der Befehle auf SQL-Anfragen durchführt, welche das DBMS der Datenbank verarbeiten kann. In einer Client-Server-Umgebung befindet sich

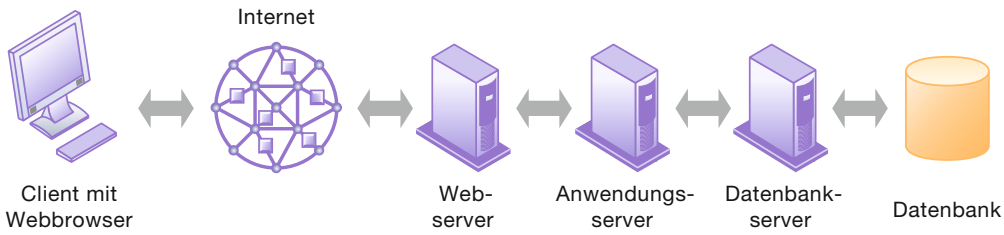

Abbildung 6.20: Verknüpfung interner Datenbanken mit dem Web
Die Benutzer greifen mithilfe z.B. eines PCs und eines Webbrowsers über das Internet auf die interne Datenbank eines Unternehmens zu.

das DBMS auf einem speziellen Computer, der als **Datenbankserver** bezeichnet wird. Das DBMS empfängt die SQL-Abfragen und stellt die erforderlichen Daten bereit. Die Middleware übermittelt die Informationen von der internen Unternehmensdatenbank zurück an den Webserver, der sie dem Anwender in Form einer Webseite zur Verfügung stellt.

Abbildung 6.20 zeigt, dass die Middleware zwischen dem Webserver und dem DBMS aus einem Anwendungsserver bestehen kann, der sich auf einem eigenen speziellen Computer befindet (vgl. *Kapitel 5*). Der Anwendungsserver verarbeitet alle Anwendungsoperationen, wie die Dialogverarbeitung und den Datenzugriff, zwischen dem Computer, auf dem der Browser läuft, und den Backend-Anwendungen oder -Datenbanken der Unternehmung. Der Anwendungsserver übernimmt die Anforderungen vom Webserver, führt die notwendigen Schritte zur Transaktionsverarbeitung auf Basis dieser Anforderungen aus und stellt die Verbindung mit den Backend-Systemen oder -Datenbanken der Organisation her. Alternativ kann es sich bei der Software zur Verarbeitung dieser Operationen um ein selbst erstelltes Programm oder ein CGI-Skript handeln. Ein CGI-Skript ist ein kompaktes Programm, das die CGI-Spezifikation (*Common Gateway Interface*) zur Datenverarbeitung auf einem Webserver verwendet.

Der Zugriff auf die internen Datenbanken eines Unternehmens über das Internet ist mit zahlreichen Vorteilen verbunden. Zunächst einmal ist die Verwendung von Webbrowsern viel einfacher als der Einsatz proprietärer Abfragetools. Darüber hinaus erfordert die Webschnittstelle nur wenige oder gar keine Anpassungen der internen Datenbank. Außerdem ist es wesentlich preisgünstiger, eine Webschnittstelle einem bewährten System hinzuzufügen, als das gesamte System neu zu entwerfen und auszubauen, wenn man den Anwenderzugriff verbessern möchte. Schließlich ermöglicht der Zugriff auf Unternehmensdatenbanken über das Web auch neue Funktionen, Chancen und Geschäftsmodelle.

Beispielsweise stellt ThomasNet.com ein Onlineverzeichnis für mehr als 650.000 Lieferanten von Industrieprodukten zur Verfügung, das z.B. Lieferanten von Chemikalien, Metallen, Kunststoffen und Automobilzubehör beinhaltet. Unter dem früheren Namen Thomas Register versendete die Unternehmung umfangreiche Papierkataloge mit diesen Informationen.

iGo.com ist eine Firma, die Batterien und Zubehör für Mobiltelefone und Computer über das Internet vertreibt. Die Website von iGo.com ist mit einer umfassenden relationalen Datenbank voller Artikeldaten über Batterien und Peripheriegeräte nahezu aller Marken und Modelle für Notebooks und portable elektronische Geräte verbunden.

Die Website der Internet Movie Database (imdb.com) ist mit einer umfangreichen Datenbank verknüpft, die Zusammenfassungen, Besetzungslisten und Biografien der Schauspieler für nahezu jeden Film enthält, der jemals produziert wurde.

Datenmanagement in der Praxis 6.6

Eine erfolgreiche Gestaltung sowie ein erfolgreicher Einsatz von Datenbanken in der Praxis erfordert viel mehr als lediglich die Auswahl eines konkreten Datenbankmanagementsystems. Der Einsatz scheitert, wenn Datenbanksysteme vom Management des Unternehmens nicht unterstützt und verstanden und keine entsprechend notwendigen organisatorischen Änderungen vollzogen werden. Daher sind folgende Einflussfaktoren von kritischer Bedeutung: die Datenverwaltung, die Datenmodelle, die Modellierungsmethoden sowie die Endanwender (▶*Abbildung 6.21*).

Copyright