Ausarbeitung zum Seminar

Business Intelligence Teil I

OLAP & Data Warehousing, eine Einführung

Betreuer: Boris Stumm

Lehrgebiet für Datenverwaltungssysteme

Universität Kaiserslautern SS 2003

Bearbeiter: Marcus Seiler

Inhalt

1. Motivation

In dieser Ausarbeitung zum Thema OLAP (On-Line Analytical Processing) und Data Warehousing soll ein Überblick über dieses Feld der Datengewinnung, -haltung und -auswertung gegeben werden. Es wird die terminologische Domäne erläutert und die historische Entwicklungsgeschichte in ihren Meilensteinen umrissen. Es werden auch die Architekturen und Anwendungsmöglichkeiten der Konzepte besprochen.

In diesem ersten Teil werden die Motivation und Anforderungen, also die Notwendigkeit und das Bedürfnis nach diesen Konzepten dargelegt, wodurch die Wichtigkeit und (auch und gerade wirtschaftliche) Relevanz dieses Themas deutlich werden wird.

1.1 Praktisch

Die heutige Situation in der **Wirtschaft**, bedingt durch Wettbewerb, Fortschritt und Technik, sowie Geschwindigkeit (Schnelllebigkeit und Aktualität) und Dimensionierung (Unternehmensausmaße) zwingt die Unternehmen und speziell die Unternehmensführung ihre operationalen und besonders ihre strategischen **Entscheidung**sprozesse diesen Gegebenheiten anzupassen. Die Ausmaße der Märkte und Unternehmungen selbst sowie die allgemeine Situation an sich führen zu einer unübersehbaren Datenflut. Die Entscheidungen müssen auf der Basis von immensen Datenmengen getroffen werden, was schon allein dadurch den Einsatz computergestützter Lösungen nahe legt. Die enorme Schnelllebigkeit und Wettbewerbsdichte der Märkte erzwingt außerdem eine detaillierte Einbeziehung und Untersuchung dieser Daten, um z.B. den Marketingmix exakt auf einzelne Marktsegmente abzustimmen. Andererseits handelt es sich aber auf Toplevelmanagementebene hauptsächlich um planerische Entscheidungen die Unternehmensstrategie betreffend, wodurch natürlich eine ganzheitliche Betrachtung meist aller (relevanten) Unternehmensbereiche nötig wird und es damit auch das Problem der Heterogenität der Teilbereiche und ihrer Daten zu lösen gilt.

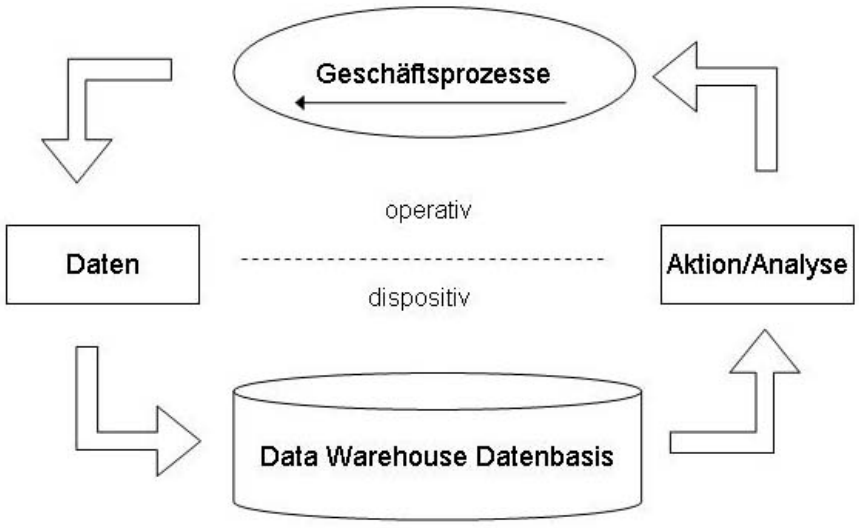

Abb. 1: Data Warehouse Einbeziehung aus Geschäftsprozesssicht.

Dadurch wird deutlich, dass es zur Unterstützung der Entscheidungsprozesse nötig ist, aus enormen Datenmengen auf möglichst benutzerfreundliche Weise (nicht alle Manager sind Informatiker) eine kompakte, einfach verständliche Darstellung der relevanten Situation zu geben [BeMu98]. Die Problematik liegt also zum einen in der Bereitstellung der Datenbasis und zum anderen in der Präsentation und (Vor-)Interpretation. Natürlich werden in diesem Bereich schon seit längerem IT-Lösungen eingesetzt. Hier sind die Decision Support Systeme (DSS, siehe Historie) als erstes zu nennen, da sie die direkten Vorläufer der hier betrachteten Systeme sind. Doch haben diese einige Schwachstellen, wie z.B. in der Integration heterogener Daten oder der Performanz aufgrund der Datenrepräsentation, die durch den Einsatz der Konzepte des On-Line Analytical Processing (OLAP) und des Data Warehousing angegangen werden konnten.

1.2 Technisch

Selbstverständlich sind auch im Datenbankbereich die Bedürfnisse und Forderungen der Wirtschaft die primäre Triebfeder der Entwicklung, doch sind in den Konzeptionen des OLAP und Data Warehousing einige Aspekte enthalten, die über die Befriedigung der speziellen Anforderungen hinaus für die Informatik selbst interessant sind. So ist z.B. eine einheitliche, standardisierte, integrierte Datenrepräsentation heterogener Daten ein allgemeiner Trend und somit auch in vielen Bereichen von Interesse [Popp99]. Auch ist das Problem der Dateninterpretation allgegenwärtig und Ansätze zu deren algorithmisierten Lösung sind für viele Anwendungsfelder interessant.

2. Anforderungen

Aus diesen oben erläuterten Bedürfnissen lassen sich nun konkrete Anforderungen formulieren, die ein OLAP-System und die gelieferten, bzw. erzeugten Informationen erfüllen müssen oder vielmehr müssen, um die Aufgaben angemessen bewältigen zu können. Bei diesen Anforderungen handelt es sich primär um **qualitative** Merkmale, da offensichtlich die Güte der Entscheidungsunterstützung die korrespondierenden Entscheidungen maßgeblich beeinflusst.

2.1 Informationen

Selbstverständlich gibt es in der Informatik je nach Anwendungsbereich sehr viele, meist leicht differierende Anforderungen an den Informationsbegriff, und andererseits wieder fast allgemeingültige Merkmale wie Präzision oder Verständlichkeit. Andersherum gibt es im hier betrachteten Bereich eine Vielzahl an Kriterien, die zu beachten sind [BeMu98]; an dieser Stelle sollen aus Platzgründen nur die Übergeordneten und für den analytischen Bereich wichtigen Eigenschaften betrachtet werden. Die wichtigsten Merkmale sind hier wohl ganz klar die Aktualität, Vollständigkeit, Relevanz und Korrektheit der Daten bzw. Informationen. Im Einzelnen erläutert bedeuten diese:

Aktualität:

Entscheidungen, welche auf Grund veralteter Informationen getroffen werden, sind schon alleine dadurch offenkundig mit hoher Wahrscheinlichkeit falsch oder zumindest nicht optimal. Darum sollten Entscheidungen stets auf aktuellen Daten fußen, um auch neueste Begebenheiten oder Veränderungen berücksichtigen zu können.

Vollständigkeit:

Hiermit ist sowohl die Gesamtheit der Daten in ihren Ausmaßen gemeint, wie auch die Detailliertheit. Es dürfen also keine Bereiche vergessen werden und innerhalb dieser darf es auch keine Lücken geben. Durch lückenhafte Informationen (hier zeitunabhängig) kann es auch hier wieder zu Fehlentscheidungen kommen.

Relevanz:

Die Datenmenge, die in einen Entscheidungsprozess einbezogen wird, sollte stets auf ein „relevantes Minimum“ reduziert werden, da durch unnötige Informationsmengen das Treffen von (richtigen) Entscheidungen stark erschwert wird, wenn nicht sogar Verschleierung und negative Beeinflussungen zu einer anderen, evtl. falschen Entscheidung führen können.

Korrektheit:

Wenn man die Korrektheit (Tatsachenkonformität) nicht schon als allgemeingültiges Merkmal oder Prinzip von Informationen im wissenschaftlichen und wirtschaftlichen Bereich betrachten will, so muss sie doch spätesten in der Analyse und demnach erst recht hier beim Entscheidungsfinden als unerlässlich konstatiert werden, denn in den seltensten Fällen werden Entscheidungen, basierend auf inkorrekten Informationen, selbst korrekt und sinnvoll sein.

2.2 Systeme

Aus den Anforderungen für die benötigten Informationen ergeben sich wiederum Anforderungen an die Systeme, die diese bereitstellen sollen. Auch hier soll es genügen die wichtigsten Punkte zu betrachten:

Kapazität:

Durch den Vollständigkeitsanspruch entsteht eine enorme Datenmenge, was wiederum eine starke Speicher- und Rechenkapazität voraussetzt, um diese Menge überhaupt halten und in sinnvoller Weise und Zeit verarbeiten zu können.

Zeitsensitivität:

Hiermit ist nicht nur Bezug auf die Aktualität genommen, sondern auch viel mehr die Relevanz des Zeitbegriffs in der analytischen Datenauswertung und Verwertung gemeint. So spielen Zeiträume und Zeitverteilungen in fast jeder Fragestellung eine Rolle und dadurch wird der Zeitbegriff sowohl zur Chance für sinnvolle und optimierte Datendarstellung und Verarbeitung, aber auch zu beachtende Fehlerquelle mit verheerendem Potenzial (z.B. Vergleich zweier Kenngrößen, jedoch aus unterschiedlichen Zeitintervallen).

Repräsentation und Darstellung:

Um aus Daten möglichst effizient Informationen für den User gewinnen zu können und diese anschließend sinnvoll darbieten zu können, ist nicht nur die Darstellung, sondern auch die **Modellierung** der zugrunde liegenden Daten (ausgerichtet auch den analytischen Verwendungszweck und Unterschiedlichkeit der Datenbasis berücksichtigend; siehe unten „Dimensionalität“) wichtig. Die Visualisierung selbst sollte neben **Verständlichkeit** auch durch **Flexibilität**, also an die jeweiligen Situationen und Bedürfnisse anpassbar, gekennzeichnet sein.

Datensammlung und Integration:

Einer der zentralen Punkte besteht in der Sammlung (wie und woher kommen die Daten) und Aufbereitung (in welcher Form werden sie benötigt) der meist **heterogenen** und **verteilten** Datenmengen, die zusammen und in Beziehung gesetzt, die Grundlage für die Analysen und Entscheidungen bilden.

3. Historie

Die Begriffe des OLAP und des Data Warehousing und die damit verbundenen Konkretisierungen und Abgrenzungen des Bereichs gibt es zwar erst seit wenigen Jahren, doch reicht das Interesse und die Behandlung grundlegender Konzepte sowie auch das Verlangen der Anwender nach diesen Möglichkeiten schon recht weit zurück.
Schon Ende der **60’er** Jahre wurde an der Modellierung und Verarbeitung von Daten in mehrdimensionalen **Matrizen** gearbeitet und die im OLAP essenzielle Multidimensionalität ist im Grunde nicht viel anderes.

Anfang der **70’er** Jahre, und spätestens mit Gründung der gleichnamigen Firma (MDS), lässt sich dann das Aufkommen erster **Management Decision Systems** festsetzen, die schon nach kurzer Zeit das Softwaretool **Express** auf den Markt brachte, worin bereits sowohl Analysefunktionen als auch das Datenmanagement integriert waren. Express war in erster Linie auf die Durchführung von Marketinganalysen ausgelegt.
Etwa zur gleichen Zeit entwickelte die Firma **Comshare** das Tool **System WDSS** welches im Gegensatz zu Express auf den Finanzbereich als Anwendungsgebiet abzielte und das Konzept der Mehrdimensionalität verwendete und stark in den Vordergrund rückte.

Zu Beginn der **80’er** Jahre kam erstmals der Begriff des „Data Supermarket“ auf und Ende der 80’er stellte IBM sogar ein internes Projekt mit dem Namen „European Business Information System“ (es wurde später in „Information Warehouse Strategie“ umbenannt) vor, in dem es um die Bewältigung von Heterogenität und Informationsexplosion ging.

Das Aufkommen der eigentlichen Produktkategorie OLAP wurde letztlich durch die Firma **Arbor Software** begründet und mit dem Produkt **Arbor Essbase** Anfang der **90’er** eingeleitet. Anfänglich und nachhaltig geprägt wurde der Begriff durch Dr. E. F. Codd, der in seinem 1993, im Auftrag von Arbor Software veröffentlichtem Buch, „Providing OLAP to User-Analysts: An IT Mandate“ Regeln festlegte, anhand derer ein Produkt als OLAP klassifiziert werden sollte [Codd93].

Aus produktklassenevolutionärer Sicht sind die OLAP Systeme auf die **Decision Support Systeme** (DSS) zurückzuführen, die sich kurz als „interaktive, EDV-gestützte, die Entschei-

dungsträger mit Modellen, Methoden und problembezogenen Daten in ihrem Entscheidungsprozess bei der Lösung in eher schlecht-strukturierten Entscheidungssituationen unterstützende Systeme“ definieren lassen [Lehn03]. Im Grunde ist das OLAP Konzept mit seinen Anforderungen eine Erweiterung der DSS’s.
Methodisch handelt es sich beim OLAP und Data Warehousing, hauptsächlich um die Weiterentwicklung von statistischen Analysesystemen. Es werden große, meist empirisch ermittelte Datenmengen analytisch und statistisch ausgewertet um Kenngrößen zu extrahieren und Zusammenhänge zu extrapolieren. Selbstverständlich stellt auch hier die betrachtete Konzeption eine enorme Erweiterung und Weiterentwicklung da, die zwar die Grundlagen der Vorgänger enthalten, doch sowohl in dem „Was“ als auch in dem „Wie“ stark darüber hinausgeht.

Heute sind diese Konzepte derart populär, dass es in der Wirtschaft für Unternehmen schon fast als schick oder gar Muss gilt, ein eigenes Data Warehouse zu betreiben.

4. OLAP

Es handelt sich beim OLAP in erster Linie um Systemanforderungen und Strategien zur Erreichung dieser. Es geht um die interne und externe (visualisierte) Modellierung und Verarbeitung bzw. Auswertung von Daten. Selbstverständlich bleibt auch die Datenbeschaffung nicht gänzlich unberührt und muss zur Erfüllung der Anforderungen dahingehend angepasst werden.

Hier werden die begriffliche Definition und Abgrenzung sowie die inhaltlichen Konzepte und Modelle des On-Line Analytical Processing genannt und erläutert.

4.1 Was ist OLAP

On-Line Analytical Processing bezeichnet (über ein Netzwerk) **verteilte** Prozesse (oder eher Systeme, die solche Prozesse enthalten), die ganz allgemein gesprochen, **umfangreiche**, **multidimensionale** Daten (-bestände) erzeugen, verwalten, interaktiv **analysieren** und **visualisieren**. Diese Daten sind in der Regel natürliche Geschäftsdaten einer Unternehmung und die Analyse soll sowohl kontrollierende als auch planerische Entscheidungen unterstützen [Fors97].

Unter dem Begriff OLAP selbst wird im Allgemeinen die Familie von Systemen verstanden, die die in Abschnitt 4.2 erklärten Kriterien erfüllt.

Im Gegensatz zum **OLTP** (**O**n-**L**ine **T**ransaction **P**rocessing), welches in der allgemeinen, operationalen Geschäftstätigkeit der Unternehmungen Anwendung findet und auf viele kleine Anfragen, die die Daten verändern, ausgelegt und spezialisiert ist, hat das OLAP einen etwas gemäßigteren Aktualitätsanspruch und beschäftigt sich mit der Analyse von (auch historischen) Daten. Eine starke Verbundenheit, die das OLTP auch innerhalb der Begriffswelt des OLAP präsent macht, ist die Tatsache, dass die Datenbasis von OLAP-Systemen meist aus

den Datenbeständen eines oder mehrer OLTP-Systeme extrahiert wird. Eine Trennung bzw. modifizierte Kopie dieser Daten ist nicht nur wegen eventueller gegenseitiger Beeinflussung notwendig, sondern wie im Folgenden näher erklärt, auch weil eine angepasste Datenmodellierung zur Verwirklichung der OLAP-Prinzipien unerlässlich ist.

4.2 Klassifizierende Anforderungen

Um eine klare Definition und die Möglichkeit einer eindeutigen Klassifizierung als OLAP-System zu geben, werden Kriterien angegeben, die ein System erfüllen muss, um als OLAP-System zu gelten. Gerade auf Grund der vielfältigen Einsatzgebiete und der damit verbundenen starken Unterschiede der Architektur, des Funktionsumfanges und der Benutzerschnittstelle, ist es schwierig, genauere und weniger allgemeine Definitionen anzugeben. Schon E. F. Codd ging mit seinen 12 Regeln [Codd93] diesen Weg der Begriffsbestimmung und auf diesem ist man auch geblieben, auch wenn diese Regeln selbst nicht mehr als absolut und alleinig korrekt betrachtet werden (nicht zuletzt wegen der sehr starken Anlehnung an Arbor Software).

Alternativ wurde mit dem FASMI-Test [CrPe94] durch den OLAP-Report, eine unabhängige Veröffentlichung, ein weiteres Kategorisierungssystem erstellt, dass mittlerweile breite Anwendung und Zustimmung findet. Mit dem Namen (**F**ast **A**nalysis of **S**hared **M**ultidimensional **I**nformation) sind auch gleich die primären Kriterien genannt, die bei dieser Klassifikation als relevant betrachtet werden.

Schnelligkeit:
: Da interaktive Analysen ein Hauptanwendungsfeld von OLAP-Systemen darstellen, müssen, um ein sinnvolles Arbeiten zu ermöglichen, die einzelnen Abfragen sehr schnell ablaufen. Allgemein gilt eine halbe Minute als Obergrenze, die nur in Ausnahmefällen erreicht werden sollte.

Multidimensionalität:
: Schon aus der Tatsachen, dass Geschäftsmodelle nahezu immer durch mehr als drei Einflussgrößen (oder eher Rahmenskalen, wie Zeit oder Abteilungszugehörigkeit) bestimmt sind, entsteht die Notwendigkeit, Daten in einem OLAP-System in einer multidimensionalen Form zu modellieren. Diese Multidimensionalität ist die Grundlage für eine sinnvolle Speicherung (da inhaltlich verwandte Daten in der Repräsentation räumlich nahe beieinander liegen), flexible Betrachtungsweise (durch die Struktur ist die Sicht auf die Daten kaum beschränkt oder gebunden) und die schnelle Verarbeitung von Analyseanfragen (folgt aus der sinnvollen Speicherung).

Intelligenz und Banalisierung:
: Da es oft um die Verarbeitung komplexer und unterschiedlich strukturierter Daten geht, ergeben sich eine Vielzahl von Rahmenbedingungen, die in einer Analyse berücksichtigt werden müssen, um zu einem richtigen Ergebnis zu kommen (so z.B. die Währungsumrechnung oder das Anpassen von Abrechnungszeiträumen). Solche Auf-

gaben sollte das System für den Benutzer erledigen können und ihm auf diese Art „intelligent“ die Arbeit erleichtern und mögliche Fehlerquellen eliminieren.

Die fertige, erarbeitete Information sollte in einer möglichst einfachen, verständlichen Form dem Benutzer dargeboten werden. Das heißt, es muss nicht nur auf wenige Werte zusammengefasst werden (was stark vom Benutzer und der jeweiligen Abfrage abhängt), sondern es muss auch verschiedene Möglichkeiten zur Darstellung geben (z.B.: Tabelle, Diagramme, Graphen, etc.), um einen optimalen Nutzen aus den Ergebnissen für Anwender durch optimales Verständlichkeit zu erzielen.

Benutzerfreundlichkeit:

Die Benutzerfreundlichkeit sei an dieser Stelle getrennt von der Banalisierung, die hauptsächlich auf Verständnis abzielt, aufgelistet, um klarer herauszustellen, dass es sich bei den meisten Nutzer nicht um Computerfachleute handelt, die die Hintergründe der Anwendung verstehen, selbstständig aus Erfahrung auf Funktionsweisen schließen oder gar mit Fehlfunktionen oder Schwachstellen im System klarkommen könnten. Das System soll den Entscheider in seiner Tätigkeit unterstützen und nicht eine weitere arbeitsaufwendige oder gar behindernde Komponente im Entscheidungsprozess werden.

Flexibilität:

Das System sollte in der Lage sein, auf die unterschiedlichen Bedürfnisse des Benutzers, sowohl in der Analyse (Sicht auf die Daten, Einbeziehung von Faktoren, usw.), als auch in der Visualisierung (z.B. Tabellenform, Diagramme, etc.) eingehen zu können. Erst diese Vielseitigkeit macht das Konzept so stark, verbreitet und brauchbar.

Mehrbenutzerbetrieb:

Für ein vernetztes System nahe liegend, doch nicht selbstverständlich ist die Verfügbarkeit der Funktionalität und Informationen für mehrere Benutzer und dies auch zur gleichen Zeit. Aufgrund der enormen Leistungsfähigkeit und der vielen Einsatzbereiche ist es nur natürlich, dass solch ein System in einer Unternehmung von möglichst vielen Personen mit Entscheidungsunterstützungsbedarf genutzt werden kann und wird.

4.3 Konzepte

Die oben skizzierten Anforderungen an Systeme der OLAP Familie sind zwar recht umfangreich und präzise in ihrer Ausrichtung, jedoch wären sie wohl alleine nicht ausreichend um dieses Forschungs- und Entwicklungsgebiet zu beschreiben und den immer noch anhaltenden Wirbel darum zu begründen. Mindestens genau so wichtig sind hier die daraus abgeleiteten und entwickelten Konzepte, die dahinter stehen und dem Ganzen die ersten Schritte Richtung Praxis ebnen. Die wichtigsten Konzepte, die in OLAP-Systemen zum Einsatz kommen, um die bereits ausgiebig dargelegten Anforderungen zu erfüllen, werden hier im Folgenden besprochen [Popp99].

4.3.1 Dimensionen

Von den bereits erwähnten Eigenschaften der multidimensionalen Datenmodellierung, so wie deren Relevanz, ist der Schritt zum direkt verwandten Konzept der Dimensionen recht einfach. Alle Begebenheiten sind primär durch Dimensionen bestimmt. Das gilt nicht nur für natürliche Dimensionen wie Raum oder Zeit, sondern auch für logische Dimensionen, die ihrerseits nicht in Zahl und Kombination beschränkt sind und auch zueinander meist unabhängig sind. Im Speziellen gilt diese Bestimmbarkeit natürlich erst recht für Geschäftsprozesse, die es ja hauptsächlich zu modellieren gilt. Betrachtet man zum Beispiel ein Produkt, welches in einem bestimmten Zeitraum in einem bestimmten Absatzgebiet verkauft wurde, so ergeben sich daraus direkt die Dimensionen Zeit, Region, Produkt. Das Schlagwort bzw. Modell in diesem Bereich ist der Hypercube (siehe Abb.2); ein mehrdimensionaler Würfel, dessen Kanten die Dimensionen repräsentieren. In den Zellen des Würfels werden die Datenwerte abgelegt, die über das entsprechende Werte-Tupel aus den Dimensionsskalen (ähnlich mehrdimensionale Koordinatensystemen) eindeutig identifiziert und angesprochen werden können.

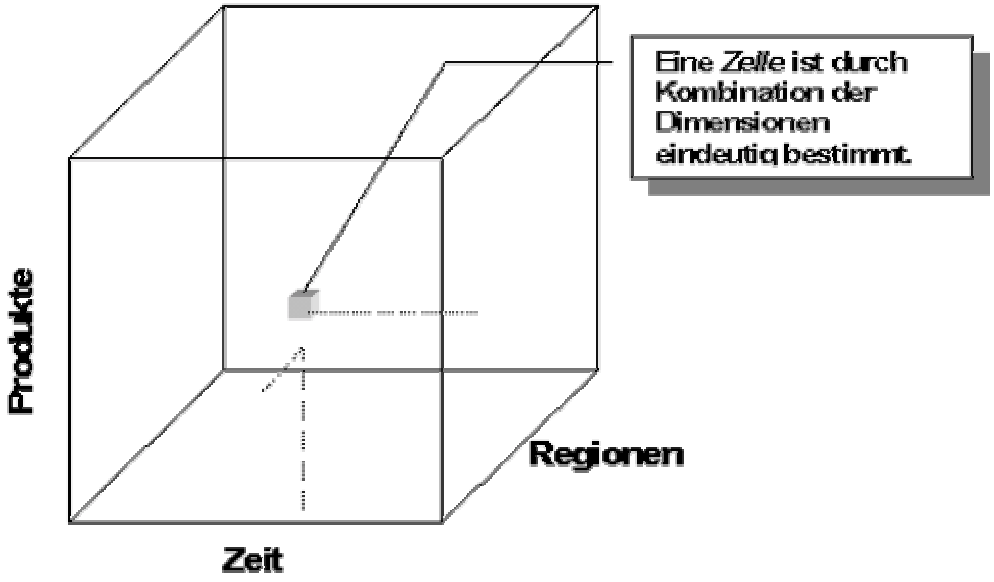

Abb.2: Vereinfachtes, räumliches Beispiel eines Hypercubes (Datenwürfels)

4.3.2 Hierarchien

Eine hierarchische Gliederung ist innerhalb der darstellenden Dimensionen was die Umsetzung angeht recht leicht zu verstehen, da Hierarchien schon im Alltag (Tag, Woche, Monat, Jahr,...) geläufig, wichtig und hilfreich sind. Durch solche Gliederungen im Datenmodell lässt sich für den Benutzer die jeweils passende Abstraktions- bzw. Detailstufe repräsentieren, die im Zusammenhang relevant ist. Außerdem ist auf Grund der engen Verbundenheit dieses Konzepts mit der Aggregation (das Zusammenfassen von mehreren Werten zu einem, z.B. die Summation) ist die Hierarchiestruktur eine gute Basis für effiziente Auswertungen in den Analysen.

4.3.3 Messwerte

Mit dem Begriff der Messwerte (Measures) sind die konkreten Daten des Modells gemeint. Somit muss formal für jeden Messwert ein Koordinatenpunkt im Modell vorhanden sein. Diese Bezeichnung lässt bereits erahnen, dass die meisten Datenbestände empirischer Herkunft sind. Viele dieser Einträge sind aus den Rohdaten direkt übernommen, andere müssen erst durch Umformung, Zusammenfassung, etc. berechnet werden. Dabei geht es nicht nur um das Säubern und Aussortieren von irrelevanten Daten, sondern hauptsächlich das Einpassen der Daten in die standardisierten, einheitlichen Strukturen der (bereits vorhandenen) Datenbasis. So müssen zum Beispiel im ROLAP (siehe Abschnitt 4.4.3) die einzelnen Daten aus den operationalen Datenrepräsentationen auf die Tabellenstruktur innerhalb des OLAP-Systems abgebildet werden.

4.3.3 Zeit

Aus der Systemanforderung der Zeitsensitivität resultiert das Konzeption des Zeitverständnisses in analytischen Systemen. Hierbei muss den maßgeblich charakteristischen Eigenschaften der Zeit, wie Sequenzialität, Totalität, etc., Rechnung getragen werden. Da die Zeit als Einflussfaktor oder Dimension in fast jeder Fragestellung eine Rolle spielt, und dies in den unterschiedlichsten Gestalten wie z.B. als Zeitpunkt oder Zeitspanne, ist es offensichtlich, dass Systeme wie die hier betrachteten für den sicheren Umgang und die effiziente Verarbeitung der Zeitkomponente entsprechende, je nach Anforderung angepasste, Funktionalitäten bereit stellen sollten.

4.4 Methoden

Die tatsächliche Umsetzung und damit Basis reeller Implementierungen der Speicherung und Verarbeitung von Daten innerhalb der obigen Konzepte lässt sich nach dem momentanen Stand in drei Gruppen einteilen:

4.4.1 ROLAP

Beim ROLAP (**r**elationales OLAP) wird der Hypercube auf Tabellen eines RDBVS abgebildet. Dadurch können bereits vorhandene relationale Datenbanken auch hierfür benutzt werden, was natürlich gerade in der Entwicklung starke Zeitersparnisse und Vereinfachungen mit sich bringt. Die Dimensionalität wird bei dieser Umsetzung durch eigene Tabellen, bzw. durch Referenzen darauf gelöst (siehe Abschnitt 5.2.2, Star/Snowflake-Scheme). Was die Performanz angeht weist diese Lösung verständlicherweise einige Schwachstellen auf, da es keine echte mehrdimensionale Darstellung ist und somit deren Vorteile auch nur bedingt genutzt werden können.

4.4.2 MOLAP

Das MOLAP (**m**ultidimensionale OLAP) ist die natürlichere und konsequentere Version, bei der die Daten in einer mehrdimensionalen Struktur abgespeichert werden. Hierzu werden meist multidimensionale Datenbanksysteme verwendet, die hinsichtlich solcher Speicherstrukturen spezialisiert sind.

Um räumliche Strukturen sinnvoll in den (physikalisch flachen) Speicher abbilden zu können, müssen die Daten in eine brauchbare Ordnung gebracht werden, so dass deren Abbildung möglichst alle Eigenschaften des multidimensionalen Modells beibehält. Hierbei kommen diverse Indexierungstechniken (z.B. B-Bäume) oder auch Verfahren wie die Z-Ordnung zum Einsatz, bei der sich benachbarte Zellen im Modell in der Regel auch im Speicher nahe sind.

4.4.3 Mischformen

Es gibt auch Implementierungen bei denen die beiden Möglichkeiten vermischt werden. Dabei wird versucht die Vorteile beider Verfahren zu kombinieren; so werden die Daten weiter in Tabellen wie beim ROLAP gehalten, bei Berechnungen in der Analyse jedoch werden z.B. Aggregationen über einen MOLAP-Server abgewickelt. Solche Mischformen fasst man unter dem Begriff der HOLAP (**h**ybrid OLAP) zusammen.

Als weitere, jedoch nicht wirklich eigenständige Form der Umsetzung ist das DOLAP (**d**esktop OLAP) bekannt, bei dem jeweils eine kleine, momentan relevante, in sich komplette Datenbasis lokal auf dem Clientrechner festsitzt. Diese Datenbasis wird über einen oder mehrere Datenbankserver auf den lokalen Client geladen und kann gegebenenfalls aktualisiert oder ersetzt werden. Als Zielgruppe für diese Form der OLAP-Applikation werden mobile Entscheider gesehen, die nicht ständig online mit den Servern verbunden sein können und deshalb kleinere lokale Datenbestände brauchen. Die Funktionalität solcher DOLAP-Systeme liegt dementsprechend weit unter dem normaler Online-Systeme.

5. Data Warehouse

Im Data Warehouse finden die Prinzipien, Richtlinien und Strategien des OLAP Anwendung und werden in konkrete Systemarchitekturen eingebunden. Ergänzend kommen hier Komponenten wie die Datenbeschaffung, -integration und -haltung hinzu, also besteht die primäre Aufgabe, als Abgrenzung zu OLAP, in der Bereitstellung von Daten (aktuelle wie historische). Besonderes Augenmerk liegt hier auf dem nahtlosen Verlauf innerhalb des Systems von den heterogenen „Rohdaten“ bis zur Präsentation der interpretierten Information.

In der Literatur wird unter Data Warehouse auch oft nur die konsolidierte und integrierte Datenbasis verstanden, auf der OLAP-Systeme aufsetzen.

5.1 Was ist Data Warehousing

Mit dem Begriff des Data Warehousing wird im allgemeinen ein vom operationalen Datenverwaltungsbetrieb eines Unternehmens unabhängiges DBVS (Datenbankverwaltungssystem) umschrieben, das als unternehmensweite Datenbasis für managementunterstützende (entscheidungsunterstützende) Systeme dient. In Analogie zu Lagerhaus werden hier also große Mengen an unterschiedlichen Informationen in einheitlichen (standardisierten) Strukturen angeboten und zur weiteren Verwendung bereitgestellt. Selbstverständlich werden, wie in jedem Datenverwaltungssystem, auch Tools zur Auslese und weiteren Verarbeitung mit angeboten. Dabei wird eine globale Sicht auf die Daten der Unternehmung ermöglicht. Hierzu werden relevante Informationen aus den operativen Systemen einer Unternehmung (z.B. OLTP) ausgelesen, anschließend in die erforderliche Struktur umgeformt, konsolidiert und in die bereits bestehende Datenmenge des Data Warehouse integriert [Lehn03].

Man benutzt als Basis für analytische Berechnungen nicht die operationalen Daten in den transaktionsorientierten Datenbankservern, weil diese auf kleine, aber sehr oft leicht verändert wiederholte Anfragen hin optimiert sind, wobei hauptsächlich kleine Datenbestände verändert oder abgefragt werden. Für die Bearbeitung von planerisch interessanten Anfragen, die große Datenmengen einbeziehen würden, sind diese Systeme nicht effizient und außerdem würden sich die Anfragen aus den beiden Bereichen gegenseitig stören und so für alle Beteiligten Performanzeinbußen verursachen [DaSu97].

Auch ist die Bereitstellung von Meta-Daten (siehe 5.2.2) über die gespeicherte Datenbasis ein charakteristisches Merkmal für ein Data Warehouse.

5.2 Architektur und Konzept

In einem Data Warehouse geht es darum, die Ideen und Konstrukte des OLAP umzusetzen und die dafür nötigen Daten zu sammeln und in brauchbarer Weise bereitzustellen. Dabei ergeben sich zwei wesentliche Schwierigkeiten, nämlich die Heterogenität der Daten, die in die Datenbasis aufgenommen werden müssen und die enorme Informationsmenge, bzw. das Verarbeiten und Bereithalten dieser Mengen. Als Überblick, jedoch allgemein bezeichnende und umfassende Architektur eines Data Warehouse, der hauptsächlich die Grenzen und Schnittstellen zu den benachbarten Komponenten darstellt sei Abb.3 angegeben.

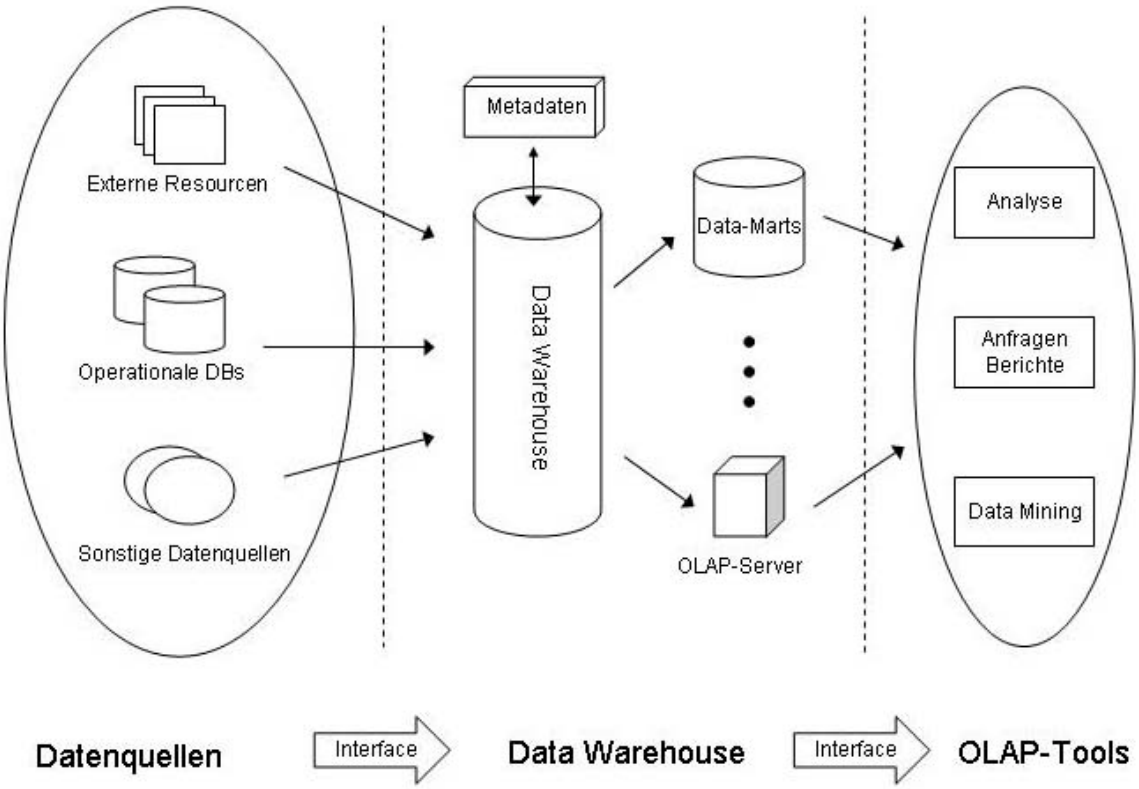

Abb. 3: Übersicht einer Data Warehouse Architektur

Allgemein lässt sich die Architektur eines Data Warehouse in drei Bereiche gliedern: Schnittstelle zu den Datenquellen, Data Warehouse (Die Datenbasis inklusive Kernfunktionalitäten) und Schnittstelle zu Endverarbeitungstools.

5.2.1 Datenquellen-Schnittstelle:

Die Ausprägung der Datenbestände eines Data Warehouses entstehen nicht wie in operationalen Datenbanken durch viele kleine Transaktionen, die Aktionen der realen Welt widerspiegeln sollen, sondern durch Extraktion aus anderen Datenquellen. Aus meist vielen verschiedenen Datenquellen werden Datenbestände ausgelesen und zur Erweiterung der Daten im Data Warehouse verwendet. Dieses Auslesen wird meist in regelmäßigen Zeitabständen wie-

derholt um das Warehouse zu aktualisieren. Dabei werden keine Daten überschrieben oder gelöscht, sondern die alten Datenbestände werden als historische Daten weiter in der Datenbank gehalten und als solche bereitgestellt.

Innerhalb dieser Ladephase (load time) werden die Daten aus den Datenquellen in die funktionalen Einflussbereich des Systems geholt und anschließend gesäubert (also von unnötigen Inhalten befreit) sowie auf existierende Restriktionen und Anforderungen geprüft.

5.2.2 Warehouse:

Die extrahierten Rohdaten müssen anschließend im Data Warehouse normalisiert und in die bereits existierenden Datenbestände integriert werden, um sie nutzbar zu machen. Innerhalb der Kernanwendung lässt sich eine detailliertere Architektur erkennen, deren Komponenten in fast allen Implementierungen auf die eine oder andere Art realisiert werden. In Abb.4 sind außerdem auch Teile der beiden Schnittstellen am oberen und unteren Ende dargestellt.

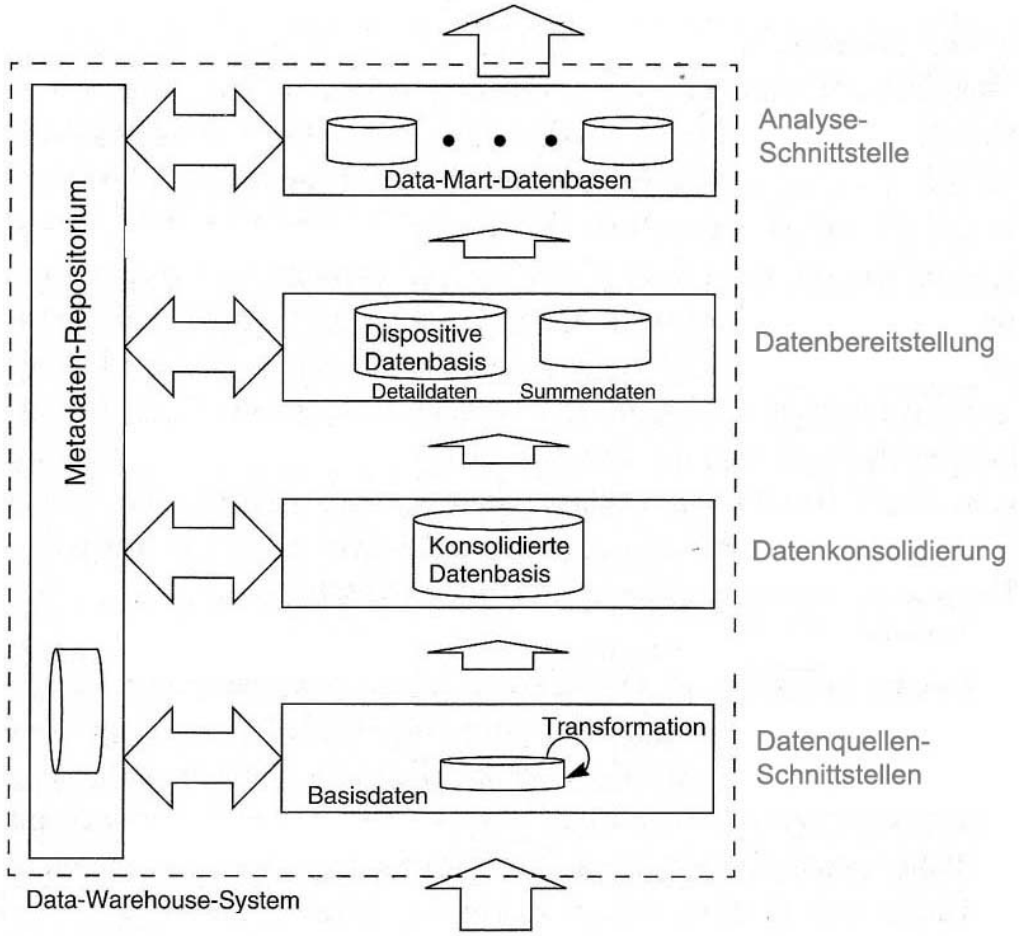

Abb. 4: Architekturmodell mit interner Gliederung des Data Warehouse Systems

Konsolidierung:

Nachdem die Daten im System eingegangen und teilweise aufbereitet wurden, werden sie nun in einer internen Aktualisierung (refresh time) dem aktuellen Data-Warehouse-Datenbestand hinzugefügt. Dadurch wird die **konsolidierte** Datenbasis aktuell gehalten, welche eine organisationsweite, anwendungsunabhängige Speicherung möglichst aller Unternehmensdaten darstellt.

Bereitstellung:

Aus der vorherigen Phase heraus wird nun ein **dispositiver** Datenbestand abgeleitet, bzw. ein bereits bestehender aktualisiert. Dabei wird diese Datenrepräsentation stark auf mögliche Anwendungsszenarien und spätere Verwendungen hin ausgerichtet und bereits in der Entwurfsphase dorthin gehend ausgelegt. In diesem Zusammenhang sollte **Snowflake-Muster** erwähnt werden, wobei in Abb.5 beispielhaft gezeigt wird, wie ein solches Snowflake-Schema zur Realisierung einer multidimensionalen Datenrepräsentation auf relationalen Strukturen (wie im ROLAP) genutzt werden kann. Hierbei handelt es sich um ein stark verbreitetes Konzept, das ist fast allen Data Warehouse Implementierungen zum Einsatz kommt.

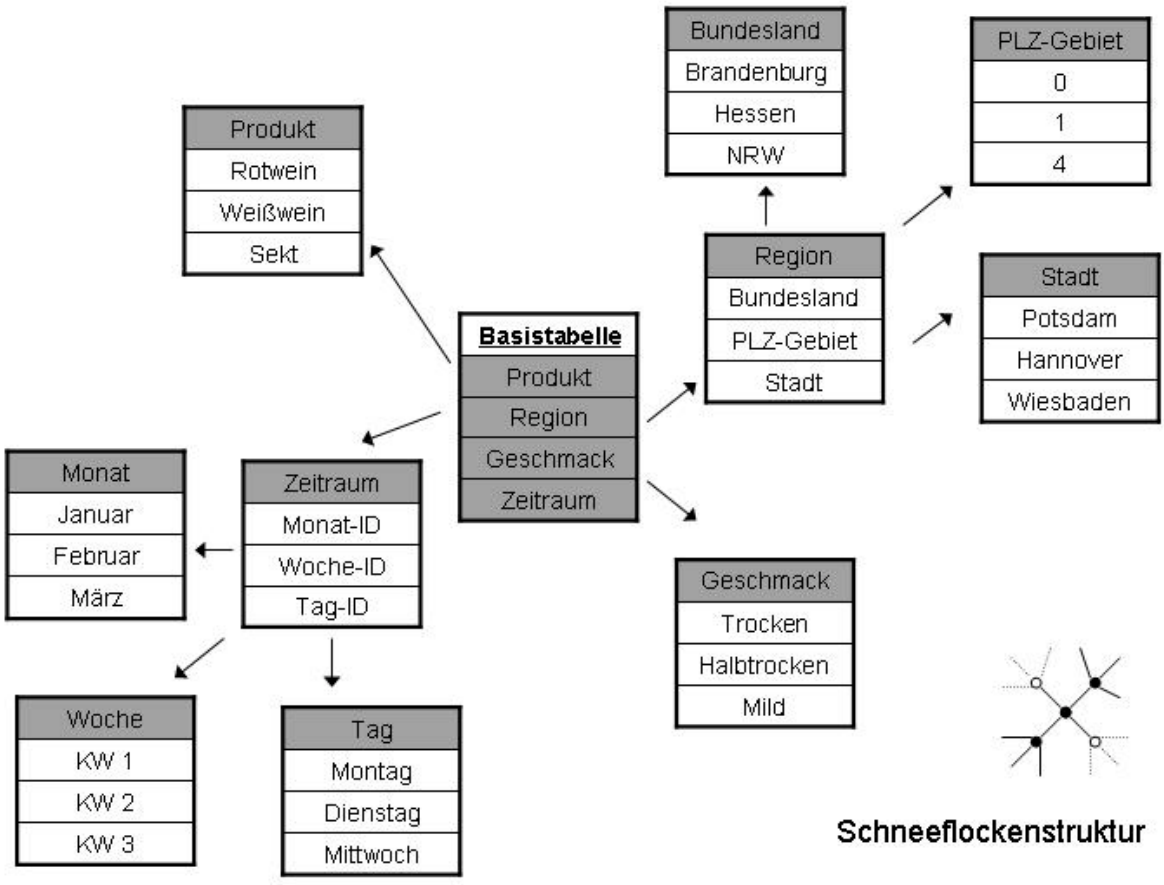

Abb. 5: Snowflake-Schema Beispiel

Ausgehend von einer Basistabelle (Faktentabelle) werden von dort aus die Dimensionstabellen (im Beispiel: Zeitraum, Produkt, Region und Geschmack) als Repräsen-

tanten für die benötigten Dimensionen referenziert. Über Fremdschlüssen in diesen Tabellen wiederum lassen sich die Attributtabellen (hier: Tag, Stadt, etc.), in denen die Messwerte gespeichert sind, oder weitere Dimensionstabellen niedriger Hierarchieebenen erreichen. Durch diese Art der Verzweigung entsteht das namensgebende Schneeflockenmuster. Verzichtet man auf die Aufspaltungen von den Dimensionstabellen aus, so reduziert sich die Struktur auf das Sternschema, aus dem das Snowflakeschema entstanden ist.

Ebenfalls in dieser Systemkomponente zu finden sind die so genannten **Summendaten**, die eine Sammlung bereits vorberechneter, wahrscheinlich in der Analyse benötigter Summenwerte sind, und aus den Detaildaten berechnet werden. Dies ist schon eine starker Optimierungsschritt in Richtung Analyseauswertung.

Metadaten-Repository:

Um den Datenbestand richtig deuten zu können, ist es, wie bei fast allen entwicklerischen Vorgängen, sehr sinnvoll möglichst alle relevanten Informationen über den Integrationsprozess der Daten in das Warehouse, sowie zu deren weiteren Transformation und Verarbeitung als Metadaten zu verwalten um eine optimale Nachvollziehbarkeit und Nutzbarkeit der Daten zu gewährleisten. Dies sind beispielsweise Informationen wie die Namen der Datenquellen mit den zugehörigen Extraktions- und Transformationsvorschriften.

Metadaten, also Daten über Daten, sind besonders in der (statistischen) Analyse ein wichtiges Thema, da es nicht ausreicht einfach nur Fakten zu kennen. Will man aus Daten Informationen und Wissen gewinnen, so muss man auch über deren Entstehung, Kontexteinordnung und Zusammenhänge möglichst viel wissen um eine richtige Interpretation möglich zu machen.

5.2.3 Analyseschnittstelle:

Bei der Schnittstelle zu den Endverarbeitungssystemen (fast ausschließlich Analysetools wie Data-Marts und OLAP-Systeme) ist die Grenze der Zuordnung zu den beiden Seiten absolut fließend. So müssen sich durch das OLAP geforderte Funktionalitäten bezüglich der Anfragen, und damit auch der Anfragesprachen und Methoden, auch im Data Warehouse niederschlagen und entsprechende Funktionen unterstützen (z.B. mächtige Aggregationen und so genannte OLAP-Funktionen). Außerdem kann diese Komponente eigene Datenrepräsentationen (separate Data-Marts) enthalten, die für spezielle Anwendungen und spezialisierte Analysetools hin ausgelegt und erstellt worden sind.

6. Zusammenfassung und Ausblick

Im Allgemeinen sind unter den Begriffen OLAP und Data Warehousing in der Wirtschaft einsetzbare Systeme und Systemkonzepte gemeint, die durch eine standardisierte, anwendugsunabhängige und verarbeitungsoptimierte Datenhaltung einen großen Schritt im Feld der Decision Support Systeme gemacht haben. Dabei soll eine organisationsweite analytisch orientierte Auswertung möglicht aller heterogenen Datenbestände mit dem Ziel der Informationsversorgung von Entscheidungsträgern ermöglicht werden.

Momentan wird oft vom „Aufarbeiten“ der stürmischen Entwicklung der letzten Jahre geredet [Lehn03], womit das wirtschaftliche und wissenschaftliche Verarbeiten und Nachbereiten gemeint ist. Es müssen klare Definitionen und Analysen der Konzepte und Methoden auf wissenschaftlicher Basis nachgeholt werden, da die Entwicklung in OLAP und Data Warehouse Beriech zu schnell von statten als das dies hätte parallel in ausreichendem Rahmen erfolgen können.

Für die Zukunft sind wohl die Punkte der Vielseitigkeit und Nahtlosigkeit von Systemen, sowie die Verfügbarkeit (z.B. Netzwerke) und Standartformatierung (z.B. XML) von Daten, die primären Entwicklungspunkte in diesem Bereich. Man wird also versuchen dem Benutzer mit seinem Warehouse eine noch breitere Funktionalität zu liefern, die im System noch selbstständiger Daten einholt, transformiert und verarbeitet.

7. Literatur

[BeMu98] Behme W.; Mucksch, H.
Das Data Warehouse-Konzept
Gabler, Wiesbaden, 1998

[Codd93] Codd, E.,F.
Providing OLAP to User-Analysts: An IT Mandate
Arbor Soft, 1993
(Elektronisch verfügbar unter:
http://www.arborsoft.com/OLAP.html)

[CrPe94] Creeth, R.; Pendse, N.
OLAP-Report
1994
(Elektronisch verfügbar unter:
http://www.olapreport.com)

[DaSu97] Dayal, U.; Surajit, C.
An Overview of Data Warehousing and OLAP Technology
1997
(Elektronisch verfügbar unter:
http://www.acm.org/sigmod/record/issues/9703/chaudhuri.ps)

[Fors97] Forsman, S.
OLAP Council White Paper
OLAP Council, 1997
(Elektronisch verfügbar unter:
http://www.olapcouncil.org/research/whtpaply.htm)

[Lehn03] Lehner, W.
Datenbanktechnologie für Data-Warehouse-Systeme, Konzepte und Methoden
Dpunkt Verlag, Heidelberg, 2003

[Popp99] Popp, G.
Knosys – Einführung in OLAP
DC Soft GmbH, 1999
(Elektronisch verfügbar unter:
http://www.dcsoft.de/knosys/html/body_olap.html)