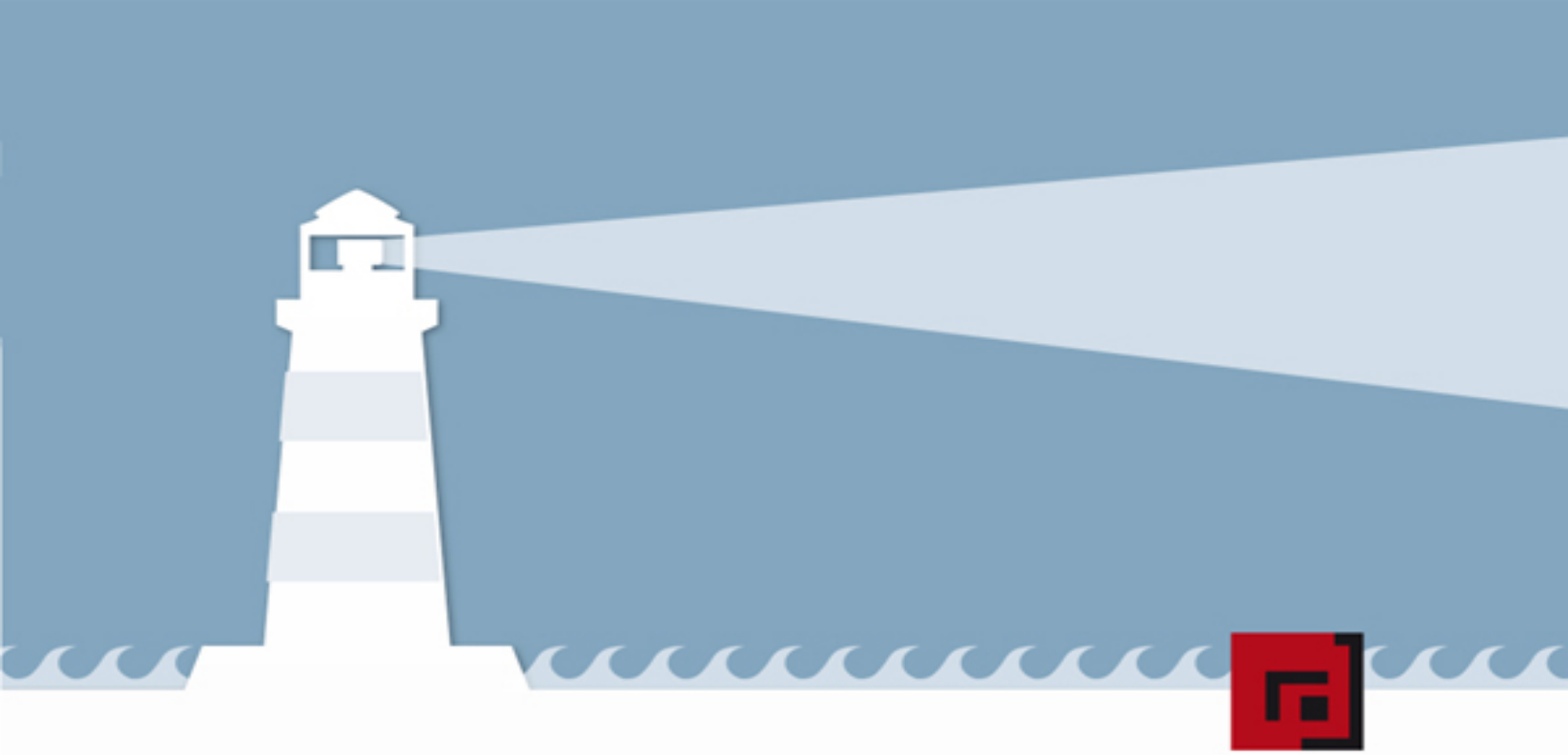


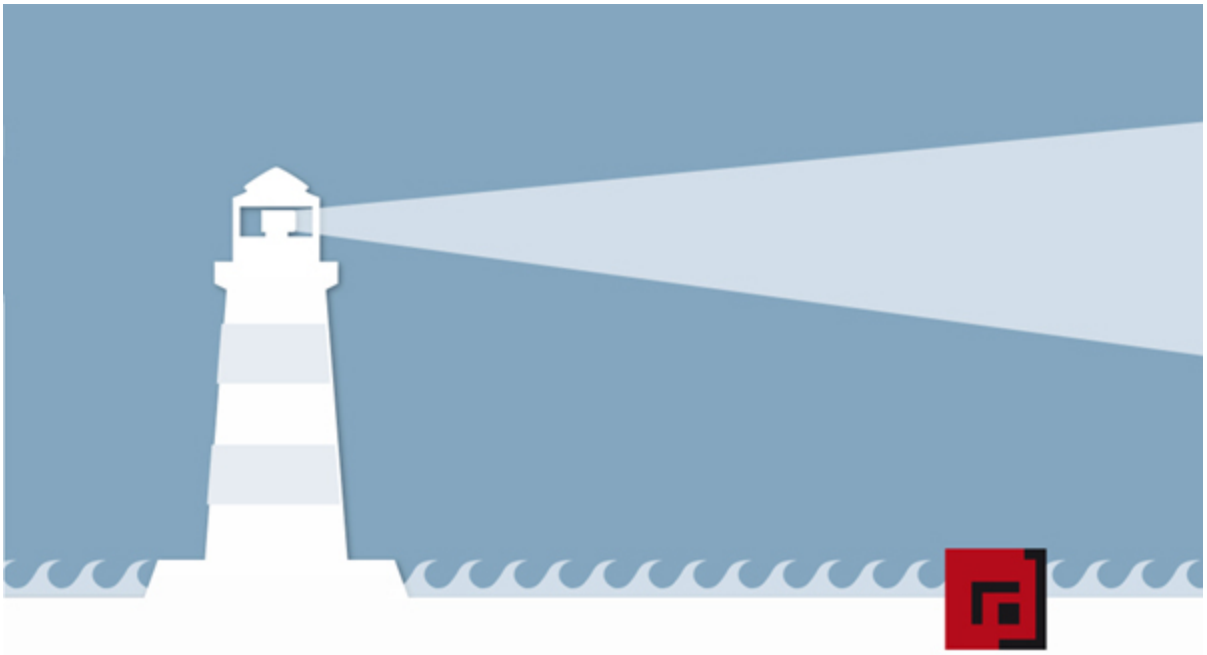
Michael Hahne

Modellierung von Business-Intelligence-Systemen

Leitfaden für erfolgreiche Projekte auf Basis flexibler Data-Warehouse-Architekturen

tdwi EUROPE

dpunkt.verlag



Michael Hahne

Modellierung von Business-Intelligence-Systemen

Leitfaden für erfolgreiche Projekte auf Basis flexibler Data-Warehouse-Architekturen

tdwi EUROPE

dpunkt.verlag

Dr. Michael Hahne ist Geschäftsführender Gesellschafter der Hahne Consulting GmbH, einem auf Business-Intelligence-Architektur und -Strategie spezialisierten Beratungsunternehmen. Zuvor war er Vice President und Business Development Manager bei SAND Technology, einem internationalen Anbieter von intelligenter Software für das Informationsmanagement spezialisiert auf Lösungen für unternehmensweite und große Data Warehouses. Darüber hinaus hat er sieben Jahre als Geschäftsbereichsleiter und CTO bei der cundus AG, einem auf Business Intelligence fokussierten IT-Dienstleistungsunternehmen, gearbeitet. Er hat mehr als 10 Jahre Erfahrung in der Implementierung und Optimierung von Data-Warehouse-Lösungen.

Zu diesem Buch – sowie zu vielen weiteren dpunkt.büchern – können Sie auch das entsprechende E-Book im PDF-Format herunterladen. Werden Sie dazu einfach Mitglied bei dpunkt.plus+:

www.dpunkt.de/plus

Modellierung von Business-Intelligence-Systemen

Leitfaden für erfolgreiche Projekte auf Basis flexibler Data-Warehouse-Architekturen

Edition TDWI

Michael Hahne

Dr. Michael Hahne
E-Mail: michael@hahneconsulting.de

Lektorat: Christa Preisendanz
Fachlektorat: Prof. Peter Gluchowski, Technische Universität Chemnitz
Copy-Editing: Ursula Zimpfer, Herrenberg
Herstellung: Frank Heidt
Umschlaggestaltung: Anna Diechtierow, Heidelberg
Druck und Bindung: M.P. Media-Print Informationstechnologie GmbH, 33100 Paderborn

Fachliche Beratung und Herausgabe von dpunkt.büchern in der Edition TDWI: Marcus Pilz · Marcus.Pilz@pilmar.com

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 978-3-89864-827-1
PDF 978-3-86491-523-9
ePub 978-3-86491-524-6

1. Auflage 2014
Copyright © 2014 dpunkt.verlag GmbH
Wieblinger Weg 17
69123 Heidelberg

Die vorliegende Publikation ist urheberrechtlich geschützt. Alle Rechte vorbehalten. Die Verwendung der Texte und Abbildungen, auch auszugsweise, ist ohne die schriftliche Zustimmung des Verlags urheberrechtswidrig und daher strafbar. Dies gilt insbesondere für die Vervielfältigung, Übersetzung oder die Verwendung in elektronischen Systemen.

Es wird darauf hingewiesen, dass die im Buch verwendeten Soft- und Hardware-Bezeichnungen sowie Markennamen und Produktbezeichnungen der jeweiligen Firmen im Allgemeinen warenzeichen-, marken- oder patentrechtlichem Schutz unterliegen.

Alle Angaben und Programme in diesem Buch wurden mit größter Sorgfalt kontrolliert. Weder Autor noch Verlag können jedoch für Schäden haftbar gemacht werden, die in Zusammenhang mit der Verwendung dieses Buches stehen.

5 4 3 2 1 0

Geleitwort

Bereits seit den Anfängen der Diskussion um die Ausgestaltung einer Data-Warehouse-Landschaft sowie die zugehörige multidimensionale Sichtweise auf betriebswirtschaftliches Datenmaterial wird intensiv über geeignete Formen und Techniken der Datenmodellierung nachgedacht. Die gewählte Art der Ablage relevanter Daten bestimmt dabei nicht zuletzt einerseits die Flexibilität bei der Reaktion auf veränderte oder erweiterte Benutzeranforderungen und andererseits die Zugriffsmöglichkeiten für den Endanwender, vor allem im Hinblick auf Navigationsfunktionalität und Antwortzeitgeschwindigkeit. Allerdings lassen sich die etablierten Verfahren und Methoden der Datenmodellierung aus dem operativen Systemumfeld nur sehr eingeschränkt nutzen, da die entsprechenden Systeme eher auf die schnelle und konsistente Erfassung zahlreicher Einzeltransaktionen ausgerichtet sind und weniger auf die umfassende Auswertung großer Datenvolumina.

Vor diesem Hintergrund nimmt das Modellierungsthema auch bei den Aktivitäten des TDWI (The Data Warehousing Institute) Germany e.V. breiten Raum ein und manifestiert sich sowohl bei den großen TDWI-Konferenzen als auch im Rahmen des Seminarprogramms. Bei diesen Gelegenheiten zeigt sich stets ein ungebrochenes großes Interesse der Teilnehmer an den unterschiedlichen Modellierungsaspekten einer umfassenden Business-Intelligence-Architektur, nicht nur hinsichtlich der verschiedenen BI-Systemkomponenten, sondern auch bezüglich verschiedener Abstraktionsebenen der Modellierung.

Somit erscheint es nicht nur folgerichtig, sondern darüber hinaus sehr begrüßenswert, dass mein langjähriger Wegbegleiter und ein ausgewiesener Experte im BI-Sektor Dr. Michael Hahne sich der Modellierung von Business-Intelligence-Systemen umfassend angenommen hat, zumal eine derartige Abhandlung zumindest im deutschsprachigen Raum bislang nicht zu finden war. Da der Autor

selbst im Projektgeschäft tätig ist und daher ständig mit den Herausforderungen in der BI-Praxis konfrontiert wird, erweist sich das vorliegende Werk als hochgradig relevant für alle Mitarbeiter aus Anwenderunternehmen und Beratungshäusern, deren Aufgaben in der Konzeption und Implementierung tragfähiger BI-Lösungen liegen. Aber auch bei der Ausbildung von Studierenden in den Bereichen Informatik und vor allem Wirtschaftsinformatik leistet das vorliegende Buch wertvolle Unterstützung.

Inhaltlich deckt das Werk ein breites Spektrum unterschiedlicher Facetten der Modellierung von BI-Systemen ab. Ausgehend von den Grundbestandteilen multidimensionaler Datenmodelle reicht die Betrachtung von semantischen Modellierungstechniken (ADAPT) über die ROLAP-Datenmodellierung (Star- und Snowflake-Schema) mit Dimensions- und Faktentabellenmodellierung bis hin zu den aktuellen Data-Vault-Konzepten für das Core Data Warehouse. Ein eigenes Kapitel widmet sich überdies dem sehr interessanten Thema der Historisierung. Die durchgängige Unterlegung der Ausführungen mit gut nachvollziehbaren Anwendungsbeispielen fördert die Verständlichkeit und stiftet unmittelbaren Mehrwert.

Ich wünsche allen Lesern bei der Lektüre viel Vergnügen und bin sicher, dass sich hilfreiche Erkenntnisse schnell einstellen werden.

Prof. Dr. Peter Gluchowski
Herdecke, im April 2014

Vorwort

Seit vielen Jahren bietet TDWI (The Data Warehousing Institute) Germany e.V.[1] mit seinem Seminar- und Konferenzprogramm umfassende Weiterbildungsangebote im Bereich Business Intelligence und Data Warehousing. Als Gründungsmitglied war ich von Anfang an dabei und bin u. a. Referent der Seminare zu Themen der Gestaltung und Architektur von Business-Intelligence-(BI-)Systemen. Und immer kam die Frage bei meinen Seminarteilnehmern auf, welche Literaturempfehlungen es denn gibt, und der Wunsch nach einem deutschsprachigen Buch zur Modellierung von BI-Systemen trat immer deutlicher zutage. Als dann das TDWI-Buchprogramm unter der Herausgeberschaft von Marcus Pilz erfolgreich startete, wurde die Idee zum Projekt, dessen Ergebnis nun in Ihren Händen liegt.

Zum Aufbau des Buches

Das Buch startet in Kapitel 1 mit der Darstellung der verschiedenen Architekturvarianten von Business-Intelligence-Landschaften. Dabei erfolgt neben der vergleichenden Betrachtung der Ansätze von Kimball und Inmon auch eine Darstellung von mehrschichtigen Architekturen.

Die Darstellung des Wesens mehrdimensionaler Datenstrukturen, Grundparadigma von Data Marts, erfolgt in Kapitel 2. Gerade für OLAP-Anwendungen ist dies ganz essenziell, sind dies doch die Objekte, auf deren Basis Analyse und Reporting vornehmlich stattfinden. Dabei stehen neben den vier konstituierenden Strukturkomponenten der Kennzahlen, Dimensionen, Hierarchien und Regeln auch Aspekte der Historisierung im Fokus. Dies beantwortet die Frage »Was ist das?«. Im anschließenden dritten Kapitel steht die Gestaltung dieser mehrdimensionalen Datenstrukturen auf fachlich konzeptioneller Ebene im Vordergrund und berücksichtigt die Fragestellung »Wie kann es gestaltet werden?«.

Eine Option der Implementierung von mehrdimensionalen Datenstrukturen auf Basis relationaler Datenbanktechnologie ist mit dem Modellierungsansatz des Star-Schemas gegeben. Kapitel 4 betrachtet zunächst die Grundlagen dieses Modellierungsansatzes. In Kapitel 5 erfolgt die eingehende Diskussion der Berücksichtigung von Historisierungsanforderungen und temporalen Aspekten. Eine eingehende Diskussion der im Kontext der Modellierung von Dimensionen relevanten Varianten des Star-Schema-Ansatzes ist Gegenstand von Kapitel 6. Für die Gestaltung von Faktentabellen sind u. a. Themen der Additivität sowie von Bestands- und Flussgrößen relevant. Neben den Gestaltungsempfehlungen für die Modellierung von Faktentabellen ist dies Gegenstand der Betrachtung in Kapitel 7. Insbesondere aufgrund der Bedeutung in den Projekten nehmen Aspekte der Star-Schema-Modellierung einen großen Teil des Buches ein, denn sie beantworten die Frage »Wie kann es implementiert werden?«.

In der Perspektive einer ganzheitlichen Data-Warehouse-Architektur stehen die Data Marts zwar dem Endbenutzer am nächsten, aber diese müssen ja auch im Rahmen eines Datenintegrationsprozesses bewirtschaftet werden. Dies erfolgt oftmals auf Basis eines zentralen Repositoriums etwa in Form eines Core Data Warehouse. Kapitel 8 widmet sich der Gestaltung dieser sehr wichtigen zentralen Komponente in einer Data-Warehouse-Architektur, denn diese ist wesentlich für die Aufgaben der Integration und Harmonisierung von Daten unterschiedlichster Quellen verantwortlich. Leitgedanke ist dabei die Frage »Woher kommt es?«.

Der rote Faden des Buches ist über die mehrdimensionalen Datenstrukturen folgendermaßen festgelegt: Beschreibung des Wesens, der Methoden der fachlichen Gestaltung, der Übertragung auf die Ebene der Implementierung und schließlich der Aspekte der Bewirtschaftung. Das Buch bietet damit eine ganzheitliche Darstellung aller Facetten der Modellierung für Business-Intelligence-Systeme, stellt adäquate Methoden dar und präsentiert Gestaltungsempfehlungen.

Für wen ist das Buch

Das Buch wendet sich an Projektleiter und Projektmitarbeiter von Data-Warehouse-und Business-Intelligence-Projekten, an Mitarbeiter aus IT und Informationsmanagement sowie an interessierte Fach- und Führungskräfte aus allen Unternehmensbereichen.

Danksagung

Wer schon mal ein Buch geschrieben hat, weiß, dass dies harte Arbeit ist und sehr viel Zeit kostet. Dieses Buchprojekt hat mehrere Jahre gedauert. Dabei blieben andere Dinge auf der Strecke liegen und die Familie musste oft zurückstecken. Ich danke meiner Frau und meinen Kindern für ihr Verständnis und die mir geschaffenen Freiräume, ohne die dieses Projekt nicht zu bewerkstelligen gewesen wäre.

Wesentlich für die Abrundung der Inhalte und auch die Erweiterung meiner vielleicht auch mal eingefahrenen Sichtweise waren die vielen fachlichen Diskussionen in Seminaren, im Kreis der Veranstaltungen des TDWI und insbesondere mit meinem Fachlektor Prof. Dr. Peter Gluchowski, mit dem mich seit nunmehr zwei Jahrzehnten die gemeinsame Leidenschaft für die Themen Data Warehousing und Business Intelligence verbindet. Viele gemeinsame Projekte und Veröffentlichungen haben sich dabei als sehr fruchtbar für die Entwicklung des Themas herausgestellt.

Mein besonderer Dank gilt allen Mitarbeitern des dpunkt.verlags und an erster Stelle Christa Preisendanz, deren Geduld immer wieder auf eine harte Probe gestellt wurde und die trotz vieler Verzögerungen das Projekt tatkräftig unterstützt hat. Ihr ist zu verdanken, dass die tief in der technischen Wolke entstandenen Satzkonstruktionen wieder zu einem lesbaren und korrekten Deutsch gefunden haben.

Über Ihre Anregungen und Kommentare als Leser des Buches freue ich mich und wünsche Ihnen viel Vergnügen sowie hilfreiche Ideen und Anregungen bei der Lektüre.

Dr. Michael Hahne

Bretzenheim, im April 2014

Inhaltsverzeichnis

1 Business-Intelligence-Architektur

1.1 Data Warehouse

1.2 OLAP und mehrdimensionale Datenbanken

1.3 Architekturvarianten

1.3.1 Stove-Pipe-Ansatz

1.3.2 Data Marts mit abgestimmten Datenmodellen

1.3.3 Core Data Warehouse

1.3.4 Hub-and-Spoke-Architektur

1.3.5 Data-Mart-Busarchitektur nach Kimball

1.3.6 Corporate Information Factory nach Inmon

1.3.7 Architekturvergleich Kimball und Inmon

1.4 Schichtenmodell der BI-Architektur

1.4.1 Acquisition Layer

1.4.2 Integration Layer

1.4.3 Reporting Layer

1.4.4 Modellierung im Schichtenmodell

2 Mehrdimensionale Datenstrukturen

2.1 Datenmodelle und Datenmodellierung

2.2 Grundbestandteile mehrdimensionaler Datenstrukturen

2.3 Hierarchische Dimensionsstrukturen

2.3.1 Strukturlose Dimensionen

2.3.2 Balancierte Baumstrukturen

2.3.3 Balancierte Waldstrukturen

2.3.4 Unbalancierte Baum- und Waldstrukturen

2.3.5 Parallele Hierarchien
2.3.6 Heterarchien (Many-Many-Beziehungen)
2.3.7 Rekursive Hierarchien und bebuchbare Knoten
2.3.8 Hierarchieattribute
2.4 Kennzahlen und deren Berechnung
2.4.1 Kennzahlen und Kennzahlensysteme
2.4.2 Kennzahlen im mehrdimensionalen Modell
2.4.3 Additivitätseigenschaft
2.5 Historisierung und Zeitabhängigkeit

3 Semantische mehrdimensionale Modellierung
3.1 Methoden auf Basis der Entity-Relationship-Modellierung
3.1.1 Grundbestandteile der ER-Modellierung
3.1.2 Erweiterte ERM-Konstrukte
3.1.3 ER-basierte mehrdimensionale Modellierung
3.1.4 Mehrdimensionales ER-Modell (ME/R)
3.2 Mehrdimensionale Modellierung mit ADAPT
3.2.1 Dimensionsmodellierung in ADAPT
3.2.2 Varianten der Hierarchiemodellierung
3.2.3 Modellierung von Würfeln
3.3 T-ADAPT: Modellierung von Zeitabhängigkeit

4 Bestandteile und Varianten des Star-Schemas
4.1 Einfaches Star-Schema
4.1.1 Grundform des Star-Schemas
4.1.2 Abbildung von Kennzahlen und Kennzahlensystemen
4.1.3 Attribute in Dimensionen
4.2 Modellierung von Dimensionshierarchien
4.2.1 Flache Strukturen

4.2.2 Balancierte Baum- und Waldstrukturen
4.2.3 Unbalancierte Strukturen
4.2.4 Parallele Hierarchien
4.2.5 Anteilige Verrechnung und Heterarchien
4.3 Normalisierung von Dimensionen
4.4 Übergang von T-ADAPT zum logischen Modell
4.4.1 Transformation von Dimensionen
4.4.2 Abbildung von Attributen
4.4.3 Transformation von Scopes
4.4.4 Behandlung spezieller ADAPT-Varianten
4.5 Modellierung von Parent-Child-Hierarchien
4.5.1 Iterative Abfrage
4.5.2 Einstufige Rekursion
4.5.3 Mehrstufige Rekursion
4.5.4 Rekursives SQL
4.5.5 Brückentabellen

5 Historisierung und Zeitabhängigkeit im Data Warehouse

5.1 Historisierung im Star-Schema
5.1.1 Keine Historisierung bei Type 0 und Type 1
5.1.2 Type-3-Attribut-Paare
5.1.3 Versionen und Zeitstempelung für as is und as of
5.2 Bewegungsdatensicht in der Historisierung
5.2.1 As-posted-Type-2-Szenario
5.2.2 Snapshot-Verfahren
5.2.3 Vollständige Zeitstempelung plus as posted
5.2.4 Varianten für hybride Historisierung
5.3 Best Practices der Historisierung

5.4 Bitemporale Historisierung

6 Dimensionsmodellierung

6.1 Dimensionstabellen

6.1.1 Degenerierte Dimensionen

6.1.2 Housekeeping und technische Dimensionen

6.1.3 Große Dimensionen

6.1.4 Mehrsprachigkeit

6.1.5 Outrigger-Tabellen

6.2 Rollen von Dimensionen

6.3 Many-Many-Beziehungen

6.3.1 Heterarchien über Faktentabellen

6.3.2 Mehrwertige Dimensionen (multi valued dimensions)

6.3.3 Many-Many-Beziehungen über Dimensionen

6.3.4 Mehrwertige Attribute

6.4 Datum- und Zeitdimension

7 Faktenmodellierung

7.1 Kennzahlen und Kennzahlensysteme

7.2 Aggregate

7.3 Snowflake-Schema

7.4 Faktenlose Faktentabellen

7.5 Granularität

7.6 Additivität und berechnete Kennzahlen

7.6.1 Transaktionsfaktentabellen

7.6.2 Bestandsmodelle

7.6.3 Prozessmodelle

7.7 Abgeleitete Schemata

8 Core-Data-Warehouse-Modellierung

8.1 Aufgaben der Data-Warehouse-Komponenten

8.1.1 Datenintegrations-Framework

8.1.2 Aufgaben und Komponenten in Multi-Layer-Architekturen

8.1.3 Eignungskriterien für Methoden der Core-Data-Warehouse-Modellierung

8.2 Star-Schema-Modellierung im Core Data Warehouse

8.2.1 Granulare Star-Schemata im Core Data Warehouse

8.2.2 Bewertung dimensionaler Core-Data-Warehouse-Modelle

8.3 3NF-Modelle im Core Data Warehouse

8.3.1 Core-Data-Warehouse-Modellierung in 3NF

8.3.2 Historisierungsaspekte von 3NF-Modellen

8.3.3 Bewertung der 3NF-Modellierung im Core Data Warehouse

8.4 Data-Vault-Ansatz

8.4.1 Hub-Tabellen

8.4.2 Satellite-Tabellen

8.4.3 Link-Tabellen

8.4.4 Zeitstempel im Data Vault

8.4.5 Harmonisierung von fachlichen Schlüsseln

8.4.6 Agilität in Data-Vault-Modellen

8.4.7 Vorgehensweise zur Data-Vault-Gestaltung

8.4.8 Bewertung der Data-Vault-Methode

Anhang

A Abkürzungen

B Literaturverzeichnis

Index

1 Business-Intelligence-Architektur

Unter dem Sammelbegriff *Business Intelligence* werden Konzepte des Data Warehouse, OLAP und Data Mining diskutiert. Durch die zunehmend strategische Ausrichtung der Informationsverarbeitung erhalten diese Konzepte einen neuen Stellenwert in der Praxis. Unter dem Begriff *Analyseorientiertes Informationssystem* werden Systemlösungen im Bereich Business Intelligence mit der Ausrichtung an der Analyseanforderung zusammengefasst.

Der Fokus analyseorientierter Informationssysteme liegt in der zeitnahen Versorgung betrieblicher Entscheidungsträger mit relevanten Informationen zu Analysezwecken. Diese Systeme zielen somit auf die Unterstützung der dispositiven und strategischen Prozesse in einem Unternehmen ab und bilden damit ein logisches Pendant zu den operativen Systemen, die zumeist in Form einer integrierten betriebswirtschaftlichen Standardsoftware wie z. B. SAP ECC (ERP Central Component) eingesetzt werden.

1.1 Data Warehouse

Allen analyseorientierten Informationssystemen gemeinsam ist eine geeignete zugrunde liegende Datenbasis. Diese bildet damit eine wesentliche Komponente, auf deren Grundlage die verschiedenen Auswertungssysteme aufsetzen. Dem Aufbau dieser zentralen Datenbasis widmet sich die Diskussion seit einigen Jahren unter dem Stichwort Data Warehouse. Hierunter soll im Folgenden ein unternehmensweites Konzept verstanden werden, dessen Ziel es ist, eine logisch zentrale, einheitliche und konsistente Datenbasis für die vielfältigen Anwendungen zur Unterstützung der analytischen Aufgaben von Führungskräften aufzubauen, die losgelöst von den operativen Datenbanken betrieben wird.

Der Begriff *Data Warehouse* geht auf Inmon zurück. Inmon beschreibt ihn mit der Aufgabe, Daten zur Unterstützung von Managemententscheidungen bereitzustellen, die die folgenden vier wesentlichen Eigenschaften aufweisen [Inmon 1996, S. 33]:

- Themenorientierung
- Vereinheitlichung
- Zeitorientierung
- Beständigkeit

Die in einem Data Warehouse abzulegenden Daten orientieren sich an dem Informationsbedarf von Entscheidungsträgern und beziehen sich demnach auf Sachverhalte, die das Handeln und den Erfolg eines Unternehmens bestimmen. Die Daten fokussieren sich daher auf die Kernbereiche der Organisation. Diese datenorientierte Vorgehensweise unterscheidet sich deutlich von den prozessorientierten Konzepten der operativen Anwendungen.

Eine wesentliche Eigenschaft eines Data Warehouse ist ein konsistenter Datenbestand, der durch eine Vereinheitlichung der Daten vor der Übernahme entsteht. Diese Vereinheitlichung bezieht sich sowohl auf die Struktur wie auch auf die Formate, häufig müssen die verwendeten Begriffe, Codierungen und Maßeinheiten zusammengeführt werden.

Für die Managementunterstützung werden Daten benötigt, die die Entwicklung des Unternehmens über einen bestimmten Zeitraum repräsentieren und zur Erkennung und Untersuchung von Trends herangezogen werden. Dazu wird der Data-Warehouse-Datenbestand periodisch aktualisiert und der Zeitpunkt der letzten Aktualisierung definiert damit einen Schnappschuss des Unternehmensgeschehens, der je nach Ladezyklus Minuten, Stunden, Tage, Wochen oder Monate zurückliegen kann.

Der Begriff *Data Warehouse* beschreibt ein unternehmensweites Konzept, dessen Ziel die Bereitstellung einer einheitlichen konsistenten Datenbasis für die vielfältigen Anwendungen zur Unterstützung der analytischen Aufgaben von Fach- und

Führungskräften ist. Diese Datenbasis ist losgelöst von operativen Datenbanken zu betreiben.

Das vierte wesentliche Charakteristikum bezieht sich auf die Beständigkeit der Daten in einem Data Warehouse. Da diese in der Regel nur einmal geladen und danach nicht mehr geändert werden, erfolgt ein Datenzugriff im Allgemeinen nur lesend. Einmal erstellte Berichte auf Basis dieses Datenbestands sind daher reproduzierbar, da auch in späteren Perioden die Datenbasis die gleiche ist. Diese Eigenschaft wird mit dem Begriff der Nicht-Volatilität umschrieben.[1] Die Beständigkeit bezieht sich aber auch auf ein verlässliches annähernd gleichbleibendes Antwortzeitverhalten.

Die Einordnung eines Data Warehouse in die IT-Struktur eines Unternehmens ergibt sich aus der in Abbildung 1–1 dargestellten Referenzarchitektur. Ausgangsbasis dieser Architektur sind die operativen Vorsysteme, aus denen periodisch Datenextrakte generiert werden. Im Rahmen des ETL-Prozesses (*extract transform load*, ETL) erfolgen die Bereinigung und Transformation der Daten aus den verschiedenen Vorsystemen sowie externen Datenquellen zu einem konsistenten einheitlichen Datenbestand und der Transport in das Data Warehouse. Hierbei sind die beiden Phasen des erstmaligen Befüllens sowie der regelmäßigen periodischen Aktualisierungen zu unterscheiden.

Dieser ETL-Komponente kommt beim Aufbau eines Data Warehouse eine zentrale Bedeutung zu, denn ein hoher Anteil des Aufwands beim Aufbau eines Data Warehouse resultiert aus der Implementierung von Zugriffsstrategien auf die operativen Datenhaltungseinrichtungen.[2]

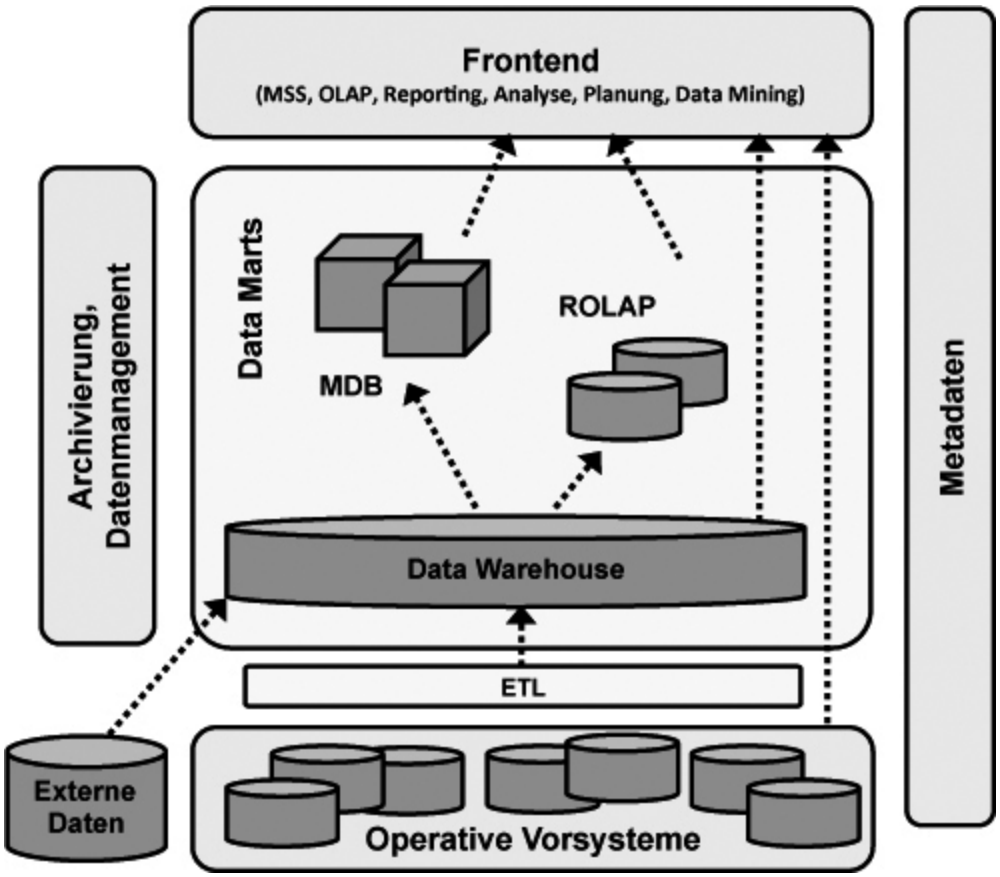


Abb. 1–1 *Data-Warehouse-Referenzarchitektur*

Aus diesem Datenbestand können des Weiteren kleinere funktions- oder bereichsbezogene Teilsichten in sogenannten *Data Marts* extrahiert werden. Diese müssen wiederum periodisch aus dem Data-Warehouse-Datenbestand aktualisiert werden. Für diese Teildatenbestände kommen im Allgemeinen sogenannte *OLAP-Datenbanken* zum Einsatz, deren Diskussion Gegenstand des nächsten Abschnittes ist.

Die Auswertung über die Frontend-Applikationen kann sowohl direkt auf dem zentralen Data Warehouse erfolgen als auch auf den einzelnen Data Marts aufsetzen. In Data-Warehouse-Konzepten können auch Applikationen wie beispielsweise Management-Support-Systeme auf diesen Datenbeständen basieren, d. h., die Datenbasis für diese Systeme kann auch in einem Data Warehouse liegen. Hier verbinden sich also bekannte Konzepte des Managementsupports mit dem neuen Konzept des Data Warehouse zu einer neuen

Systemkategorie. Eine weitere wesentliche Erweiterung ergibt sich aus dem Ansatz des Online Analytical Processing (OLAP), der im folgenden Abschnitt dargestellt wird.

1.2 OLAP und mehrdimensionale Datenbanken

Der Begriff OLAP beschreibt ein Leitbild für eine endanwenderorientierte Analysetechnik und wird häufig konträr zum sogenannten Online Transaction Processing (OLTP) gesehen. Online Analytical Processing (OLAP) ist ein mittlerweile anerkannter Bestandteil für eine angemessene DV-Unterstützung betrieblicher Fach- und Führungskräfte und bietet einen endanwenderorientierten Gestaltungsrahmen für den Aufbau von Systemen zur Unterstützung dispositiver bzw. analytischer Aufgaben [Gluchowski/Chamoni 2009, S. 197 ff.].

Als zentrales Charakteristikum gewährleisten multidimensionale Sichtweisen auf unternehmensinterne und -externe Datenbestände brauchbare Näherungen an das mentale Unternehmensbild des Managers. Betriebswirtschaftliche Variablen bzw. Kennzahlen (wie z. B. Umsatz oder Kostengrößen) werden entlang unterschiedlicher Dimensionen (wie z. B. Kunden, Artikel, Regionen) angeordnet, und diese Strukturierung gilt als geeignete entscheidungsorientierte Sichtweise auf betriebswirtschaftliches Zahlenmaterial. Bildlich gesprochen werden die quantitativen Kenngrößen in mehrdimensionalen Würfeln gespeichert, deren Kanten durch die einzelnen Dimensionen definiert und beschriftet sind.

OLAP soll es Benutzern ermöglichen, flexible komplexe betriebswirtschaftliche Analysen wie auch Ad-hoc-Auswertungen mit geringem Aufwand eigenständig durchführen zu können. Um dieses Ziel zu erreichen, wurden von Codd, Codd und Sally 12 Regeln als Anforderung an OLAP-Lösungen definiert:[3]

1. Die *mehrdimensionale konzeptionelle Sicht* auf die Daten wird als elementarstes Wesensmerkmal für OLAP postuliert. Diese

Darstellungsform ermöglicht eine Navigation in den Datenwürfeln mit beliebigen Projektionen und Verdichtungs- und Detaildarstellungen.

2. *Transparenz* beschreibt die nahtlose Integration in Benutzerumgebungen.
3. Eine offene Architektur gewährleistet *Zugriffsmöglichkeiten* auf heterogene Datenbasen, eingebunden in eine logische Gesamtsicht.
4. Ein *gleichbleibendes Antwortzeitverhalten*, selbst bei vielen Dimensionen und sehr großen Datenvolumina, ist ein wesentlicher Aspekt.
5. Postuliert wird auf Basis einer *Client-Server-Architektur* die Möglichkeit verteilter Datenhaltung sowie der verteilten Programmausführung.
6. Aufgrund der *generischen Dimensionalität* stimmen alle Dimensionen in ihren Verwendungsmöglichkeiten überein.
7. Betriebswirtschaftliche mehrdimensionale Modelle sind oft sehr gering besetzt. Das *dynamische Handling »dünnbesetzter Würfel«* ist elementar für eine optimale physikalische Datenspeicherung.
8. Unter *Mehrbenutzerfähigkeit* in OLAP-Systemen wird der gleichzeitige Zugriff verschiedener Benutzer auf die Analysedatenbestände, verbunden mit einem Sicherheits- und Berechtigungskonzept, verstanden.
9. Der Kennzahlenberechnung und Konsolidierung dienen *unbeschränkte dimensionsübergreifende Operationen* innerhalb einer vollständigen integrierten Datenmanipulationssprache.
10. Eine ergonomische Benutzerführung soll *intuitive Datenmanipulation* und Navigation im Datenraum ermöglichen.
11. Auf Basis des mehrdimensionalen Modells soll ein leichtes und *flexibles Berichtswesen* generiert werden können.
12. Die Forderung nach einer *unbegrenzten Anzahl an Dimensionen und Aggregationsebenen* ist in der Praxis schwer realisierbar.

Dieses Regelwerk ist nicht unumstritten und erfuhr verschiedene Erweiterungsvorschläge u. a. von der Gartner Group [Gartner 1995].[4] Eine etwas pragmatischere und technologiefreie Variante zur Definition der konstituierenden Charakteristika von OLAP stammt von Pendse und Creeth, die ihren Ansatz mit FASMI benennen [Pendse/Creeth 1995]:

1. *Fast*: Ganz konkret wird für das Antwortzeitverhalten ein Grenzwert von zwei Sekunden für Standardabfragen und 20 Sekunden für komplexe Analysen festgelegt.
2. *Analysis*: Benutzern muss es ohne detaillierte Programmierkenntnis möglich sein, analytische Berechnungen und Strukturuntersuchungen auf Basis definierter Verfahren und Techniken ad hoc zu formulieren.
3. *Shared*: Für den Mehrbenutzerbetrieb werden Berechtigungsmöglichkeiten bis auf Datenelementebene sowie Sperrmechanismen bei konkurrierenden Schreibzugriffen gefordert.
4. *Multidimensional*: Die mehrdimensionale Sichtweise ist ein elementares Wesensmerkmal analytischer Systeme.
5. *Information*: Für OLAP-Systeme ist die verwaltbare Informationsmenge bei stabilem Antwortzeitverhalten ein kritischer Bewertungsfaktor.

Verschiedene Ansätze zur Definition dessen, was OLAP ausmacht, resultieren in der Anforderung nach Vereinheitlichung und dem Setzen von Standards. Dieses hat sich der OLAP-Council zum Ziel gesetzt.[5] Diese Diskussion ist losgelöst von Implementierungsaspekten, für die es jedoch gleichermaßen verschiedenste Architekturansätze gibt.

Online Analytical Processing (OLAP) als Grundprinzip für den Aufbau von Systemen zur Unterstützung von Fach- und Führungskräften in ihren analytisch geprägten Aufgaben basiert im Kern auf einer mehrdimensionalen konzeptionellen Sicht auf die Daten mit

Möglichkeiten der Navigation in den Würfeln mit beliebigen Projektionen und auf verschiedenen Verdichtungsstufen.

1.3 Architekturvarianten

Die Architektur von BI-Systemen dient der Beschreibung der wesentlichen Komponenten mit ihren Eigenschaften und Funktionen sowie deren Beziehung untereinander. Dabei sind in der Praxis sehr unterschiedliche Formen und Ausgestaltungen anzutreffen, die nicht immer aufgrund proaktiver Entscheidungen etwa aus einer BIOrganisation heraus entstehen, sondern oft das Ergebnis historisch gewachsener Landschaften sind. Jedoch zeigt sich, dass eine saubere Architektur als Grundlage für die BI-Systeme in Unternehmen Vorteile für Entwicklung und Betrieb mit sich bringt. Dies drückt sich auch in der zunehmend bedeutenden Rolle des BI-Architekten aus.[6]

Bekannte Architekturvarianten unterscheiden sich deutlich hinsichtlich der Anzahl der Komponenten, der gesamten Komplexität, des Aufwands für Entwicklung und Betrieb sowie der Performance und Skalierbarkeit. Aber auch die Fähigkeit, effizient und agil mit neuen Anforderungen umzugehen und den Wandel zu unterstützen, ist ein wesentliches Erfolgskriterium für verschiedene Ansätze.[7]

1.3.1 Stove-Pipe-Ansatz

In den Fällen, in denen im Unternehmen keine übergreifenden Auswertungen erforderlich sind, kann eine dezentrale Architektur mit unabhängigen Data Marts wie in Abbildung 1–2 dargestellt durchaus sinnvoll sein. Bei diesem Ansatz, auch unter dem Namen »Stove Pipe« (Ofenrohr) bekannt [Kimball et al. 2008, S. 249], entstehen einzelne Silos bzw. Inseln, da die Daten für jeden Anwendungsbereich isoliert aus den Quellsystemen extrahiert und aufbereitet werden [Kemper et al. 2010, S. 22 f.]. Dabei erfolgt die Transformation der Daten redundant,

was auch zu entsprechenden Aufwänden und potenziellen Inkonsistenzen im Falle einer Änderung führt.

Diese Datensilos sind nur eingeschränkt in einem anderen Kontext nutzbar und bieten damit eine sehr schlechte Unterstützung für bereichsübergreifende Auswertungen. Für autonome Organisationseinheiten kann dieser Ansatz aufgrund der leichteren Berücksichtigung fachlicher Anforderungen jedoch durchaus geeignet sein [Sinz/Ulbrich-vom-Ende 2009, S. 189 f.]. Oftmals handelt es sich aber um eine rein historisch gewachsene Struktur, die den im Zeitablauf gestiegenen Anforderungen nicht mehr gerecht wird.

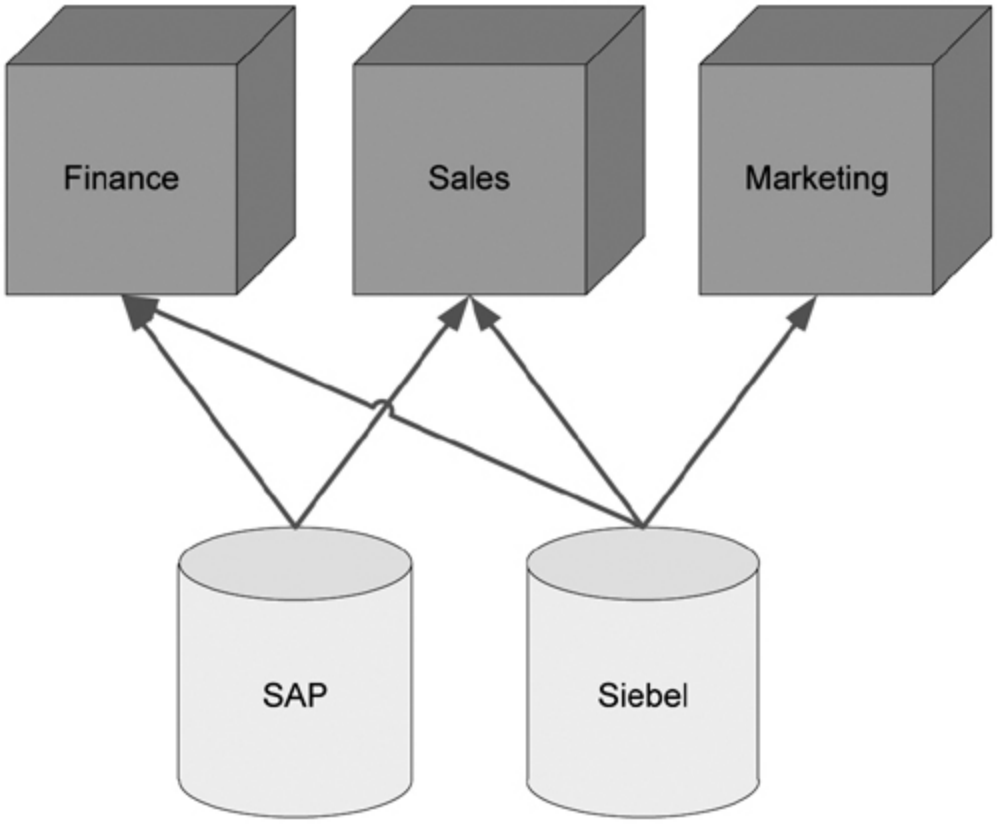
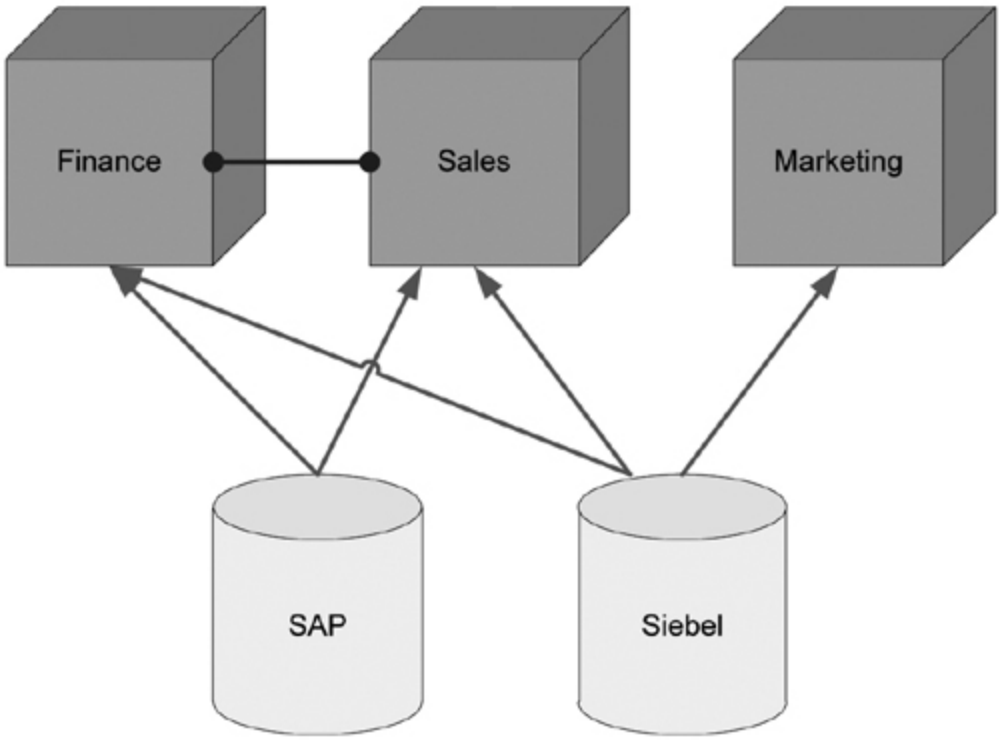


Abb. 1–2 *Unabhängige Data Marts*

Der Übergang von einer derartigen Struktur mit unabhängigen Data Marts zu einer besser geeigneten Architektur mit logischer Integration der Daten gestaltet sich im Allgemeinen recht schwierig, da es auch keine standardisierten bewährten Migrationskonzepte gibt. Dies wird auch durch empirische Untersuchungen untermauert.[8]

Der Stove-Pipe-Ansatz ist oftmals historisch bedingt und liefert keine Integration, sondern isolierte Data Marts.

1.3.2 Data Marts mit abgestimmten Datenmodellen

Eine erste Möglichkeit zur Entschärfung der Probleme des Stove-Pipe-Ansatzes besteht in der Abstimmung der Data-Mart-Datenmodelle bzw. Dimensionsstruktur (*conformed dimensions*, siehe Abb. 1–3). Diese erleichtern die Gewährleistung von Konsistenz und Integrität der dispositiven Daten. Neben den Dimensionen sind auch die Kennzahlen abgestimmt, man spricht hier von *conformed facts*.

Die Abstimmung und der Aufbau eines konsolidierten Datenbestands erfolgt dabei virtuell durch die Abstimmung und Koordination zwischen den Unternehmensbereichen ohne den Aufwand für dessen Entwicklung. Andererseits geht dies einher mit einem erhöhten Aufwand für die Abstimmung [Sinz/Ulbrich-vom-Ende 2009, S. 191].