→ 4., überarbeitete und erweiterte Auflage

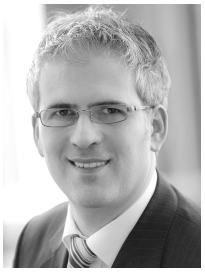
Andreas Bauer · Holger Günzel (Hrsg.)

Data Warehouse Systeme

Architektur, Entwicklung, Anwendung

dpunkt.verlag

Prof. Dr.-Ing. Holger Günzel studierte Informatik an der Universität Erlangen-Nürnberg. Danach war er dort als wissenschaftlicher Mitarbeiter am Lehrstuhl für Datenbanksysteme tätig. Von 2001 bis 2007 war er Berater und zuletzt Führungskraft bei der IBM Business Consulting Services in den Themen »unternehmensweite Architekturen«, »Business Intelligence« und »Serviceorientierte Architekturen«. Seit 2007 ist er Professor an der Fakultät für Betriebswirtschaftslehre der Hochschule München für das Lehrgebiet »Prozess- und Informationsmanagement«. Seit 2013 ist er Studiengangsleiter des Masterstudiengangs Betriebswirtschaft – European Business Consulting. Er ist Mitgründer der GI-Arbeitskreise »Konzepte des Data Warehousing« und »Enterprise Architecture«.



Dr.-Ing. Andreas Bauer studierte Informatik an der Universität Erlangen-Nürnberg. Danach war er als wissenschaftlicher Mitarbeiter am Fachgebiet Wirtschaftsinformatik der TU Darmstadt und am Lehrstuhl für Datenbanksysteme der Universität Erlangen-Nürnberg tätig. Von 2003 bis 2008 war er Berater bei der T-Systems sowie Siemens IT Solutions and Services im Bereich Data Warehousing und Business Intelligence. Seit 2008 ist er bei Capgemini, Service Line Business Information Management, aktuell als Geschäftsbereichsmanager tätig. Er ist Mitgründer und war Sprecher des GI-Arbeitskreises »Konzepte des Data Warehousing«.

Andreas Bauer · Holger Günzel (Hrsg.)

Data-Warehouse-Systeme

Architektur · Entwicklung · Anwendung

4., überarbeitete und erweiterte Auflage

dpunkt.verlag

Andreas Bauer
bauer@data-warehouse-systeme.de

Holger Günzel
guenzel@data-warehouse-systeme.de

Lektorat: Christa Preisendanz
Copy-Editing: Annette Schwarz, Ditzingen
Satz: Andreas Bauer, Holger Günzel
Herstellung: Birgit Bäuerlein
Umschlaggestaltung: Helmut Kraus, www.exclam.de
Druck und Bindung: M.P. Media-Print Informationstechnologie GmbH, 33100 Paderborn

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

ISBN:
Buch 978-3-89864-785-4
PDF 978-3-86491-300-6
ePub 978-3-86491-301-3

4., überarbeitete und erweiterte Auflage 2013

Ringstraße 19 B
69115 Heidelberg

5 4 3 2 1 0

Vorwort

Mit der vierten Auflage geht das Buch in das zwölfte Jahr – einige Data-Warehouse-Systeme sind in diesem Zeitraum entstanden und auch bereits wieder durch andere ersetzt worden. Warum geht es diesem Buch nicht auch so? Wir denken, das liegt an mehreren Dingen:

- Das Buch hat einen stabilen »Kern« – die Referenzarchitektur. Alle Kapitel orientieren sich an dem Konstrukt – alle Abschnitte können sich »anlehnen« oder »reiben«.
- Das Buch wurde bereits vor zwölf Jahren aus einer Community heraus getrieben – viele Autoren haben sich mit ihrem Spezialgebiet zusammengetan, um gemeinsam einen »Standard« hervorzubringen.
- Das Buch wurde größtenteils zeitneutral und idealtypisch geschrieben. Es wurde auf explizite Spezifika und Bezeichnungen von Herstellern und Dienstleistern verzichtet.

Natürlich muss die Frage erlaubt sein – wie bereits in der letzten Auflage –, ob das Thema überhaupt noch eine Relevanz besitzt. Ist Data Warehousing nicht obsolet oder gar ein »Commodity-Produkt«, also etwas, über das man sich keine Gedanken machen muss? Wir können das aus tiefster Überzeugung verneinen: Interessanterweise kommen immer neue Einsatzgebiete hinzu, die eine derartige Dateninfrastrukturplattform wie das Data-Warehouse-System einsetzen. Aktuelle Stichworte, die erst in ein paar Jahren großflächig zum Einsatz kommen werden, sind beispielsweise »Embedded Analytics« oder »Identity and Access Intelligence« ([Rayn10], [KrBl11]). Wiederum positiv überrascht hat uns das große Interesse an der Weiterentwicklung des Buches – sowohl von bestehenden als auch von neuen Autoren.

Was hat sich geändert zur 3. Auflage? Eine grundsätzliche Veränderung liegt in der Weiterentwicklung der Referenzarchitektur: Die aktuelle Auflage verzichtet vollständig auf den Begriff des »Data Warehouse«, da sich gezeigt hat, dass dieser Begriff immer zu Missverständnissen führt. In der vorliegenden Auflage wird entweder über das gesamte System gesprochen (Data-Warehouse-System)

oder über dessen funktionale und Datenhaltungskomponenten, die jetzt eindeutige Namen, abgeleitet von deren Aufgabe, besitzen.

Weiterhin wurde an vielen Details gefeilt und Erweiterungen wurden vorgenommen: Data Mining, Datenschutz, Integration von unstrukturierten Daten, neue Technologien wie InMemory, Aspekte des Projektmanagements, Reifegradmodell, Open-Source-Software sowie Ergänzungen im Vorgehensmodell wie Anforderungs- und Testmanagement oder organisatorische Aspekte wie BICC.

Unser Dank gilt wie in jeder Auflage den bestehenden und neu gewonnenen Autoren, die Sie zahlreich im Autorenverzeichnis aufgeführt finden und kontaktieren können. Unser besonderer Dank bei dieser Auflage geht wieder an Thomas Zeh, der durch seine kritischen Anmerkungen und inhaltlichen Beiträge das Buch vorangetrieben hat.

Andreas Bauer und *Holger Günzel*
München, Februar 2013

Vorwort zu 3. Auflage

Vier weitere Jahre sind seit dem Erscheinen der 2. Auflage vergangen. In Zeiten des Internets ist das ein Zeitraum, in dem das Wissen – vor allem über Informationstechnologie – oftmals vollständig veraltet. Auch im Bereich der Data-Warehouse-Systeme ist die Zeit nicht stehen geblieben. Eine gute Gelegenheit für einen kleinen (unvollständigen und subjektiven) Rückblick:

Der Markt der Data-Warehouse-Werkzeuge hat sich konsolidiert. Zwischenzeitlich existieren zahlreiche ETL-Anbieter nicht mehr, sie wurden aufgekauft oder sind aus anderen Gründen verschwunden. Weiterhin war gerade bei den Anbietern von Analysewerkzeugen ein Trend hin zum Vollsortimenter zu verzeichnen. Die Anbieter haben ihr Produktportfolio erweitert, um alle relevanten Bereiche wie ETL, Data Quality sowie Analyse, Reporting, Planung oder Prognose (oft auch anzutreffen unter dem Schlagwort »Business Performance Management«) abzudecken. Im letzten Jahr wurden schließlich einige der führenden Produkthersteller von den großen Anbietern IBM, Microsoft, Oracle oder SAP übernommen. Es ist ungewiss, wie sich der Markt in diesem Bereich weiter entwickelt.

Neben dem Trend zur Konsolidierung der Anbieter wächst die Bedeutung der Open-Source-Werkzeuge. Neben Einzelwerkzeugen für ETL, Analyse oder Datenhaltung sind auch Komplettlösungen (Frameworks) im Kommen. Damit erwächst wie in anderen Bereichen auch eine Konkurrenz zu den kommerziellen Anbietern mit allen Vor- und Nachteilen. Aus Gründen der (weitgehenden) Produktneutralität des Buches wird im Weiteren nicht näher darauf eingegangen.

Im Zuge der Aktualisierung des Buches entstand die Diskussion, ob der Begriff »Business Intelligence« den Begriff »Data Warehouse« bzw. »Data Warehousing« ablösen wird oder sogar bereits abgelöst hat. In der einschlägigen Fachpresse stößt man allerorts auf »Business Intelligence«. Wir haben uns in der Autorenschaft bewusst dagegen entschieden, das Buch umzutitulieren, da im Begriff »Business Intelligence« mehr steckt und er darüber hinaus eine andere Ausrichtung besitzt. Unter Business Intelligence wird der Bereich der analytischen Anwendungen im Unternehmensumfeld verstanden. Hierzu zählen häufig auch weiterführende Anwendungsgebiete wie beispielsweise Planung und Wissensmanagement. Die besagten Anwendungsgebiete benötigen zwar wiederum ein Data Warehouse als gemeinsame Datenbasis. Der Fokus des Buches liegt aber genau auf diesem Data Warehouse, das die Grundlage für verschiedene Anwendungen bildet.

Vor diesem Hintergrund ist festzustellen, dass sich der Hype um das Thema Data Warehouse in den letzten Jahren zwar weiter abgeflacht bzw. auf angrenzende Themen wie Business Performance Management verlagert hat, aber in der Praxis nicht an Bedeutung verloren hat. Viele Unternehmen haben das Data Warehouse (und natürlich auch Business Intelligence) als ein zentrales Thema identifiziert, vielerorts wird bereits eine Konsolidierung der gewachsenen Data-

Warehouse-Landschaft durchgeführt. Auch neue Tendenzen wie »Serviceorientierte Architekturen« (SOA) führen indirekt wieder zu einer Rückbesinnung auf diese Mechanismen.

Einige Neuerungen im Buch betreffen auch gerade diese aktuellen Markttendenzen bzw. Anforderungen. Die systematische Istanalyse und Weiterentwicklung des Themas Data Warehouse in einem Unternehmen werden durch Reifegradmodelle unterstützt. In vielen Anwendungsbereichen werden immer kürzere Aktualisierungszyklen gefordert, was das Thema »Realtime Data Warehouse« adressiert. Die neu aufgenommenen bzw. grundlegend aktualisierten Praxisbeispiele vermitteln einen Überblick über aktuelle Einsatzformen von Data-Warehouse-Systemen.

Das Data-Warehouse-System gehört also inzwischen zum festen Bestandteil im Unternehmen bzw. der Organisation, »neudeutsch« würde man dies wohl als »Commodity« bezeichnen. Passend zu dieser Entwicklung hat sich der Arbeitskreis »Konzepte des Data Warehouse« in der Gesellschaft für Informatik aufgelöst. Nichtsdestotrotz gibt es weiter aktiv Interessierte an dem Thema, die sich in anderen Communities wieder zusammengetan haben. Es war interessant zu erleben, wie viele ehemalige und auch neue Autoren sofort für die 3. Auflage zugesagt haben, wenn auch die Umsetzung eines solchen Vorhabens neben der alltäglichen Arbeit schwierig zu bewerkstelligen ist.

Analog zur 2. Auflage gilt unser besonderer Dank den Autoren und Koordinatoren, die vor etlichen Jahren die Weitsicht und den Einblick in ein Thema hatten, um diesen lang anhaltenden Erfolg zu erzielen. Im Besonderen sind dies die Autoren der 3. Auflage: Jens Albrecht, Carsten Bange, Wolfgang Behme, Carsten Dittmar, Heiko Gronwald, Otto Görlich, Holger Heinze, Claudio Jossen, Christian Koncilia, Achim Langner, Stefan Mueck, Roland Pieringer, Torsten Priebe, Christoph Quix, André Scholz, Steffen Stock, Andreas Totok, Hermann Völlinger, Mirjam Wedler und Thomas Zeh. Ohne ihren Einsatz neben der täglichen beruflichen Herausforderung wäre auch diese Auflage nicht möglich gewesen. Herzlichen Dank an Euch alle! Einen treuen Mitstreiter, der uns auch durch diese Auflage mit Engagement, Ideen und Optimierungsvorschlägen begleitet hat, wollen wir hier hervorheben: Thomas Zeh.

Andreas Bauer und *Holger Günzel*
Erlangen/München, September 2008

Vorwort zur 2. Auflage

Auch drei Jahre nach der ersten Auflage dieses Buches hat sich wenig an der Bedeutung und Aktualität des Themas »Data-Warehouse-Systeme« geändert. Daten werden noch immer auf unterschiedlichsten Datenbanken redundant und unbeabsichtigt verteilt in einem Unternehmen gehalten, die Datenqualität ist ungenügend, und eine Analyse dieser Daten soll immer noch beliebig schnell möglich sein. Diese Situation wurde durch das steigende Datenvolumen eher noch verschärft.

Die weiterhin bestehende Aktualität des Themas, verbunden mit den konstruktiven und positiven Reaktionen zur ersten Auflage des Buches, hat uns bewogen, eine zweite Auflage herauszugeben. Die damalige Entscheidung für ein Grundlagenbuch mit einer Abstimmung unter vielen Experten des Themengebietes hat sich als richtig erwiesen, da die beschriebenen Konzepte weiterhin Bestand haben. Positiv, aber nicht ohne manche Kritik, fiel der Verzicht bzw. die kritische Betrachtung von »blumigen« Begrifflichkeiten mancher Software- und Hardwarehersteller und Beratungsfirmen auf. Die Diskussion von Firmenspezifika wurde deshalb bewusst vermieden und auf kurzlebigere Beschreibungen, wie sie häufig im Internet zu finden sind, verwiesen.

Nichtsdestotrotz gibt es einige Neuerungen, die in dieser Auflage erwähnt oder diskutiert werden. Neben den vielen Änderungen, die aus Anregungen der Leserschaft entstanden sind, wurden vor allem neue technologische Trends aufgegriffen und der Bereich der Methodik verfeinert. Größere Änderungen liegen deshalb im Anwendungsteil, für den mit der strikten Trennung zwischen Methodik und Projekt eine geeignetere Struktur gefunden wurde.

Auch in der zweiten Auflage sei nochmals den Autoren und Koordinatoren gedankt, die eine perfekte Grundlage für dieses Buch geschaffen haben. Weiterhin möchten wir den Autoren und Unterstützern der zweiten Auflage danken, die durch ihre Mitarbeit das Buch erst möglich gemacht haben. An dieser Stelle sind Wolfgang Behme, Holger Hinrichs, Wolfgang Hümmer, Christian Koncilia, Jürgen Meister, Martin Rohde und Thomas Zeh persönlich zu nennen.

Andreas Bauer und *Holger Günzel*
Erlangen/Nürnberg, Juni 2004

Vorwort zur 1. Auflage

Ein weiteres Data-Warehouse-Buch? Es gibt ein Buch über die Data-Warehouse-Architektur, ein anderes über Data-Warehouse-Entwicklung, ein weiteres ist ein Erfahrungsbericht. Es sind somit schon viele Bücher zum Thema Data Warehouse auf dem Markt, nur fehlt ein Buch aus der Datenbanksichtweise. Eine zusätzliche Motivation zu einem neuen Buch liegt darin begründet, dass das Themengebiet nicht nur unter technischen Gesichtspunkten, sondern gleichzeitig auch aus der Anwendungssicht heraus betrachtet werden muss. Eine Integration dieser beiden Seiten ist aber nur möglich, wenn einheitliche Begriffsdefinitionen und Termini geschaffen werden. Das Buch bietet durch diesen Informationsgehalt und die konsolidierten Begriffe eine ideale Basis für Fachleute aus der Entwicklung, dem Consulting und der Anwendung.

Die Grundgedanken zu diesem Buch sind in den Diskussionsrunden des Arbeitskreises »Konzepte des Data Warehousing« der Gesellschaft für Informatik (GI) entstanden. Er ist dem Fachbereich 2.5.1 (Datenbanksysteme) zugeordnet, wurde im Frühjahr 1999 als Treffpunkt von Forschung, Anwendung und Industrie gegründet und bietet seitdem die Möglichkeit des Austausches und der Diskussion über das Themengebiet »Data Warehousing«.

Am Anfang dieses Buchprojekts im Rahmen des Arbeitskreises wurden die Ziele hoch gesteckt; viele haben diese schlichtweg als unmöglich bezeichnet: Das Buch soll ein wissenschaftliches Standardwerk werden, das aber trotzdem in der Praxis verwendbar ist. Das Buch wird von nahezu 50 Autoren und Autorinnen geschrieben; es soll aber dennoch aus »einem Guss« erscheinen.

Bei der Diskussion über den Inhalt und vor allem über das Glossar stellte sich heraus, dass auch in unserem Arbeitskreis, den wir von der dort verwendeten Begrifflichkeit als homogen einschätzten, leicht unterschiedliche Begriffsdefinitionen benutzt wurden. Thomas Zeh, einer der Koordinatoren, hat einmal den Vergleich verwendet: »Dieses Buch ist ein Data Warehouse.« Über 50 Quellen mussten integriert und der Inhalt bereinigt werden, um das Ziel zu erreichen. Das Ziel war klar; der Weg aber keineswegs eindeutig vorgezeichnet. Die größte Herausforderung lag im Aufbau einer eindeutigen Begrifflichkeit und Abstimmung untereinander. Nachdem diese Herausforderung überwunden war, haben wir es dann geschafft, dass alle dasselbe Begriffsverständnis hatten und dieselben Bezeichner verwendeten.

Einige Hinweise zu Konventionen: Das Buch soll verständlich sein. Deshalb wurden so weit wie möglich deutsche Bezeichnungen verwendet und die englischen Bezeichner in Klammern gesetzt. Es wurde aber nicht zwanghaft nach einer deutschen Entsprechung gesucht. Außerdem wurde aus Gründen der Lesbarkeit die Verwendung der explizit femininen Form weggelassen. Natürlich sollen sich Frauen und Männer gleichermaßen angesprochen fühlen. Uns ist weiterhin bewusst, dass Internetadressen zwar interessant, aber meist nicht länger gültig sind, als bis das Buch im Druck ist. Aus Gründen der Allgemeingültigkeit haben

deshalb alle Autoren darauf weitgehend verzichtet. Wir haben uns auf einige exemplarische Angaben und Produkte beschränkt.

Unser ganzer Dank gilt allen beteiligten Autoren – auch denen, die aus Zeitmangel abspringen mussten –, ohne deren Wissen niemals dieses Buch entstanden wäre. Besonders hervorzuheben sind die Abschnittskoordinatoren Steffen Stock, Jens Albrecht, Wolfgang Hümmer und Thomas Zeh, die uns immer durch neue Kritik zu Verbesserungen des Werkes herausgefordert haben und uns durch ihre aktive Hilfe viel Arbeit abgenommen haben. In diesem Zusammenhang danken wir auch den Autoren, die durch mehrere Reviews zur Verbesserung des Inhaltes beigetragen haben. Unser besonderer Dank gilt Thomas Vetterli, der durch seine initiale Idee den Entwurf der Referenzarchitektur vorantrieb.

Weiterhin gibt es viele, die im Hintergrund dieses Projekts mitgewirkt haben, ohne die es aber nicht zustande gekommen wäre. Hierbei ist vor allem Frau Preisendanz vom dpunkt.verlag zu nennen. Sie war es, die an uns und dieses Projekt von Anfang an geglaubt hat. Für die konstruktive Kritik und Anmerkungen sei auch Frau Professor Gerti Kappel, Herrn Dr. Kai-Uwe Sattler und Frau Ursula Zimpfer gedankt. Last, but not least, dürfen alle diejenigen hilfreichen Geister, die sich um die Infrastruktur wie Mailverteiler oder gemeinsame Dokumentenablage (BSCW-Server) gekümmert haben, nicht vergessen werden. Ohne diese Hilfsmittel wäre dieses Werk nicht so reibungslos vonstatten gegangen. Außerdem bedanken wir uns – stellvertretend für alle Chefs – bei Professor H. Wedekind, der uns die Zeit gab, dieses Buch zu schreiben.

Andreas Bauer und *Holger Günzel*
Erlangen, Oktober 2000

Inhaltsverzeichnis

Teil I

Architektur

Koordinator:

Steffen Stock

Autoren:

C. Bange
A. Bauer
W. Behme
C. Dittmar
R. Düsing
H. Frietsch
M. Frisch
S. Gatziu
O. Görlich
H. Günzel
C. Heidsieck
H. Heinze
O. Herden
H. Hinrichs
W. Hümmer
C. Jossen
C. Pohl
T. Priebe
C. Quix
C. Sapia
H. Schinzer
S. Stock
J. Tako
P. Tomsich
A. Totok
A. Unterreitmayer
A. Vaduva
A. Vavouras
H. Völlinger
T. Zeh
K. Zimmermann

Der Begriff »Architektur«, eigentlich seit jeher im Bereich von Bauwerken geläufig, definiert die Struktur eines Gegenstands. Diese Struktur muss gleichzeitig drei Aufgaben übernehmen [Ency78]: Primär müssen die geforderten Anforderungen erfüllt werden, weiterhin muss sie ausreichend robust gegen Änderungen sein und noch eine gewisse »Ästhetik« aufweisen. Diese Struktur ist sowohl durch statische, strukturbildende als auch durch dynamische Aspekte, also durch das Zusammenspiel der statischen Anteile, geprägt.

Auch wenn diese Auffassungen über und Ansprüche an eine Architektur für andere Disziplinen verwirrend klingen, kann das auch auf Systeme in der Informationstechnologie angewendet werden. Das Data-Warehouse-System, von außen betrachtet ein monolithisches Informationssystem zur Auswertung von Daten, kann im Detail durch eine spezifische Architektur beschrieben werden. Auch das Data-Warehouse-System kann einerseits in einzelne, statische Komponenten zerlegt werden, die andererseits durch Datenflüsse verbunden sind und durch Kontrollflüsse (Dynamik) gesteuert werden.

Die Architektur von Data-Warehouse-Systemen dient deshalb im Teil I als Kommunikationsmittel zur strukturierten Einführung in das komplexe Themengebiet. Im Gegensatz dazu wird im Teil III die Architektur als Grundlage zum Aufbau und Pflege eines Data-Warehouse-Systems verwendet. Teil I beschreibt eine idealtypische Referenzarchitektur (Kap. 2), die in einer statischen Sichtweise durch mehrere Einzelkomponenten geprägt ist. Aus einer prozessorientierten Sichtweise wird in Kapitel 3 der Datenfluss von den Datenquellen bis hin zu den Auswertungskomponenten durch eine Zerlegung in mehrere Phasen dargestellt. Die dynamische Sichtweise ermöglicht die detaillierte Betrachtung des Zusammenspiels der einzelnen Komponenten, der nach dem eigentlichen Aufbau zu der dauerhaften Beladungs- und Auswertungstätigkeit im Data-Warehouse-System führt. Kapitel 4 rundet den Teil I durch die Untersuchung der physischen Architektur und der spezifischen Ausprägungen eines Data-Warehouse-Systems ab.

1 Abgrenzung und Einordnung

Im Bereich der auswertungsorientierten Informationssysteme gibt es nur wenige Begriffe, die seit den 90er Jahren häufiger und andauernder erwähnt und diskutiert wurden als der des Data-Warehouse-Systems. Viele Zeitungsartikel, Forschungsbeiträge und Produktinformationen propagieren zwar die Notwendigkeit eines Data-Warehouse-Systems, es geht aber selten eindeutig hervor, worin die Charakteristika und der Nutzen eines Data-Warehouse-Systems liegen. Die Verwendung des Begriffes ist derart vielseitig, dass es notwendig ist, nicht nur die Eigenschaften eines Data-Warehouse-Systems aufzuzeigen, sondern auch eine einheitliche Begriffsverwendung im Sprachgebrauch zu erreichen.[1]

Die Vielseitigkeit des Data-Warehouse-Begriffes resultiert aus zwei grundlegenden Bereichen, die diesen Begriff geprägt haben: Auf der technischen Seite stehen die Grundlagen der Datenintegrationsmöglichkeiten und Datenbanksysteme, auf der Anwendungsseite finden sich die betriebswirtschaftlichen, wissenschaftlichen und technischen Anforderungen aus einer Nutzungsperspektive. Weiterhin ist der Einfluss der Marktanalysten, Beratungshäuser und Softwarefirmen nicht zu gering zu erachten. Es ist somit ein Muss, diese oft gegensätzlichen Gebiete in diesem Buch gleichermaßen zu betrachten.

Eine Lösung dieses Dualismus von Informatik und Betriebswirtschaft kann nur in der Kombination liegen. Das Data-Warehouse-System wird deshalb als ein System aus Datenbanken und Komponenten gesehen, das aus der technischen Sicht Daten aus verschiedenen Datenquellen integriert und aus der betriebswirtschaftlichen Sicht dem Anwender diese Daten zu Auswertungszwecken zur Verfügung stellt.

Weiterhin soll an dieser Stelle auch angemerkt werden, dass der Begriff »Data-Warehouse-System« zunehmend durch den Begriff »Business Intelligence« ergänzt, überlagert oder oft auch ersetzt wird. Business Intelligence ist als Erweiterung zum Data Warehousing zu sehen, da insbesondere die Anwendungen sowie die anwendungsseitigen Prozesse darunter verstanden werden. Das Data-

1. Leider muss auch nach drei Auflagen dieses Buches festgehalten werden, dass sich an dieser Tatsache nur wenig geändert hat.

Warehouse-System dient weiterhin der Integration eines zentralen, auswertungsorientierten Datenbestandes.

In Abschnitt 1.1 werden für das weitere Verständnis wichtige Definitionen und Abgrenzungen zu verwandten Bereichen gegeben. In Abschnitt 1.2 wird die lange Historie des Themengebietes sowohl von Anwendungs- als auch von Datenbankseite skizziert. Nachfolgend wird der Begriff »Business Intelligence« aufgegriffen (Abschnitt 1.3), um sowohl aktuelle Tendenzen als auch die Fokussierung dieses Buches herauszuarbeiten. Im daran anschließenden Abschnitt folgt ein Überblick über die Vielfältigkeit der möglichen Einsatzbereiche eines Data-Warehouse-Systems. Abschnitt 1.5 umreißt abschließend den Inhalt des Buches und schafft mit einem speziellen Anwendungsbeispiel die Basis für ein durchgängiges Beispiel.

1.1 Begriffliche Einordnung

Zu den drei konventionellen Produktionsfaktoren Boden, Arbeit und Kapital wird immer häufiger die Information als vierte Säule hinzugenommen. Informationen basieren auf Daten, die entweder aus einem Unternehmen selbst stammen oder extern zugekauft werden. Die Tatsache, dass Daten eine besondere Bedeutung zukommt, ist aber nicht allein im betriebswirtschaftlichen Kontext zu finden, sondern gilt ebenso für statistische, wissenschaftliche oder technische Anwendungen.

Verschiedenen Informationssystemen ist gemein, dass Daten erfasst und verwaltet werden. Daten in einem Datenbanksystem zu erfassen, ist an sich nichts Neues. In jedem Unternehmen werden Personaldaten eingegeben oder Verkäufe durch Scannerkassen erfasst. Die Verarbeitung und Verwaltung der Daten geschieht in der Regel aber autonom unter Verantwortung der jeweiligen Abteilung. Interessant wird es erst, Daten aus autonomen Quellen zu vereinen. Dieser Vorgang ist besonders schwierig, wenn heterogene Daten unterschiedlichster Qualität, in verschiedenen Datenformaten, in heterogenen Datenmodellen und Datenbanksystemen gehalten werden.

Zur Vollständigkeit soll an dieser Stelle ein Exkurs zu verschiedenen Integrationsmöglichkeiten gegeben werden. Grundsätzlich wird eine *Backend-Integration* von einer *Frontend-Integration* unterschieden. Während bei der Frontend-Integration die Daten und Applikationen aus verschiedenen Systemen nur durch eine gemeinsame Oberfläche integriert werden (Stichwort: Portal, [Lint00b]), werden bei der Backend-Integration entweder die Daten physisch oder virtuell integriert oder Applikationen über eine gemeinsame Schnittstelle zusammengebracht (Stichwort: Enterprise Application Integration (EAI), [Lint00a]). Das Data-Warehouse-System kann in dieser Klassifikation der Backend-Integration zugeordnet werden. Eine pauschale Bewertung kann leider nicht erfolgen, da die Vorteile jeder Integrationsart von dem jeweiligen Zweck und der Strategie abhängen. Anzustreben ist grundsätzlich eine möglichst frühzeitige Integration der Daten und Anwendungen, was jedoch immer einen mitunter beträchtlichen Aufwand nach sich zieht.

Ein Data-Warehouse-System ist aber nicht nur von diesem integrativen Aspekt geprägt, sondern zusätzlich vom Aspekt der Auswertung. Die Verwendung von Daten in operativen Anwendungen war lange Zeit geprägt von einer *transaktionalen* Verarbeitung mit vielen kurzen Lese- und Schreiboperationen. Im Gegensatz dazu steht beim Data-Warehouse-System eine eher vergleichende oder auswertende Verwendung der Daten im Vordergrund, bei der auf große Datenmengen lesend zugegriffen wird.

Einige Fragen müssen jetzt erlaubt sein: Was ist eigentlich ein Data-Warehouse-System und was zeichnet es aus? Was bedeutet der Begriff Data-Warehouse-System? Ist ein Data-Warehouse-System eine integrierte Datenbank oder eine Datenbasis zu Auswertungszwecken? Wo liegen die Gemeinsamkeiten der Einsatzbereiche? Die vielen Interpretationsmöglichkeiten machen es notwendig, einige Begriffe zu definieren.

1.1.1 Definitionen

Die Tatsache, dass dieses Themengebiet sowohl von der Anwendungsseite als auch der Informatikseite durch eigene Fachtermini geprägt ist, impliziert ein unterschiedliches Begriffsverständnis. Verschiedene Normungsgremien versuchen, diese Begriffe zu standardisieren. Diese Bestrebungen waren aber bislang wenig erfolgreich.

Eine der ersten Definitionen aus dem Umfeld des Begriffes Data-Warehouse-System wurde von Inmon geprägt:

> *»A data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions.« [Inmo96]*

Ein »Data Warehouse« hat seiner Ansicht nach also vier Eigenschaften, die alle der Entscheidungsunterstützung dienen. Die Eigenschaften sollen hier kurz skizziert werden:

- *Fachorientierung* (engl. subject orientation):
  Der Zweck der Datenbasis liegt nicht mehr auf der Erfüllung einer Aufgabe wie z.B. der Lohn- und Gehaltsabrechnung, sondern in der Möglichkeit, ganze Themenbereiche wie Produkte und Kunden auszuwerten.
- *Integrierte Datenbasis* (engl. integration):
  Die Datenverarbeitung findet auf integrierten Daten aus mehreren Datenbanken statt.
- *Nicht flüchtige Datenbasis* (engl. non-volatile):
  Die Datenbasis ist als stabil zu betrachten. Daten, die einmal in das Data-Warehouse-System eingebracht wurden, werden nicht mehr entfernt oder geändert.

- *Historische Daten* (engl. time variance):
  Die Verarbeitung der Daten ist so angelegt, dass vor allem Vergleiche über die Zeit stattfinden. Es ist dazu unumgänglich, Daten über einen längeren Zeitraum zu halten.

Diese Definition ist einerseits nicht aussagekräftig genug, um sie in der Praxis oder der Theorie verwenden zu können, andererseits ist sie so einschränkend, dass viele Anwendungsgebiete und Ansätze herausfallen. Eine neue Definition ist notwendig, um dieses Manko zu überwinden:

> *»Ein Data-Warehouse-System ist ein physisches Informationssystem, das eine integrierte Sicht auf beliebige Daten zu Auswertungszwecken ermöglicht.«*

Aus der vermeintlich trivialen Forderung nach einer »physischen Integration zu Auswertungszwecken« entstehen Fragestellungen wie die Integration von Schemata und Daten aus unterschiedlichen Quellen. Diese Thematik ist zwar u.a. auch in *föderierten Datenbanksystemen* [Conr97] anzutreffen. Im Unterschied zu diesen bestehen zusätzliche Forderungen nach der *physischen* Integration und dem *Auswertungsaspekt*, der Kenntnisse und Denkweisen des Nutzers mit einbezieht.

Häufig wird diese Anforderung durch ein *multidimensionales Modell* [KRRT98] erreicht, das die Denkweise des Anwenders in *Dimensionen* und *Klassifikationshierarchien* widerspiegelt. Das multidimensionale Modell stellt im Gegensatz zu anderen Modellen besondere Strukturen und Auswertungsmöglichkeiten zur Verfügung, die schon bei der Modellierung einen Auswertungskontext schaffen.

Ein in diesem Zusammenhang wichtiger Treiber ist das *Online Analytical Processing* (OLAP, [CoCS93]), das eine explorative, interaktive Datenauswertung auf der Grundlage des konzeptuellen multidimensionalen Datenmodells darstellt. Weiterhin fällt oft das Stichwort *Data Mining*, darunter versteht man eine Suche nach unbekannten Mustern oder Beziehungen im Datenbestand des Data-Warehouse-Systems (Abschnitt 3.5.3). Obwohl ein Data-Warehouse-System keine notwendige Voraussetzung für Data Mining darstellt, kann ein Data-Warehouse-System als Ausgangspunkt für Data Mining verwendet werden.

Ein weiterer Unterschied eines Data-Warehouse-Systems gegenüber einem anderen Informationssystem liegt darin, dass die Daten in der Regel *nicht modifiziert* werden. Daten, die einmal in das Data-Warehouse-System übernommen wurden, dürfen nicht mehr verändert werden. Es können aber neue Daten in das Data-Warehouse-System aufgenommen werden, ohne die bereits vorhandenen zu überschreiben.

Da eine Datenbankinstanz diese Eigenschaften in der Regel alleine nicht zur Verfügung stellen kann, werden mehrere Datenbanken mit spezifischen Verwendungszwecken für ein Data-Warehouse-System benötigt. Weiterhin umfasst das Data-Warehouse-System alle für die Integration und Auswertung notwendigen