

# Data Warehouseing

## IBAI-Report 7/2001

Analyseorientierte Informationssysteme = Datawarehouse ?

### 1. Einleitung

Durch die Einführung von CRM in Unternehmen bekommen auch die unternehmenseigenen Daten eine neue Bedeutung. Vorbei sind die Zeiten, in denen Datenmanagement nur die richtige Adresse im Mailing zu benutzen bedeutete. Die heute angestrebte individuelle Kundenkommunikation benötigt mehr und besser qualifizierte Informationen zum Kunden und Kundenverhalten. Das Datenmanagement bildet dabei die Grundlage, um mit intelligenten Methoden z.B. des Data Minings dieses umfassende Wissen bereit zustellen und den Kunden in differenzierte Zielgruppen einzuteilen. Daraus resultierend ist eine optimale Kommunikation zum Kunden aufzubauen. Es geht also darum, die unterschiedlichen Techniken zur Informationsspeicherung, -verwaltung und Analysen im Unternehmen so einzusetzen, daß Kunden optimal bedient werden können.

### 1.1. Wissen, Information und Daten

In vielen Unternehmen wird kaum zwischen den Begriffen unterschieden. Information ist als Betrachtungsgegenstand unter anderem in der Informatik und in der Wirtschaftswissenschaft geläufig, wobei in der Literatur stark divergierende Begriffsauffassungen erkennbar sind und auch innerhalb beider Fachbereiche unterschiedliche Ansätze existieren. In der Informatik werden die Begriffe Information und Daten häufig synonym verwendet, da hier eine explizite Abgrenzung nicht unbedingt erforderlich scheint (Lehner et. al 1994). Dabei werden die Daten gleichgesetzt mit den Informationen, die sie repräsentieren. Die Wirtschaftswissenschaft dagegen sieht Information sowohl als bedeutsamen Produktionsfaktor wie auch als Zwischen- oder Endprodukt des betrieblichen Transformationsprozesses an, deshalb wird hier eine deutliche Abgrenzung zwischen Wissen und Daten vorgenommen. Ausgangspunkt ist häufig die Definition von Information als zweckorientiertem Wissen.(Wittmann 1959). So unterschiedlich der Umgang mit den Begriffen ist, um so wichtiger ist es hier die verwendete Sicht zu erläutern. Wissen stellt das begriffliche Dach dar, unter dem sich sowohl Daten als auch Information als Teilmenge subsumieren läßt. Wissen besteht aus Wahrnehmung, Erfahrung und Kenntnissen einer Person über ihre Umwelt.

### 1.2. Arten von betrieblichen Informationssystemen

Im Weiteren werden nur computergestützte Informationssysteme betrachtet, andere innerbetriebliche Informationssysteme z.B. Besprechungen, Aushänge und ähnliches werden hier nicht weiter berücksichtigt. In der Praxis haben sich zwei grundverschiedene Arten von Informationssystemen in den Unternehmen etabliert: die sogenannten operativen und analyseorientierten Informationssysteme. Informationssysteme zur Unterstützung der operativen, leistungserstellenden Prozesse sind wohl in praktisch allen Unternehmen in unterschiedlicher Ausprägung anzutreffen. Sie automatisieren, standardisieren Arbeitsvorgänge und können Nutzeffekte erzielen, die sich ursächlich auf die ökonomische Effizienz der elektronischen Informationsverarbeitung zurückführen läßt. Der Einsatz der operativen Systeme zielt zunächst auf die Rationalisierung der standardisierten, administrativen Abläufe, die durch den Anfall großer Datenmengen charakterisiert sind, damit geht einher eine Verkürzung der Durchlaufzeiten von Prozessen.(Müller, 2000) Die operativen Informationssysteme bestehen aus der Summe aller Einzelsysteme, die zur Erledigung des Tagesgeschäftes notwendig sind. Teil der operativen Informationssysteme sind neben den klassischen Warenwirtschaftssystemen, auch die Finanzbuchhaltung oder aber auch die Datenerfassungssysteme. Alle diesen Anwendungen ist gemeinsam, das sie auf die jeweiligen Teilaspekte der operativen Arbeit sowohl in der Datenverarbeitung als auch in der Menge der zu speichernden Daten optimiert wurden. Dadurch und durch die Tatsache, das solche Systeme häufig sehr heterogen zusammengesetzt sind, werden Datenbestände redundant geführt und gepflegt. Dies birgt das Risiko das Inkonsistenzen in den Datenbeständen entstehen. Ein weiteres Merkmal der operativen Systeme bildet die Art und Weise der Datenverarbeitung, hier werden große Menge von Einzelsätze verarbeitet. Die zweite Art der Informationssysteme, die analysebasierten Informationssystemen ist Gegenstand der ausführlichen Betrachtung in Kapitel 2.

### 1.3. Quelle und Qualität der Daten

Daten, die im Unternehmen zur Informations- und Wissensgewinnung genutzt werden sollen, können entweder aus interne oder aus externen Quellen stammen. Die größte Zahl an Daten entsteht intern, das heißt in den operativen Informationssystem beim Durchlaufen der unterschiedlichen Prozesse. Da diese Daten in erster Linie zur Abwicklung des Tagesgeschäftes dienen, wirkt sich das Fehlen einer umfassenden Historie in den operativen Systemen wenig aus. Durch die teilweise redundante Datenhaltung in den sehr unterschiedlichen Teilsystemen entstehen unter Umständen Inkonsistenzen. Sowenig sich viele Qualitätsmängel in den Daten auf die Arbeit mit und in den operativen Systemen auswirken, um so größer ist diese Auswirkung auf die analyseorientierten

Informationssysteme. Die Qualität der Daten hat einen maßgeblichen Einfluß auf die Qualität der darauf basierenden Analysen. Externe Daten entstehen außerhalb der unternehmenseigenen Prozesse, häufig werden Sie als Zusatzinformationen (z.B. eine Schuldnerbeauskunftung) oder als Vergleichswerte (z.B. Zahlen des Statistischen Bundesamtes) hinzugezogen. Für die diversen Fragestellungen z. B. im Database Marketing (DBM) und Customer Relationship Management (CRM) werden den analyseorientierten Informationssystemen häufig auch speziell für diese Fragestellung eingekaufte Zusatzinformationen zum Kunden oder zur Adresse des Kunden zugefügt. Grundsätzlich ist festzustellen das häufig die Qualität der internen Daten besser ist als die der externen Bestände nicht zuletzt deshalb, da sich die Entstehung/Erhebung der internen Daten einfacher kontrollieren und verbessern läßt.

## **2. Analyseorientierte Informationssysteme**

Als analyseorientierte Informationssysteme versteht man die Menge aller Systeme, die notwendig sind, die Daten zu speichern und für Analysen bereit zu stellen wie auch die Endbenutzerwerkzeuge, mit deren Hilfe man aus den Informationen und Wissen gewinnen kann. Im Sinne einer schematisierten Architekturbetrachtung sind die angeschlossenen Front-Endwerkzeuge, mit deren Hilfe die Benutzer auf die Datenbestände zugreifen können, von den datenspeichernden Komponenten zu trennen und somit keine Bestandteile des Data Warehouse im engeren Sinne. Idealisiert liegt auch hier Daten-Programm Unabhängigkeit vor, die es erlaubt, mit verschiedenen Tools über offene Schnittstellen auf die Daten zuzugreifen, so daß für die Back-End- wie für die Front-End-Seite jeweils besonders geeignete und unabhängig voneinander austauschbare Hard- und Softwarekomponenten zum Einsatz kommen können. Trotz dieser begrifflichen Trennung darf jedoch nicht übersehen werden, daß in der Praxis häufig eine enge technologische Verzahnung der Endbenutzerwerkzeuge mit den datenspeichernden Komponenten vorliegt. Im einzelnen setzt sich damit ein analyseorientierte Informationssystem aus folgende Bereichen Data Warehouses, Data Marts und als Endbenutzerwerkzeuge (OLAP und Data Mining Tools) zusammen.

### **2.1. Ziele und Nutzen**

Für die Wettbewerbsfähigkeit eines Unternehmens spielt der Faktor Information eine entscheidende Rolle. Es werden in Zukunft diejenigen Unternehmen Wettbewerbsvorteile erzielen können, die durch den Einsatz innovativer Technologien schnell und flexibel auf sich rasch verändernde Marktfaktoren und Kundenbedürfnisse reagieren. In den operativen Datenbanken des Tagesgeschäftes steht eine Flut von Daten zur Verfügung. Dieses Kapital wird jedoch vielfach nur unzureichend genutzt bzw. liegt brach. Wesentliche Informationen stehen den Entscheidern auf Management- und Sachbearbeiterebene zum entscheidenden Zeitpunkt nicht oder zumindest nicht in der Form zur Verfügung, die für eine kreative Analyse und Entscheidungsfindung notwendig wäre. Über entsprechende Untersuchungen konnte für die Datenbanken dieser Systeme die bekannte 80:20-Regel bestätigt werden, die besagt, daß 20% der Daten 80% der für die Entscheider relevanten Informationen enthalten und umgekehrt. Als fundierte Basis für unternehmensrelevante Entscheidungen müssen aussagefähige Daten schnell und gezielt zur Verfügung gestellt werden können. So stellte es kein Problem dar, die Anzahl oder das Durchschnittsalter neuer Kunden oder Interessenten auszugeben. Während die Nutzung der Daten zur Beantwortung entscheidungsrelevanter Fragen einen großen strategischen Vorteil darstellt: Beispiele für solche Fragen sind:

- Welchen Kunden sollte wann welches Angebot unterbreitet werden?
- Welche Kunden sind gefährdet?
- Wie hoch ist das Cross-Selling-Potential für ein neues Produkt?
- Welcher Lifetime-Profit läßt sich mit welchem Kunden erzielen?
- Wie lassen sich Top-Interessenten mit hohen Lifetime-Values gewinnen?
- Welcher Umsatz wird im nächsten Jahr erzielt?

Kernfrage ist: Warum blieben gerade diese, für das Management entscheidende, Fragen häufig unbeantwortet? Betrachtet man den Charakter der Fragestellungen, so liegt die Antwort auf die aufgezählten Fragen nicht in einem einzelnen Datenfeld oder einem Kundenmerkmal, sondern in der richtigen Kombination unterschiedlicher Kundeninformationen. So kann bspw. die Angebotsaffinität eines Kunden von einer Vielzahl von Merkmalen wie Alter, Geschlecht, Familienstand, demographischen Typologien, bisher gekauften Produkten, gezeigtem Produktinteresse, Zahlungsmoral und einer Reihe weiterer Eigenschaften abhängen.

## **1 Data Warehouse**

Grundsätzlich wird ein Data Warehouse als eine Datensammlung zum Zweck der Analysen verstanden. Aus allen teilweise sehr unterschiedlichen operativen Systemen eines Unternehmens werden die relevanten Daten für das Data Warehouse auf ihre Qualität geprüft und dann gegebenenfalls aufbereitet in das Data Warehouse geladen. Im Gegensatz zu den operativen Systemen spielt die Abbildung von Vergangenheitsdaten/Historien und externen Daten im Data Warehouse eine große Rolle. Unter dem Begriff Data Warehouse wird heute im

allgemeinen eine Datenbank verstanden, die als unternehmensweite Datenbasis für das gesamte Spektrum der Anwendungen zur Unterstützung der analytischen Aufgaben von Fach- und Führungskräften dient (Gluchowski 1997). Diese wird getrennt von den operativen Informationssystemen betrieben und aus diesen und aus externen Quellen mit Daten gefüllt. Hierbei wird eine logische Zentralisierung angestrebt. Unter einem Data Warehouse versteht man im allgemeinen kein konkretes Datenbanksystemprodukt, sondern eine unternehmensindividuelle Anwendung des zugrundeliegenden Konzepts der Separierung der entscheidungsbezogenen Daten, deshalb wird vielfach auch von Data Warehouse-Konzept gesprochen. Die Inhalte eines Data Warehouses lassen sich durch vier wesentliche Merkmale charakterisieren, die bereits deutliche Unterschiede zu den operativen Daten erkennen lassen: (Immon et.al, 1994)

- Themenorientierung
- Logische Integration und Homogenisierung
- Zeitraumbezug
- Geringe Volatilität
- Themenorientierung:

Im Gegensatz zu den operativen Systemen, die sich nach Organizationseinheiten, Aufgabengebieten und Arbeitsprozessen gliedern, orientieren und organisieren sich die Inhalte des Data Warehouses nach Sachverhalten, welche die Entscheidungsgegenstände im Unternehmen betreffen. Typische Themen sind dabei die Kunden, die Produkte und die Werbe- bzw. Vertriebsmaßnahmen. Neben den Inhalten hat dieser Umstand natürlich auch großen Einfluß auf das logische Datenmodell eines Data Warehouse.

#### **Logische Integration und Homogenisierung:**

Ein Data Warehouse lebt von einheitlichen Datenstrukturen. Es wird eine unternehmensweite Integration aller relevanten Daten zu einem konsistenten Datenbestand in einem durchgängig modellierten System angestrebt. Dieses Ziel ist zugleich Ausdruck des genannten Merkmals der Themenorientierung, denn dies impliziert z.B. die funktionsübergreifende Verwendung der Daten. (Müller, 2000)

#### *Zeitraumbezug:*

Die Informationen zur Entscheidungsunterstützung sollen zwar schnell und zeitnah bereitgestellt werden, aber eine zeitpunktgenaue Bearbeitung von Daten wie Sie in den operativen Systemen stattfindet ist eher unwichtig. Viel wichtiger ist es das Problemlos unterschiedliche Zeiträume in die jeweiligen Analysen einfließen können. Im operativen System hat der Faktor Zeit nur eine beschreibende Rolle, im Data Warehouse ist er ein wichtiger struktureller Bestandteil. Die Besonderheit des Data Warehouse ist unter anderem die Tatsache, das auch historische Daten gespeichert werden, die in den operativen Systemen schon lange archiviert worden sind oder bei einer Reorganisation vernichtet wurden.

#### **Geringe Volatilität**

Daten, die einmal in einem Data Warehouse abgelegt wurden, sollten sich nach Möglichkeit nicht mehr verändern, Ausnahmen bilden Fehler in den Daten z.B. durch einen fehlerhaften Ladeprozeß. Ansonsten ändern sich Daten im Data Warehouse kaum, neue Datensätze eines Vorgangs werden eingefügt, die alten werden nicht überschrieben. Dies ist ein klarer Gegensatz zu den operativen Systemen. Als Beispiel könnte man den Kauf eines Produktes mit anschließender Stornierung betrachten. Im operativen System würde der Datensatz mit der Bestellung mit dem Stornodatensatz überschrieben oder gelöscht. Im Data Warehouse hätte man zwei Datensätze einen mit der Bestellung und einen mit der Stornierung. Dadurch könnte man im Data Warehouse die Aktion des Kunden nachvollziehen und in den unterschiedlichen Analysen berücksichtigen. Im operativen System wäre im externen Fall keine Information mehr vorhanden, da sie ja für die weitere operative Verarbeitung nicht mehr relevant ist.

Zusammenfassend kann man also sagen, daß es sich beim Data Warehouse um eine zentrale Datenbank handelt, die mit den genannten Eigenschaften der Speicherung aller analyserelevanten Daten im Unternehmen dient. Allgemein kann daher unter dem Data Warehouse-Konzept -unabhängig von der konkreten Architektur oder Realisierungsform- der Gedanke verstanden werden, die operativen Daten in der genannten Form integriert und unabhängig von den operativen Systemen zu speichern, um sie dann zu entscheidungsunterstützenden bzw. analytischen Zwecken zu nutzen.(Müller 2000).

Anders als im operativen System wird ein Data Warehouse für den effizienten lesenden Zugriff auf große Datenmengen in komplexen Strukturen konzipiert. Ein besonderes Augenmerk ist dabei auf den sich ändernden Informationsbedarf gerichtet. Durch diesen wird es notwendig, die Strukturen so auszulegen, daß auch komplexe Queries, die große Datenmengen betreffen, sowie umfangreiche Join- und Aggregationsoperationen bewältigt werden können.

Diese typische Form der Nutzung eines Data Warehouse führt dazu, das sich die Auslastung eines Data Warehouse deutlich von der der operativen Systeme unterscheidet. Im Data Warehouse unterliegt die Auslastung erheblichen Schwankungen mit ausgeprägten Lastspitzen, die im direkten Zusammenhang mit der gerade getätigten Abfrage/Analyse steht. Im Gegensatz dazu ist die Auslastung eines operativen Systems nahe zu konstant und bewegt sich auf einem gleichmäßig hohen Niveau. Eben dieses führt in Unternehmen, die auf ein Data Warehouse und/oder auf Data Marts verzichten, aber gleichzeitig die komplexe Analysen zur entscheidungsunterstützung benötigen, zu einem massiven Ressourcenkonflikt zwischen der Durchführung der Analyse und der Abwicklung des Tagesgeschäfts.

Zur Realisierung eines Data Warehouses werden im allgemeinen drei Organisationsformen diskutiert.

- zentrale Datenbankanwendung
- verteilte Datenbankanwendung
- virtuelles Data Warehouse

Die häufigste Realisierungsform stellt das zentrale Data Warehouse dar, bei dem die Verwaltung aller Datenbestände für die verschiedenen Front-End-Anwendungen auf einem einzelnen Rechner erfolgt. (Mucksch et al, 1998)

Versteht man ein Data Warehouse als einen Anwendungsfall eines Datenbanksystems, dann kann auf das allgemeine Konzept der Aufteilung eines Datenbanksystems in die drei Komponenten Datenbankverwaltungssystem, Datenbank und Datenbankkommunikationsschnittstelle zurückgegriffen werden. Das Datenbankverwaltungssystem ist im Data Warehouse die zentrale Instanz zur Verwaltung der analyseorientierten Datenbestände. Es stellt die Funktionen zur Datendefinition und -manipulation bereit, hierbei werden an das Datenbankverwaltungssystem eines Data Warehouses andere Anforderungen gestellt als an das eines operativen Systems.

Die Fragestellung nach Integrität und Konsistenz in den Datenbeständen ist unter Analysegesichtspunkten anders zu bewerten als im operativen System. Auch das Thema Datensicherheit und Verfügbarkeit des System sind anders zu bewerten, so führt die Tatsache, das es sich bei den Daten eines Data Warehouse um Kopien aus operativen Datenbeständen handelt, dazu, das die Anstrengungen zur Verfügbarkeit nicht so groß sein müssen wie im operativen System. Gleichzeitig haben die Daten aber durch die andere Form der Speicherung einen Wert erhalten, der die aus den Daten zu generierenden Informationen auch für Unbefugte "leichter" zur Verfügung stellt, so daß erhöhte Anforderungen an das Thema Sicherheit und daraus resultierenden Rollenkonzepten gestellt werden. Damit die Nutzer kurze Antwortzeiten auch bei komplexen Anfragen und Analysen erhalten, muß die Speicherung und der Zugriff auf die Daten entsprechend optimiert werden, was aber nicht zu einer Einschränkung der Flexibilität bei den Analysen führen darf. Viele Unternehmen helfen sich hier mit Data Marts um dem Nutzer voraggregierte Informationen zur Verfügung stellen. Die Data Warehouse-Datenbank bildet mit den dort gespeicherten aktuellen und historischen Daten aus allen Unternehmensbereichen in den unterschiedlichen Verdichtungsstufen den Kern des analyseorientierten Informationssystems. Hier ist aus der Sicht der Nutzer ein Zielkonflikt erkennbar: aggregierte Daten sind für Auswertungen von großem Interesse, allerdings erlauben nur detaillierte Einzelwerte die Flexibilität, die Daten den neuen Anforderungen entsprechend zu verknüpfen. Auch in diesem Fall befreit einen die Erstellung von Data Marts aus dem Dilemma, so werden in Data Marts die aggregierten und ggf. auch transformierten Daten bereitgestellt, während im Data Warehouse die Daten auf der feinsten zur Verfügung stehenden Granularität gespeichert sind. Für das Data Warehouse hat sich mit den Jahren die relationale Speicherung der Daten als Quasi-Status heraus kristallisiert, Während Data Marts je nach Verwendungszweck sowohl relational als auch multidimensional erstellt werden. Die Datenbankkommunikationsschnittstellen spielen im analyseorientierten Informationssystem eine herausragende Rolle, da ohne sie die Nutzung der sich im Datawarehouse befindlichen Daten nur schwer möglich ist. Gleichzeitig entsteht hier die Situation, das abhängig von den eingesetzten Fornt-Endwerkzeugen und deren internen Datenhaltungssystemen ganz unterschiedliche Anforderungen an die Schnittstellen gestellt werden. Ein unverzichtbarer Bestandteil der analyseorientierten Infromationssysteme und besonders des Data Warehouse sind ausführliche Meta-atenbanksysteme. Anders als im operativen System, wo ihre Rolle eher untergeordnet ist, sind sie besonders für die Nutzer/Anwender des Data Warehouses und der Data Marts unverzichtbar um effektiv Analysen auf den Daten durchführen zu können. Man kann eine gut gepflegte und weitgehend mit den Geschäftsbegriffen versehene Meta-Datenbank als einen der kritischen Erfolgsfaktoren für ein analyseorientiertes Informationssystem sehen.

### **2.3. Data Mart**

Im folgenden soll ein gemeinsames Verständnis von Data Marts geschaffen werden, da dieser Begriff teilweise sehr unterschiedlich benutzt und gegen ein Data Warehouse abgegrenzt wird. Die Begriffe Data Warehouse und

Data Mart bezeichnen innerhalb eines analyseorientierten Informationssystems die Bausteine, die der Datenspeicherung dienen. Unter einem Data Mart versteht man eine spezifische Datensammlung, in der nur die Bedürfnisse der jeweiligen Datensicht bzw. Nutzung abgebildet werden und häufig in sehr unterschiedlicher Weise in das Gesamtkonzept analyseorientierter Informationssysteme eingeordnet. So wird ein Data Mart einerseits als Teilmenge des Data Warehouse gesehen, in dem ein Ausschnitt der Datenbestände nochmals redundant gehalten wird. Dies läßt sich aus der Größe und Struktur des Data Warehouse begründen. Es beinhaltet sehr große Datenbestände, die auf relationalen Datenbanksystemen basieren und somit bezogen auf Anfragen in nicht vollständig problemadequaten Strukturen organisiert sind. Insbesondere beim interaktiven Zugriff auf die Datenbestände erweist sich daher das Data Warehouse hinsichtlich der Repräsentation der Daten und auch der Antwortzeiten möglicherweise als nicht anforderungsgerecht. Zur Lösung dieses Problems können z.B. funktions- oder bereichsspezifische Extrakte aus der Data Warehouse-Datenbank entnommen und redundant in einem Data Mart gespeichert werden. Dieser kann mit der gleichen Technologie und einem Datenmodell realisiert sein, das einer echten Teilmenge des Data Warehouse entspricht, so daß eine leichte Pflegbarkeit des Data Mart vorliegt. Alternativ erscheint es aber auch zweckmäßig, für den Data Mart mit seinem überschaubaren Datenvolumen im Gegensatz zum relational basierten Data Warehouse ein mehrdimensionales Datenbanksystem zu nutzen, um die potentiell besseren Modellierungs- und Abfragemöglichkeiten dieser Technologie ausnutzen zu können. Aufgrund der dann notwendigen Transformation der Daten in das neue Modell wird allerdings die Pflege solcher Data Marts aufwendiger, so daß hier eine Abwägung der Vor- und Nachteile heterogener Datenmodelle erforderlich ist. Die Anwender erhalten mit dem Data Mart einen auf ihre Informationsbedürfnisse zugeschnittenen Ausschnitt aus der unternehmensweiten Datenbasis. So können bei sorgfältiger Abgrenzung dieser Ausschnitte wesentliche Teile der Anfragen aus dem Data Mart bedient werden, was gegenüber einem direkten Zugriff auf die Data Warehouse Vorteile hinsichtlich der Zugriffsgeschwindigkeit erwarten läßt. Grundsätzlich lassen sich Data Marts neben der Form der Datenhaltung (relational und multidimensional) auch dahin gehend unterscheiden ob sie nur einmalig erstellt werden müssen oder ob ganz oder teilweise regelmäßig neubefüllt werden müssen.

### **2.3.1. Regelmäßig befüllte Data Marts**

Data Marts die regelmäßig nach ihrer Initialladung upgedatet werden sollen, werden häufig im Rahmen des Berichtswesen/OLAP benötigt oder Sie bilden im Data Mining einen Rahmen von ständig zur Verfügung stehenden aggregierten Basisinformationen. Typische Beispiele für solche Data Marts sind Tabellen/Dateien mit auf unterschiedlichen Leveln verdichteten Verkaufszahlen für das aktuelle Jahr. Je nach Definition sollen diese Informationen dann täglich, wöchentlich oder monatlich zur Verfügung stehen. Je kürzer der Updatezeitraum ist, um so wichtiger ist es, das der Prozess des Neubefüllens oder Updatens vollautomatisiert nach festen Regeln und innerhalb von vorgegebenen Abläufen durchgeführt wird. Bei einem täglichen Befüllungszyklus findet dieser meistens in der Nacht statt, nach dem im Data Warehouse aus den operativen Systemen die aktuellsten Informationen zur Verfügung stehen.

### **2.3.2. Einmalig zu erstellende Data Marts**

Die große Gruppe der einmalig zu erstellenden Data Marts muß nochmal unterschieden werden nach permanent zur Verfügung stehenden Data Marts und solchen die im Rahmen einer einmaligen Analyse erstellt worden sind.

#### *Permanente Data Marts*

In diesen Bereich fallen alle Data Marts, die in verdichteter Form historische, d. h. sich nicht mehr verändernde Daten zur Verfügung stellen. Wenn diese Data Marts nicht aus regelmäßig zu erstellenden Data Marts entstehen, so hat man hier die Situation, das sie nur einmal initial erstellt werden müssen und ansonsten (außer im Falle eines Fehlers in den Daten) nicht mehr verändert werden. Im Rahmen des zu definierenden Workflows spielen sie nur noch eine Rolle als Informationslieferant für diverse Analysen.

#### *Data Marts als Basis von komplexe Analysen*

Besonders im Bereich der Ad hoc Anfragen und Data Mining Analysen stellt sich oft die Notwendigkeit ein, Daten des Data Warehouse in Bezug auf unterschiedliche Sichten und Zusammenhänge zu verdichten. Für die Umsetzung gibt es grundsätzlich zwei Möglichkeiten: zu einem schreibt der Poweruser oder Analyst eine entsprechende Anforderung an die Data Warehouse Administration mit der Bitte um Erstellung eines Data Marts nach folgenden Regeln. Oder man setzt den Poweruser/Analysten durch eine entsprechende Softwarelösung in die Lage, das er sich fast immer die entsprechenden Data Marts selbst zusammenstellen kann.

## **2.4. Endbenutzerwerkzeuge**

Als fundierte Basis für unternehmensrelevante Entscheidungen müssen aussagefähige Daten schnell und gezielt zur Verfügung gestellt werden können. Dieses wird in der Regel mit OLAP und Data Mining und den dazugehörigen Tools bewerkstelligt.

### **2.4.1. OLAP- "Online Analytical Processing"**

"Online Analytical Processing" (OLAP) hat sich in den letzten Jahren als die Lösung herausgestellt, mittels der man große Datenmengen analytisch bearbeiten kann, ohne über Programmierkenntnisse oder umfangreiche Statistikenkenntnisse zu verfügen. Die Stärken der OLAP-Werkzeuge liegen im Berichtswesen. Ihre Aufgabe ist es Daten anzuzeigen. Das kann als statisch vorgefertigter Bericht oder dynamisch, also adhoc/online, geschehen. Die Darstellung der Daten erfolgt wahlweise als Tabelle oder Grafik. In diesem Sinne sind OLAP-Werkzeuge hauptsächlich Darstellungswerkzeuge, mit deren Hilfe aber durchaus Wissen erzeugt werden kann. Die mit OLAP-Werkzeugen bearbeitete Datenmenge stammt dabei entweder aus dem Data Warehouse oder aus Data Marts, die zum Teil extra für das OLAP-Werkzeug beziehungsweise dadurch erstellt worden sind. Das heißt OLAP ist beides: ein Darstellungswerkzeug, aber auch eine bestimmte Form der Datenbankarchitektur. Die META Group definiert OLAP als ein "Bündel von Softwarewerkzeugen, die der Erstellung von Applikationen zur Entscheidungsfindung dienen und sich deshalb für mehrdimensionale Analysen komplexer Daten eignen müssen." Dazu bedarf es entweder einer physischen oder einer virtuellen multidimensionalen Datenhaltung. Laut Pendse und Creeth, den Autoren des OLAP-Reports, ist es äußerst schwierig zu entscheiden, wann man von einem OLAP-Produkt sprechen kann. Diese Definition sollte einfach, unvergeßlich und produktunabhängig sein; so kam es zu dem "FASMI"-Test. Die OLAP-Definition von Codd aus dem Jahr 1993 wurde in fünf Schlüsselwörtern zusammengefaßt:

#### *Fast*

Analysis of Shared Multidimensional Information (FASMI). Diese Definition wurde 1995 zuerst genutzt und bisher war keine Überarbeitung notwendig. Fast Ein OLAP-Tool soll einen möglichst schnellen Zugriff auf die Daten ermöglichen und zwar im Schnitt innerhalb von 5 Sekunden. Selbst aufwendige Abfragen sollten nicht länger als 20 Sekunden benötigen.

#### *Analysis*

OLAP soll mit Geschäftslogik und statistischen Analysen umgehen können, die für den Endbenutzer im Rahmen von Datenanalysen relevant sind. Dies alles sollte ohne Einsatz einer Programmiersprache möglich sein. Typische Analyseformen sind Zeitreihenvergleiche, Was-wäre-wenn-Simulationen, das Berichtswesen, usw.

#### *Shared*

Die OLAP-Datenbasis muß von mehreren Benutzern gleichzeitig zu nutzen sein. Während dies bei lesenden Zugriffen unproblematisch ist, ergeben sich bei Schreibvorgängen zum Teil erhebliche Schwierigkeiten, da durch den großen Anteil an verdichteten Daten unter Umständen eine Neuberechnung eines Großteils des Datenbestandes erforderlich wird. Weiterhin muß ein benutzerabhängiger Zugriffsschutz eingerichtet sein, da es nicht sinnvoll ist, jedem Endbenutzer die gleichen Rechte auf dem Datenbestand zuzugestehen.

#### *Multidimensional*

Mit OLAP soll es darüber hinaus möglich sein, multidimensionale betriebliche Kennzahleninformationen effizient zu speichern und den Endanwendern bei Bedarf direkt für unterschiedliche Analysen zur Verfügung zu stellen.

#### *Information*

Ein OLAP-Tool wird danach bewertet, wieviele Inputdaten es verwalten kann, nicht wieviele Gigabyte zu deren Speicherung notwendig sind. Es sollen die gesamten, von den Benutzern benötigten Informationen bereitgestellt werden, unabhängig von der Datenmenge und der Datenherkunft.

### **2.4.2. Data Mining**

Unter Data Mining versteht man eine Reihe von Analysen, mit deren Hilfe neue häufig unerwartete Zusammenhänge festgestellt werden. Der Übergang zwischen klassischen statistischen Methoden und Data Mining ist fließend. Grundsätzlich unterscheidet man beim Data Mining zwei Vorgehensweisen:

#### *Validieren von Hypothesen auf den Daten*

Entdecken und Entwickeln von bisher unbekanntem Muster/Regeln in den Daten Während für die erste Vorgehensweise schon vorab eine entsprechende Hypothese entwickelt werden muß, versucht man in der zweiten Vorgehensweise neue Muster/Regeln in den Daten mit Hilfe der geeigneten Verfahren zu finden. Wobei natürlich grundsätzlich gilt, nur Informationen, die ich den Analyseverfahren zur Verfügung stelle, können auch bei der Entwicklung der neuen Regeln berücksichtigt werden. Im Data Mining lassen sich folgende Anwendungskategorien zusammenstellen:

- Klassifikation
- Clustering
- Statistik und Prognosen

- Zusammenhangsanalysen
- Textmining
- Webmining
- Klassifikation

Unter der Anwendungskategorie Klassifikation versteht man meistens Entscheidungsbäume oder -regeln. Typische Anwendungen sind Kundensegmentierung, Bonitäts-, Kredit- oder Responsescoring.

#### *Clustering*

Durch die verschiedenen Clustering Verfahren wird versucht, möglichst ähnliche Gruppen von Kunden oder Produkten oder Verhaltensweisen zu entwickeln, die sich von den anderen gefundenen Gruppen stark unterscheiden. Eine Anwendungsmöglichkeit hier ist das Finden von neuen Zielgruppendefinitionen, Entwicklung von neuen Lifestylemerkmalen oder Analyse des Schadensverhalten in der Kraftfahrzeugversicherung.

#### *Statistik und Prognosen*

In diesen Bereich fallen die meisten statistischen Verfahren wie z.B. Regressionsanalysen, um z.B. die Kaufwahrscheinlichkeit für ein bestimmtes Produkt innerhalb einer bestimmten Zielgruppe zu bestimmen oder Zeitreihenanalysen um Umsatzentwicklungen für das nächste Jahr zu planen. Darüber hinaus beinhaltet dieser Bereich auch alle notwendigen Testverfahren, um z.B. die Wirksamkeit einer neuen Marketingkampagne zu untersuchen.

#### *Zusammenhangsanalysen*

Ein typisches Beispiel für die Analysen, die unter Zusammenhangsanalysen fallen sind die sogenannten Warenkorbanalysen. Hier wird mit verschiedenen Algorithmen untersucht, welche Produkte zusammengekauft werden. Besonders häufig wird dieses verwendet, wenn sich die einzelnen Transaktionen nicht mehr auf einen einzelnen Kunden zurückführen lassen, sondern z.B. nur noch als Kassenbon vorliegen. Die sehr bekannten

#### *Data Mining*

Beispiele aus den USA sind häufig Zusammenhangsanalysen. Weitere Anwendungsmöglichkeiten bestehen in der Entwicklung von neuen Artikelsets oder z.B. der Analyse des Konsumverhaltens.

#### *Textmining*

Bei Textmining geht es darum, dass hier nicht Daten in Form von Zahlen ausgewertet werden, sondern Texte. Die verschiedenen Verfahren versuchen Texte automatisch zu klassifizieren und zu Verschlagworten. Typische Beispiel wären Emails, die automatisch weitergeleitet oder verarbeitet werden, oder aber auch interne oder externe Dokumentationen, Projektberichte oder ähnliches.

#### *Webmining*

Unter Webmining versteht man die Anwendung von Data Mining aufs Web, so daß nach Möglichkeit jeder Klick und Verweildauer ausgewertet werden kann, sogenannte Click-Stream-Analysen.

Der Einsatz der unterschiedlichen Methoden hängt stark vom Charakter der Aufgabenstellung ab. Eine eindeutige Zuordnung der Instrumente nach Aufgabenstellung ist jedoch nicht möglich. Oftmals werden mehrere Data Mining-Lösungen für dieselbe Aufgabenstellung entwickelt und gegeneinander ausgetestet. Auch die Kombination unterschiedlicher Methoden innerhalb einer Lösung ist möglich. Data Mining ist als iterativer Prozess zu verstehen.

### **3. Zusammenfassung**

Zusammenfassend kann man sagen, das erst der unternehmens- und branchentypische Einsatz von Data Warehouses, Data Mart und den entsprechenden Analysetools OLAP und Data Mining ein analyseorientes Informationssystem aufspannt, das hilft schnell und effektiv Antworten auf entscheidungsrelevante Fragen zu finden. Dabei bildet das Data Warehouse den entscheidenen Grundstock, dafür welche Daten in welcher Qualität zur Verfügung stehen. Doch ohne eine entsprechende Vorverdichtung und die Unterstützung der Analysetools würden viele Fragen nicht in kurzer Zeit und absehbarem Arbeitsaufwand beantwortbar sein.

#### Literaturverzeichnis

1. Anahory, Sam; Murray, Dennis (1997): Data Warehouse, Planung, Implementierung und Administration, Bonn et al. 1997.
2. Appelrath, Hnas-Jürgen (Hrsg.) 1991: datenbanksysteme in Büro, technik und wissenschaft, berlin et al 1991 Barquin, Ramon C.; Edelstein, Herbert A. (Hrsg.) (1997a): Buildng, Using andmanaging the Data Warehouse, Upper Saddle River 1997

3. Barquin, Ramon C.; Edelstein, Herbert A. (Hrsg.) (1997b): Planning and Designing the Data Warehouse, Upper Saddle River 1997
4. Chamoni, Peter; Gluchowski Peter (Hrsg.) 1998a): Analytische Informationssysteme: Data Warehouse, on-line Analytical processing, Data Mining, Berlin et al. 1998
5. Codd, Edgar F. (1970): A relational Model for Large Shared Data Banks, in: Communications of the ACM, Vol. 13, No. 6, June 1970
6. Codd, Edgar F. (1990): The Relational Model for Database Management. Version 2, Reading et al. 1990.
7. Codd Sharon B.; Salley, Clynch T. (1993): Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate, White Paper, E.F. Codd & Associates 1993
8. Corey, Michael J.; Abbey, Michael (1997): Oracle data Warehousing, Berkeley 1997
9. Eicker, Stefan; Schüngel, Martin (1998): Stand der Unternehmensdaten Modellierung in der Praxis, in. Information Management & Consulting, 13. Jg. Nr. 4, 1998
10. Fayyad, Usama M. et al (1996): Advances in Knowledge Discovery and data Mining, Menlo Park 1996  
Gabriel, Roland; Röhrs Heinz-Peter (1995): Datenbanksysteme: Konzeptionelle Datenmodellierung und Datenbankarchitekturen, 2. Auflage, Berlin 1995)
11. Gluchowski, Peter (1997): Data Warehouse, in: Informatik Spektrum, 20-Jg., Nr. 1 Februar 1997
12. Gluchowski, Peter; Gabriel, Roland, Chamoni, Peter (1997): Management Support Systeme. Computergestützt Informationssysteme für Führungskräfte und Entscheidungsträger, Berlin et al. 1997
13. Inmon, William H. (1996): Building the Data Warehouse, 2nd Edition, New York et al. 1996
14. Inmon, William H.; Hackathorn, Richard D. (1994): Using the Data Warehouse, New York et al. 1994
15. Lehner, Franz; Maier, Roland (1994) Information in Betriebswirtschaftslehre, Informatik und Wirtschaftsinformatik, Forschungsbericht Nr. 1.. der Schriftenreihe des Lehrstuhls für Wirtschaftsinformatik und Informationsmanagement, Wissenschaftliche Hochschule für Unternehmensführung, Koblenz 1994
16. Martin, Wolfgang (Hrsg) (1998): Data Warehouseing, Bonn et al 1998
17. Mucksch, Harry; Behme, Wolfgang (Hrsg.) (1998) das Data Warehouse-Konzept. Architektur-Datenmodelle-Anwendungen, 3. Auflage, Wiesbaden 1998
18. Müller, Jochen (2000): Transformation operativer Daten zur Nutzung im Data Warehouse, Wiesbaden 2000
19. OLAP Council (1995): OLAP and OLAP server definitions, The OLAP Council, 1995 <http://www.olapcouncil.org>
20. Scheer, August-Wilhelm (1988): Wirtschaftsinformatik: Informationssysteme im Industriebetrieb Berlin et al 1988
21. Wedekund, Hartmut (1991): Datenbanksysteme, 3. Auflage, Mannheim et al 1991
22. Wittmann, Waldemar (1995): Unternehmung und unvollkommene Information, Köln, Opladen 1995