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

Merkmale	Data Mart	Data Warehouse
Philosophie	anwendungsorientiert	anwendungsneutral
Adressat der Datenbereitstellung	Abteilung	Unternehmen
Vorherrschende Datenbanktechnologie	multidimensional	relational
Granularität (Detaillierungsgrad der gespeicherten Daten)	niedrig	hoch
Datenmenge	niedrig	hoch
Menge historischer Daten	niedrig	hoch
Optimierungsziel	Abfragegeschwindigkeit	Datenmenge
Anzahl	mehrere	eins bzw. sehr wenige
Datenmodell	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?
- Wir 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.

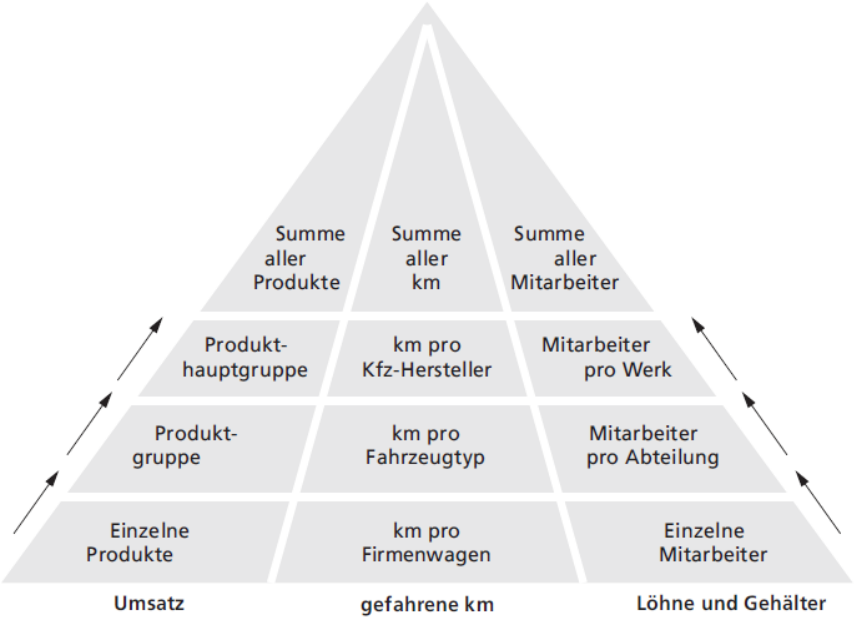
Bereinigung	1. Klasse: Automatische Erkennung und Korrektur	2. Klasse: Automatische Erkennung und manuelle Korrektur	3. Klasse: Manuelle Erkennung und manuelle Korrektur
Syntaktische Mängel	Formatanpassungen	Erkennbare Formatinkompatibilitäten	–
Semantische Mängel	Fehlende Datenwerte	Ausreißerwerte, unstimmige 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 gegeben Kennzahlensystem (z.B. DuPont-Schema → 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

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	