Kapitel 2
Terminologie und Definition

In zahlreichen Publikationen und Fachzeitschriften tauchen die Begriffe Data Warehouse, Data Warehousing, Data-Warehouse-System, Metadaten, Dimension, multidimensionale Datenmodellierung und OLAP (*On-Line Analytical Processing*) zum Teil mit unterschiedlichen Interpretationen und Auslegungen auf, was manchmal sehr verwirrend sein kann.

Um eine gemeinsame Begriffswelt mit einheitlichen Fachtermini zu schaffen, werden in diesem Kapitel essenzielle Begriffe genauer erläutert und definiert. Dabei wurde einerseits bewusst auf die in der Literatur am häufigsten verwendeten und zitierten Definitionen zurückgegriffen, um unterschiedlichen Begriffsverständnissen vorzubeugen, andererseits sind einige wichtige und geläufige Interpretationen dieser Begriffe der Vollständigkeit halber erwähnt, um ein genaueres Verständnis des Themenbereichs zu erzielen. Hierbei sind die Definitionen und Erläuterungen größerer Softwarefirmen, die durch ihre am Markt vorhandenen zahlreichen Produkte zum Teil diese Begriffe mitgeprägt haben, berücksichtigt worden.

2.1 Data Warehouse

Ein Data Warehouse integriert Informationen aus vielen unterschiedlichen Quellen in einer für die Entscheidungsfindung optimierten Datenbank.

> a data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data in support of managements decisions.[1]

W. H. (Bill) Inmon[2] gilt als der Vater des Konzepts des Data Warehouse. Er definiert und charakterisiert ein Data Warehouse als „themenorientierte, integrierte, zeitbezogene und nicht-flüchtige Datenbank zur Unterstützung von Managemententscheidungen".

Themenorientiert besagt hier, dass beispielsweise nicht die einzelnen Transaktionen, Aufträge oder Buchungen, sondern Kennzahlen wie Kosten und Umsätze

[1] Vgl. Inmon (1996, S. 33).

[2] W.H. (Bill) Inmon soll den Begriff Data Warehouse geprägt haben.

K. Farkisch, *Data-Warehouse-Systeme kompakt*, Xpert.press,
DOI 10.1007/978-3-642-21533-9_2, © Springer-Verlag Berlin Heidelberg 2011

eines bestimmten Geschäftsbereichs von Interesse sind. Das heißt, der Zweck der Datenbank liegt nicht in der Erfüllung einer Aufgabe wie etwa eine Auftragsdatenverwaltung, sondern auf der Modellierung eines spezifischen Themenbereichs. Ausgehend von einer zweckneutralen Darstellung werden die Daten für spezifische Auswertungen sowohl logisch als auch physisch organisiert. Hierbei wird für eine Vielzahl der Anwendungen eine multidimensionale Sichtweise mit einer Unterscheidung in quantifizierbare Kennzahlen und beschreibende Informationen erwünscht.[3]

Integrierte Daten können aus vielen unterschiedlichen Quellen und Fremdsystemen stammen, werden aber in einheitliche Formate umgesetzt und folgen gleichen Regeln und Konventionen. Die *Integration* ist die Funktionalität und die Mächtigkeit Daten aus unterschiedlichen operativen Datenbanken miteinander in Verbindung zu bringen. Daten werden aus unterschiedlichen inkompatiblen Systemen mit jeweils unterschiedlichen Datenmodellen extrahiert, in einem oft komplexen Prozess bereinigt und anschließend zu einem integrierten Datenbestand verknüpft. Hierbei stellt sich die Frage nicht nur nach der Integration von Daten, sondern auch nach der Integration von Schemata aus heterogenen multiplen inkompatiblen Quellen.

Zeitlich veränderlich bezieht sich auf die Zeitabhängigkeit der Daten. In operativen Systemen sind diese so aktuell wie im Moment der Eingabe und reichen in der Regel 30–90 Tage zurück. In einer Data Warehouse Umgebung stellen vorhandene Daten immer eine Art ‚Schnappschuss' eines bestimmten Zeitpunkts oder Raumes dar. Sie werden periodisch (je nach Anwendungsfall stündlich, täglich, wöchentlich oder monatlich) aktualisiert und nicht in Echtzeit verarbeitet. Im Unterschied zu operativen Datenbanken, in denen die Daten meist den aktuellen Stand repräsentieren, werden die Daten im Data Warehouse *zeitbezogen* abgelegt. Die Verarbeitung der Daten ist so angelegt, dass Vergleiche über die Zeit ermöglicht werden. Hierzu ist es notwendig, Daten über einen längeren Zeitraum zu erfassen und zu speichern.

Unter dem Begriff *nicht-flüchtige* Datenbank wird hierbei verstanden, dass einmal im Data Warehouse eingebrachte und gespeicherte Daten weder modifiziert noch entfernt werden dürfen, sodass im Laufe der Zeit eine Historisierung der extrahierten Zustände und Daten der Quellesysteme erreicht wird. Die Historisierung kann aus systemtechnischen Gründen wie z. B. Speicherkapazität optional eingeschränkt werden.

Ergänzend zu dieser Definition von Data Warehouse wird in Bauer und Günzel (2004) eine neue und erweiterte Definition vorgeschlagen:

> Ein Data Warehouse ist eine physische Datenbank, die eine integrierte Sicht auf beliebige Daten zu Analysezwecken ermöglicht.[4]

K. U. Sattler und G. Saake halten an der Definition von W.H. Inmon fest und beschreiben Data Warehouse als Sammlung von Technologien zur Unterstützung von Entscheidungsprozessen.[5]

---

[3] Vgl. Lehner (2003, S. 10).

[4] Vgl. Bauer und Günzel (2004, S. 7).

[5] Vgl. Sattler und Saake (2006/2007, S. 9).

Die Softwarefirma Oracle orientiert sich ebenfalls an der von W.H. Inmon eingeführten Definition und beschreibt Data Warehouse wie folgt:

> A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but can include data from other sources. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources.[6]

IBM liefert folgende Beschreibung für den organisatorischen Aspekt eines Data Warehouse:

> A *data warehouse* is an organization's data with a corporate wide scope for use in decision support and informational applications.[7]

Das Kernelement eines jeden Data Warehouse stellen eine oder auch mehrere autonome Datenbanken dar, die aufbereitete Daten aus den inkompatiblen Datenbeständen inklusive der Datenmodelle enthalten. Alle relevanten Informationen werden hier gesammelt und archiviert. Neue Informationen kommen hinzu, während die alten Informationen erhalten bleiben. Dadurch entsteht eine Historie der Daten. Zusätzlich zu den im Unternehmen anfallenden Daten, die die Basis für Auswertungen sind, können Daten aus externen Quellen genutzt werden, etwa aus dem *World Wide Web* oder von beliebigen externen Dienstanbietern.

In diesen Datenbanken entstehen somit sehr schnell umfangreiche Datenmengen. Um in den großen Mengen anfallender Daten nicht den Überblick zu verlieren und damit es nicht zum Datenchaos kommt, ist es wichtig, vernünftige Granularitätsebenen und Verdichtungen zu planen und einzuführen. So kann das Data Warehouse auch den sehr unterschiedlichen Informationsbedürfnissen eines gesamten Unternehmens dienen bzw. gerecht werden.

Der große Vorteil beim Einsatz eines Data Warehouse ist es, dass die Daten aus unterschiedlichen Datenquellen bereinigt, integriert und anschließend analysiert werden können, ohne diese Quellen selbst in ihrer Funktion zu beeinträchtigen. Diese physische Trennung zwischen Data Warehouse und den Quelldaten erlaubt speziell die Modellierung der Daten hinsichtlich analytischer Anwendungen.

2.2 Data-Warehouse-System

Während der Begriff Data Warehouse nur die eigentliche Datenbank bezeichnet, beschreibt ein Data-Warehouse-System die gesamte technische Infrastruktur zur Beschaffung, Speicherung und Auswertung der Daten.[8]

W. Lehner definiert ein Data-Warehouse-System als eine Sammlung von Systemkomponenten und einzelnen Datenbanken, deren Daten auswertungsorientiert

---

[6] Vgl. Oracle (September 2007, S. 29, Abschn. 1-1).

[7] Vgl. Bruni et al. (July 2008, S. 32).

[8] Vgl. Albrecht (2001, S. 3).

organisiert sind und in einem mehrstufigen Prozess, basierend auf einer Vielzahl von Quellsystemen, abgeleitet werden.[9]

2.3 Data Warehousing

Data Warehousing ist ein dynamischer Vorgang bzw. ein Prozess, angefangen beim Datenbeschaffungsprozess über das Speichern bis hin zur Analyse der Daten, d.h., es beschreibt den Fluss und die Verarbeitung der Daten aus den Datenquellen bis zum Analyseergebnis beim Anwender.[10]

Data Warehousing beschreibt den Prozess, der notwendig ist, um ein Data Warehouse System zu planen, aufzubauen und insbesondere zu betreiben. Dazu zählen vor allem die Extraktion der relevanten Daten aus Quellsystemen, die Transformation und ggf. die Datenbereinigung,[11] Integration, Modellierung, Analyse und Auswertung dieser Daten. Data Warehousing ist der Prozess der Beschaffung und der Auswertung der Daten eines Data Warehouse und umfasst folgende Aktivitäten, die in Abb. 2.1 dargestellt werden:

1. Die Extraktion der relevanten Daten aus Quellsystemen
2. Die Transformation und Bereinigung der Daten der Quellsysteme
3. Laden der bereinigten, konsistenten Daten in das Data Warehouse
4. Dauerhafte Speicherung der Daten im Date Warehouse
5. Bereitstellung der zu Analysezwecken benötigten Datenbestände aus dem Data Warehouse
6. Auswertung und Analyse der Datenbestände

Die Aktivitäten in Punkt 1–3 werden zusammengefasst auch als ETL-Prozess[12] bezeichnet, der später (siehe Abschn. 8.3) genauer erörtert wird. Unter Data Warehousing wird teilweise aber auch die Technologie, die mit dem Einsatz von Data-Warehouse-Systemen verbunden ist, verstanden.[13]

2.4 Herkunft des Data Warehousing

Die Herkunft des Data Warehousing geht auf föderierte Datenbanksysteme zurück. Um eine einheitliche unternehmensweite Datenbasis als Grundlage für Analyse und Auswertung der Daten zu schaffen, müssen die Probleme der Daten-,

[9] Vgl. Lehner (2003, S. 9–10).

[10] Vgl. Bauer undGünzel (2004, S. 8).

[11] Im Englischen Data Cleaning oder Data Cleansing genannt.

[12] ETL steht für *Extracion, Transformation and Load* (Extraktion, Transformation und Laden).

[13] Vgl. Albrecht (2001, S. 3).

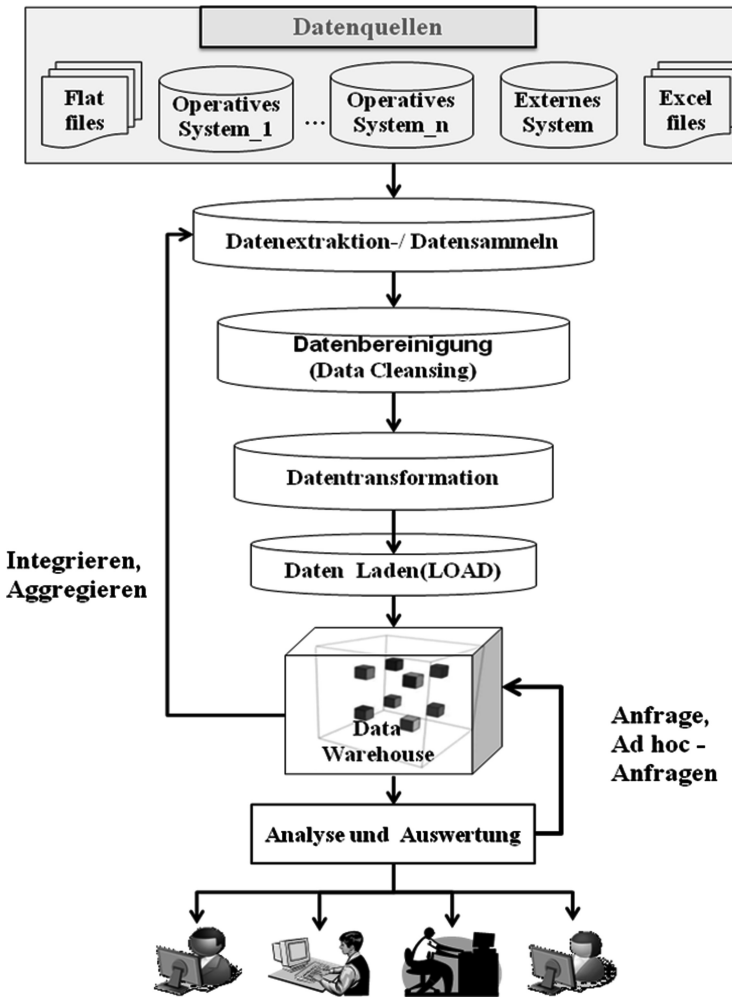

Abb. 2.1 Aktivitäten beim Data Warehousing

Schemata-, und Systemheterogenität beseitigt werden. Die Erfüllung dieser Anforderungen war das Ziel der so genannten föderierten Datenbanksysteme,[14] die Konzepte und Methoden zur Datenintegration zur Verfügung stellen. Zur Verarbeitung einer Analyseanfrage an so ein föderiertes Datenbanksystem wird die Analyseanfrage in Teilen zerlegt und an die entsprechenden Teilnehmersysteme der Föderation weitergereicht. Die an der Föderation beteiligten Systeme verarbeiten die Analyseanfrage anhand ihrer eigenen Datenmodelle und liefern jeweils ihre berechneten Ergebnismengen an das föderative Datenbanksystem zurück, wo

[14] Vgl. Conrad (1997, S. 31–50).

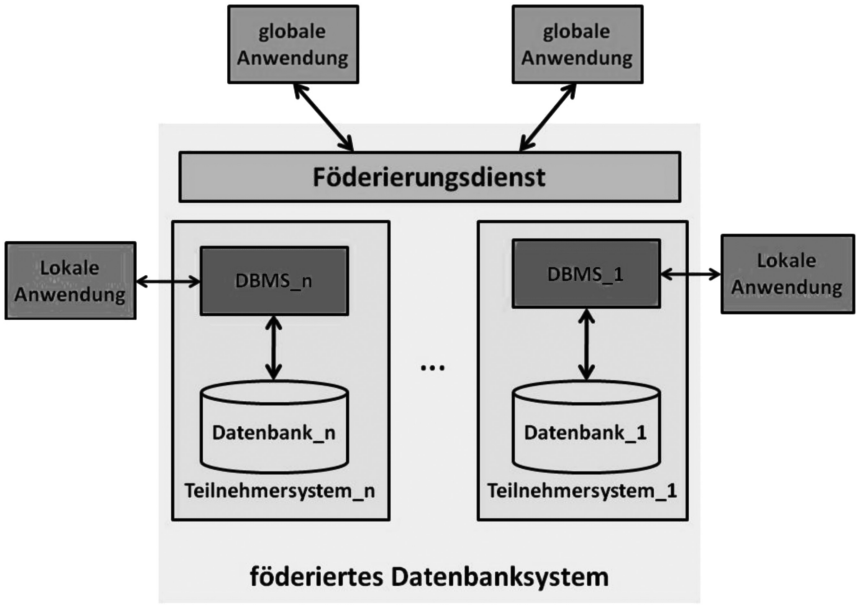

Abb. 2.2 Allgemeine Architektur föderierter Datenbanksysteme (nach Conrad)[15]

die jeweiligen Teilergebnismengen zusammengeführt werden. Sollte im Rahmen der föderierten Datenbanksysteme eine Datenintegration auf Ebene der einzelnen Datensätze erfolgen, so besteht ein großes Problem in der Ermittlung der Abbildungsregeln des globalen Datenmodells auf die lokalen Datenschemata beim Vorhandensein semantischer Mehrdeutigkeiten. Ein Data Warehouse umgeht dieses Problem dadurch, dass die Daten extrahiert, transformiert werden, um die Konsistenz der Daten zu sichern.[16]

[15] Vgl. Conrad (1997, S. 49).
[16] Vgl. Lehner (2003, S. 7f).

http://www.springer.com/978-3-642-21532-2

Data-Warehouse-Systeme kompakt
Aufbau, Architektur, Grundfunktionen
Farkisch, K.
2011, XI, 122 S. 31 Abb., 15 Abb. in Farbe., Hardcover
ISBN: 978-3-642-21532-2