

## Definition Data Warehouse

Ein Data Warehouse dient dazu, Daten aus unterschiedlichen internen und externen Quellen zusammenzuführen und zu speichern, um anschließend mithilfe unterschiedlicher Abfrage-, Analyse- und Auswertungsprogrammen neue Informationen zu gewinnen.

Worin besteht der Unterschied zwischen operativen & analytischen Daten?

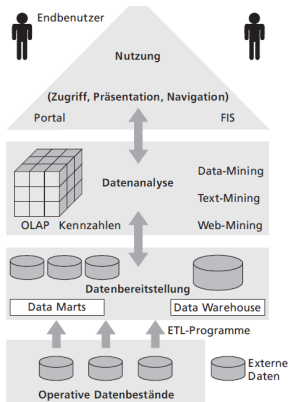
Kriterien	Daten für <i>operative</i> Anwendungen	Daten für <i>analytische</i> Anwendungen
<b>Zweck</b>	Unterstützung und Abwicklung operativer Geschäftsvorfälle	Informationen für das Management, Unterstützung von Entscheidungen, themenorientiert
<b>Inhalt</b>	detaillierte, aktuelle Daten über Geschäftsvorfälle, zeitpunktorientiert	verdichtete und bereinigte Daten, historische und zum Teil zukünftige Daten, zeitraumorientiert
<b>Aktualität</b>	hoch (Online, Realtime)	meist keine Tagesaktualität
<b>Modellierung</b>	Altdatenbestand oft nicht modelliert	themenbezogen modelliert, standardisiert, endbenutzertauglich
<b>Zustand</b>	redundant, inkonsistent, i. d. R. normalisiert	konsistent modelliert, kontrollierte Redundanzen, denormalisiert
<b>Änderungen</b>	laufend	automatische Fortschreibung, Beständigkeit des einmal übernommenen Datenbestandes
<b>Abfragen</b>	strukturiert, vordefiniert	Ad-hoc-Abfragen für komplexe, ständig wechselnde Fragestellungen, vordefinierte Standardauswertungen

Ziele eines Data Warehouse?

- Informationen für das Management
- Unterstützung von Entscheidungen
- zusammenführen unterschiedlicher Daten aus operativen Anwendungssystemen
- es werden (un-)strukturierte Daten übernommen
- Veränderung, Aggregation der Daten

# Aufbau analytischer Informationssysteme

- **Zentrales DWH** enthält eine von den operativen Systemen isolierte Datenbank
- **Data Mart** ist ein subjektspezifisches oder abteilungsspezifisches DWH; entweder Datenbestände gleichzeitig an mehreren Orten schneller bereitzustellen oder einzelne Fachabteilungen spezifische Daten zu liefern





## Unterschied Data Warehouse & Data Mart

<b>Merkmale</b>	<b>Data Mart</b>	<b>Data Warehouse</b>
<b>Philosophie</b>	anwendungsorientiert	anwendungsneutral
<b>Adressat der Datenbereitstellung</b>	Abteilung	Unternehmen
<b>Vorherrschende Datenbanktechnologie</b>	multidimensional	relational
<b>Granularität (Detaillierungsgrad der gespeicherten Daten)</b>	niedrig	hoch
<b>Datenmenge</b>	niedrig	hoch
<b>Menge historischer Daten</b>	niedrig	hoch
<b>Optimierungsziel</b>	Abfragegeschwindigkeit	Datenmenge
<b>Anzahl</b>	mehrere	eins bzw. sehr wenige
<b>Datenmodell</b>	je nach Data Mart verschieden	unternehmensweit

Ablauf ETL?

- Analyse und Dokumentation operativer und externer Datenquellen
- Extrahieren der ausgewählten Daten
- Transformation operativer Daten
- Bereinigung transformierter Daten
- periodisches Laden der Daten ins DWH

# Extraktion

Unter Extraktion versteht man die Selektion der Daten aus den (zumeist) operativen Datenquellen und ihre anschließende Speicherung in einen Arbeitsbereich des DWH (Staging Area). Hier werden die Daten zwischengespeichert und transformiert bzw. bereinigt und im Anschluss in das DWH übertragen.

Wann wird die Extraktion durchgeführt?

- Periodisch
- Anfrage
- Ereignisgesteuert (wenn z.B. Werte unterschritten werden)
- Sofort (DWH hat die gleiche Aktualität wie die operativen Systeme)



# Transformation

Transformation findet in der s.g. Staging Area statt und bereinigt bzw. transformiert die Quelldaten in das gewünschte Zielformat.

## Qualitätsmängel der Quelldaten

- inkorrekte Daten (Eingabe-/Verarbeitungsfehler)
- logisch widersprüchliche Daten
- unvollständige, ungenaue, zu grobe Daten
- redundante Daten
- uneinheitliche Daten
- veraltete Daten
- irrelevante Daten
- unverständliche Daten (wegen qualitativ mangelhafter Metadaten)

**Verfahren:**

- Bereinigung
- Harmonisierung (betriebswirtschaftlich: Codierung, Schlüssel, Attribute)
- Verdichtung (für Analysezwecke aggregiert werden → Regionalzahlen usw.)
- Anreicherung (Ergänzung um errechnete Kennzahlen)

Bereinigung - Was ist zu beachten?

- Muss-Feld?
- Plausibilitätsprüfung bei der Eingabe?
- Wird das Feld gemäß der ursprünglichen Bestimmung genutzt?
- Wurde das Datenfeld nachträglich aufgenommen? (fehlt bei älteren Daten dann)
- Existieren konkrete Änderungspläne für die operativen Daten?

## Daten-Mängel

Es werden **syntaktische** und **semantische** Mängel unterschieden.

<b>Bereinigung</b>	1. Klasse: Automatische Erkennung und Korrektur	2. Klasse: Automatische Erkennung und manuelle Korrektur	3. Klasse: Manuelle Erkennung und manuelle Korrektur
<b>Syntaktische Mängel</b>	Formatanpassungen	Erkennbare Formatinkompatibilitäten	–
<b>Semantische Mängel</b>	Fehlende Datenwerte	Ausreißerwerte, unstimmgige Werte	Unerkannte semantische Fehler in operativen Datenquellen

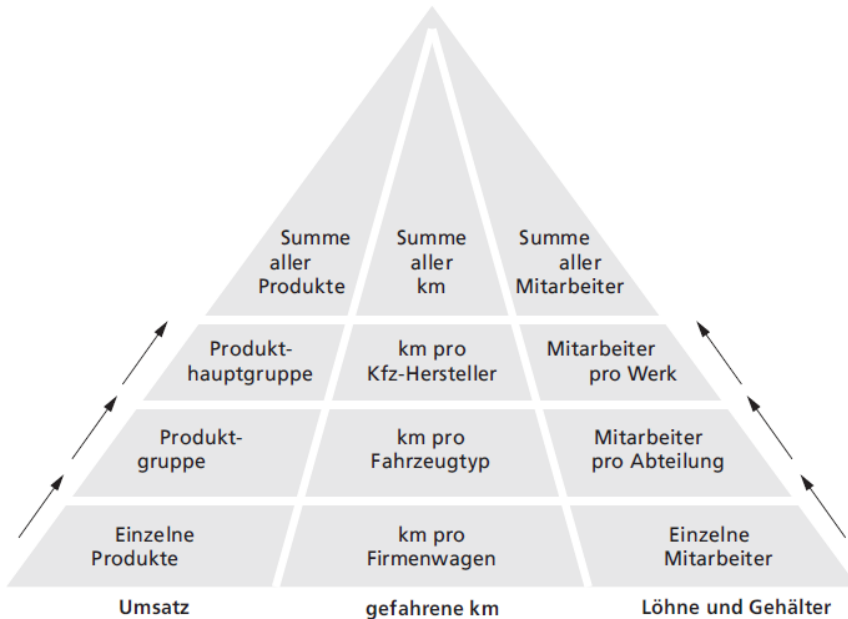


Harmonisierung - Was wird getan?

- Vereinheitlichung unterschiedlicher Codierungen (z.B. männlich, m, 1, weiblich, w, 0)
- Synonyme und Homonymen (unterschiedliche Attributnamen mit gleicher Bedeutung z.B. vorname, vname, firstname)
- Harmonisierung von Schlüsseln und Kennzahlen

Verdichtung

Es werden Daten im DWH (Staging Area) auf verschiedenen Stufen aufsummiert.



Anreicherung

Es werden Berechnungen durchgeführt, die zusammen mit den übrigen analytischen Daten gespeichert werden, d.h. es werden konkrete Kennzahlen ermittelt basierend auf einem gegebenen Kennzahlensystem (z.B. DuPont-Schema  $\rightarrow$  ROI)

Vorteile der Anreicherung sind:

- kürzere Antwortzeiten bei späteren Anfragen da es sich um vorberechnete Werte handelt
- hohe Datenkonsistenz, da sie nach einem einheitlichen Algorithmus berechnet werden

Laden

Es wird unterschieden zwischen:

- Initialem Füllen aus den operativen Datenbanken
- Zyklischer Aktualisierung, neue Werte werden ergänzt, alte archiviert

Wenn die Daten zyklisch übernommen werden kann dies als:

- Kompletter Abzug (einfach aber zeitaufwendig)
- jeweilige Änderungen (geringe Datenmenge, aufwendig das Delta zu ermitteln, nur der letzte Stand wird ermittelt)
- Auswahl protokollierter Datenbanktransaktionen (auch Änderungen innerhalb des Deltas erfasst werden)

geschehen.



Metadaten

Metadaten sind Daten über Daten und enthalten Hintergrundinformationen über die im DWH gespeicherten Werte. Sie geben Aufschluss über:

- Umfang der verfügbaren Daten
- Datenstruktur und Beziehungen (Relationen)
- Herkunft der operativen Daten
- Speicherort im DWH
- Formate
- Zugriffsberechtigungen

In der Metadatenbank wird festgehalten:

- Welche Daten woher kommen
- Wie sie aufbereitet und verdichtet werden
- Wo sie gespeichert werden
- Welcher Anwender auf welche Daten Zugriff erhält

Archivierung

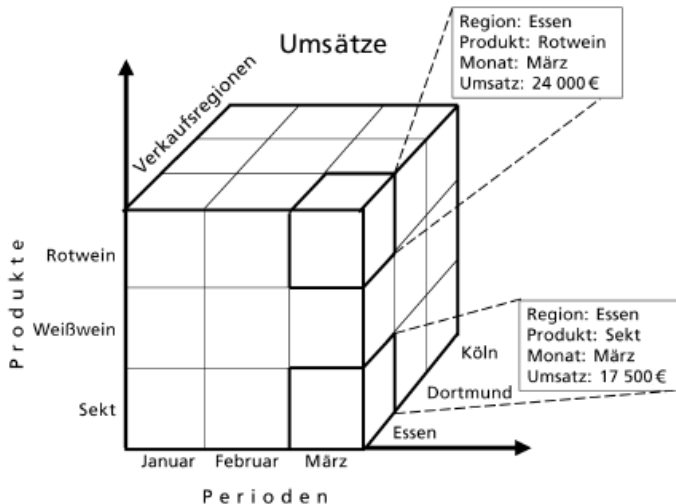
Es wird zwischen der:

- Datenarchivierung (auslagern auf Offlinedatenträger nach Zeit)
- Datensicherung (Dienen zur Wiederherstellung des DWH)

unterschieden.

OLAP

Der Begriff steht für Online Analytical Processing und umfasst alle Formen der *mehrdimensionalen Datenanalyse*. Im Focus stehen betriebswirtschaftliche Kennzahlen. Die Mehrdimensionalität wird durch s.g. *Datenwürfel* veranschaulicht.



Unterschied OLAP/TLTP

<b>Merkmal</b>	<b>OLTP Operative IT-Systeme</b>	<b>OLAP Data Warehouse</b>
Typische Datenstruktur	flache Tabellen	multidimensionale Strukturen
Datenmanipulation	<ul style="list-style-type: none"><li>– Erfassung einzelner Datensätze</li><li>– Update/Einfügen von Datensätzen zulässig</li></ul>	<ul style="list-style-type: none"><li>– spezifische Analyse großer Datenmengen</li><li>– nur lesender Zugriff möglich</li></ul>
Aktualisierung durch	Transaktionen	Batch-Läufe
Datenquelle	intern	intern und extern
Datenmenge pro Transaktion	klein	sehr umfangreich
Typische Betrachtungsebene	detailliert	aggregiert
Systemlast	vorhersehbar	ad hoc
geforderte Antwortzeit	2 – 3 Sekunden	mehrere Sekunden bis Minuten



## Anforderungen an OLAP

- Mehrdimensionale konzeptionelle Sicht auf die Daten (Zeit, Produktgruppe, Region, Person, usw.)
- Transparenz (Anwender müssen keine technischen Details kennen)
- Zugriffsmöglichkeiten (auf möglichst viele heterogene und interne/externe Datenquellen)
- Stabile Antwortzeiten (möglichst schnell und vor allem gleichbleibend)
- Client-/Server-Architektur
- Gleichrangige Dimensionen
- Dynamische Handhabung dünn besetzter Matrizen (effiziente Speicherung trotz Lücken)
- Mehrbenutzerfähigkeit
- Unbeschränkt dimensionsübergreifende Operationen
- Intuitive Datenanalyse
- Flexibles Berichtswesen (Dokumentation in Form von Berichten und Grafiken)
- Unbegrenzte Anzahl von Dimensionen/Aggregationsstufen

**Kritik:** Die unscharfe Trennung zwischen fachlich-konzeptionellen Anforderungen und technischer Realisierung.

Alternativ **FASMI**:

- Fast (Antwortzeit max. 20s)
- Analysis (Anwender ohne technisches Wissen müssen auswerten können)
- Shared (Mehrbenutzer)
- Multidimensional
- Information (sämtliche benötigten Informationen können geliefert werden)



# Mehrdimensionalität

Die Anzahl der Dimensionen lässt sich mit der Fakultät der Dimensionen berechnen:

$$2 \text{ Dim} = 1 \times 2 = 2 \text{ Sichten}$$

$$3 \text{ Dim} = 1 \times 2 \times 3 = 6 \text{ Sichten}$$

$$4 \text{ Dim} = 1 \times 2 \times 3 \times 4 = 24 \text{ Sichten}$$

Die verschiedenen Betrachtungsmöglichkeiten werden auch als **Slice and Dice** bezeichnet. *Slice* bedeutet das Herausschneiden von Scheiben aus dem Würfel. *Dice* bedeutet die Bildung von kleinen Würfeln aus dem Gesamtwürfel zur Einschränkung auf einen Wert bzw. Wertebereich.

Mittels **Drill down** ist es möglich von einer bestehenden Verdichtungsebene auf eine detaillierte Ebene zu wechseln. **Drill up** wechselt von einer Ebene auf eine verdichtere Ebene. **Drill across** ermöglicht zu einem anderen Wert auf der selben Ebene zu wechseln.

Wie können Daten verdichtet werden?

Wie die Daten verdichtet werden können hängt unmittelbar vom Typ ab.

- **Additive Daten** lassen sich beliebig aufsummieren (Umsatz in Kombination mit Produkten und Regionen)
- **Semiadditive Daten** lassen sich nicht über alle Dimensionen aufaddieren (z.B. bei Zeiträumen und Lagerbeständen)
- **Nichtadditive Daten** lassen keine sinnvolle Aufsummierung zu (Anteilswerte)



# Modellierung

Die **multidimensionale Modellierung** ist in der Lage die konzeptionellen Datenstrukturen in physische Datenbank-/Speicherstrukturen umzusetzen. Im Falle eine *dünn besetzten Matrix* wird Speicherplatz vergeudet, da für nicht existente Werte gleich viel Speicher verbraucht wird, als wenn der Wert existiert.

Die **relationale Modellierung** setzt die mehrdimensionale Datenstruktur in einer relationalen Datenbankstruktur ab. Es wird dabei zwischen:

- Fakten (quantitative Kerndaten)
- Dimension (beschreibende Daten, kann beliebig viele Ausprägungen haben)

unterschieden.

Star-Schema

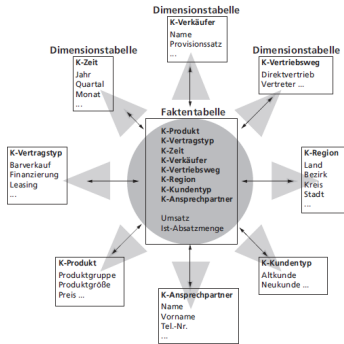
Für jede Dimension wird eine Tabelle eingerichtet. Die Dimensionstabellen sind nicht miteinander verknüpft, sondern stehen nur über die Faktentabelle miteinander in Beziehung. Der Primärschlüssel in der Faktentabelle setzt sich zusammen aus den Primärschlüsseln aller Dimensionstabellen.

Vorteil:

- einfaches, intuitive Datenmodell
- benötigt nur wenige Join-Operationen
- wenige Tabellen benötigt

Nachteil:

- durch die große Anzahl der Verknüpfungsmöglichkeiten kann es zu Performan-  
ceproblemen kommen





## Snowflake-Schema

Dieses Schema ist eine Erweiterung des Star-Schemas und wird durch die Normalisierung entwickelt.

Vorteile:

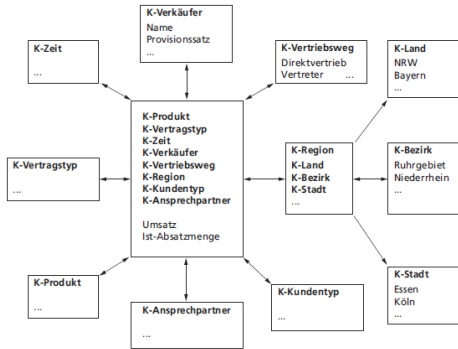
- erleichtert die Aggregation
- keine Redundanzen

Nachteile

- höhere Anzahl JOINS
- höhere Anzahl Tabellen
- komplexere SQL-Abfragen



## Achtung: Rückseite von # 26 ist zu voll.





Was ist Data-Mining?



## Kartenübersicht ANS08

#	Karte	Notizen
1	Definition Data Warehouse	
2	Worin besteht der Unterschied zwischen operativen & analytischen Daten?	
3	Ziele eines Data Warehouse?	
4	Aufbau analytischer Informationssysteme	
5	Unterschied Data Warehouse & Data Mart	
6	Ablauf ETL?	
7	Extraktion	
8	Wann wird die Extraktion durchgeführt?	
9	Transformation	
10	Qualitätsmängel der Quelldaten	
11	Bereinigung - Was ist zu beachten?	
12	Daten-Mängel	
13	Harmonisierung - Was wird getan?	
14	Verdichtung	
15	Anreicherung	
16	Laden	
17	Metadaten	
18	Archivierung	
19	OLAP	
20	Unterschied OLAP/TLTP	
21	Anforderungen an OLAP	
22	Mehrdimensionalität	
23	Wie können Daten verdichtet werden?	
24	Modellierung	
25	Star-Schema	
26	Snowflake-Schema	
27	Was ist Data-Mining?	