

# DATA WAREHOUSING – LESS DATA AN MORE BUSINESS INFORMATION<sup>1</sup>

*Willhild Angelika Kreitel*



**Prof. Dr. Willhild Angelika Kreitel**

Professorin für Datenverarbeitung und Informationswirtschaft am Fachbereich Betriebswirtschaft der Fachhochschule Würzburg-Schweinfurt

In der heutigen Zeit des starken Wettbewerbs ist es für die Manager besonders wichtig, die richtigen Entscheidungen in einer angemessenen Zeit treffen zu können.

Dieser Prozess der Entscheidungsfindung erfordert eine selektierte, konsistente und temporale Datensammlung, die dynamische Auswertungen und flexible Zugriffe gestattet.

### Was braucht ein Manager in der heutigen Wettbewerbssituation?

Er braucht Geschäftsinformationen, Schlüsselfaktoren und die dazugehörigen Schlüsselindikatoren zur Entscheidungsunterstützung.

- Die richtigen Informationen
- zur richtigen Zeit
- am richtigen Ort
- für die richtigen Personen

das ist die Forderung des Managements angereichert mit folgenden Attributen:

### Was besitzt ein Manager in der heutigen Wettbewerbssituation?

Er verfügt über eine Menge statischer Berichte aus verschiedenen operativen Quellsystemen, über eine Flut von Daten, über viele separate Dateien, zum Beispiel die so beliebten Excel-Tabellen mit Daten und Geschäftsgraphiken, und über eine Menge von Assistenten, die die Informationen manuell aufbereiten.

Um diese Schere zwischen den dringenden Erfordernissen und der momentanen, aus der historischen Entwicklung erkläraren Situation zu verringern, empfiehlt sich die Anwendung einer Datenaufbereitungs- und Datenhaltungstechnologie, das Data Warehousing mit dem Data Warehouse als Kernstück.

### Data Warehouse

Entsprechend der Definition von W.H. Inmon, der als Vater des Data Ware-

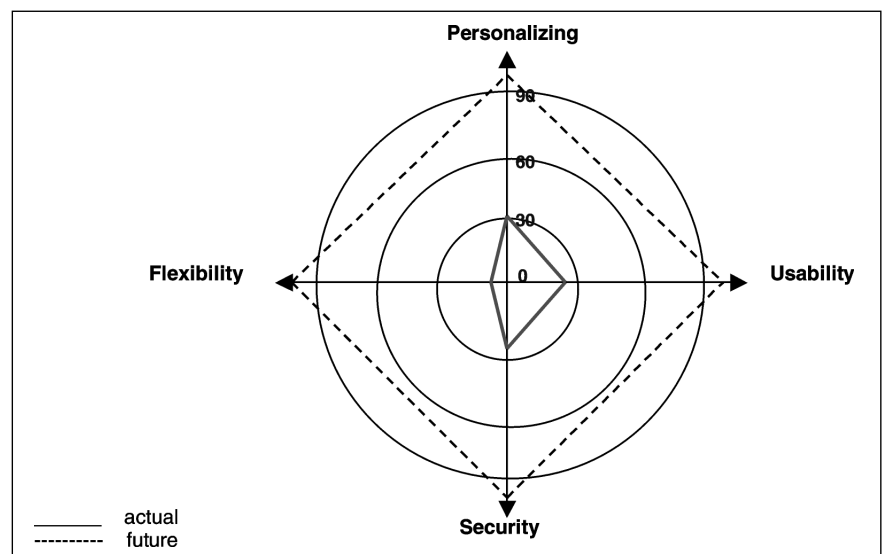


Abb. 1: Nutzeranforderungen

housing gilt, ist ein Data Warehouse eine

- subjektorientierte,
- integrierte und
- temporale

<sup>1</sup> Unter Nutzung der Diplomarbeit von Herrn Bernd Schwab.

Sammlung von Daten zur Unterstützung von Managemententscheidungen.<sup>2</sup>

**Subjektorientiert** bedeutet, dass sich die Strukturierung der Daten im Data Warehouse an der Analyseproblematik orientiert. Dabei werden die Daten entsprechend der unterschiedlichen Betrachtungsweisen und Auswertungswünsche der Anwender mit Hilfe mehrdimensionaler Datenmodellierungstechniken organisiert, um eine flexible und zeitkritische Auswertung zu garantieren.

**Integriert** bedeutet: Die Extraktion von Daten aus den operativen Quellsystemen inhaltlich konsistent und in einheitlicher Form. Zum Beispiel darf dann das Attribut Geschlecht für Kunden oder Mitarbeiter entweder durch 0 / 1 als numerischer oder durch m / w als alphabetischer Datentyp zugelassen werden.

**Temporal** bedeutet eine fortgeschriebene historische Datensammlung mit entsprechenden Zeitstempeln, deren Aktualitätsgrad von der Frequenz der Übernahme der Daten aus den operativen Systemen abhängt.

Dieses Konzept der redundanten Datenhaltung getrennt von den operativen Systemen schliesst sowohl Software-, Hardware- und Internetkomponenten als auch zahlreiche spezielle Data Warehousing-Prozesse ein.

### Data Warehousing-Architektur

Die Grundlage für eine Data Warehousing-Architektur bildet ein 3-Layer-Referenzmodell, welches auf die operativen Quelldaten aufsetzt und den data layer, den application layer und den presentation layer beinhaltet, die in der Praxis aus Gründen der Performance eigenständigen Servern bzw. Clients zugeordnet werden.

Die folgende Abbildung zeigt diese 3 Layer einschliesslich der operationalen Datenbasis als eine mögliche Data Warehousing-Architektur:

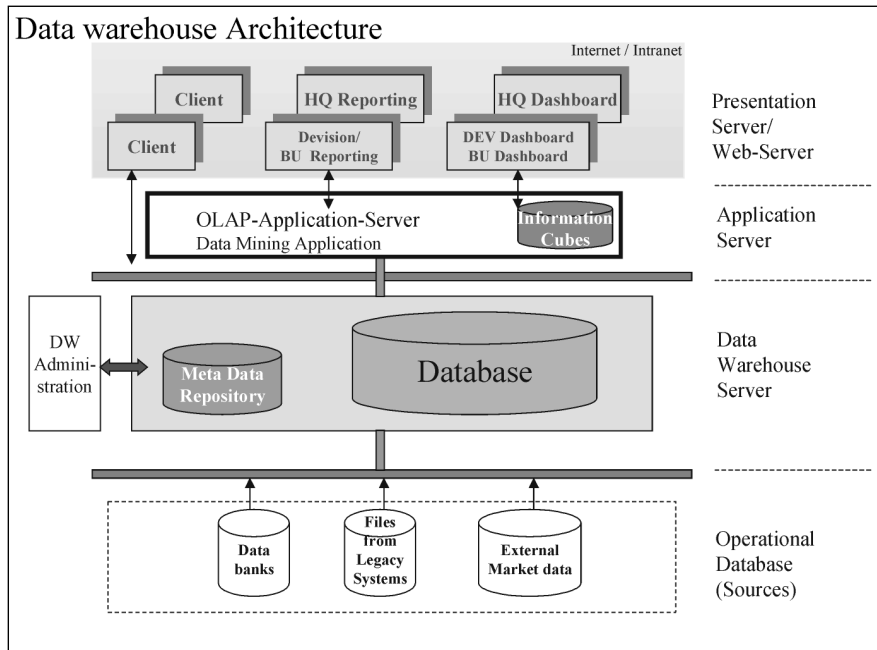


Abb. 2: Data Warehousing-Architektur

Die Data Warehousing Architektur weist dabei folgende wesentliche Merkmale auf:

- Integration jeglicher Datenquellen innerhalb und ausserhalb des Unternehmens zu einem konsistenten Datenbestand mit Hilfe von Extraktions- und Transformationsprozessen (ETL).
- Dynamische Einbeziehung von neuen Daten zu historischen Datenbeständen (im Gegensatz zu statischen Management Information Systems).
- Multidimensionale Analyse im Datenbestand ohne Beeinflussung der zeitkritischen, transaktionsorientierten, operativen Systeme mit Hilfe von Analyseprozessen.

### Data Extraction, Data Transformation and Data Load Processes (ETL)

Der Grundstein für eine qualitativ hochwertige Datenanalyse und die Erzeugung von Informationen wird durch die korrekte Übernahme der Daten aus den unterschiedlichen Datenquellen gelegt. Die Datenübernahme erfolgt durch die ETL-Software, die nach speziellen Methoden und Verfahren den kompletten Prozess der Datenübernahme steuert.

Nach Kimball gestaltet sich der Prozess der Datenmigration zusammengefasst in 10 Schritten:<sup>3</sup>

1. Datenextraktion aus den Quellsystemen.
2. Identifizierung der Datensätze, die sich seit dem letzten ETL-Prozess geändert haben.
3. Anpassung der Verknüpfungsschlüssel für neu hinzukommende und geänderte Dimensionsdaten im Meta data Repository.
4. Neustrukturierung der extrahierten Quelldaten auf der Grundlage des logischen Datenmodells. Die extrahierten Daten werden den entsprechenden Spalten in den Zieltabellen zugeordnet und ggf. in ein neues Format konvertiert.
5. Übernahme der Daten in das Data Warehouse.
6. Sortierung der Daten und Bildung von Aggregaten (kann auch ausserhalb des Data Warehouses geschehen).
7. Anpassung der Verknüpfungsschlüssel für neu hinzukommende und geänderte Dimensionsdaten.
8. Laden und Indizieren der aufbereiteten Daten in die Fakt- und Dimensionstabellen gemäss dem logischen Datenmodell. Nach Abschluss des Ladevorgangs wird der neue Datenbestand auf Vollständigkeit geprüft (Test der referenziellen Integrität).
9. Qualitätskontrolle. Prüfung, ob die Daten korrekt übernommen wurden.
10. Benachrichtigung der Data Warehouse Anwender über die Beendigung des ETL-Prozesses und die Integration des neuen Datenbestandes in das Data Warehouse.

<sup>2</sup> Vgl. Inmon, W.H., Building the Data Warehouse, New York 1994, S.25f.

<sup>3</sup> Vgl. Kimball, R., The Data Warehouse Toolkit, New York 1996, S. 217.

Die Qualität dieser ETL-Applikationen wird gemessen an:

- der Minimierung von Ausfallzeiten während des Extraktionsprozesses,
- der Flexibilität der Integration unterschiedlicher Quelldatensysteme,
- der Identifizierung der geänderten Daten (Change data capture support),
- der Sicherung der Datenintegrität und Konsistenz (Integration von restart/recovery/logging-Verfahren).

Um die Ausfallzeit des Data Warehouses zu minimieren, werden die Datenbestände gespiegelt, d.h. es wird eine zweite redundante Datenbasis parallel betrieben. Während der Datenübernahme in das Data Warehouse werden die Anfragen an das gespiegelte Data Warehouse weitergeleitet. So stehen, wenn notwendig, die Daten 24 Stunden zur Verfügung.

Die Übernahme aus den unterschiedlichen Quellsystemen in das Data Warehouse erfordert die Konvertierung, Vereinheitlichung und Umstrukturierung der extrahierten Daten zu neuen einheitlichen Datentypen. Um dies zu ermöglichen, sind neben den eigentlichen Rohdaten beschreibende Daten erforderlich, die auch Meta data genannt und in einem Meta data Repository des Data Warehouses abgelegt werden. Meta data beschreiben den logischen Aufbau der Daten, deren Beziehungen untereinander und Sicherheitsmechanismen, wie Zugriffsrechte auf bestimmte Daten.

### Online Analytical Processing (OLAP)

Geprägt wurde die OLAP-Technologie durch E.F. Codd, einer der Väter der relationalen Datenmodellierung und des Online Transaction Processings (OLTP).

OLAP bezeichnet eine Technologie, die durch eine dynamische und multidimensionale Datenanalyse auf Basis historischer und aktueller Daten gekennzeichnet ist. Im Gegensatz zu traditionellen Analysesystemen muss bei OLAP-Anwendungen der Analysepfad nicht vollständig vorgegeben werden, sondern wird vom Benutzer im Laufe der Analyse vervollständigt.

In traditionellen Analysesystemen formuliert der Anwender eine Datenbank-

abfrage und erhält eine Ergebnismenge vom Datenbanksystem zurück. Der Prozess der Datenanalyse ist damit abgeschlossen. Möchte der Nutzer nun weitergehende Informationen über die erhaltenen Ergebnisse, so muss eine komplett neue Abfrage gestartet werden. Die OLAP-Technologie ermöglicht es, auf bereits erzielte Abfrageergebnisse aufzubauen, indem diese als Grundlage für weitere Datenanalysen dienen. Dadurch entsteht ein flexibler und dynamischer Analysepfad, der dem Nutzer gestattet, sich bis zur gewünschten Ergebnismenge zu navigieren.

Der Zugriff auf die Daten des Data Warehouse geschieht mittels Business-Intelligence (BI)-Applikationen innerhalb des OLAP-Systems.

Business Intelligence (BI) wurde von der Gartner Group im Jahre 1990 als Prozess der Transformation von operativen Daten in Information und anschließender Ableitung von Wissen definiert.<sup>4</sup>

Demzufolge beinhaltet die BI-Software die gesamte Logik der Datenanalyse und generiert dynamische Data Warehouse-Abfragen, indem aus den Benutzeranfragen datenbankspezifische Befehlsfolgen erzeugt werden.

Ausserdem stellt sie eine Funktionalität bereit, von Codd als „speculative what-if and/or why data model scenarios“ bezeichnet, die dem Manager die Simulation zukünftiger Szenarien gestattet.

### Data Mining

Eine erweiterte Form der Datenanalyse im Data Warehouse-Umfeld stellt das Data Mining dar.

Unter Data Mining (Knowledge Discovery in Database) werden Analyseverfahren verstanden, die unbekannte Strukturen aus bisher ungeordneten Datenhalten automatisiert aufdecken und Trends und Verhaltensmuster auf der Basis von bekannten Verhaltenschemata aus der Vergangenheit vorhersagen. Die Zielsetzung ist, neues Wissen aus dem vorhandenen Datenbestand zu extrahieren. Data Mining Programme durchsuchen dabei selbständig den Datenbestand des Data Warehouse, um Triviale Zusammenhänge, bekannte und unbekannt Korrelationen zwischen den gespeicherten Daten aufzuzeigen. Mögliche Anwendungsgebiete sind z.B. Hilfestellung zur Früherkennung von Kundenbewegungen, Betrugswahrscheinlichkeiten und Risiken.

Damit bietet die OLAP-Technologie dem Management nicht nur die Möglichkeit der Wiedergabe der gespeicherten Daten als Informationen in dynamischen Strukturen, sondern auch das Schöpfen neuer Erkenntnisse aus einem vorhandenen Datenbestand über intelligente Ursache-Wirkungs-Forschung.

### Implementierung

Das Data Warehouse als aufbereitete Datenbasis für die Online-Analyse

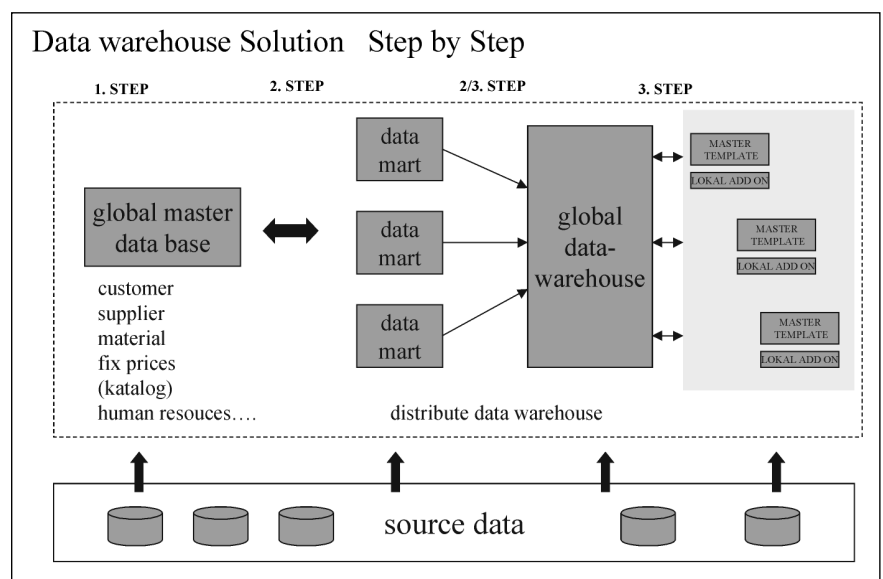


Abb. 3

<sup>4</sup> Vgl. Kurz, A., Data Warehousing Enabling Technology, Landsberg 1990, S. 597.

---

kann in zwei unterschiedlichen Ansätzen implementiert werden. Der erste Ansatz ist die zentrale Datenhaltung als Enterprise Data Warehouse für das gesamte Unternehmen (siehe Abb. 2). Die zweite Möglichkeit besteht in der Dezentralisierung der Datenbestände nach verschiedenen Kriterien, wie *business unit* oder *workgroups*, und demzufolge in der Entwicklung eines verteilten Data Warehouse. Ein Teilbereich dieser verteilten Datenwelt wird auch Data Mart genannt. Für grosse Unternehmen mit einer Anzahl von Geschäftsstellen im In- und Ausland bringt dieses verteilte Datenhaltungskonzept Vorteile bezüglich einer schrittweisen Implementierung der Data Warehouse-Konzeption und damit der Möglichkeit des schnelleren Teilerfolges bereits im ersten Schritt.

Die Abb. 3 zeigt die stufenweise Annäherung an ein Enterprise Data Warehouse über die Modellierung und Implementierung von Data Marts, die sich durch Skalierbarkeit auf das globale Data Warehouse auszeichnen.

Im zweiten Schritt erfolgt dann die Übernahme in das Enterprise Data Warehouse, um dann im 3. Schritt die

Anbindung sämtlicher Geschäftsstellen mit den möglichen Ergänzungen (add ons) für lokale Besonderheiten zu realisieren.

### Resümee

Data Warehousing ist eine ausgereifte Technologie, die unter Einbeziehung verschiedener oben dargestellter Methoden und der Internet/Intranet-Technologie folgenden Nutzen für das Management unterstützt:

- Hilfe im täglichen Entscheidungsprozess durch
- aktuelle Schlüsselinformationen und Schlüsselindikatoren, die für den Unternehmenserfolg entscheidend sind,
- sofortige Zugriffe,
- analytische Methoden und Simulationsmöglichkeiten für qualitativ bessere Vorhersagen,
- zeitabhängige, historische Daten zur flexiblen Auswertung und
- reduzierte Suchprozesse bei komplexen und dynamischen Abfragen.
- Konsistente Datenbasis für ein hochwertiges Controlling zur Steuerung

und Verbesserung der Leistungsfähigkeit der Unternehmen.

- Möglichkeiten für eine Wissensbasis zur Unterstützung des strategischen Lernprozesses.
- Realistische Basis für ein effizientes Customer Relationship Management (CRM) und Risk Management.

Der Erfolg des Data Warehousing hängt jedoch letztlich von drei Hauptfaktoren ab:

- Von einer effizienten Implementierung des Data Warehouse in die vorhandene Softwareumgebung des Unternehmens,
- von der permanenten Datenpflege, die keine Datenhalden, sondern qualitativ hochwertige historische Datenbestände garantiert, und
- vom Bewusstsein des Managements von der Notwendigkeit einer konsistenten, temporalen und redundanten Datenhaltung und dem Commitment des Managements für die Implementierung und Pflege.