Veit Köppen

Gunter Saake

Kai-Uwe Sattler

2. Auflage

Data Warehouse Technologien

1

Einführung in Data-Warehouse-Systeme

Die Verwaltung großer Datenbestände ist seit vielen Jahren im Bereich der Datenbanken angesiedelt. Datenbanksysteme ermöglichen eine integrierte Speicherung sehr großer Datenbestände. Zudem können mehrere Nutzer und Anwendungen gleichzeitig auf die Daten zugreifen. Im betrieblichen Umfeld haben sich für die spezifischen Anforderungen diverse Systeme hinsichtlich der konzeptionellen Ebene, aber auch technischer Implementierungen ergeben. Hierzu zählen beispielsweise Kunden- und Adressdatenbanken, Buchhaltungssysteme, Wissensdatenbanken, Lieferanten- und Produktkataloge und Prozessdatenbanken. Häufig kann es in einem Unternehmen daher vorkommen, dass in den einzelnen Abteilungen und Fachbereichen eine Vielzahl von Systemen eingesetzt wird.

Aus der Perspektive eines zentralisierten Managements, aber auch aus bereichsübergreifenden Anforderungen heraus ergibt sich die Notwendigkeit, diese Systeme in einer Systemlandschaft zu verbinden. Im Hinblick auf einen holistischen Entscheidungsansatz, wie er im Unternehmensmanagement erfolgen soll, bedeutet dies auch die Integration, Konsolidierung und die Aufbereitung des Datenbestandes. Aufgrund der stetigen Nutzung im operativen Geschäft sind jedoch Einschränkungen unerwünscht, die aber durch die komplexen Analyseanfragen auf den Datenbestand erfolgen müssen.

Als Ausweg für die Nichtbeeinträchtigung der operativen Systeme bietet sich hier eine redundante Datenhaltung an, d.h. Daten müssen aus den operativen Quellen in ein zentralisiertes System überführt werden. Zudem sind in diesem System insbesondere Fragen der Konsistenz zu berücksichtigen. Dieses zentrale Datenlager wird als *Data Warehouse* bezeichnet. Es stellt im Unter-

nehmen häufig den qualitativ hochwertigen Analysepunkt (*Single Version of Truth*) dar. Dies erfolgt neben der Zusammenführung der heterogenen Quelllandschaft auch durch Bereinigung und Transformationen der Daten. Das Data Warehouse ermöglicht dann komplexe Analysen, ohne das betriebliche Umfeld hinsichtlich der Datenbanken negativ zu beeinflussen. Zudem ermöglichen Optimierungen hinsichtlich der Analysen einen effizienteren Einsatz im Data Warehouse. Ziel dieses Buches ist die Vermittlung der wichtigsten Technologien für den Einsatz betrieblicher analytischer Informationssysteme, deren Grundlage das Data Warehouse legt.

1.1 Anwendungsszenario Getränkemarkt

Wir wollen uns im vorliegenden Buch einem durchgängigen Beispiel widmen, um die Probleme und Lösungen zu illustrieren. Das Szenario wird an einem fiktiven Getränkemarkt illustriert, der sich auf den Verkauf von Bier, Wein und Softdrinks spezialisiert hat. In Abbildung 1.1 ist unser Getränkemarkt schematisch abgebildet.

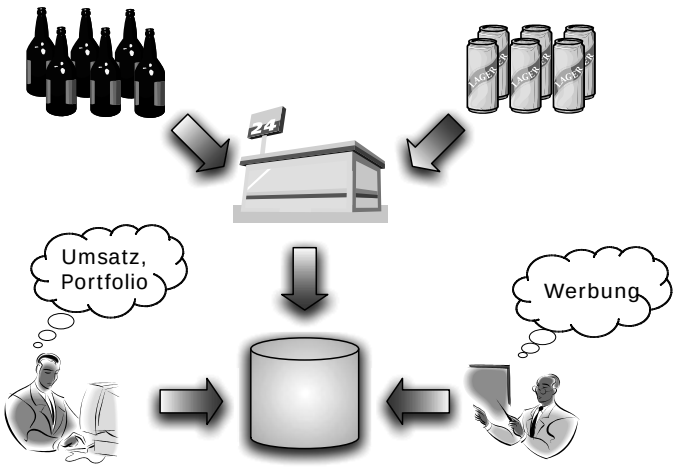

Abbildung 1.1: Szenario Getränkemarkt

Sowohl der Wein als auch das Bier werden von Lieferanten bzw. direkt von den Herstellern bezogen. Die Verkäufe im Getränkemarkt werden in einer Datenbank erfasst. Unser Getränkemarkt betreibt Filialen in verschiedenen Orten der Bundesländer Thüringen und Sachsen-Anhalt; dies ist in Abbildung 1.2 exemplarisch dargestellt.

Das Management des Getränkemarkts versucht die Gesamtunternehmensentwicklung anhand von Kennzahlen zu steuern. Zugleich werden spezia-

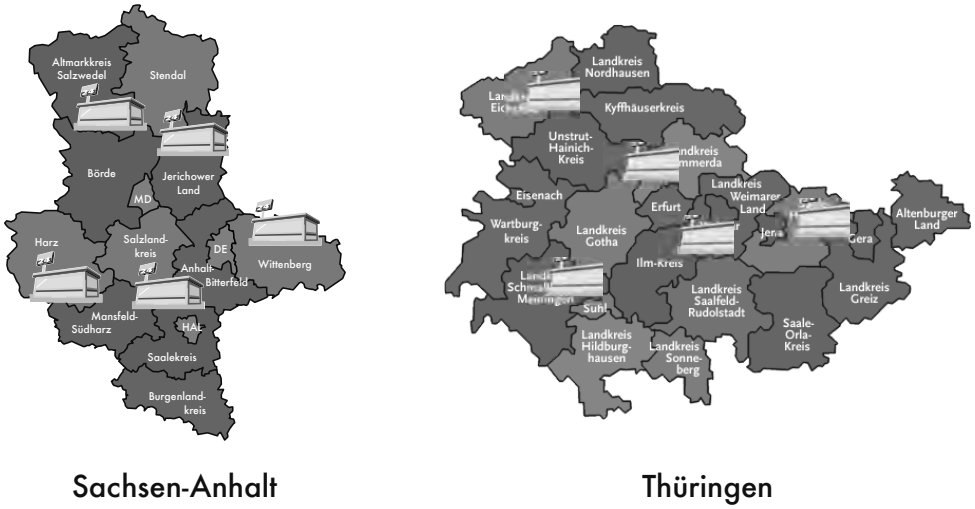

Abbildung 1.2: Szenario Getränkemarkt (Standortübersicht)

lisierte Analysen in den Fachabteilungen erforderlich. Hierzu gehört beispielsweise die Analyse des Kaufverhaltens für Marketingkampagnen.

Auf der Filialebene bzw. dem operativen Geschäft ergibt sich ein vereinfachtes Relationenschema, wie in Abbildung 1.3 dargestellt. Kunden können Produkte (Bier und Wein) kaufen. Die Lieferung der Produkte in den Getränkemarkt erfolgt durch Lieferanten. Beide Informationen werden in der Datenbank abgespeichert. Zudem wird pro Kundeneinkauf die verkaufte Menge in der Datenbank erfasst.

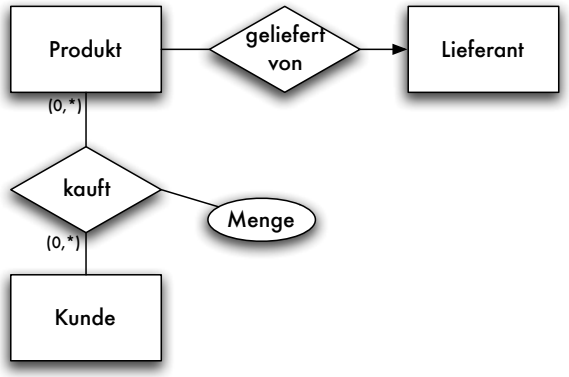

Abbildung 1.3: Datenbankschema Getränkemarkt

Typische Fragen, die innerhalb des betrieblichen Kontextes im Filialbetrieb auftreten, lauten beispielsweise:

- Wie viele Flaschen Bordeaux wurden letzten Monat verkauft?
- Wie hat sich der Verkauf von Rotwein im letzten Jahr entwickelt?
- Wer sind unsere Top-Kunden?
- Von welchem Lieferanten beziehen wir die meisten Kisten?

Zur Beantwortung dieser Fragen besteht aber auch die Notwendigkeit, externe Quellen einzubinden. So ist es erforderlich, externe Quellen wie Kundendatenbanken oder Lieferantenkataloge einzubeziehen. Zudem ist der zeitliche Bezug zu berücksichtigen. Dies ist im betrieblichen Kontext und den operativen Datenbanken jedoch meist unzureichend unterstützt.

Für Entscheidungen auf Führungsebene ist jedoch eine andere Perspektive notwendig. Hier kann nicht mehr nur allein eine einzelne Filiale betrachtet werden, sondern es müssen regionale Aspekte und ein ganzheitlicher Managementansatz in der Analyse unterstützt werden. Filialübergreifende Fragestellungen können beispielsweise lauten:

- Verkaufen wir in Ilmenau mehr Bier als in Erfurt?
- Wie viel Roséwein wurde im Sommer in ganz Thüringen verkauft?
- Sind die Verkaufszahlen von Roséwein höher als von Rotwein?

Zur Beantwortung dieser Fragen müssen die einzelnen Filialdatenbanken herangezogen werden. Als Lösungsansätze bieten sich dann einerseits *verteilte Datenbanken* an. Dies bedeutet, dass die einzelnen Datenbanken angefragt werden und eine Sicht mittels der Vereinigung der einzelnen Ergebnismengen (**`UNION`**) aufgebaut wird. Die resultierende Anfrageausführung ist dabei sehr aufwendig. Andererseits ist es möglich, alle Filialen auf eine *zentrale Datenbank* zugreifen zu lassen. Nachteilig ist hierbei die erhöhte Bearbeitung der Datenbankanfragen im operativen Geschäft. Als dritte Alternative hat sich der Data-Warehouse-Ansatz herausgestellt. Diesen wollen wir in den folgenden Kapiteln näher betrachten.

1.2 OLTP versus OLAP

Online Transactional Processing (kurz OLTP) wird in Datenbanken als Konzept eingesetzt, um die Anforderungen einer zentralisierten Datenhaltung zu erfüllen. Dies erfolgt ohne Zeitverzug sowohl in der Verarbeitung als auch der Verbuchung. Hierbei spielen die Anfragen ebenso wie die Datenänderungen

eine wichtige Rolle. Im Bereich der operativen Datenbanken sind vor allem die Operationen **SELECT**, **INSERT**, **DELETE** und **UPDATE** im Einsatz. Dabei hat sich das Konzept der transaktionalen Anfragen unter Berücksichtigung der ACID-Eigenschaften durchgesetzt.

Für analytische Anfragen stellt dies jedoch einen unzureichenden Ansatz dar. Denn während im Bereich von OLTP oftmals kurze Transaktionen und sehr viele Nutzer und Anwender auf die Daten gleichzeitig zugreifen, sind die komplexen und lang andauernden Anfragen im Bereich der Analyse besonders häufig. Für Data-Warehouse-Systeme hat sich daher der Begriff *Online Analytical Processing* (kurz OLAP) herausgebildet. Hier stehen komplexe und lang andauernde Lesetransaktionen im Fokus und sollen durch das Data Warehouse effizient unterstützt werden.

Im betrieblichen Umfeld müssen Entscheidungen auf einer konsistenten und belastbaren Basis gründen. Daher müssen die unterschiedlichen Quellen im Data Warehouse integriert werden. Für diese *Integration* wird der *ETL*-Prozess (Extraktions-, Transformations- und Ladeprozess) genutzt. Dieser ermöglicht die Vereinigung von Daten aus verschiedenen, meist heterogenen Quellen. Dabei muss die Heterogenität auf verschiedenen Ebenen (System, Schema, Daten) überwunden werden. Für die effiziente *Analyse* müssen die Daten nicht nur für den Anwender bereitgestellt, sondern bezogen auf das Analysegebiet in die gewünschte Form transformiert werden. Dies erfordert neben der Vorauswahl geeigneter Daten unter anderem auch die Herstellung eines Zeitbezuges und die notwendigen Aggregationen hinsichtlich unterschiedlicher Eigenschaften.

1.2.1 OLAP- versus OLTP-Transaktionen

Eine typische OLTP-Transaktion ist mit dem folgenden Beispiel aus unserer im Getränkemarkt vorhandenen Kundendatenbank aufgezeigt. Hier werden aus der Relation der Kunden der Vorname und Nachname des Kunden mit der *ID 0800* gewählt.

Kunde					
ID	Nachname	Vorname	PLZ	Ort	Straße
4711	Saake	Gunter	01234	Irgendwo	Am Berg 3
42	Sattler	K.	12345	Hier	Zufahrt 18
0800	Köppen	Veit	60701	Dort	Weg 9A

```
SELECT Vorname, Nachname
FROM Kunde
WHERE ID = 0800
```

Ergebnis:

Vorname	Nachname
Veit	Köppen

Eine analytische Anfrage, die eine lang andauernde Transaktion im Sinne von OLAP darstellt, ist hingegen im folgenden Beispiel dargestellt. Dabei wird der durchschnittliche Umsatz für die Jahre, Artikel und Verkaufsgebiete angefragt. Aufgrund der Komplexität müssen mehrere Joins durchgeführt werden. Die einzelnen Gruppierungselemente werden im Kontext des Data Warehouse als Dimensionen bezeichnet, und die Kennzahlen, die zur Analyse und Entscheidungsfindung herangezogen werden, heißen Fakten. Das multidimensionale Schema wird in Kapitel 3 näher behandelt.

```
SELECT DISTINCT ROW  Zeit.Dimension AS Jahr,
                     Produkt.Dimension AS Artikel,
                     AVG(Fact.Umsatz) AS Umsatzdurchschnitt,
                     Ort.Dimension AS Verkaufsgebiet

FROM (Produktgruppe INNER JOIN Produkt ON Produktgruppe.
      [Gruppen-Nr] = Produkt.[Gruppen-ID]) INNER JOIN
      ((((Produkt INNER JOIN [Fact.Umsatz] ON Produkt.[Artikel-Nr]
      = [Fact.Umsatz].[Artikel-Nr]) INNER JOIN Order ON
      [Fact.Umsatz].[Bestell-Nr]= Order.[Order-ID]) INNER JOIN
      Zeit.Dimension ON Orders.[Order-ID] =
      Zeit.Dimension.[Order-ID]) INNER JOIN Ort.Dimension ON
      Order.[Order-ID] = Ort.Dimension.[Order-ID]) ON
      Produktgruppe.[Gruppen-Nr] = Produkt.[Gruppen-ID]

GROUP BY Produkt.Dimension.Gruppenname, Ort.Dimension.Bundesland,
         Zeit.Dimension.Jahr;
```

Durch die beiden exemplarischen Anfragen in den unterschiedlichen Systemumgebungen wird deutlich, dass eine unterschiedliche Funktionalität und Ausrichtung in beiden Systemwelten vorliegt. Wir wollen im Folgenden die wichtigsten Unterschiede zwischen den operativen Datenbanksystemen, die auf OLTP-Basis arbeiten, und den für die Analyse ausgerichteten Data-Warehouse-Systemen kennenlernen.

1.2.2 Vergleich von OLTP und OLAP

In klassischen operativen Informationssystemen ist das Online Transactional Processing vorherrschend. Hier werden große Datenbestände sowohl erfasst als auch verwaltet. Eine Verarbeitung der Daten erfolgt dabei unter Verantwortung der jeweiligen Abteilung. Die transaktionale Verarbeitung bedeutet, dass kurze Lese-/ Schreibzugriffe auf einigen wenigen Datensätzen stattfinden. Im Gegensatz dazu wird im Data Warehouse mit dem Online Analytical Processing die Analyse auf dem Datenbestand in den Vordergrund gestellt. Dies bedeutet viele lange Lesetransaktionen auf vielen Datensätzen. Zudem soll das Data

	OLTP	OLAP
Fokus	Lesen, Schreiben, Modifizieren, Löschen	Lesen, periodisches Hinzufügen
Transaktionsdauer und -typ	kurze Lese-/Schreibtransaktionen	lange Lesetransaktionen
Anfragestruktur	einfach strukturiert	komplex
Datenvolumen einer Anfrage	wenige Datensätze	viele Datensätze
Datenmodell	anfrageflexibel	analysebezogen
Datenquellen	meist eine	mehrere
Eigenschaften	nicht abgeleitet, zeitaktuell, autonom, dynamisch	abgeleitet/konsolidiert, historisiert, integriert, stabil
Datenvolumen	MByte ... GByte	GByte ... TByte ... PByte
Zugriffe	Einzeltupelzugriff	Tabellenzugriff (spaltenweise)
Anwendertyp	Ein-/Ausgabe durch Angestellte oder Applikationssoftware	Manager, Controller, Analyst
Anwenderzahl	sehr viele	wenige (bis einige Hundert)
Antwortzeit	msecs ... secs	secs ... min

Tabelle 1.1: Vergleich von OLTP und OLAP nach [BG04]

Warehouse als Entscheidungsunterstützungssystem dienen und muss daher eine Integration, Konsolidierung und Aggregation der Daten gewährleisten.

In Tabelle 1.1 haben wir die wichtigsten Punkte hinsichtlich der Unterschiede zwischen OLTP und OLAP aufgeführt. Dies erfolgt in Anlehnung an [BG04]. An dieser Stelle unterscheiden wir die Anfragen, die Daten und die Anwender.

Bei Betrachtung der aufgeführten Eigenschaften wird ersichtlich, dass es aufgrund der großen Unterschiede notwendig ist, sowohl das Data Warehouse als eigenständiges System zu betreiben als auch innerhalb dieses Systems Technologien einzusetzen, die sich von Datenbanktechnologien teilweise stark unterscheiden. Nur so ist es möglich, den Anforderungen der Anwender gerecht zu werden und eine effiziente Analyseplattform bereitzustellen.

1.2.3 Abgrenzung: DBMS-Techniken

Auch im Datenbankkontext gibt es teilweise ähnliche Anforderungen. Die Entwicklung spezieller Techniken und die Weiterentwicklung im Datenbankbereich führen jedoch häufig zu einer notwendigen gekoppelten Betrachtung von Datenbankforschung und Data-Warehouse- bzw. Business-Intelligence-

Forschung. Wir wollen uns in diesem Buch insbesondere der relationalen Umsetzung des Data Warehouse widmen. Somit sind Techniken aus dem Datenbankbereich einfacher transferierbar.

Das Konzept der parallelen Datenbanken ist eine Technik zur Realisierung eines Data Warehouse. Hier werden Multiprozessoren eingesetzt, um die Verarbeitung der Daten effizient zu gestalten. Dadurch werden Transaktionen und Queries schneller bearbeitet.

Verteilte Datenbanken hingegen nutzen in der Regel keine redundante Datenhaltung. Die Verteilung des Datenbestandes in diesem Konzept erfolgt maßgeblich zur Lastverteilung. Zudem wird in verteilten Datenbanken keine inhaltliche, d.h. auf Analysezwecke ausgerichtete Integration und Aggregation vorgenommen.

Bei *föderierten Datenbanken* ist eine höhere Autonomie gegeben. Damit geht eine größere Heterogenität einher. Auch in diesem Fall ist der Analysezweck weder spezifiziert noch steht er im Kontext. Außerdem werden föderierte Datenbanken eingesetzt, ohne eine spezielle Optimierung hinsichtlich des Lesezugriffs zu implementieren.

Mit der Entwicklung von *In-Memory-Datenbanken* und dem kostengünstigen Einsatz von Hauptspeichern ergeben sich nicht nur für die analytischen Aufgaben zahlreiche Effizienzsteigerungen, sondern dies führt ebenfalls zu einer Zusammenführung von operativen Datenbanken und dem Data Warehouse. Hierbei zeigt die SAP-Sanssouci-DB [PZ12] ein mögliches Konzept, wie ein holistischer Ansatz funktionieren könnte. Für eine Betrachtung von Hauptspeicherdatenbanken siehe auch Abschnitt 6.4.

1.3 Charakteristika und Begriffe

Die wichtigsten Begriffe wollen wir hier kurz vorstellen. Eine Vertiefung erfolgt dann in den entsprechenden weiterführenden Kapiteln.

Der Begriff des Data Warehouse geht auf W. H. Inmon aus dem Jahr 1996 zurück. Er definiert:

> A *Data Warehouse* is a **subject-oriented**, **integrated**, **non-volatile**, and **time variant** collection of data in support of managements decisions. [Inm96]

Data Warehouse (DW) bezeichnet nach Inmon also eine *themenorientierte*, *integrierte*, *zeitbezogene* und *dauerhafte* Sammlung von Informationen zur *Entscheidungsunterstützung*. Die Themenorientierung bzw. Fachorientierung wird dabei verstanden als Unterstützung bereichsübergreifender Auswertungsmöglichkeiten für unterschiedliche Domänen. So erfolgt im Data Warehouse eine zentralisierte Bereitstellung der Daten über Geschäftsobjekte (Themen). Die

integrierte Datenbasis ermöglicht die Verarbeitung von Daten aus mehreren verschiedenen (internen und externen) Datenquellen, z.B. operationalen DB oder dem Web. Die Datenbasis selbst ist hierbei nicht-flüchtig, d.h. sie ist über die Zeit stabil und persistent. Daten innerhalb des Data Warehouse werden somit im Normalfall nicht mehr gelöscht oder verändert. Zudem sind die Daten im Data Warehouse zeitbezogen. So sind Zeitreihenanalysen möglich, also der Vergleich der Daten über zeitliche Aspekte. Auch wird im Kontext des Data Warehouse von einem Historisierungskonzept der Daten gesprochen, die Daten werden über einen längeren Zeitraum gesammelt und gespeichert.

Unter *Data Warehousing* verstehen wir den Data-Warehouse-Prozess, d.h. alle Schritte von der Datenbeschaffung (Extraktion, Transformation, Laden) über die Speicherung bis hin zur Analyse. Die Daten im Data Warehouse sind multidimensional und werden in einem Datenwürfel zusammengeführt. Dieser Datenwürfel stellt dabei ein mehrdimensionales Konstrukt zur Datendarstellung dar. Die Informationen zum Datenzugriff werden als *Dimension* bezeichnet und die Daten selbst als *Kennzahlen*. Es ist aber auch oft notwendig, anwendungsspezifische Analysedaten zu erstellen. Diese spezifische Sicht auf den Datenwürfel wird als *Data Mart* bezeichnet und erfolgt durch Kopieren der notwendigen Daten aus dem Datenwürfel bzw. Transformationen dieser Daten. Die explorative und interaktive Analyse auf Basis des konzeptionellen Datenmodells wird als Online Analytical Processing (OLAP) bezeichnet. Das Schlagwort *Business Intelligence* umspannt die Aktivitäten im Data Warehouse und zielt zudem auf die Managementunterstützung hin. Somit besteht Business Intelligence aus dem Data Warehousing, Reportingaktivitäten für das Management und Analysen zur Wissensentdeckung aus den Data-Warehouse-Daten. Dies beinhaltet ebenfalls die automatisierte Erstellung von Berichten in Unternehmen.

1.4 Big Data und Data Warehousing

Seit ca. 2010 hat sich das Thema „Big Data“ zu einem großen Trend entwickelt. Big Data ist zunächst ein eher unscharfer Begriff, der Datensammlungen bezeichnet, die für klassische Techniken der Datenverarbeitung zu groß sind, so dass neue Techniken benötigt werden. Allerdings gibt es hier keine konkrete Größenangabe – je nach Bedarf fallen unter „Big Data“ Datenmengen im Bereich von Terabyte bis Exabyte. Ein wesentlicher Faktor für das Interesse an Big Data bildet die massive Zunahme an (maschinell) erzeugten Daten, die von Sensoren, Kameras, Mobilgeräten etc. produziert werden. Dies betrifft nicht nur Geschäftsprozesse und unser tägliches Leben, sondern auch den Finanzbereich (z.B. Börsentransaktionen), Telekommunikationssysteme (z.B. Verbindungsdaten), die Energieversorgung (z.B. Smart Metering) und natürlich den Bereich der Naturwissenschaften, Astronomie, Klima- und Umweltforschung.

So hat Jim Gray in [HTT09] ein „viertes Paradigma“ der Wissenschaften formuliert, wonach nach der empirischen Forschung durch Beobachtung, den theoretischen Wissenschaften mit mathematischen Modellen, dem wissenschaftlichen Rechnen (Computational Sciences) mit (numerischen) Simulationen ein neuer Trend zu datenintensiven Naturwissenschaften zu erkennen ist. Hierbei liegt der Schwerpunkt auf der Analyse großer Datenbestände wie etwa Klima-, Satelliten- oder Teleskopdaten. Aber auch für Unternehmen und Behörden eröffnen sich durch die Nutzung und Auswertung großer Datenbestände neue Möglichkeiten, etwa durch die Analyse von Nutzerverhalten, Bewegungs- oder Verbrauchsdaten. Einige Beispiele aus der jüngeren Vergangenheit haben aber auch die Probleme und Gefahren von Big Data aufgezeigt. Neben Fragen des Datenschutzes ist auch zu berücksichtigen, dass mehr Daten nicht gleichzeitig bessere Daten bedeuten.

Big Data wird in Anlehnung an eine Studie der META Group [Lan01] oft auch durch die 3V beschrieben: Neben dem naheliegenden *Volume* zählen dazu noch *Variety*, um auszudrücken, dass strukturierte sowie unstrukturierte Daten, Texte und sogar Bilder und Videos zu verarbeiten sind, sowie *Velocity* zur Charakterisierung des Wechsels von einer Batch- zur Echtzeitverarbeitung.

Klassische relationale Datenbanksysteme und damit auch Data-Warehouse-Systeme können diese Anforderungen offensichtlich nicht vollständig erfüllen. Daher wurden für die Verarbeitung von Big Data eine Reihe neuer Technologien entwickelt. Beispiele hierfür sind Systeme wie Apache Hadoop auf Basis des MapReduce-Paradigmas zur verteilten (datenparallelen) Verarbeitung großer Datenmengen in großen Rechenclustern von hunderten oder mehr Knoten. Auch einige NoSQL-Systeme, die auf die strenge relationale Strukturierung, leistungsfähige Anfrageoperatoren und häufig auch strikte Konsistenzgarantien zugunsten einer besseren horizontalen Skalierung über viele Knoten hinweg verzichten, werden unter Big-Data-Technologien eingeordnet. Beispiele hierfür sind der Amazon-Dienst DynamoDB, Google's Spanner sowie Systeme wie Cassandra, MongoDB oder CouchDB.

Allerdings ist die Grenze zwischen Big Data und Data Warehousing fließend. So existieren durchaus Data-Warehouse-Installationen, die Datenmengen im Petabyte-Bereich verwalten sowie Text- und Bilddaten integrieren können. Auch das Problem der Analyse in „Echtzeit“ wird beispielsweise durch In-Memory-Techniken adressiert. Schließlich haben einige DBMS-Hersteller inzwischen auch MapReduce in ihre SQL-Systeme integriert, so dass ETL- und Analyseprozesse in MapReduce-Programmen formuliert werden und somit externe Daten auf einfache Weise integriert werden können. Ein Beispiel hierfür ist u.a. TeraData Aster.

Eines der wesentlichen Unterscheidungsmerkmale ist jedoch, dass ein Data Warehouse eine integrierte, dauerhafte Datenbasis für Reports und Analysen bildet und somit auch sorgfältige Planung, Entwurf und Betrieb erfordert. Demgegenüber sind MapReduce-Technologien wie Hadoop auf die

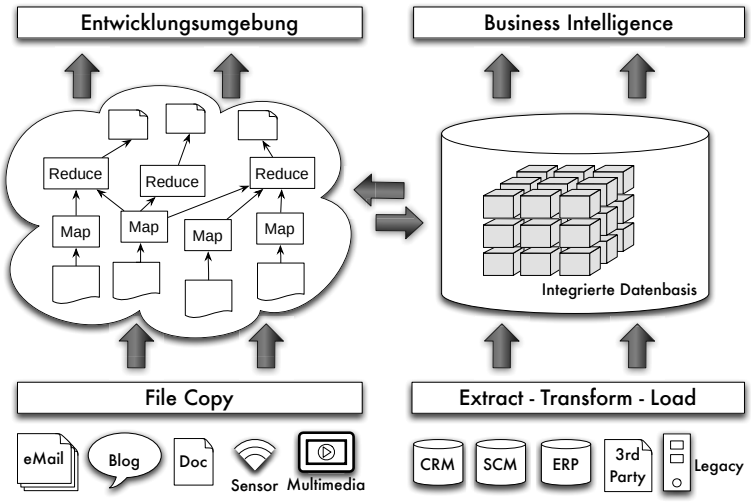

Abbildung 1.4: Anwendungsszenarien Big Data und Data Warehouse

Analyse von schwach oder wenig strukturierten Datenmengen ausgerichtet: Anstelle umfassender Datenmodellierung werden Datenextraktions- und -transformationsschritte implementiert. So bieten sich derartige Techniken als Ergänzung von Data Warehousing z.B. für die Implementierung von ETL-Prozessen oder die Analyse nicht-integrierter Datenbestände an.

1.5 Aufbau des Buches

Im vorliegenden Buch wollen wir uns Techniken des Data Warehouse in acht Kapiteln widmen. Hierzu können die Kapitel zwar einzeln gelesen werden, wir empfehlen jedoch, der Struktur des Buches zu folgen.

Im Kapitel 2 widmen wir uns den Fragestellungen der *Architektur* eines Data Warehouse. Hierbei wollen wir unter anderem die Anforderungen analytischer Informationssysteme betrachten und den damit einhergehenden Datenfluss innerhalb des Data-Warehouse-Systems vorstellen. Dies führt uns zu einer Referenzarchitektur. Darüber hinaus diskutieren wir auch noch weitere Architekturen.

In Kapitel 3 stellen wir das dem Data Warehouse zugrunde liegende *multidimensionale Datenmodell* vor. Dabei gehen wir auf die wichtigsten Konzepte ein, die den Datenwürfel repräsentieren. Zudem zeigen wir, wie eine Umsetzung des multidimensionalen Datenmodells in relationalen Datenbanken möglich ist.

Der *Extraktions-, Transformations- und Ladeprozess* (ETL) führt die heterogenen Quellen im Data Warehouse zusammen und steht im Mittelpunkt von Kapitel 4. Da für die Entscheidungsgrundlage die Qualität der Daten eine wichtige Rolle spielt, gehen wir auf wichtige Daten- und Schemaaspekte im Data-Warehouse-Kontext ein. Zudem gehen wir auf die typischen Aufgaben hinsichtlich der Extraktion der Daten aus den Quellen ein. Die Transformation stellt ein ganzheitliches Datenschema und vorbereitende Maßnahmen für die Analysezwecke her. Um den Analyseprozess zu unterstützen, bieten sich für Data Warehouse spezielle Techniken an, die Daten effizient zu laden.

Das Kapitel 5 zeigt typische *Data-Warehouse-Anfragen*. Dabei gehen wir auf die typischen OLAP-Operationen (Online Analytical Processing) im Datenwürfel ein. Bei den relationalen Umsetzungen und SQL-Unterstützungen stellen wir den Star-Join vor. Auch die im SQL-Standard definierten Operationen `CUBE` und `ROLLUP` werden präsentiert. Neben weiteren SQL:2003 OLAP-Funktionen gehen wir auf die multidimensionalen Ausdrücke von Microsoft (MDX) ein.

Wie der Datenwürfel *gespeichert* werden kann, ist Gegenstand von Kapitel 6. Dabei stellen wir sowohl die relationale Umsetzung wie auch die Speicherung in multidimensionalen Datenbanken vor. Fragen der Partitionierung des großen Data-Warehouse-Datenbestandes sowie Speicherungen für einen optimierten Datenzugriff werden in diesem Kapitel ebenfalls beantwortet. Neuen Trends im Bereich der Hauptspeicherbanken und ihren Einfluss auf das Data Warehouse widmen wir uns ebenfalls.

Ein effizienter Datenzugriff kann ebenfalls über *Indexstrukturen* erfolgen. Daher stehen diese im Mittelpunkt von Kapitel 7. Bereits in Datenbanken werden Indexstrukturen häufig genutzt. Diese sind jedoch zumeist eindimensional und somit für die analytischen Anfragen im Data Warehouse überwiegend ungeeignet. Aufgrund ihrer weitgehenden und effizienten Implementierung sind sie aber ein guter Ausgangspunkt, insbesondere der B-Baum. Für Daten mit wenigen Attributausprägungen stellen Bitmap-Indexstrukturen eine geeignete und effiziente Zugriffsstruktur dar. Außerdem stellen wir typische Vertreter der mehrdimensionalen Indexstrukturen vor und gehen auf Hierarchien ein.

In Kapitel 8 gehen wir auf die Anfrageverarbeitung im Data Warehouse ein. Hierzu gehört auch die Anfrageplanung inklusive der Star-Join-Optimierung. Auch die Berechnung des `CUBE`-Operators wird in diesem Kapitel adressiert. Außerdem stellen wir das Konzept der materialisierten Sichten im Data Warehouse-Konzept vor.

Im letzten Kapitel widmen wir uns typischen Anwendungsfällen. Diese werden unter dem Begriff *Business Intelligence* zusammengefasst. Die Hauptaufgabe im Data Warehousing ist die Erstellung von Berichten. Daher widmet sich Kapitel 9 auch dem Reporting. Häufig müssen aber auch Muster in den Daten erkannt werden, die für die Entscheidungsfindung herangezogen werden. Der Wissensentdeckungsprozess und insbesondere Data Mining sind

typische Analysen auf Data-Warehouse-Datenbeständen. In Abhängigkeit des Fachgebietes kommen dabei Untersuchungen wie Warenkorbanalyse, Kundensegmentierungen, Klassifikationen oder Prognosen vor.

1.6 Vertiefende Literatur

Der Begriff des Data Warehouse wurde von Inmon in den 90er Jahren geprägt [Inm92]. Bereits 1988 hatten Devlin und Murphy aufgrund der fortschreitenden Entwicklungen einen Information-Warehouse-Ansatz vorgestellt [DM88]. Kimball gibt einen praxisnahen Einblick in den Aufbau und die Nutzung von Data-Warehouse-Systemen [Kim08]. Auch die anderen Bücher von Kimball sind empfehlenswert, z.B. [KR02]. Kimball wird vor allem als Vater der Data Marts angesehen, während Inmon als treibende Kraft für das Data Warehouse gilt. Obwohl auch die Veröffentlichung des Buchs von Devlin [Dev96] shon einige Jahre her ist, sind viele der dortigen Konzepte für das Data Warehouse auch heute noch relevant.

Im deutschsprachigen Raum eignen sich für einen Überblick die Werke von Bauer und Günzel [BG04], Lehner [Leh03] und Chamoni und Gluchowski [CG10]. Zusätzlich empfiehlt sich für die Einbettung in den Themenbereich Business Intelligence das Lehrbuch von Kemper et al. [KBM10].

Da wir uns insbesondere der relationalen Umsetzung von Data-Warehouse-Systemen widmen, sind Kenntnisse im Bereich der Implementierung von relationalen Datenbanken notwendig. Ein Überblick hierzu findet sich beispielsweise in Saake, Sattler und Heuer [SSH11] und in [SSH13].

Herausforderungen des Data Warehousing stellt Widom [Wid95] übersichtlich dar. Chaudhuri und Dayal [CD97] geben einen Überblick unter anderem zu OLAP-Technologien.

1.7 Übungen

Übung 1-1 Erläutern Sie die konzeptionellen Unterschiede zwischen transaktionsbasierten Datenbanken (OLTP) und Data-Warehouse-Technologien (OLAP). Erklären Sie dazu den Begriff einer Transaktion.

Übung 1-2 Definieren Sie die Begriffe Data Warehouse, verteilte und föderierte Datenbanken und grenzen Sie die Systeme von einander ab. Wie sind in diesem Zusammenhang Data Marts einzuordnen?

Übung 1-3 Was ist unter Dimensionen und Fakten und Kennzahlen im Zusammenhang mit Data Warehouse zu verstehen? Geben Sie ein selbstgewähltes Beispiel.

Übung 1-4 Informieren Sie sich über das Benchmarking von Datenbanken. Welche Benchmarks gibt es? Gehen Sie besonders auf die TPC-H Benchmarks ein. Informationen hierzu finden Sie beispielsweise unter `http://tpc.org/`.

Übung 1-5 Verschaffen Sie sich einen Überblick über Data-Warehouse- und NoSQL-Systeme. Wodurch grenzen sich die Hersteller von der jeweiligen anderen Domäne ab?