

Todsünden des Data Warehousing

Das Data-Warehouse-Konzept hat sich als unverzichtbarer Bestandteil moderner Systeme zur Entscheidungsunterstützung am Markt etabliert. Der Synergieeffekt eines Data Warehouse als einheitlicher und integrierter Datenpool für fortschrittliche Data-Mining-Applikationen, CRM-Systeme, OLAP-Analysen oder auch klassische Berichtswerkzeuge ist unumstritten. Im vorliegenden Artikel werden als Ergebnis mehrjähriger Projekterfahrung wesentliche Fallstricke („Todsünden“) bei der Entwicklung eines Data-Warehouse-Systems skizziert, um Projektbeteiligten wie auch Entscheidern ein kompaktes Hilfsmittel an die Hand zu geben, Rahmenbedingungen für ein erfolgreiches Data-Warehouse-Vorhaben abzustecken.

1 Einführung

Um den unternehmerischen Erfolg und damit den Bestand einer Unternehmung auf Dauer zu sichern, spielen nicht nur kurzfristige Lenkungsaktivitäten innerhalb des Unternehmens, sondern vor allem auch Gestaltungsaktivitäten mit häufig mittel- bis langfristiger Wirkung eine zunehmend wichtige Rolle. Die hierzu notwendigen Entscheidungen setzen eine adäquate Informationsgrundlage voraus, um rechtzeitig und flexibel auf Veränderungen und sich daraus ergebenden zukünftigen Chancen bzw. Bedrohungen in der Wettbewerbsumwelt reagieren zu können.

Ein derzeit viel diskutierter Ansatz zur Verbesserung der unternehmensweiten Informationsversorgung stellt das *Data-Warehouse-Konzept* dar. Ein Data Warehouse erhebt den Anspruch, eine unternehmensweite, integrierte und konsistente Datenbasis getrennt von den Datenbeständen operativer Systeme des jeweiligen Tagesgeschäfts zu schaffen.

Laut Aussagen der Meta Group führen jedoch nur 30 Prozent aller Data-Warehouse-Projekte zum Erfolg, während 50 Prozent die gestellten Erwartungen nur zum Teil erfüllen. Die verbleibenden 20 Prozent hingegen scheitern völlig. Diese unbefriedigende Situation wird dahingehend weiter verschärft, dass Data-Warehouse-Vorhaben i.d.R. sehr kostenintensiv sind. Einer Umfrage des Data Warehousing Institute (TDWI) zufolge sind Durchschnittskosten in Höhe von ca. 1,2 Millionen Dollar für das erste Projektjahr aufzuwenden.

Ziel des vorliegenden Artikels ist es, Problembewusstsein für mögliche Bedrohungen im Rahmen eines Data-Warehouse-Projektes zu wecken. Diese „Todsünden“ im Sinne von Projektrisiken sollten bereits im Vorfeld eines Data-Warehouse-Vorhabens, soweit möglich, eingehend beleuchtet und reflektiert werden. Abbildung 1 vermittelt einen Überblick über die sieben zentralen Themenbereiche für Todsünden, die uns in unserer langjährigen Projekterfahrung im Data-Warehouse-Umfeld immer wieder begegnet sind. Diese Problembereiche werden in der betrieblichen Praxis leider häufig zu Unrecht stark unterschätzt. Streng genommen kann man anstelle von Todsünden auch von „Unterlassungssünden“ sprechen.

Wie Abbildung 1 zeigt, überwiegen sog. „weiche Faktoren“ bei den Problembereichen, wie z.B. das Projektteam oder die Informationsbedarfs-

analyse. Technologische Probleme sind meist leichter beherrschbar.

Ziel des Artikels ist es nicht, allgemeingültige Lösungsvorschläge für die Todsünden innerhalb der jeweiligen Problembereiche zu skizzieren. Vielmehr soll Grundwissen zur Schärfung des eigenen Problembewusstseins vermittelt werden, um die vorgestellten Bereiche bereits selbst kompetent und frühzeitig in die eigenen Überlegungen einbeziehen zu können. Darauf aufbauend werden Lösungsstrategien vorgeschlagen, um die Chancen für ein erfolgreiches Data-Warehouse-Vorhaben zu erhöhen.

Im folgenden werden die einzelnen Themenbereiche und die mit ihnen verbundenen Todsünden detailliert betrachtet.

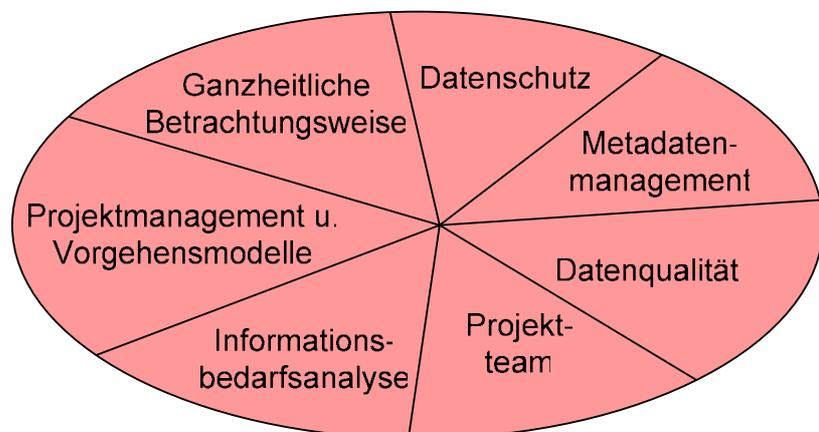


Abbildung 1: Todsünden des Data Warehousing

2.1 Ganzheitliche Betrachtungsweise

Die Vernachlässigung einer ganzheitlichen Betrachtungsweise mit einem mittleren bis langfristigen Fokus bei der Konzeption ihrer Data-Warehouse-Vorhaben führt bei vielen Organisationen zu Fehlentwicklungen in ihrer Entscheidungsunterstützungs-Landschaft. Diese „Second Best“-Lösungen sind oft nur durch einen verhältnismässig hohen Kosten- und Zeitaufwand nachträglich zu korrigieren. Vor allem drei Arten von unzureichenden Herangehensweisen sind besonders oft zu beobachten:

- Um schnelle Erfolge zu realisieren werden sukzessive *Entscheidungs-„Inseln“* in Form von voneinander isolierten Data Marts aufgebaut. Um Inselübergreifende Auswertungen zu ermöglichen, werden ggf. mehrstufig Daten zwischen diesen Lösungen ausgetauscht. Inmon, der Schöpfer des „Data-Warehouse“-Gedankens, spricht in diesem Zusammenhang von „*Spinnennetzen*“, deren gegenseitige Abhängigkeiten kaum mehr entflochten werden können.
- Aus Unsicherheit, welche konkreten Ziele im Fokus des Projekts stehen sollen, wird der große „Wurf“ versucht, der alle Probleme der Unternehmung auf einmal löst. Diese „Big Bang“-Herangehensweise führt in der Regel zu einer enormen Komplexität, die das Projektteam gänzlich überfordert.
- Eine Unternehmung sollte darüber hinaus nicht der Illusion unterliegen, ihr Data-Warehouse-Vorhaben auf der „grünen Wiese“ durchführen zu können. Vielmehr muss das Data Warehouse in die bestehende Anwendungssystemlandschaft des Unternehmens integriert werden.

Entscheidend beim Aufbau einer umfassenden Data-Warehouse-Lösung ist eine *ganzheitliche Konzeption unter Berücksichtigung einer langfristigen Perspektive*. Hierbei ist das viel zitierte Motto „*think big, start small*“ zu berücksichtigen.

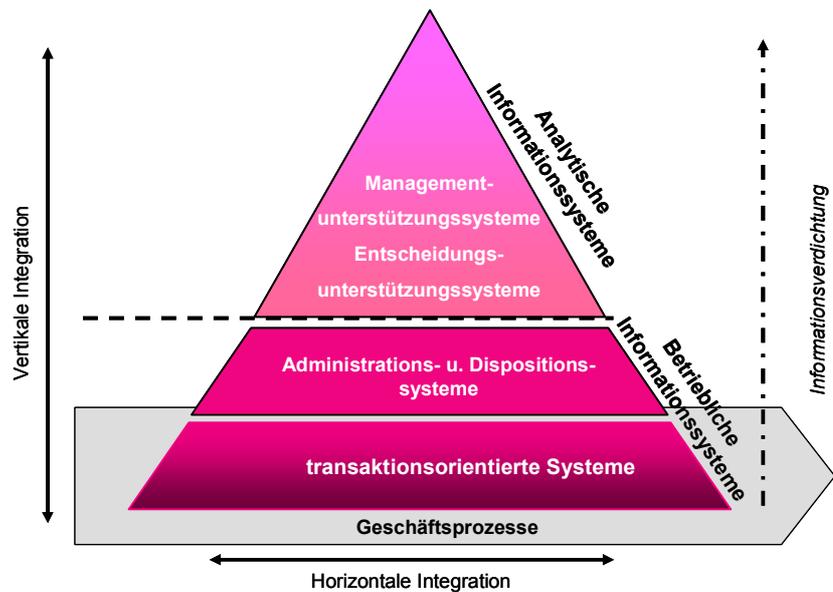


Abbildung 2: Informationssystemarchitektur einer Unternehmung

- **Think big:** Bei der Planung des Data-Warehouse-Vorhabens ist die Informationssystemarchitektur des Unternehmens einzubeziehen (vgl. Abbildung 2). Zur Unterstützung der Geschäftsprozesse eines Unternehmens existieren etablierte, meist heterogene betriebliche Informationssysteme. Zur Unterstützung des täglichen Geschäftsablaufs kommen transaktionsorientierte Systeme zum Einsatz. Diese werden durch Administrations- und Dispositionssysteme unmittelbar gesteuert. Im Data-Warehouse-Umfeld spricht man bei diesen betrieblichen Informationssystemen von operativen Systemen. Die horizontale Integration der Geschäftsprozesse vieler Unternehmungen ist durch eine Vielzahl von Projekten unter dem Schlagwort „Business Process Reengineering“ weit fortgeschritten. Das Data-Warehouse-Konzept kann in Ergänzung hierzu einen wesentlichen Beitrag zur vertikalen Integration der Informationen in einer Unternehmung leisten. Eine ganzheitliche Konzeption auf der Grundlage einer zentralen Data-Warehouse-Datenbasis ist ein elementarer Bestandteil eines jeden fortschrittlichen analytischen Informationssystem mit dem Ziel der Management- bzw. Entscheidungsunterstützung. Abbildung 3 vermittelt einen ersten Eindruck einer einfachen Data-Warehouse-Architektur. Das Data Warehouse bildet das Fundament für OLAP-Technologien und „State of the Art“ Business Intelligence-Frontends. Die Data-Warehouse-Architektur muss sich zu einem integralen Bestandteil der gesamten Informationssystemarchitektur entfalten.
 - **Start small:** Der Aufbau eines umfassenden Data Warehouse wird aufgrund von Kosten-, Zeit- und Risikoaspekten typischerweise nicht als ein Projekt, sondern als Sequenz von mehreren (Teil-)Projekten gestaltet. Ein langfristiger Horizont ist dabei unabdingbar. Besser (zügig) mit einer 80-prozentigen Lösung beginnen als nach einer 100-Prozent-Lösung zu streben, die nie erreicht wird.
- Bei der ganzheitlichen Konzeption einer Data-Warehouse-Lösung ist eine Reihe von zentralen Erfolgsfaktoren zu berücksichtigen:
- **Zweck:** Im Mittelpunkt einer jeden Data-Warehouse-Konzeption steht der Zweck und die mit diesem verbundenen Ziele. Der idealtypische Zweck eines Data Warehouse ist es, grundsätzlich alle Auswertungen (auch noch nicht bekannte) zu ermöglichen. Bei der Festlegung des Zwecks sollte

beachtet werden, dass ein Data Warehouse sicherlich nicht das Allheilmittel für alle betrieblichen wie organisatorischen Probleme in einem Unternehmen ist. Vielmehr ist eine realistische Erwartungshaltung gefordert.

- **Data-Warehouse-Prozess:** „Data-Warehousing ist kein Produkt, sondern der Prozess der Zusammenführung und des Managements von Daten aus verschiedenen Quellen mit dem Zweck, eine einheitliche, detaillierte Sicht auf einen einzelnen Geschäftsbereich oder das gesamte Unternehmen zu erhalten.“ (Gardner, 1998). Data-Warehouse-Werkzeuge sollten nach dem „Best of Breed“-Ansatz entsprechend der spezifischen Anforderungen ausgewählt werden. Bei einer Unterstützung durch externe Berater sollte vor allem auf Herstellerunabhängigkeit und Systemintegrationsfähigkeiten geachtet werden.
- **Unternehmensspezifische, differenzierte Gestaltung:** Das „One Size Fits All“-Konzept erscheint problematisch, da eine Vielzahl von unternehmensindividuellen Faktoren zu berücksichtigen sind. Dazu gehören u.a. die historische Informationssystemlandschaft sowie rechtliche oder ökonomische Rahmenbedingun-

- **Wirtschaftlichkeitsanalyse:** Eine Wirtschaftlichkeitsanalyse bildet die Grundlage zur Beurteilung und Einschätzung der Vorteilhaftigkeit eines Data-Warehouse-Projektes unter Anwendung geeigneter Methoden der Investitionsrechnung. Im Rahmen von Wirtschaftlichkeitsanalysen wird der Versuch unternommen, den Nutzen mit monetären Größen zu bewerten. Grundsätzlich sollte mit dem Teilprojekt mit dem größten messbaren (quantifizierbaren) bzw. zumindest identifizierbaren Nutzen begonnen werden. Typischerweise ist das ein bereichsspezifischer Data Mart mit begrenzter Entwicklungsdauer und Kosten. Wichtig ist es hierbei die Gesamtkonzeption, z.B. eines zentralen Data Warehouse, zu berücksichtigen und schrittweise parallel weiter auszubauen. Somit kann eine kurzfristige Realisierung eines Nutzens nachgewiesen werden, die häufig entscheidend für eine Genehmigung weiterer Mittel im Rahmen des gesamten Data-Warehouse-Vorhabens ist.

Bei der Wirtschaftlichkeitsanalyse sollten strategische Leistungs- und Nutzenpotentiale als qualitative Faktoren neben quantitativen Messgrößen Berücksichtigung finden. Diese lassen sich auch als

gung eines Data-Warehouse-Projektes einsetzen:

- Verbesserung der Informationsqualität zur Entscheidungsunterstützung
- Erhöhung der Effektivität der Informationsversorgung für dispositive, entscheidungsunterstützende Aufgaben
- Erhöhung der Wettbewerbsfähigkeit
- Erhalt bestehender bzw. Erwerb neuer Marktanteile
- Schaffung oder Erschließung neuer Märkte
- Erhöhung der Kundenzufriedenheit
- Verbesserung der Produkt- oder Dienstleistungsqualität

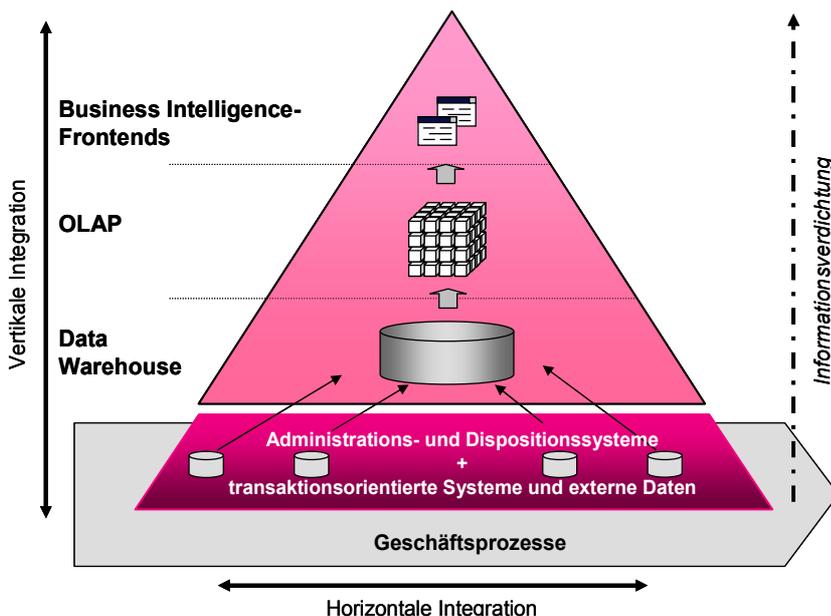
Die Betrachtung einer ganzheitlichen Konzeption mit einer langfristigen Perspektive unter Beachtung der genannten Erfolgsfaktoren ist essentiell für erfolgreiche Data-Warehouse-Vorhaben.

Da Data-Warehouse-Vorhaben grundsätzlich projektorientiert durchgeführt werden, wird als nächstes der Themenbereich Projektmanagement betrachtet.

2.2 Projektmanagement und Vorgehensmodelle

Die mit Data-Warehouse-Vorhaben verbundenen Aufgaben sind sowohl im Bezug auf die beanspruchten Ressourcen als auch hinsichtlich der Inhalte umfangreich und komplex und erfordern ein hohes Maß an Koordinations- und Steuerungsaktivitäten, um eine termin-, inhalts- und kostengerechte Zielerreichung zu gewährleisten. Somit liegen die besten Voraussetzungen vor, diese Aufgaben in ein klar definiertes Projekt zu überführen und strukturiert abzarbeiten.

Projekte zeichnen sich durch die Merkmale Neuartigkeit, d.h. es kann nur bedingt auf vorhandenes unternehmensinternes Know-how oder vorliegende Planungen zurückgegriffen werden, zeitliche Begrenztheit, Komplexität und Beteiligung mehrerer Stellen aus. Die Planung, Steuerung und Kontrolle, das sog. *Projektmanagement*, begleitet den gesamten Entwicklungsprozess, der in ein im Vor-



gen.

Argumente für die Rechtferti-

Abbildung 3: Einfache Data-Warehouse-Architektur

aus definiertes, klar abgegrenztes Ergebnis mündet. Dieses kann im Data-Warehouse-Kontext beispielsweise eine Data-Mart-Entwicklung sein. Der Projektleiter kann mit weiteren Aufgaben, wie z.B. Entscheiden, Delegieren, Koordinieren und Organisieren konfrontiert sein. Entscheidend für die erfolgreiche Durchführung eines Data-Warehouse-Projektes im Kontext des Projektmanagements sind vor allem zwei Faktoren:

- **Projektleiter:** Erfolgskritisch ist die Auswahl eines *erfahrenen und qualifizierten Projektleiters*, der alle Aufgaben im Rahmen des Projektmanagements konsequent wahrnehmen kann.
- **Vorgehensmodell:** Neben der Person des Projektleiters ist zu beachten, dass ein auf Data-Warehouse-Vorhaben abgestimmtes Vorgehensmodell für die effiziente Unterstützung des Projektmanagements zum Einsatz kommt.

Um Softwareprojekte besser plan- und kontrollierbar zu machen, strebt man nach einer Strukturierung des Ablaufs von Softwareprojekten, worunter man die Aufteilung des Projektes in wohldefinierte Teilaktivitäten versteht. Diese *Vorgehensmodelle* legen einen Rahmen für die Softwareentwicklung mit dem Ziel eines „in-

genieurmäßigen“ Vorgehens fest.

Klassische Projektmodelle, wie z.B. das Wasserfallmodell, weisen bei einer Anwendung im Data-Warehouse-Kontext eine Vielzahl von Defiziten auf, da die Entwicklung von Data Warehouses in der Regel von ganz anderen Anforderungen an die Projektdurchführung geprägt ist. Folgende für Data-Warehouse-Projekte zentrale Anforderungen sind im Vorgehensmodell besonders zu berücksichtigen:

- Der Auftraggeber und die Endbenutzer sind oft nicht in der Lage ihre Anforderungen an das neue System in den frühen Projektphasen explizit und/oder vollständig zu formulieren. Dabei ist das Motto „*I can't tell you what I want, but I'll know it when I see it*“ häufig anzutreffen.
- Laufende Anpassung bzw. Änderung von Anforderungen nach dem Motto „der Appetit kommt beim Verzehr“ sind zu berücksichtigen.
- Im Projekt müssen Projektrisiken frühzeitig erkannt und anschließend ausgeschaltet oder überwacht werden.
- Ein linearer, streng sequentieller Projektablauf ist eine ideale Annahme, die in realen Projekten meist nicht erfüllbar ist.

- Rücksprünge in frühere Phasen und Schleifen (Iterationen) sind notwendig.

Die Softwareindustrie hat eine Vielzahl von Vorgehensmodellen hervorgebracht. Jedes Vorgehensmodell hat besondere Vorzüge, aber auch Schwachstellen. Beim Data Warehousing bietet sich eine Kombination von neuartigen Vorgehensweisen an, um möglichst viele dieser Vorzüge zu nutzen. Im Folgenden werden vor allem die *Merkmale* aufgezeigt, die ein Vorgehensmodell im Data-Warehouse-Kontext aufweisen sollte. Zunächst wird jedoch ein dazu passendes exemplarisches Vorgehensmodell (vgl. Abbildung 4) beschrieben.

Jedes Vorgehensmodell/Projektmodell besteht aus dem Durchlaufen mehrerer Phasen, wobei Phasen ggf. wiederum in *Teilphasen* oder Aktivitäten weiter untergliedert werden. Am Ende ausgewählter Phasen existieren sog. *Phasenlinien* bzw. *Meilensteine*, die im Projektverlauf definierte Zwischenergebnisse liefern.

Die Kernphasen des Projektes werden durch vor- und nachgelagerte Phasen eingeschlossen. In der *Initialisierungsphase* sind grundsätzliche Rahmenbedingungen, wie z.B. Anwendungszweck, Data-Warehouse-Architektur oder Toolauswahl zu klä-

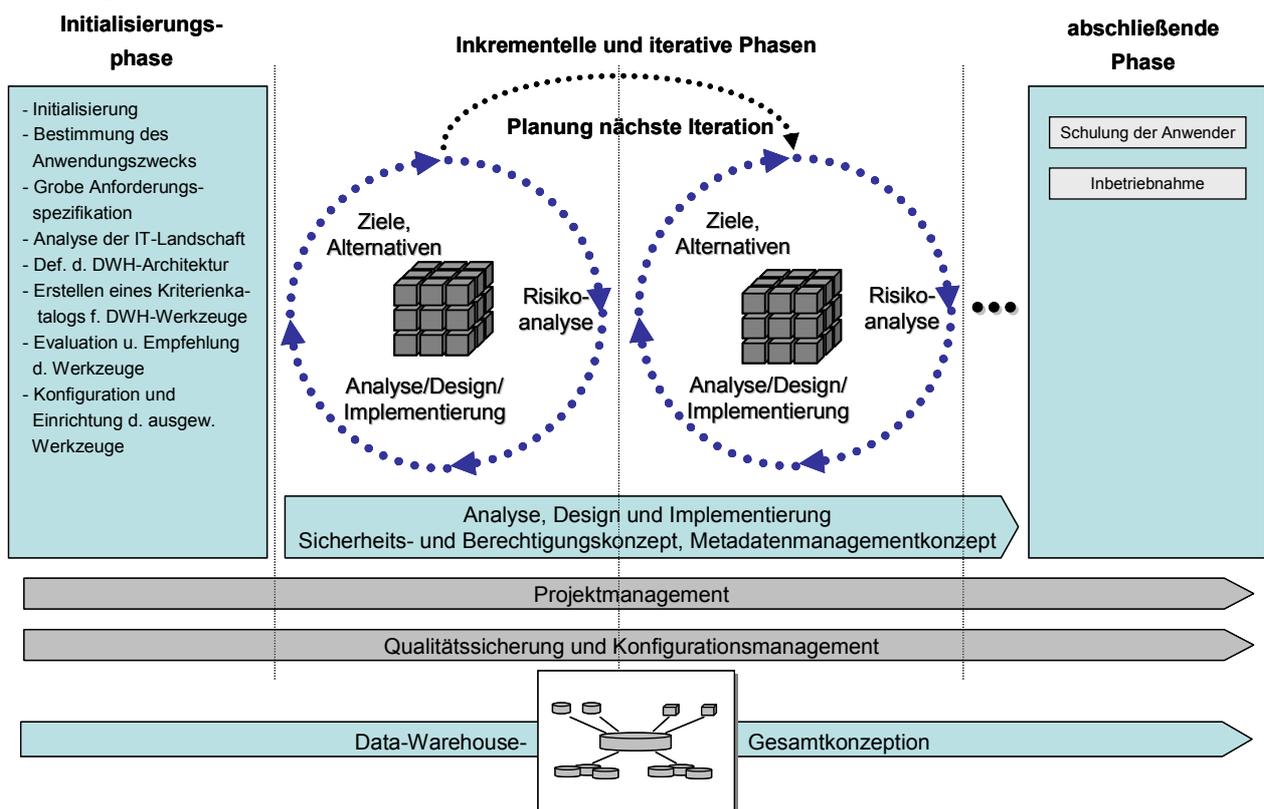


Abbildung 4: Exemplarisches Data-Warehouse-Vorgehensmodell

ren während sich die *abschließende Phase* mit Aktivitäten wie Schulung oder Inbetriebnahme beschäftigt.

Die Aufgabe der Kernphasen, die sog. iterativen und inkrementellen Phasen besteht in der Entwicklung bzw. Integration des Data Warehouse in einzelnen kleinen und überschaubaren Schritten von ca. 3 bis 6 Monaten Dauer. Häufig entspricht eine Einheit einer Data-Mart-Entwicklung. Neben der Analyse (Informationsbedarfsanalyse, Quelldatenanalyse), dem Design (Fachkonzeptentwurf, Datenmodellentwurf/-modifikation) spielt hierbei vor allem die Implementierung eine wichtige Rolle.

Diese Phasen der Softwareentwicklung werden ergänzt durch eine Reihe von projektbegleitenden Aktivitäten, wie z.B. *Projektmanagement*, *Qualitätssicherung* und *Konfigurationsmanagement*. Sie finden parallel zum übrigen Projektablauf statt. Eine besonders wichtige projektbegleitende Aktivität ist die Weiterfortführung der *Data-Warehouse-Gesamtkonzeption*, durch die die langfristige Perspektive des Data-Warehouse-Vorhabens immer im Fokus bleibt.

Das exemplarisch vorgestellte Vorgehensmodell weist folgende *Merkmale* auf, die für ein erfolgreiches Data-Warehouse-Vorhaben essentiell sind:

- **Iterativ:** Die einzelnen Phasen laufen nicht streng sequentiell ab, sondern können sich bei Bedarf auch wiederholen. Die dabei vorgesehene Möglichkeit der Rückkopplung erlaubt auch Abläufe, die nicht streng „wasserfallartig“ verlaufen, sondern Schleifen enthalten. Im Rahmen der Rückkopplung sollte die bisherige Sammlung der Anforderungen ggf. modifiziert werden kann.
- **Inkrementell:** Die Entwicklung erfolgt in kleinen überschaubaren Einheiten. Die Inkremente werden normalerweise im Rahmen der jeweiligen Iteration erstellt.
- **Prototyping:** Die Entwicklung von Prototypen trägt zur besseren Kommunikation der fachlichen Anforderungen mit dem Auftraggeber bei. Es werden unterschiedliche Arten von Prototypen unterschieden. Während der *Wegwerfprototyp* im Rahmen einer Quick

& Dirty-Realisierung für Data-Warehouse-Vorhaben nicht geeignet ist, erlaubt das sog. *evolutionäre Prototyping* eine Softwareentwicklung, die über eine Folge von Prototypen im produktiven Programmsystem gipfelt.

Das Prototyping bringt vor allem folgende Vorteile mit sich:

- Identifikation von weiteren fachlichen Anforderungen nach dem WYSIWIG-Prinzip (*what you see is what you get*).
- Risikominimierung bezüglich der fachlichen Anforderungen.
- Verbesserung der Anforderungsanalyse und -definition hinsichtlich ihrer Konsistenz und ihrer Detaillierungstiefe.
- Da man in kürzeren Abständen fertige Teilergebnisse mit dem Auftraggeber abstimmen kann, erhöht sich im Allgemeinen auch die Akzeptanz des Systems.
- **Risikogesteuert:** Es bietet sich an, eine Risikoanalyse vorzuschalten (Auflösung der aktuell sichtbaren Projektrisiken).

Data Warehousing ist kein einmaliges Vorhaben mit klar definiertem Abschluss. Sowohl die Informationsbedürfnisse der Endbenutzer wie auch die operativen Applikationen als Datenquellen sind vielen Änderungen unterworfen, die sich permanent auf ein Data Warehouse auswirken. Vielmehr führen diese zu weiteren kleineren Teilprojekten. Deshalb sind die genannten Merkmale besonders wichtig.

2.3 Informationsbedarfsanalyse

Der Aktivität Informationsbedarfsanalyse ist im Rahmen von Data-Warehouse-Vorhaben ein sehr hoher Stellenwert beizumessen, da der Gewinnung von Informationen und Wissen als Grundlage für die Entscheidungsfindung eine zentrale Bedeutung zukommt.

Innerhalb eines Vorgehensmodells (vgl. Abbildung 4) ist sie der Analyse zuzuordnen und somit das Fundament

für die darauf aufbauenden Design- und Implementierungsschritte. Dem hohen Stellenwert der Informationsbedarfsanalyse wird dadurch Rechnung getragen, dass sie hier noch einmal explizit hervorgehoben wird.

Zwei Zitate unterstreichen die hohe Relevanz der Erhebung von Informationen und Wissen im Rahmen der Informationsbedarfsanalyse anschaulich:

“*Knowledge is the only meaningful economic resource*” (Drucker, 1995)

“*In an economy where the only certainty is uncertainty, the one sure source of lasting competitive advantage is knowledge*” (Nonaka, 1991)

Mit der *Informationsbedarfsanalyse* werden die Informationsbedürfnisse des Managements und der jeweiligen Fachbereiche ermittelt.

Spätere Phasen in der Entwicklung des Data-Warehouse-Systems bauen auf den Ergebnissen der Informationsbedarfsanalyse auf. Somit sind Fehler und Unzulänglichkeiten in darauf folgenden Entwicklungsphasen nur mit großem Aufwand korrigierbar.

Bevor der Begriff Informationsbedarf eingehend betrachtet und darauf aufbauend ein Lösungsansatz vermittelt wird, werden im Folgenden eine Reihe von *Problemen*, die bei der Erhebung des Informationsbedarf auftreten können, skizziert:

- Die Phantasie der zukünftigen Anwender reicht in der Praxis in der Regel nicht aus, den wirklichen Informationsbedarf zu Beginn eines Projekts überhaupt und abschließend zu definieren. Dies liegt teilweise daran, dass der Fachbereich bzw. das Management seinen jeweiligen Informationsbedarf selbst nicht gut genug kennt.
- Der Informationsbedarf ist in vielen Fällen nur vage bestimmbar und hängt vor allem von der zugrunde liegenden Aufgabenstellung, den angestrebten Zielen und den psychologischen Eigenschaften der Entscheidungsträger ab.

- Zudem sind sich die Anwender oft nicht bewusst, welche neuartigen Informationen mit dem zu entwickelnden Data-Warehouse-System bereitgestellt werden können.
- Häufig sind falsche Erwartungen des Managements bzw. der Endanwender im Hinblick auf den gewünschten Informationsbedarf anzutreffen. Dies kann ein weitreichendes Projektrisiko sein.

Abbildung 5 zeigt den Zusammenhang zwischen den wesentlichen Begrifflichkeiten bei der Informationsbedarfsanalyse. Das Verständnis von Informationsangebot, -nachfrage, subjektiven vs. objektiven Informationsbedarf erlaubt es, die genannten Probleme besser einzuordnen:

- **Informationsraum:** Die den ein-

onsbedarf beinhaltet jene Informationen, die für den Entscheidungsträger zur Erfüllung seiner Aufgaben relevant sind.

- **Subjektiver Informationsbedarf:** Der *subjektive Informationsbedarf* umfasst jene Informationen, die dem Entscheidungsträger als relevant erscheinen.
- **Informationsnachfrage:** In der Regel artikulieren Entscheidungsträger nur einen Teil ihres subjektiven Informationsbedarfs. Deshalb wird die *Informationsnachfrage*, d.h. die konkret nachgefragte Informationsmenge, als Teilmenge des subjektiven Informationsbedarfs beschrieben.
- **Informationsangebot:** Das *Informationsangebot* wird definiert als die Gesamtheit der Informati-

- **Informationsstand:** Der Teil der Informationsversorgung, der durch den objektiven Informationsbedarf geprägt ist, ist der *Informationsstand*. Der Informationsstand ist die für die Aufgabe des Entscheiders relevante Informationsmenge, die letztendlich für eine Entscheidung sinnvoll verwertet werden kann.

Ein Königsweg für die Informationsbedarfsanalyse gibt es nicht. Beispielsweise können Eigenschaften der Vorgehensmodelle, wie Prototyping sowie inkrementelle und iterative Phasen helfen, da die späteren Anwender unmittelbar und frühzeitig eingebunden werden. Entscheidend ist eine umfassende Interaktion mit dem betroffenen Personenkreis (Fachleuten bzw. Entscheidern). Das System soll Teil der Tagesarbeit dieser Mitarbeiter werden, von diesen akzeptiert und ständig genutzt werden. Wichtig ist die Kommunikation, d.h. zum Beispiel sind die Mitarbeiter des Unternehmens frühzeitig über das Projekt, dessen Intention, Inhalt und interne Organisation zu informieren.

Damit die Anwenderwünsche erkannt und erfasst werden können, werden im Folgenden mehrere Teilschritte als Lösungsansatz für die Informationsbedarfsanalyse vorgestellt:

- **Identifikation von Informationsbedarfen:** Data Warehousing soll durch die Mitarbeiter „gelebt“ und als dynamischer Prozess verstanden werden. Dafür ist es jedoch zwingend erforderlich, die Mitarbeiter mit den Methoden der Informationsbedarfsanalyse vertraut zu machen, um die erforderlichen Ergebnisse zu erzielen. Viele Methoden der Informationsbedarfsanalyse aus der klassischen Softwareentwicklung lassen sich problemlos im Data-Warehouse-Umfeld anwenden:

(1) Deduktive Methoden: Zu den deduktiven Methoden zählen vor allem die Modellanalyse, die Analyse relevanter Gesetzestexte oder von Literaturquellen, die Anwendung von Kreativitätstechniken, wie z.B. Brainstorming sowie die Aufgabenanalyse.

(2) Induktive Methoden: Induktive Methoden sind z.B. die Dokumentenanalyse, datentechni-

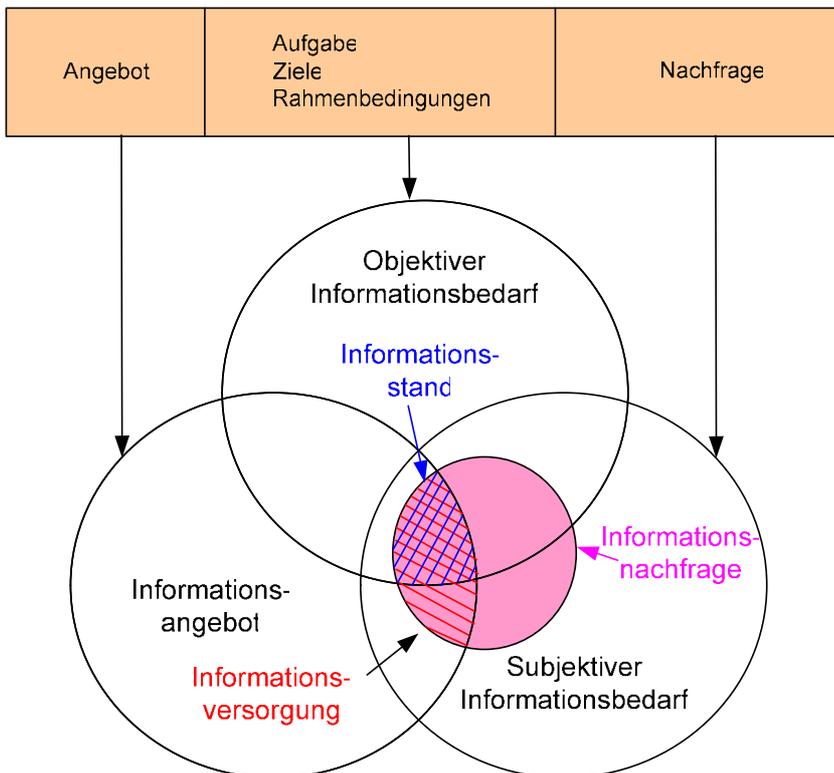


Abbildung 5: Informationsangebot, -nachfrage u. -bedarf

zelen Begriffen jeweils zugeordneten Informationsmengen sind Bestandteil eines umfassenden *Informationsraums* und werden hier anschaulich durch unterschiedlich große sich teilweise überlappende Kreise symbolisiert.

- **Objektiver Informationsbedarf:** Der *objektive Informati-*

onen, die einem Nachfrager zu einem bestimmten Zeitpunkt an einem bestimmten Ort zur Verfügung stehen.

- **Informationsversorgung:** Die Schnittmenge zwischen Nachfrage und Informationsangebot wird als *Informationsversorgung* bezeichnet.

sche Analysen, Organisationsanalyse sowie Befragungen durch Interviews oder Fragebögen.

Diese Methoden sollten als Grundlage für die Strukturierung von Anwenderwünschen eingesetzt werden. Wichtig ist eine

im Hinblick auf eine spätere Integration in die Data-Warehouse-Umgebung zu bewerten.

- **Priorisierung zusätzlicher Informationsbedarfe:** Die durch das Data-Warehouse-System bereitzustellenden Informationen

Projektteams sind besonders folgende Problembereiche zu beachten:

- Rollen der Mitarbeiter bei der Zusammensetzung des Projektteams
- Auswahl des Projektteams
- Einbettung der Projektorganisation in die Unternehmensorganisation
- Sponsoring durch das Management

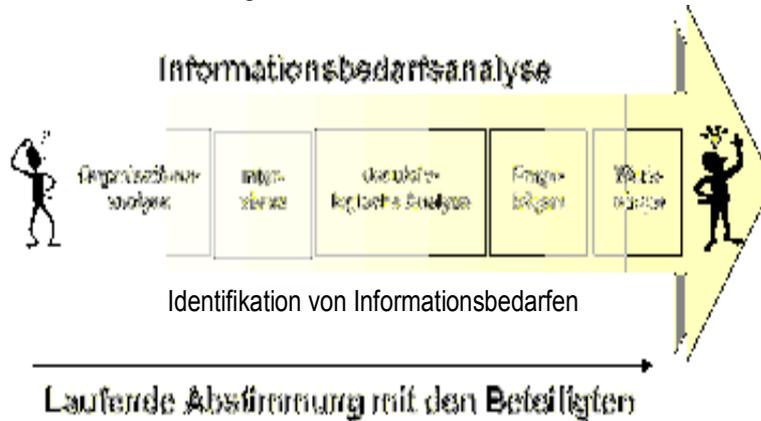


Abbildung 6: Methoden der Informationsbedarfsanalyse

standardisierte, flächendeckende Erhebung, sowie eine systematische Auswertung.

Im Data-Warehouse-Umfeld bietet sich vor allem eine Kombination induktiver und deduktiver Elemente an (vgl. Abbildung 6).

- **Synchronisation von Informationsbedarf und -angebot:** Der ermittelte Informationsbedarf ist mit dem vorhandenen Informationsangebot abzustimmen. Dies erfolgt in den meisten Fällen über eine Quelldatenanalyse, in der die Wünsche der Anwender mit den Daten in den operativen Systemen abgeglichen werden. Ein großer Teil der Bedürfnisse nach strategischen Managementinformationen kann jedoch häufig nicht unmittelbar befriedigt werden, da die notwendigen Daten nicht vorhanden sind. Hier gilt es abzuschätzen, ob eine Integration externer Datenquellen den zu erwartenden Nutzen rechtfertigt. Allenfalls gilt es die Erwartungen bzw. die Informationsbedürfnisse zu korrigieren.
- **Bewertung von Informationslücken:** Es ist ein Vergleich des Ist-Zustands der Informationsversorgung mit einem möglichen zukünftigen Soll-Zustand der Informationsversorgung durchzuführen. Informationslücken sind

müssen aufgrund von beispielsweise zeitlichen oder finanziellen Restriktionen und ihrer Wichtigkeit für die betriebliche Aufgabenerfüllung beurteilt und priorisiert werden. Dies kann in Workshops erfolgen, in denen eine Rückkopplung zum späteren Endanwender stattfindet.

Als Ergebnis der Informationsbedarfsanalyse sollte zumindest ein konzeptuelles, semantisches Datenmodell, das die Anwenderwünsche erfasst, eine Menge möglicher Abfragen/Berichte sowie eine Auswahl der hierzu notwendigen Datenquellen.

Es ist grundsätzlich illusorisch zu glauben, dass die Anforderungen der Nutzer vollständig erfasst werden können. Eine laufende Abstimmung mit allen Beteiligten fördert die Akzeptanz und hilft die Anforderungen ggf. anzupassen bzw. zu erweitern.

2.4 Projektteam

Neben eher formalen Faktoren, wie Vorgehensmodell, Informationsbedarfsanalyse und Konzeption spielt im Projektalltag der *menschliche Faktor* eine nicht zu unterschätzende Rolle. Vor allem das Projektteam und seine Einbettung in das betriebliche Umfeld bestimmen maßgeblich den Erfolg eines Data-Warehouse-Vorhabens. Bei der Gestaltung eines schlagkräftigen

Rollen: Die mit der Entwicklung eines Data Warehouse verbundene Komplexität erfordert das Einbinden einer Vielzahl spezialisierter Fähigkeiten. Diese Fähigkeiten lassen sich wiederum zu Rollen zusammenfassen. Neben Rollen, die im Rahmen der klassischen Softwareentwicklung weitgehend bekannt sind, sind im Data-Warehouse-Umfeld auch neuartige Rollen, wie z.B. die des ETL- oder OLAP-Spezialist zu besetzen. Im Folgenden wird eine mögliche Spannweite der benötigter Rollen aufgezeigt, die im Projekt als Checkliste herangezogen werden kann.

- **Projektleiter/Teamleiter:** Dies ist die zentrale Schlüsselrolle für das Projekt. Der Projektleiter ist einerseits für Projektplanung, -steuerung und -kontrolle und andererseits für die Kommunikation und das Projektmarketing im Sinne einer gezielt durchgeführten Werbekampagne für das Vorhaben zuständig. Dies setzt neben Führungsqualitäten ein weitreichendes Verständnis der betrieblichen Zusammenhänge voraus. Der Projektleiter sollte eine starke Persönlichkeit sein, die eine straffe Planung und eine konsequente Anpassung an Abweichungen vornimmt.
- **Data-Warehouse-Architekt:** Die zentrale Aufgabe des Data-Warehouse-Architekten ist die Konzeption einer ganzheitlichen Data-Warehouse-Architektur losgelöst von den operativen Systemen.
- **Datenbank-Administrator:** Umfassende Kenntnisse der zugrunde liegenden Datenbank sind für ein Data-Warehouse-Vorhaben genauso essentiell wie bei der klassischen Softwareentwicklung.

Neue Schwerpunkte sind dabei z.B. Aggregationsbildung und die Beachtung von Zeitfenstern beim Datenladen.

- **Spezialist für Datenmigration/ETL-Spezialist:** Die Extraktion, die Datentransformation und das Datenladen unter Beachtung der Datenqualität und unter Performanzgesichtspunkten stellt eine wesentliche Aufgabe dar.
- **Spezialist für Altanwendungen (Legacy Data):** Da die für das Data Warehouse bestimmten Daten oftmals aus einer Vielzahl von heterogenen Altanwendungen extrahiert werden, sind fundierte Kenntnisse dieser Systeme entscheidend.
- **Spezialist für Qualitätssicherung:** Im Rahmen des Data-Warehouse-Vorhabens ist ebenfalls für eine qualitätsgesicherte Entwicklung Sorge zu tragen.
- **Tester:** Alle Entwicklungsschritte sind vor Überführung in den Wirkbetrieb analog zur klassischen Softwareentwicklung intensiv zu testen.
- **Data-Mart-Entwickler/OLAP-Spezialist:** Neben der Data-Warehouse-Datenbasis sind vor allem bereichs- bzw. abteilungsspezifische Sichten, die sog. Data Marts, von besonderer Bedeutung. Die diesen zugrunde liegende Verarbeitungsform wird als Online Analytical Processing (OLAP) bezeichnet. Sie bilden die Basis für das neu zu gestaltende Data-Warehouse-Berichtswesen.
- **Infrastrukturspezialist/Systemadministrator:** Da Data-Warehouse-Projekte immer in die bestehende Landschaft des Unternehmens zu integrieren sind, ist die Rolle eines Infrastrukturspezialisten wichtig. Dieser wird ergänzt durch Systemadministratoren für die bestehenden Systeme sowie für das neu zu konzipierende Data-Warehouse-System.
- **Power User und Standard User:** Für den gesamten Data-Warehouse-Entwicklungsprozess muss eine geeignete Kooperationsform zwischen IT-Abteilung

und den Fachabteilungen gefunden werden. Die Einbindung der Fachbereiche in die Entwicklungsteams ist hier von besonderer Bedeutung, da nur auf diese Weise zu gewährleisten ist, dass das Data Warehouse später auch den Anforderungen der Benutzer entspricht. Hierbei sind die verschiedenen späteren Nutzertypen, wie z.B. Power User und Standard User, zu berücksichtigen. Die Zusammenarbeit mit den Fachanwendern bringt verstärkt fachliche Kompetenz in das Projekt ein. Diese Personen können im Rahmen der Einführung zur Information der übrigen Anwender bzw. als Multiplikatoren eingesetzt werden.

- **Verantwortlicher für Public Relations:** Er vermittelt den Nutzen des Data Warehouse und „verkauft“ das Projekt nach „oben“. Insbesondere bei großen Data-Warehouse-Vorhaben wird hierfür eine eigenständige Rolle losgelöst von den Aufgaben des Projektleiters benötigt.
- **Sponsor:** Der „Schirmherr“, der das Data-Warehouse-Vorhaben überwacht. Auf diese Rolle wird später im Abschnitt noch detaillierter eingegangen.

Eine konkrete Projektsituation kann die Zuordnung einer Rolle zu mehreren Mitarbeiter genauso wie eine Zuordnung mehrerer Rollen zu einem Mitarbeiter erfordern. Bei der Besetzung des Data-Warehouse-Teams sollte vor allem auch auf eine angemessene Teamgröße und eine Mitarbeiterkonstanz im Projektverlauf besonderes Augenmerk gelegt werden. Häufig ist es sinnvoll ein sog. Kernprojektteam zu bilden, das abhängig vom Projektverlauf durch Teilprojektteams jeweils ergänzt wird.

Teamauswahl: Bei der Zusammenstellung des Projektteams können Engpässe offensichtlich werden:

- Ist genügend *Personal* für die erforderliche Aufgabe vorhanden?
- Sind die benötigten *Fähigkeiten* innerhalb der Organisation verfügbar?

Data Warehousing stellt spezifische Anforderungen an die einzusetzende Methodik und Systematik. Das dafür

benötigte Wissen ist in den Unternehmen in aller Regel nicht oder nicht ausreichend vorhanden. Häufig besitzt das erste Data-Warehouse-Vorhaben in einer Unternehmung eine Pionierfunktion.

Dies führt zunächst zu einer *Make-oder-Buy-Entscheidung*. Diese beiden Extrempole, die vollständige Eigenrealisierung bzw. das komplette Outsourcing der Entwicklung kommen in den seltensten Fällen in ihrer Reinform zur Anwendung. Dazwischen existiert eine Vielzahl von sinnvollen Möglichkeiten, die abhängig von den unternehmerischen Rahmenbedingungen gestaltet werden sollte.

Objektivität, Wissenstransfer und Ausnutzung des vorhandenen Erfahrungsschatzes sind oft die zentralen Gründe, sich externe Unterstützung in Form von Beratungsleistungen zuzukaufen. Folgende Formen der externen Unterstützung finden in den meisten Fällen Anwendung:

- **Coaching Modell:** Der Berater stellt sein methodisches und fachliches Wissen im Rahmen der einzelnen Arbeitsschritte zur Verfügung und führt durch die Einzelphasen ohne an der inhaltlichen Erarbeitung der Informationen unmittelbar beteiligt zu sein.
- **Begleitende Funktion:** Der Berater erhebt gemeinsam mit den Mitarbeitern des Unternehmens die benötigten Informationen und wirkt aktiv in den Teilschritten mit.
- **Wissenserweiterung:** Schulung, Seminare, Vorträge bieten weitere Möglichkeiten, fehlendes Know-How aufzubauen.

Bei der Auswahl und Einbindung externer Berater sollte vor allem auf folgende Faktoren geachtet werden:

- Die externen Berater sind in das Projektteam geeignet zu integrieren, damit sie nicht als Eindringlinge wahrgenommen werden.
- Bei der Auswahl der Berater sollte insbesondere auf ausreichende Erfahrung im Aufbau und der Integration von Data-Warehouse-Systemen geachtet werden. Dies lässt sich häufig durch entsprechende Referenzprojekte in der jeweiligen Branche überprüfen.

- Externe Berater sollten möglichst frühzeitig in das Projekt eingebunden werden.

Projektorganisation: Obwohl das Data Warehouse längerfristig zu einem permanenten Bestandteil der täglichen Arbeit der Betroffenen wird, ist eine Integration der Projektarbeit in das operative Tagesgeschäft in der Regel zum Scheitern verurteilt. Der „Vorrang des Tagesgeschäfts“ stempelt das Data-Warehouse-Projekt zu einer unwichtigen Nebentätigkeit ab. Daher sollte eine eigenständige Projektorganisation etabliert werden, wobei deren Mitglieder von der klassischen Linienorganisation während der Projektlaufzeit weitgehend abgetrennt werden müssen. Ob hierbei eine permanente oder eine temporäre Projektorganisation gewählt wird, ist nicht entscheidend. Wichtig ist es jedoch, die Mitglieder dieser Projektorganisation funktions- bzw. bereichsübergreifend zu besetzen, um möglichst viele Facetten ausleuchten zu können.

Sponsoring: Jedes Data-Warehouse-Projekt ist Chefsache und sollte von einer einflussreichen Persönlichkeit aus einer höheren Managementebene begleitet werden. Dieser Mentor, häufig aus der Unternehmensleitung, ist verantwortlich für das Vorhandensein des Data-Warehouse-Projektes, ermöglicht den initialen Startschuss und fördert das Projekt während der Laufzeit. Idealerweise ist das Commitment aller verantwortlichen Führungskräfte zu sichern und auf eine hierarchieübergreifende Projektunterstützung vom Top Management bis zu den jeweiligen Process Ownern hinzuarbeiten. Ohne zumindest einen mächtigen Mentor wird das Projekt bei den ersten auftretenden Problemen häufig bereits in Frage gestellt. Da bei einer ganzheitlichen Data-Warehouse-Konzeption in den meisten Fällen Personen aus den unterschiedlichsten Unternehmensbereichen mit häufig unterschiedlichsten „subjektiven Mikrointeressen“ beteiligt sind, ist eine einflussreiche „Entscheidungsinstanz bzw. Mentor“ zwingend erforderlich.

2.5 Datenqualität

Eine unzureichende Qualität der Daten stellt häufig alle noch so intensiven Bemühungen zum Aufbau eines

Data Warehouse in Frage. Nach dem Leitspruch „*Garbage in, garbage out*“ stoßen mangelhafte Daten auf Akzeptanzprobleme beim Entscheidungsträger. Dessen einmal erworbenes gesundes Misstrauen gegenüber der bereitgestellten Datengrundlage verurteilt so manches Data-Warehouse-Projekt nachhaltig zum Scheitern, da niemand die enthaltenen Daten nutzt.

Folgende Aussagen belegen die Bedeutung einer angemessenen Datenqualität anschaulich:

“A smaller warehouse with cleaner data can be better than a large warehouse with bad data.”

„Nothing is more likely to undermine the performance and business value of a data warehouse than inappropriate, misunderstood, or ignored data quality.“

Werden Probleme der Datenqualität zu Beginn eines Projektes als trivial betrachtet oder schlicht ignoriert, so rächt sich dies im weiteren Projektverlauf.

Um eine besseres Verständnis für die Probleme und Auswirkungen unzureichender Datenqualität zu vermitteln, werden zwei Definitionen vorgestellt, die die wesentlichen Facetten ausleuchten:

- Nach der Norm DIN ISO 8402 ist Qualität durch die „*Gesamtheit von Merkmalen einer Einheit bezüglich ihrer Eignung, festgelegte und vorausgesetzte Erfordernisse zu erfüllen*“ definiert. Diese eher technische Festlegung stellt die Übereinstimmung der Eigenschaften der „Daten“ mit vorher festgelegten Anforderungen in den Vordergrund.
- Für L. English sind aber vor allem die Anforderungen der Endanwender entscheidend. Deshalb versteht er unter Datenqualität: „*Consistently meeting all knowledge worker and end-customer expectations*“.

Diese anwenderbezogene Sichtweise im Sinne von „*fitness for purpose*“ bzw. „*fitness for use*“ zeigt, dass die Datenqualitätsanforderungen grundsätzlich immer abhängig vom jeweiligen Kontext festzulegen sind.

Die Beurteilung der Datenqualität kann anhand von Dimensionen vorgenommen werden. Huang et al. 1999 haben vier Kategorien mit jeweils

mehreren Merkmalen vorgeschlagen, die eine weitreichende Analyse des vorliegenden Datenbestands ermöglichen. Während die *innere Datenqualität* die Qualität der Daten an sich zu bewerten hilft, stellt die *kontextabhängige Datenqualität* die Daten in einen Interpretationszusammenhang. Die Betrachtung der *Darstellungsqualität* und die *Qualität des Zugangs* runden die beiden anderen ab.

Um die vorgestellten Dimensionen besser verstehen zu können, ist es nützlich, die Ausprägungen zu kennen, anhand derer die Dimensionen beurteilt werden können:

- **Innere Datenqualität:** Genauigkeit, Objektivität, Glaubwürdigkeit, Vertrauenswürdigkeit. Diese Ausprägungen manifestieren sich häufig in inkorrekten Werten, syntaktischen Fehlern, Schlüsselanomalien und dem Auftreten von Duplikaten.
- **Kontextabhängige Datenqualität:** Unvollständigkeit, Nichtrelevanz, Nichtaktualität (zeitlicher Bezug) und Anwendungsbezug.
- **Darstellungsqualität:** ungenügende Dateninterpretationen, wie z.B. nicht interpretierbare Schlüsselssysteme oder frei verwendete Textfelder, Widerspruchsfreiheit, knappe Darstellung, Verständlichkeit.
- **Zugangsqualität:** Schutzverletzungen, fehlende Zugriffsmöglichkeiten.

Bei der Analyse der Datenqualität anhand der vorgestellten Qualitätsdimensionen zeigt sich, dass vor allem folgende drei **Problemfelder** entscheidend für Data-Warehouse-Systeme sind:

- Nicht korrekte Daten: 1-5% aller Datenfelder sind fehlerbehaftet.
- Inkonsistenzen zwischen Daten in unterschiedlichen Datenbanken.
- Nicht verfügbare Daten sind notwendig für Entscheidungen.

Diese Datenqualitätsprobleme führen neben unmittelbaren Auswirkungen auf der operativen Ebene auch zu mittelbaren Auswirkungen auf der Ebene der Entscheidungsunterstützung (dispositive Ebene).

Bezogen auf die operative Ebene sind dies vor allem folgende Auswirkungen:

- Unkorrekte Daten führen zur Verärgerung der betroffenen Kunden und damit zu einer geringeren Kundenzufriedenheit. Auch die Mitarbeiter des Unternehmens verspüren durch die damit verbundenen Konflikte häufig eine geringere Arbeitszufriedenheit.
- Die Kosten für schlechte Datenqualität sind für Unternehmen sehr hoch. Nach Schätzungen liegen sie bei 8 bis 12 % des Umsatzes. Diese Kosten entstehen, weil falsche und fehlende Daten zu einem Mehraufwand führen, der ansonsten vermieden werden könnte. Die Korrektur qualitativ unzureichender Daten ist im Vergleich zu qualitätssichernden Maßnahmen bei der Datenerfassung um den Faktor 5 bis 10 teurer.

Die aufgezeigten Folgen werden auf der dispositiven Ebene durch weitere Auswirkungen verstärkt:

- Daten bilden die Basis für Analysen und Auswertungen eines jeden Data-Warehouse. Qualitativ hochwertige und somit geeignete Auswertungen sind nur auf Grundlage korrekter und den Anforderungen entsprechender Daten zu erreichen. Eine unzureichende Datenqualität führt zu

Falschbeurteilungen und in letzter Konsequenz auch zu schlechteren Entscheidungen.

- Diese Wirkung wird verstärkt durch massive Akzeptanzprobleme bei den Entscheidungsträgern, die sich im fehlenden oder unzureichenden Vertrauen in die Daten widerspiegeln.
- Wird die Korrektur der Daten nicht bereits weitgehend auf der Ebene der operativen Systeme vorgenommen, führt dies zu erheblichen Kosten für das nachträgliche Bereinigen der Daten beim Laden in die Data-Warehouse-Landschaft.

Der Zusammenhang zwischen Zeit, Kosten und Qualität lässt sich darüber hinaus anschaulich im sog. „Teufelsdreieck“ oder auch „magischen Dreieck“ veranschaulichen (vgl. Abbildung 7). Das Dilemma besteht darin, dass eine Verbesserung des Qualitätsaspekts zu einer Erhöhung der Kosten und Zeitaufwände und umgekehrt führt. Umso wichtiger ist es, sich bereits frühzeitig im Data-Warehouse-Prozess mit möglichen Problemen der Datenqualität auseinanderzusetzen.

English (English, 1999) schlägt folgende Schritte vor, um sich Problemen in der Datenqualität zu nähern:

- Definition und Festlegung von Datenqualitätsanforderungen aus Nutzersicht für die Entscheidungsunterstützung. „... organi-

zations have to distinguish between what is vital and what is merely desirable.“

- Datenqualitätsbestimmung: Es ist zunächst ein Bewusstsein über den aktuellen Stand der Datenqualität im Unternehmen zu schaffen. *“This understanding is mandatory in data warehousing environments characterized by multiple users with different needs for data quality.”*
- Auf Basis eines Soll-/Ist-Abgleichs mit den Ergebnissen der beiden ersten Schritte ist eine Kostenbewertung für nicht qualitativ geeignete Daten aufzustellen.
- Auf einer auf Schritt drei aufzusetzenden Priorisierung ist eine Datenbereinigung und Datenumstrukturierung anzustoßen.

Für die Sicherung der Datenqualität lässt sich eine Vielzahl von Maßnahmen, Konzepten, Abläufen und Methoden heranziehen. Im Data-Warehouse-Umfeld ist als Schwerpunkt die Auswahl geeigneter Werkzeuge für den ETL-Prozess zu berücksichtigen. Nicht zuletzt wegen der komplexen Bereinigungsprozesse gehen Schätzungen davon aus, dass der Aufwand für den ETL-Prozess bis 80 Prozent der gesamten Data-Warehouse-Projekt-dauer betragen kann.

Viele der am Markt verfügbaren ETL-Werkzeuge besitzen bereits umfangreiche Möglichkeiten zur Bereinigung der Datenqualität. Darüber hinaus haben sich spezialisierte Werkzeuge für den Bereinigungsprozess etabliert. Die folgenden Kategorien zeigen die mögliche Bandbreite von Unterstützungsmöglichkeiten auf:

- **Data Migration Tools:** Diese Werkzeuge unterstützen die Extraktion, den Transport und die Transformation von Daten auf ihrem Weg in das Data-Warehouse. Vor allem die Transformationsregeln, Filtereinstellungen, Vorgaben für Wertebereiche in der Transformationsphase verbessern Plausibilitäts- und Konsistenzprüfungen
- **Data Scrubbing/Profiling Software:** Diese Werkzeuge analysieren und lesen die Daten ein. Sie

Das magische Dreieck

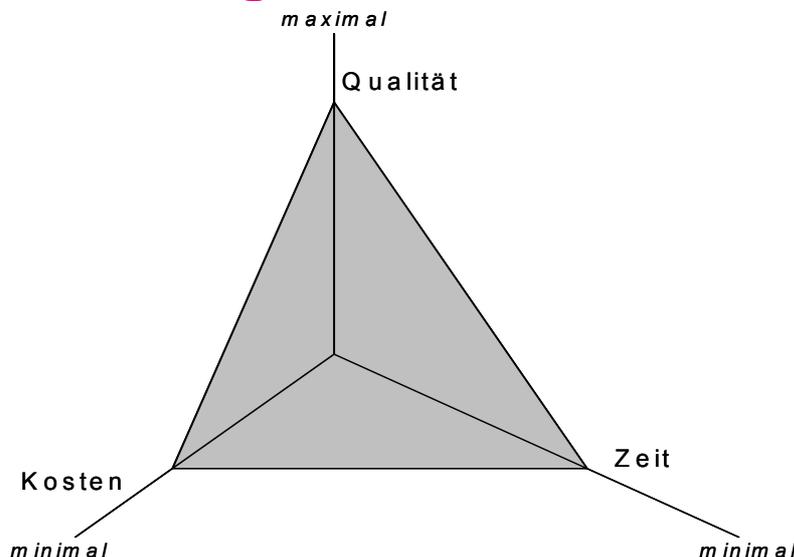


Abbildung 7: Das magische Dreieck

entdecken Zusammenhänge auf Basis von Regeln, neuronalen Netzen oder durch Fuzzy Logik.

- **Data Auditing/Profiling Tools:** Diese Werkzeuge entdecken die Datenregeln durch die Datenanalyse selbst. Sie stellen darüber hinaus Analysen auf Basis statistischer Stichproben zur Verfügung.

Das Aufdecken von Fehlern und Ungereimtheiten in den geladenen Daten führt häufig auch zum Aufspüren von Mängeln in der operativen Vorsystemen. Diese Mängel sollten, wenn möglich, immer auch am Ort ihrer Entstehung beseitigt werden.

Nach einem weiter gefassten Begriffsverständnis lässt sich auch der *Total Quality Management* (TQM)-Gedanke aus der Produktion auf Aspekte der Datenqualität übertragen. In diesem Zusammenhang kann von einem *Total Data Quality Management* (TDQM) gesprochen werden. Dieser rundet die vorgestellten Aspekte der Datenqualität durch eine Übertragung der Sicherstellung der Datenqualität in allen Data-Warehouse-Phasen – von der Planung über die Implementierung bis zur Wartung – ab.

2.6 Metadatenmanagement

In nahezu allen Bereichen einer Data-Warehouse-Architektur entstehen Metadaten bzw. werden Metadaten benötigt. Durch Metadaten soll die Komplexität des Data Warehouse beherrschbar gemacht werden. Folgende Aussage belegt die hohe Relevanz von Metadaten anschaulich:

„*Metadata is the data warehouse’s Gordian knot, and Alexander and his sword are nowhere in sight.*“
(Kimball et al. 1998)

Der Nutzen von Metadaten im Data-Warehouse-Umfeld ist vielfältig:

- **Allgemeiner Nutzen:** Metadaten sollen ein Verständnis der Zusammenhänge im Data Warehouse für Anwender wie auch Entwickler vermitteln.
- **Nutzen aus Anwendersicht:** Aus Anwendersicht steht der Aufbau eines einheitlichen Begriffs- und Definitionssystems, die Transparenz und Nachvollziehbarkeit sowie die Vermeidung von In-

konistenzen des Data-Warehouse-Prozesses im Vordergrund.

- **Nutzen aus Entwicklersicht:** Aus Entwicklersicht sollen mit Metadaten die Zusammenhänge des Administrationsprozesses automatisiert werden.

Generell gilt es anzumerken, dass die Menge aller Daten, die man unter dem Begriff Metadaten kategorisiert, sehr vielschichtig und umfangreich sein kann. *Metadaten* werden als Daten über Daten definiert. Sie beschreiben den inhaltlichen und strukturellen Aufbau und die Bedeutung von Daten. „*Metadata is like the roadmap to data.*“ Genauso wie ein Katalog einer Bibliothek sowohl den Inhalt als auch den Bücherstandort beschreibt, ebenso zeigen Metadaten auf den Ort und die Bedeutung der verschiedensten Informationen in einer Data-Warehouse-Umgebung. „*Metadata can be described as a description of what you wish were in your data fields.*“

Folgende zwei Beispiele veranschaulichen den Zusammenhang zwischen Daten und Metadaten:

<i>Daten:</i>	<i>Metadaten:</i>
26.10.01	Datum im Format TT.MM.JJ
IC 253	Zugnummer oder Elektrisches Bauteil

Metadaten lassen sich grundsätzlich in *fachliche* und *technische* Metadaten differenzieren. Während technische Metadaten der Entwicklung und dem Betrieb des Data Warehouse, und damit den Entwicklern zugeordnet werden, unterstützen die fachlichen Metadaten die Nutzung des Data Warehouse durch die Endanwender.

Darüber hinaus lässt sich eine *aktive* und *passive* Metadatenhaltung unterscheiden. Eine passive Metadatenhaltung dient ausschließlich der Beschreibung der Daten (Dokumentation), Strukturen und Prozesse einer Data-Warehouse-Umgebung. Als aktiv wird eine Metadatenhaltung bezeichnet, die es nicht nur erlaubt Beschreibungen zu speichern, zu verändern und auszugeben, sondern darüber hinaus aus den gespeicherten Metadaten Aktionen in Form einer Steuerung des Data-Warehouse-Prozesses abzuleiten. Aktive Metadaten bewirken

meist eine Veränderung in den Daten, Strukturen und Prozessen der Data-Warehouse-Umgebung.

Die Metadatenhaltung in heutigen Data-Warehouse-Umgebungen ist durch eine Vielzahl von *Problembereichen* gekennzeichnet, deren Verständnis für die darauf aufbauenden Lösungsansätze essentiell sind:

- Analog zu den Problemen der Datenintegration existieren vergleichbare Probleme bei der *Metadatenintegration*. Beim Aufbau eines umfassenden Metadaten-systems sind Vagheiten, Synonyme, Homonyme anzutreffen, die den Entwicklungsprozess behindern.
- Im Rahmen des Data-Warehouse Prozesses sind häufig viele spezialisierte Werkzeuge mit separater Metadatenhaltung, die als Standardsoftware angeboten werden nach dem *Best of Breed*-Ansatz zu integrieren.
- Die Verwendung von verschiedenen produktspezifischen, proprietären Speicherungsformaten oder der Einsatz proprietärer *Application Programming Interfaces* (APIs) bei der Metadatenhaltung erschwert die Integration.
- Weiterhin stellen die große Anzahl von Komponenten, aus denen ein Data-Warehouse bestehen kann, und die Tatsache, dass diese Komponenten u.U. von unterschiedlichen Herstellern stammen, hohe Anforderungen an die Metadateninteroperabilität.

Ein umfassendes zentrales Metadatenmanagement in einem Data Warehouse ist heutzutage noch nicht durch ein entsprechendes Werkzeug vollständig abgedeckt. Vielmehr ist ein dezentraler Ansatz vorherrschend, der die Berücksichtigung folgender Problembereiche mit sich bringt:

- Die Gewährleistung des Austausches zwischen den verschiedenen Herstellerwelten bzw. die Integration in ein einheitliches Modell ist zu beachten.
- Zudem sind begleitende organisatorische Maßnahmen aufzusetzen, die eine kontinuierliche Pflege der Metadaten fördern.

Um eine integrative Sichtweise auf die Metadaten zu fördern, sollten die

ausgewählten Werkzeuge einem einheitlichen „Metadaten“-Standard genügen. In den letzten Jahren haben sich zwei Ansätze zur Standardisierung herausgebildet:

- Das *Open Information Model* (OIM) der *Meta Data Coalition* (MDC)



sam genutzten Metadaten für Data-Warehouse-Umgebungen. Es ist zukunftssicher und offen durch Verwendung von objektorientierten Schlüssel- und Internettechnologien.

CWM integriert drei Schlüsseltechnologien und Industriestandards:

- *XML* — die eXtensible Markup

schreibung von Tätigkeiten, die im Rahmen von Analysen durchgeführt werden.

- **Management:** Der Management-Bereich umfasst Metamodelle, die zur Beschreibung der Prozesse innerhalb eines Data Warehouse genutzt werden.

Der Einsatz von Werkzeugen auf der Basis eines weithin akzeptierten Standards im Metadatenumfelds garantiert weitgehend Investitionssicherheit in die Zukunft.

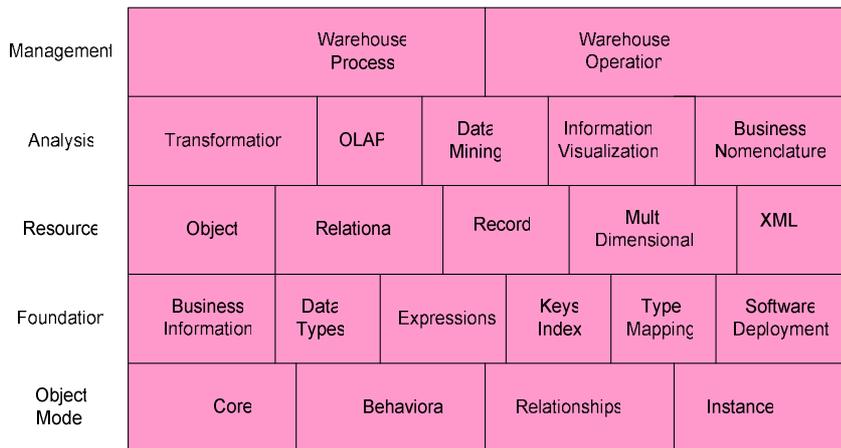


Abbildung 8: Architektur des Common Warehouse Metamodel

- *Common Warehouse Metamodel* (CWM) der *Object Management Group* (OMG)

Die MDC hat sich mittlerweile aufgelöst und arbeitet in der OMG am CWM weiter. Der OIM-Ansatz ist somit gescheitert. Damit hat sich CWM als der industrieweite de facto Standard für den Austausch von Metadaten im Data-Warehouse-Umfeld etabliert.

Das CWM definiert einen Metadatenstandard für alle wichtigen Anwendungsbereiche eines Data Warehouse, so dass es einen Grundstein für die Operabilität von Metadaten in Data Warehouses legt.

Durch die Unterstützung seitens vieler namhafter Hersteller ist zu erwarten, dass das CWM sich zukünftig als das Austauschformat in der Praxis durchsetzen wird. Die OMG wurde 1989 von 11 Unternehmen gegründet und umfasst mittlerweile 800 Unternehmen, Forschungsinstitute und Organisationen.

Das CWM umfasst eine komplette Spezifikation der Syntax und Semantik zum Im- und Export von gemein-

- *Language*, ein W3C-Standard
- *UML* — Unified Modeling Language, ein OMG-Standard
- *MOF* — Meta Object Facility, ein OMG-Standard

Die CWM-Architektur (vgl. Abbildung 8) besteht aus den Schichten *Management*, *Analysis*, *Resource*, *Foundation* sowie *Object Model* und deckt damit alle Bereiche eines Data Warehouse ab. Für das Data Warehousing stellen die einzelnen Ebenen aufeinander aufbauende, aber voneinander logisch abgrenzbare Bereiche dar.

- **Foundation:** Der Foundation-Bereich enthält grundlegende Metamodelle, wie zum Beispiel Datentypen, die als Basis anderer Bereiche dienen.
- **Resource:** Im Resource-Bereich finden sich Metamodelle zur Beschreibung von Quell- und Zieldatenstrukturen, wie z.B. relationalen Datenbanken oder XML-Dateien.
- **Analysis:** Der Analysis-Bereich beinhaltet Metamodelle zur Be-

2.7 Datenschutz

Schlagworte, wie *analytisches Customer Relationship Management* (aCRM), *kundenzentrisches Data Warehouse* und *Informationsflut* führen zu einem „Warnsignal“ beim Datenschutzbeauftragten (Möller 1998). Das Data-Warehouse-Konzept bewegt sich im Spannungsfeld zwischen der gewinnbringenden Nutzung von Kundendaten einerseits und der Angst des individuellen Anwenders andererseits, als „gläserner“ Kunde ausgebeutet zu werden.

Deshalb sind die den Datenschutz betreffende Fragestellungen zwingend im Vorfeld zu klären und daraus resultierende Einschränkungen ggf. zu berücksichtigen:

- Was soll, was darf gespeichert werden?
- Für welchen Zeitraum dürfen Daten gehalten werden?
- Ist der Wahrheitsgehalt der Daten sichergestellt?
- Werden gespeicherte Daten richtig interpretiert?
- Ist der Zugang zu den Daten abgesichert?
- Wann verletzt die Datenanalyse die Privatsphäre des Kunden?

Der *Datenschutz* beschäftigt sich mit den rechtlichen Regelungen zum Schutz von Persönlichkeit, Privatsphäre und der informationellen Selbstbestimmung.

Wer Datenschutz betreibt, will somit in erster Linie nicht Daten schützen, sondern die Privatsphäre von Personen. Aufgabe des Datenschutzes ist es daher, das Recht des Einzelnen zu schützen, grundsätzlich selbst über die

Preisgabe und Verwendung seiner Daten zu bestimmen.

In Abgrenzung dazu befasst sich *Datensicherheit* mit dem Schutz der gespeicherten Daten vor unbefugtem Zugriff.

Die gesetzlichen Regelungen für den Datenschutz sind national und regional äußerst unterschiedlich ausgeprägt. Während in Deutschland auf Landes- sowie auf Bundesebene (*Bundesdatenschutzgesetz*, BDSG) detaillierte Vorschriften existieren, gibt es z.B. in den USA kein allgemeines Datenschutzgesetz. Für spezielle Ausprägungen, wie z.B. dem Telekommunikationsbereich sind in Deutschland weitere Sonderregelungen vorhanden.

Verstöße gegen das Bundesdatenschutzgesetz können zu Straf- und Bußgeldverfahren mit bis zu einem Jahr Gefängnis und ca. 25.000 Euro Geldbuße führen.

Wesentlich für den Datenschutz ist es, ob es sich bei den gespeicherten Daten um „Personendaten“, d.h. von eindeutig identifizierten und somit bestimmbar Personen, handelt, denn nur sie sind vom Datenschutz erfasst. Bestimmbarkeit ist gegeben, wenn die Daten mittels zusätzlicher Informationen einer bestimmten Person zugeordnet werden können.

Die Verarbeitung personenbezogener Daten in einem Data Warehouse ist mit einer Vielzahl von Problemen verbunden:

- Eine Verarbeitung und Nutzung personenbezogener Daten ist demnach nur zulässig, wenn sie durch das BDSG oder eine andere Rechtsvorschrift erlaubt oder die betroffene Person eingewilligt hat. Sie unterliegt demnach grundsätzlich einem Verbot mit *Erlaubnisvorbehalt*. Die Wahrung des Rechts auf informationelle Selbstbestimmung überlässt dem einzelnen Menschen grundsätzlich die Entscheidung, ob und wie seine Person betreffende Daten verarbeitet werden (Einwilligung). Für den Betroffenen muss jederzeit *transparent* sein, welche Daten über ihn gespeichert sind, d.h. er kann z.B. eine Auskunft bzw. Bestätigung über seine gespeicherten Daten einfordern.

- Die personenbezogenen Daten können grundsätzlich nur für einen zuvor vereinbarten *Zweck* genutzt werden. Dieser Zweck ist meist mit einem konkreten Vertrag verknüpft. Diese Form der Zweckbestimmung ist jedoch im Data-Warehouse-Umfeld problematisch. Durch die natürliche Evolution eines Data Warehouse werden die Daten häufig später anders genutzt als zu Anfang festgelegt. Dies liegt an den zu Beginn noch nicht vorgesehenen Benutzern mit zum Teil nicht klar festgelegten Zielen. Diese Loslösung vom ursprünglichen Verwendungszweck bedarf wiederum der Einwilligung der Betroffenen. Beispielsweise müssten eine Klassifikation von unterschiedlichen Kundentypen und darauf aufbauende gezielte Marketing- und Werbemaßnahmen im Hinblick auf CRM-Aktivitäten als Zweck formuliert sein. Analog ist das Aufspüren von Zusammenhängen mittels Data-Mining-Aktivitäten zu betrachten.
- Grundsätzlich unterscheidet sich der Zweck im Data-Warehouse-Umfeld klar von der Datenerhebung in den operativen Systemen. Der Zweck „Data Warehousing“ ist zu pauschal und nicht konkret genug. Der Zweck kann zudem auch nicht „Vertrag“ sein. Erschwerend kommt hinzu, dass die Navigationsmöglichkeiten durch OLAP-Analysen ein nahezu beliebiges in Beziehung setzen von Merkmalen ermöglichen. Es werden gerichtete Analysen auf unterschiedlichen Navigationspfaden im Sinne von „Surfen“ in den Datenbeständen erlaubt.
- In Konzernen besteht großes Interesse daran, Data-Warehouse-Systeme konzernweit einzusetzen. Ein Konzern stellt jedoch keine einheitliche datenverarbeitende Stelle dar. Vielmehr ist jedes juristisch selbständige Tochterunternehmen eine solche Stelle. Daher ist jede Datenweitergabe an bzw. der Datenabruf durch andere Tochterunternehmen des Konzerns bzw. an die Konzernmutter eine Übermittlung an Dritte und somit wiederum an die

Einwilligung der Betroffenen gebunden.

- Trotz des Weglassens der identifizierenden Merkmale und des Namens kann der Personenbezug in der Regel trotzdem mit vertretbarem Aufwand an Zeit, Kosten und Arbeitskraft wiederhergestellt werden und somit liegen keine anonymisierten, sondern wiederum personenbezogene Daten vor.

Aufgrund der dargestellten Probleme ist streng genommen zu argumentieren, dass für das Betreiben eines Data Warehouse mit personenbezogenen Daten derzeit keine Rechtsgrundlage existiert. Somit besteht das latente Risiko einer Stilllegung des gesamten Data-Warehouse-Prozesses auf zunächst unbestimmte Zeit, wenn personenbezogene Daten in größerem Umfang gespeichert werden. Eine Verarbeitung personenbezogener Daten im Data Warehouse und deren Verwendung zum Data Mining ist deshalb nur eingeschränkt möglich.

Folgende Lösungsansätze sind im Rahmen des Datenschutzes zu beachten, um die dargestellte Problematik abzumildern:

- Da ein Data Warehouse zur Entscheidungsunterstützung und nicht zur operativen Steuerung eingesetzt wird, ist die Speicherung von personenbezogenen Daten nur eingeschränkt oder überhaupt nicht nötig. Hier bietet sich das Konzept der *Anonymisierung* an. Sie dient der Veränderung personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbar Person zugeordnet werden können. Kein unverhältnismäßig großer Aufwand liegt z.B. vor, wenn eine Re-Identifizierung über Kundennummern vorgenommen werden kann. Der Aufwand für die Entschlüsselung muss im Bezug auf den Wert der Information als sehr hoch einzustufen sein (unverhältnismäßiger Aufwand). In diesem Zusammenhang spricht man von

faktischer Anonymisierung bzw. von Pseudoanonymisierung.

- Sind nicht anonyme Daten zwingend nötig, ist eine Einwilligung des Kunden mit einem klar definierten Zweck nötig. Hierbei ist zu beachten, dass der Betroffene über die Bedeutung seiner Einwilligung aufgeklärt werden muss, um rechtskonform in die Datenverarbeitung einwilligen zu können. Hierzu gehört die Information des Betroffenen über den Zweck der Verarbeitung. Allgemein gehaltene Erläuterungen, wie solche, dass die Daten zu Werbezwecken verarbeitet werden, sind insoweit unzureichend.
- Um den Kunden zu einer Einwilligung zu bewegen, können Grundprinzipien des *Permission Marketings* eingesetzt werden. In diesem Zusammenhang sind vor allem folgende Prinzipien zu respektieren:
 - **Kundennutzen**
 - **Qualität:** Die Daten sind immer richtig und auf dem neusten Stand.
 - **Konsens:** Die Daten werden nur zu den Zwecken eingesetzt, denen der Kunden zugestimmt hat.
 - **Transparenz:** Der Kunde weiß, was mit seinen Daten geschieht.
 - **Sicherheit:** Mit den Daten wird sorgfältig und vertrauensvoll umgegangen.
 - **Weitergabe:** Eine Weitergabe der Daten findet ohne Einwilligung des Kunden nicht statt.
 - **Zugriff:** Der Kunde hat die Option, seine Daten einzusehen und zu verändern, ggf. auch zu löschen.
- Um sich bereits frühzeitig im Data-Warehouse-Prozess abzustimmen und späteren Problemen aus dem Weg zu gehen, ist sowohl die Einbindung des Betriebsrats als auch die Hinzuziehung des jeweiligen Datenschutzbeauftragten anzuraten.

Zusammenfassend lässt sich festhalten, dass bei der Planung von Data-Warehousing- und Data-Mining-

Projekten zunächst immer geprüft werden sollte, ob die Verarbeitung personenbezogener Daten wirklich notwendig ist, oder ob die Verarbeitung von anonymen Daten ausreicht.

Als Konsequenz sollte weiterhin der Datenschutz nicht mehr nur als notwendiges Übel verstanden werden, sondern als wichtiges und unverzichtbares Merkmal der Vertrauensbildung und Differenzierung gegenüber den Kunden und Konkurrenten.

3. Zusammenfassung und Ausblick

Die vorgestellten Themenbereiche und die damit verbundenen Todsünden sind nach unserem Ermessen die häufigsten Ursachen für Probleme bei der Abwicklung von Data-Warehouse-Vorhaben, die sogar zu einem Scheitern des gesamten Projektes führen können. Wir hoffen, dass wir durch die Problembeschreibungen sowie die Skizzierung von Lösungsansätzen einen Denkanstoß bei der Beurteilung des eigenen Data-Warehouse-Vorhabens geliefert haben. Zur Vertiefung der Problemstellung sind im Folgenden, gegliedert nach den Themenbereiche, weiterführende Literaturquellen genannt.

4. Literatur

Ganzheitliche Konzeption:

- [Gard98] Gardner, S. R.: *Building the Data Warehouse*, in: Communications of the ACM, 41. Jg., Nr. 9, 1998, S. 52-60.
- [Kach99] Kachur, R.J.: *Planning and Project Management For the Data Warehouse*, DM Direct, Februar 1999.
- [Kurz99] Kurz, A.: *Data Warehousing – Enabling Technology*, MITP, Bonn, 1999.
- [KRRT98] Kimball, R., Reeves, L., Ross, M., Thornthwaite, W.: *The Data Warehouse Lifecycle Toolkit*, Wiley, New York, 1998.

Projektmgmt und Vorgehensmodelle:

- [Boe88] Boehm, B.W.: *A Spiral Model of Software Development and Enhancement*, in: IEEE Computer, 21. Jg., Nr. 5, 1988, S. 61-72.

- [BKKZ92] Budde, R., Kuhlenkamp, K., Mathiessen, L., Züllighoven, H.: *Ap-*

proaches to Prototyping, Springer, Berlin, 1992.

- [Balz98] Balzert, H.: *Lehrbuch der Software-Technik – Software-Management, Software-Qualitätssicherung, Unternehmensmodellierung*, Springer, Berlin, 1998.

Informationsbedarfsanalyse:

- [BoUT99] Böhnlein, M., Ulbrich-vom Ende, A., Tropp, G.: *CEUS- Entwicklung eines Data Warehouse-Systems für die bayerischen Hochschulen*, 1. Workshop des GI-Arbeitskreises Modellierung und Nutzung von Data Warehouse-Systemen auf der MoBIS'1999, Bamberg, 1999.

- [Druc95] Drucker, P.F.: *Managing in a Time of Great Change*, Truman Talley Books, New York, 1995.

- [Nona91] Nonaka, I.: *The Knowledge-Creating Company*, in: Harvard Business Review, 69. Jg., Nr. 6, November-Dezember 1991, S. 96-104.

- [Schi01] Schirp, G.: *Anforderungsanalyse im Data-Warehouse-Projekt: Ein Erfahrungsbericht aus der Praxis*, HMD, Heft 22, 2001.

Projektteam:

- [DeLi99] DeMarco, T., Lister, T.: *Wien wartet auf DICH – Der Faktor Mensch im DV-Management*, 2. Aufl., Hanser, München, 1999.

- [CoAb96] Corey, M.J., Abbey, M.: *Oracle Data Warehousing*, Oracle Press, Berkeley, 1996.

Datenqualität:

- [Engl99] English, L.: *Improving Data Warehouse and Business Information Quality – Methods for reducing costs and increasing profits*, Wiley, New York, 1999.

- [Engl02] English, L.: *Essentials for Effective Information Quality Management*, 8th Data-Warehouse-Forum, St. Gallen, Schweiz, 2002.

- [Gard97] Gardyn, E.: *A Data Quality Handbook for a Data Warehouse*, in: Strong, D.M., Beverly, K.K. (Hrsg.): *Proceedings of the 1997 Conference on Information Quality*, Massachusetts, 1997, S. 267-290.

- [HuLW99] Huang, J., Lee, Y.W., Wang, R.Y.: *Quality Information and Knowledge*, Prentice Hall, Upper Saddle River, 1999.

Metadatenmanagement:

- [DSMR00] Do, H.; Stöhr, T.; Müller, R.; Rahm, E.: *Metadaten-Management im Data Warehouse-Bereich*, 2. Workshop des GI-Arbeitskreises Modellierung und Nutzung von Data Warehouse-Systemen, Freiburg, 2000.
- [StVV99] Staudt, M., Vaduva, A., Metterli, T.: *Metadata Management and Data Warehousing*, Technical Report 21, Institut für Informatik, Universität Zürich, 1999.
- [OMG02] Object Management Group: *Data Warehousing, CWM™ and MOF™ Resource Page*, <http://www.omg.org/cwm>, 2002.

Datenschutz:

- [Moel98] Möller, F.: *Data Warehouse als Warnsignal an den Datenschutzbeauftragten*, in: *Datenschutz und Datensicherung*, 22. Jg., Nr. 10, 1998, S. 555-560.
- [o.V.00] *Data Warehouse, Data Mining und Datenschutz - Entschließung der 59. Konferenz der Datenschutzbeauftragten des Bundes und der Länder* vom 14./15. März, <http://www.datenschutz-berlin.de/doc/de/konf/59/datawa.htm>, 2000.
- [Schw99] Schweizer, A.: *Data Mining, Data Warehousing – Datenschutzrechtliche Orientierungshilfen für Privatunternehmen*, Orell Füssli, Zürich, 1999.