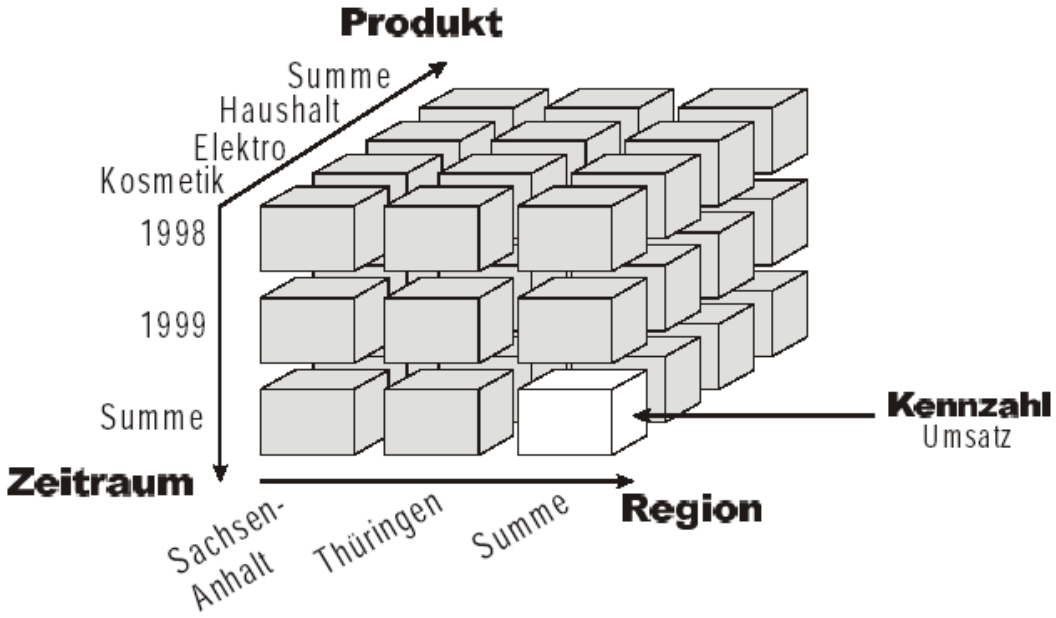
Data Engineer

Requirement Engineering
Aufgaben,

Multidimensionales Datenmodell

- **Datenmodell zur Unterstützung der Analyse**
  - Fakten und Dimensionen
  - Klassifikationsschema
  - Würfel
  - Operationen



- **Notationen zur konzeptuellen Modellierung**
- **Relationale Umsetzung**
  - Star-, Snowflake-Schema
- **Multidimensionale Speicherung**

Eine Einführung in OLAP (Online Analytical Processing)

OLAP Überblick

- Einführendes Beispiel
- Begriffsdefinition
- Charakteristika
- Architektur
- Funktionalität
- OLAP & SQL
  (insb. ROLLUP & CUBE)

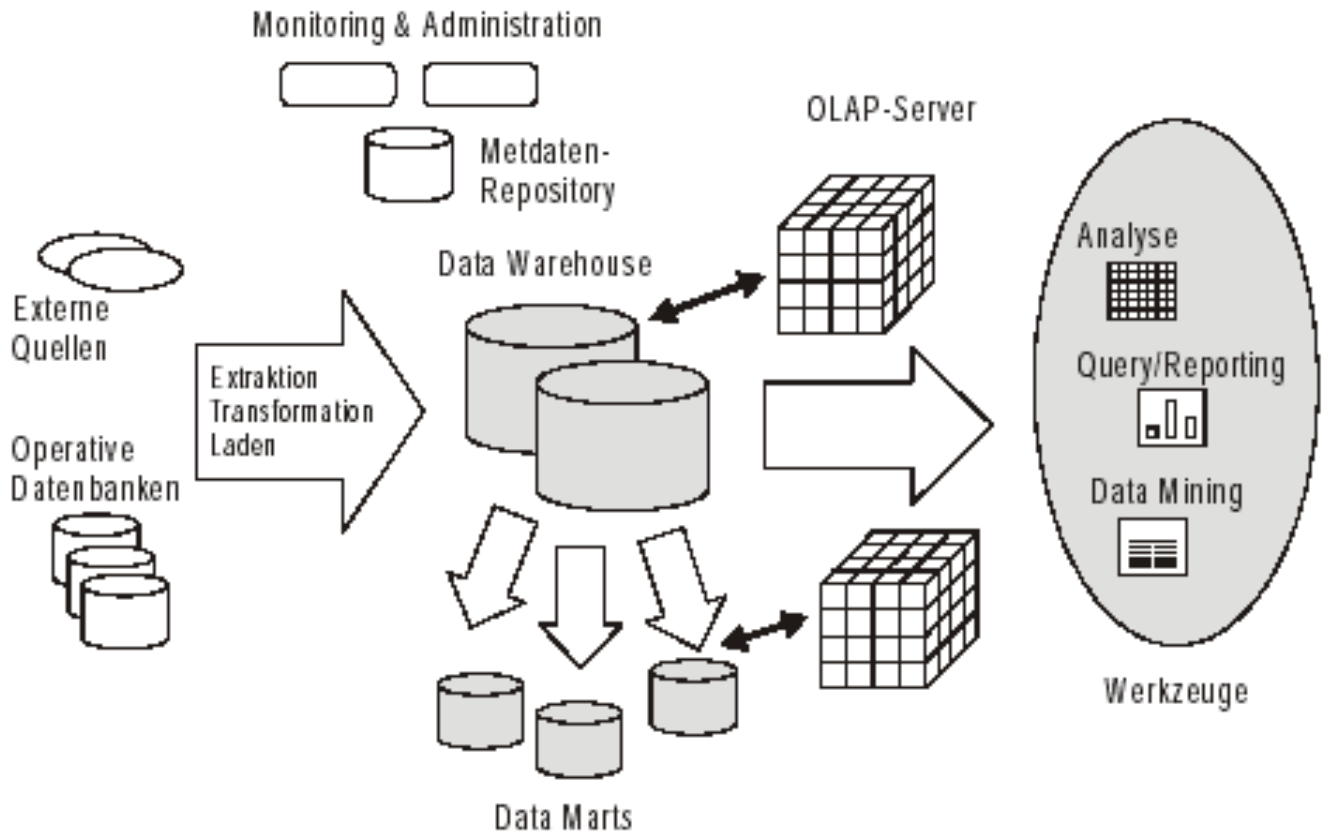
Warum?

- Daten einer Firma verfügbar machen für Entscheidungsprozesse
  - Umsetzung schwierig

- neue Konzepte notwendig zur analytischen Informationsverarbeitung
  - OLAP
  - Data Warehousing
  - Data Mining
  - Process Mining
  - Task Mining

OLAP Einleitung

DSS: Decision Support System



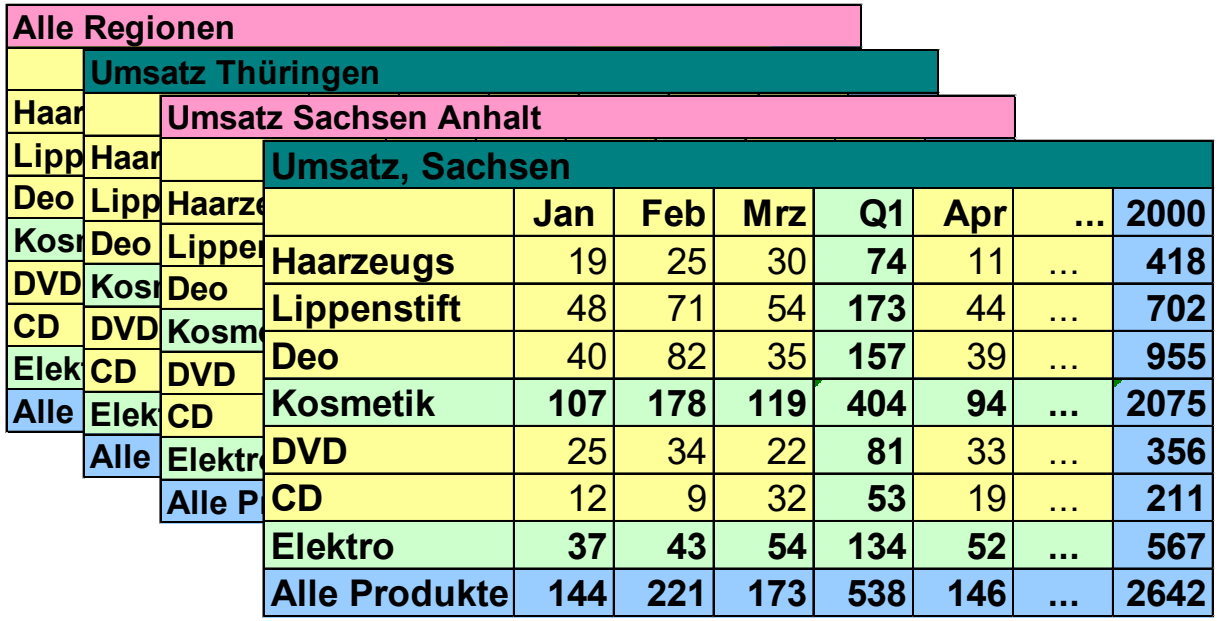
Einführungsbeispiel

Umsatz pro Zeit und Produkt

Umsatz							
	Jan	**Feb**	**Mrz**	**Q1**	**Apr**	**...**	**2000**
Haarzeugs	33	55	56	144	18	...	760
Lippenstift	72	136	117	325	74	...	1338
Deo	85	128	99	312	92	...	1662
Kosmetik	**190**	**319**	**272**	**781**	**184**	**...**	**3760**
DVD	55	69	99	223	84	...	1051
CD	22	17	47	86	39	...	493
Elektro	**77**	**86**	**146**	**309**	**123**	**...**	**1544**
Alle Produkte	**267**	**405**	**418**	**1090**	**307**	**...**	**5304**

Einführungsbeispiel

Umsatz pro Zeit, Produkt und Region



	Jan	Feb	Mrz	Q1	Apr	...	2000
Haarzeugs	19	25	30	74	11	...	418
Lippenstift	48	71	54	173	44	...	702
Deo	40	82	35	157	39	...	955
Kosmetik	107	178	119	404	94	...	2075
DVD	25	34	22	81	33	...	356
CD	12	9	32	53	19	...	211
Elektro	37	43	54	134	52	...	567
Alle Produkte	144	221	173	538	146	...	2642

Einführungsbeispiel

Umsatz, Sachsen Anhalt, Telefon

Umsatz, S-A, Homepage

Umsatz,Sachsen Anahlt , Fax

Umsatz,S-A, Alle Distributionskanäle

	Jan	Feb	Mrz	Q1	Apr	...	2000
Haar...	11	26	22	59	4	...	299
Lippenstift	16	54	49	119	18	...	480
Deo	29	34	35	98	18	...	402
Kosmetik	56	114	106	276	40	...	1181
DVD	19	18	53	90	27	...	482
CD	6	5	12	23	15	...	202
Elektronik	25	23	65	113	42	...	684
Alle Produkte	81	137	171	389	82	...	1865

Umsatz, Alle Regionen, Telefon

Umsatz, Alle Regionen, Telefon,Homepage

Umsatz, Alle Regionen, Fax

Umsatz, Alle Regionen, Telefon, Alle Distributionskanäle

	Jan	Feb	Mrz	Q1	Apr	...	2000
Haar...	33	55	56	144	18	...	760
Lippenstift	72	136	117	325	74	...	1338
Deo	85	128	99	312	92	...	1662
Kosmetik	190	319	272	781	184	...	3760
DVD	55	69	99	223	84	...	1051
CD	22	17	47	86	39	...	493
Elektronik	77	86	146	309	123	...	1544
Alle Produkte	267	405	418	1090	307	...	5304

Umsatz, Sachsen, Telefon

Umsatz, Sachsen, Homepage

Umsatz Sachsen, FAX

Umsatz, Sachsen, Alle Distributionskanäle

	Jan	Feb	Mrz	Q1	Apr	...	2000
Haar...	19	25	30	74	11	...	418
Lippenstift	48	71	54	173	44	...	702
Deo	40	82	35	157	39	...	955
Kosmetik	107	178	119	404	94	...	2075
DVD	25	34	22	81	33	...	356
CD	12	9	32	53	19	...	211
Elektronik	37	43	54	134	52	...	567
Alle Produkte	144	221	173	538	146	...	2642

Umsatz, Thüringen, Telefon

Umsatz, Th, Homepage

Umsatz,Thüringen , Fax

Umsatz, Th, Alle Distributionskanäle

	Jan	Feb	Mrz	Q1	Apr	...	2000
Haar...	3	4	4	11	3	...	43
Lippenstift	8	11	14	33	12	...	156
Deo	16	12	29	57	35	...	305
Kosmetik	27	27	47	101	50	...	504
DVD	11	17	24	52	24	...	213
CD	4	3	3	10	5	...	80
Elektronik	15	20	27	62	29	...	293
Alle Produkte	42	47	74	163	79	...	797

OLAP

- OLAP erleichtert die Analyse von Kennzahlen unter verschiedenen Gesichtspunkten (Dimensionen)
  - z.B. Produktmanager, Bereichsleiterin
  - Kennzahlen
  - graphische Darstellung (Diagramme)

- Dynamische, multidimensionale Geschäftsanalyse mit Simulationskomponente

Was ist OLAP?

alfatraining

OLAP ist ...

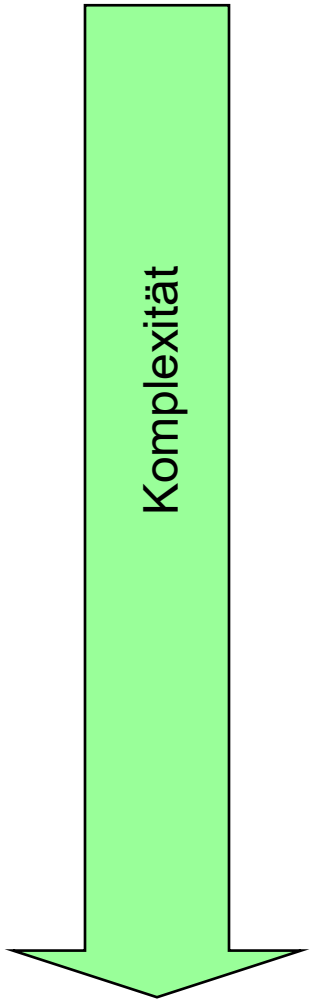
... ein Überbegriff für Technologien, Methoden und Tools zur Ad-hoc-Analyse multidimensionaler Informationen

... bietet verschiedene Sichtweisen

... eine Komponente der entscheidungsorientierten Informationsverarbeitung

Analyse-Datenmodelle

- kategorisches (beschreibendes) Modell
  - statisches Analysemodell zur Beschreibung des gegenwärtigen Zustands
  - Vergleich von historischen mit aktuellen Daten
- exegetisches (erklärendes) Modell
  - zur Erklärung der Ursachen für Zustand durch Nach-vollziehen der Schritte, die ihn hervorgebracht haben (durch einfache Anfragen)
- kontemplatives (bedenkendes) Modell
  - *Simulation* von „What If“Szenarios für vorgegebene Werte oder Abweichungen innerhalb einer Dimension oder über mehrere Dimensionen hinweg
- formelbasiertes Modell
  - gibt Lösungswege vor: ermittelt für vorgegebene Anfangs- und Endzustände, welche Veränderung für welche Kenngröße bzgl. welcher Kenngröße für ange-strebtes Ergebnis notwendig



OLAP Charakteristika*

12 Regeln nach E. F. Codd

- Multidimensionale konzeptionelle Sichten
- funktionale Transparenz
- unbeschränkter Zugriff auf operative und/oder externe Datenquellen
- gleichbleibende Berichtsleistung
- Client-/Server Architektur
- gleichgestellte Dimensionen
- dynamische Behandlung dünn besetzter Datenwürfel
- mehrere Anwender
- unbeschränkte, dimensionsübergreifende Operationen
- intuitive Datenmanipulation
- flexibles Berichtswesen
- unbegrenzte Dimensions- und Aggregationsstufen

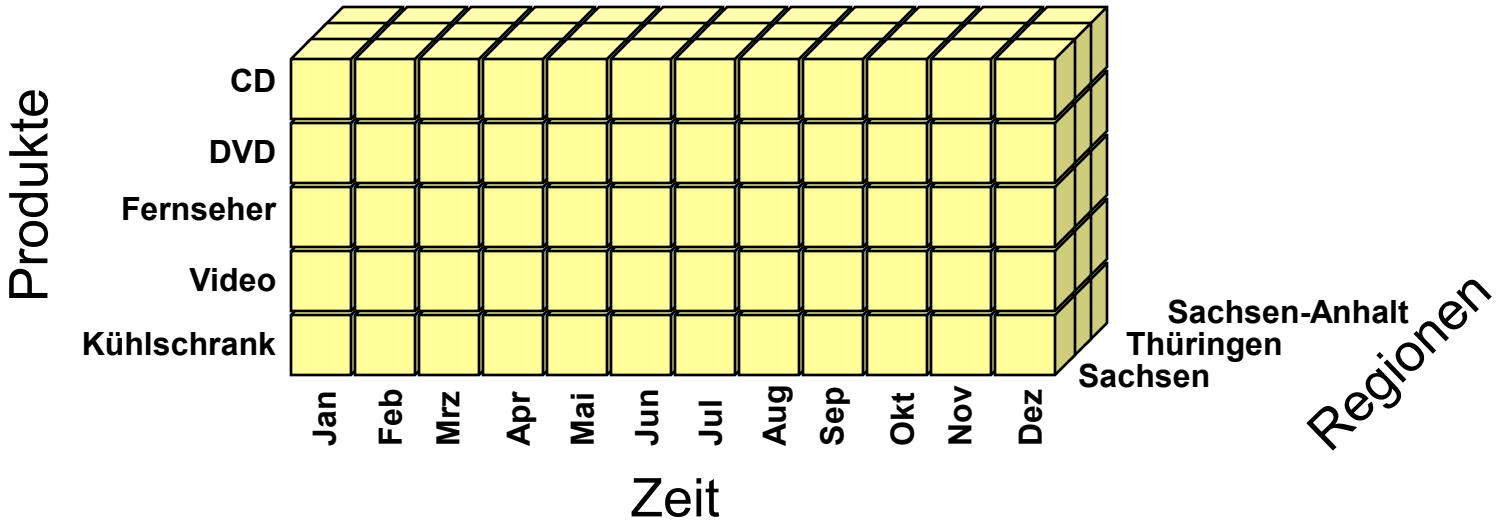
OLAP Charakteristika - FASMI

FASMI = Fast Analysis of Shared Multidimensional Information

- **Fast**: 1-2 Sekunden als Antwortzeit bei einfachen Anfragen bis maximal 20 Sekunden für komplexe Datenanalysen
- **Analysis**: Verfahren und Techniken zu einfachen mathematischen Berechnungen und Strukturuntersuchungen
- **Shared**: Schutzmechanismen für den Zugriff im Mehrbenutzerbetrieb
- **Multidimensional**: Multidimensionale konzeptionelle Sicht auf Informationsobjekte, d.h. freier Zugriff auf einen Datenwürfel und multiple Berichtshierarchien über die Dimensionen

OLAP Charakteristika

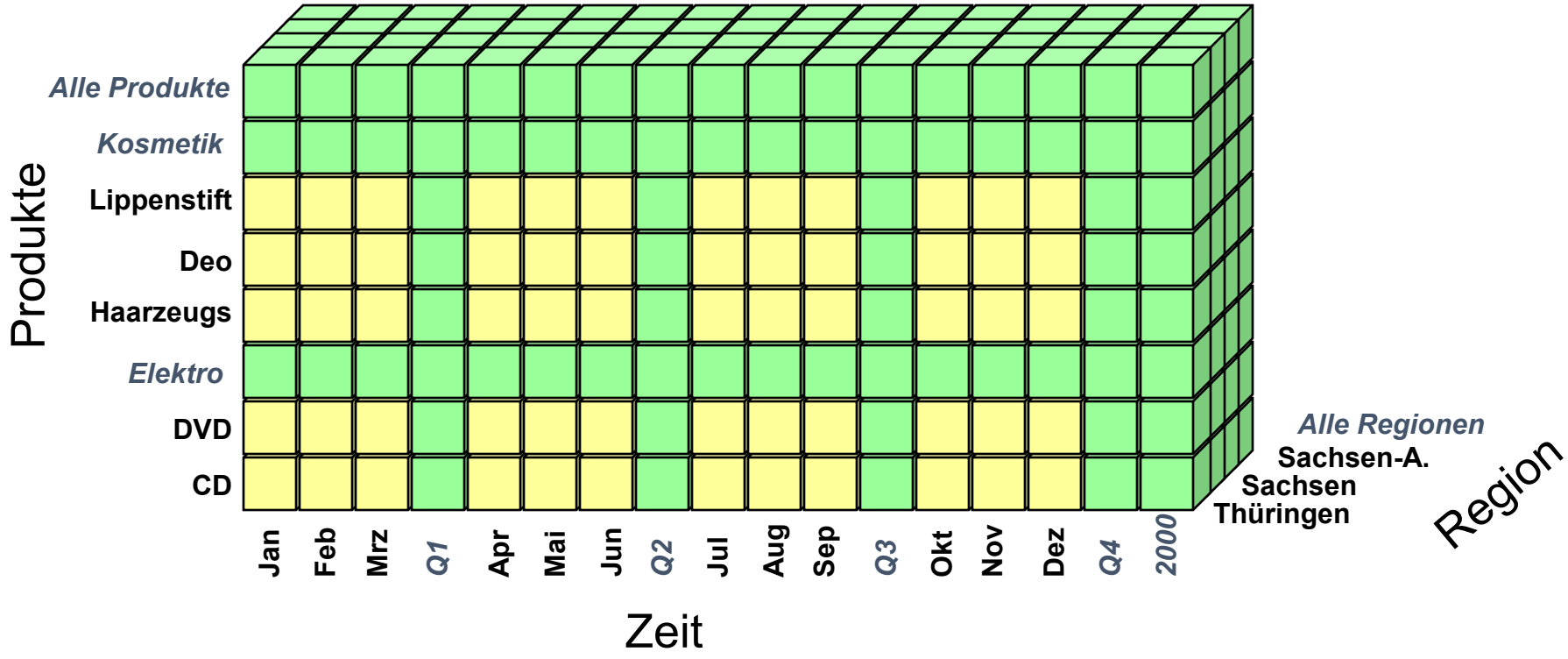
Daten werden über **Dimensionen** beschrieben.



Begriffe: Multidimensionalität, Hypercubes,
Ausprägungen (Members), Zellen

OLAP Charakteristika

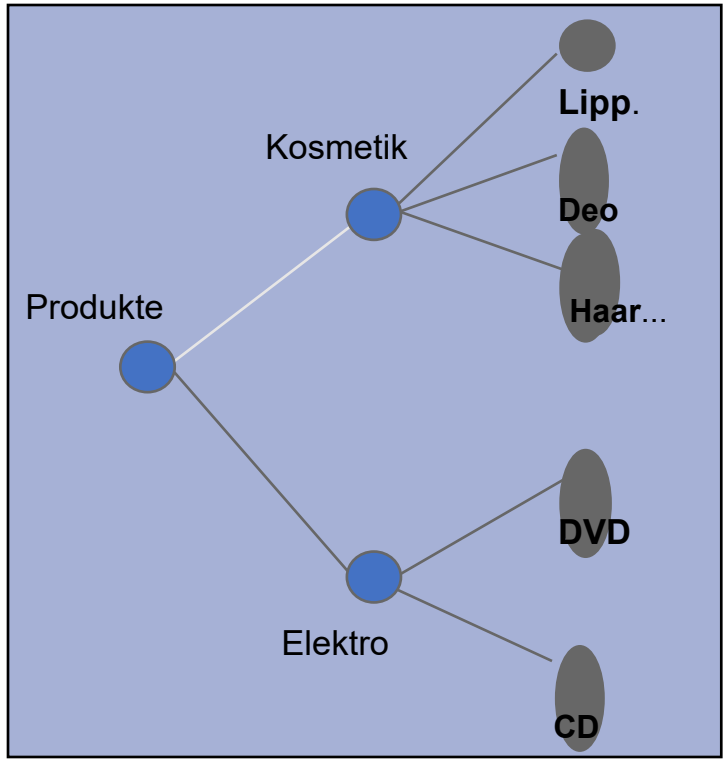
Dimensionen können **Hierarchien** haben.



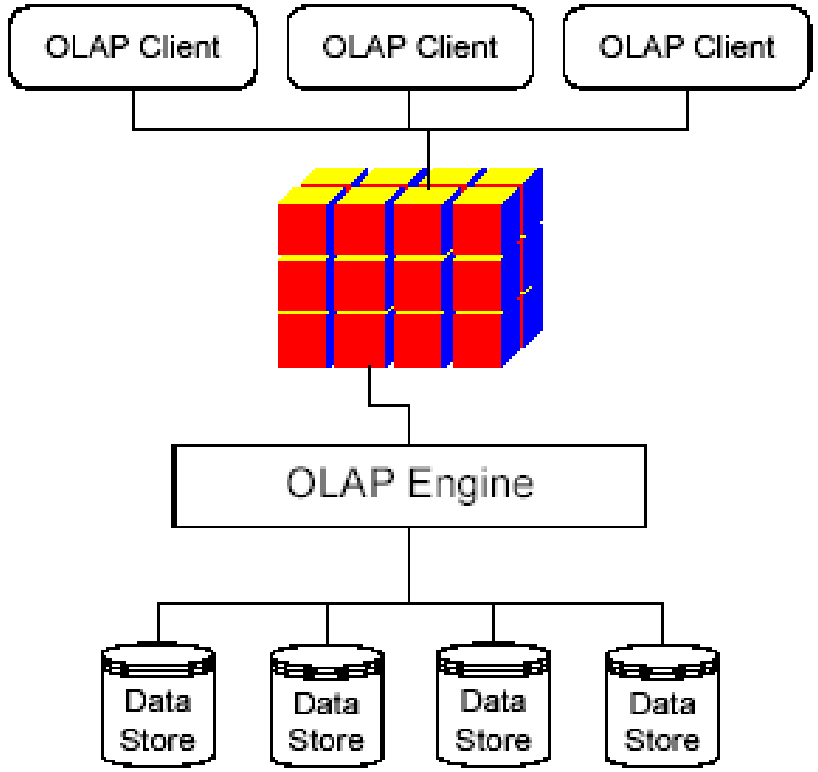
Zu Hierarchien

- Hierarchie
  - **Hierarchische Aufteilung der Dimension**

Dimension Produkte



OLAP Grobarchitektur

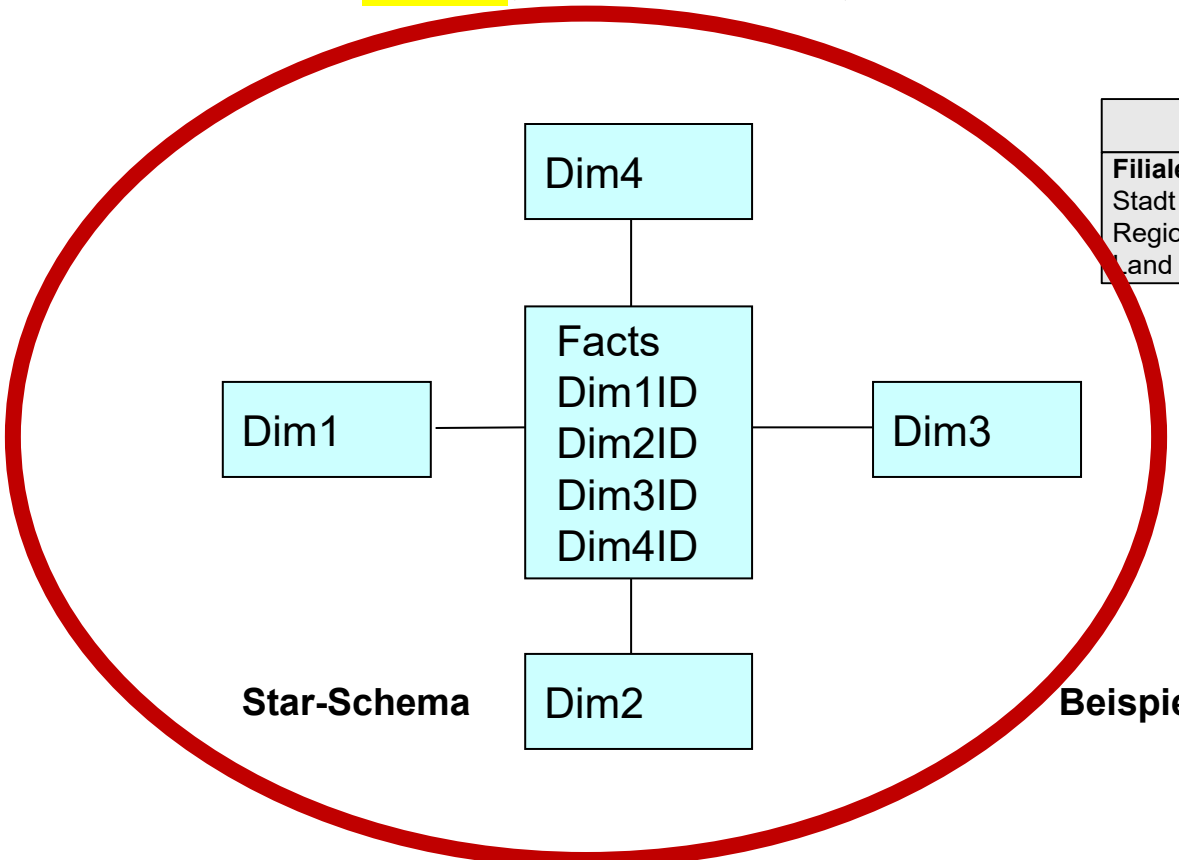


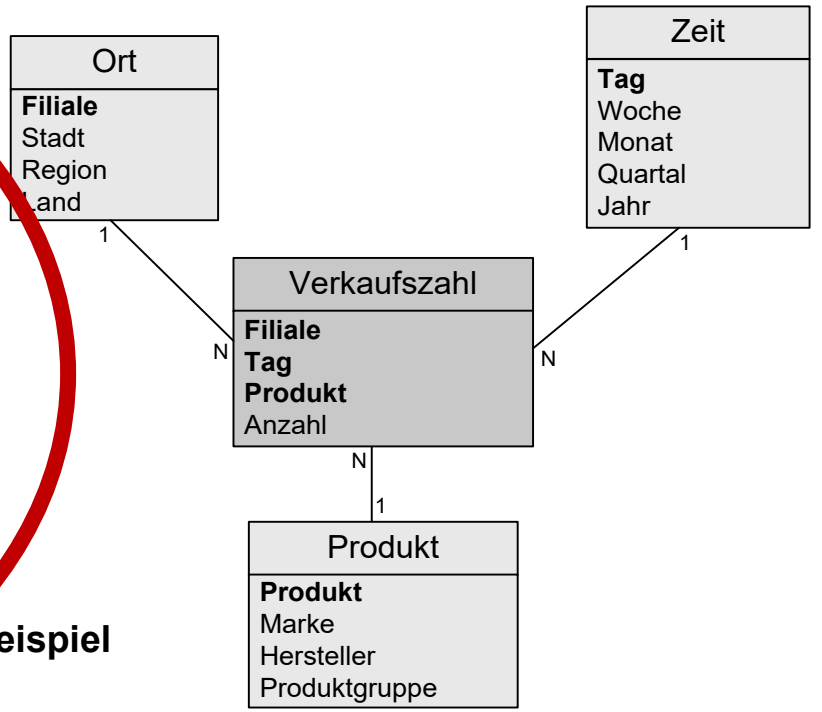
OLAP Architekturkonzepte

- ROLAP = Relational OLAP
  - bei Abbildung in Relationen: möglichst wenig Verlust von Semantik, die im multidimensionalen Modell enthalten
  - Effiziente Übersetzung und Abarbeitung von multidimensionalen Anfragen
  - Einfache Wartung (z.B. Laden neuer Daten)
- MOLAP = Multidimensional OLAP
  - direkte Speicherung multidimensionaler Daten in multidimensionalen DBMS
- HOLAP = Hybrid OLAP
  - Kombiniert Vorteile von relationaler und multidimensionaler Realisierung

Architekturkonzept ROLAP

- SQL zur Datentransformation
- Multidimensionale Datenmodelle werden in 2-dimensionalen Tabellen gespeichert
- **Star-**, Snowflake, Starflake-Schema





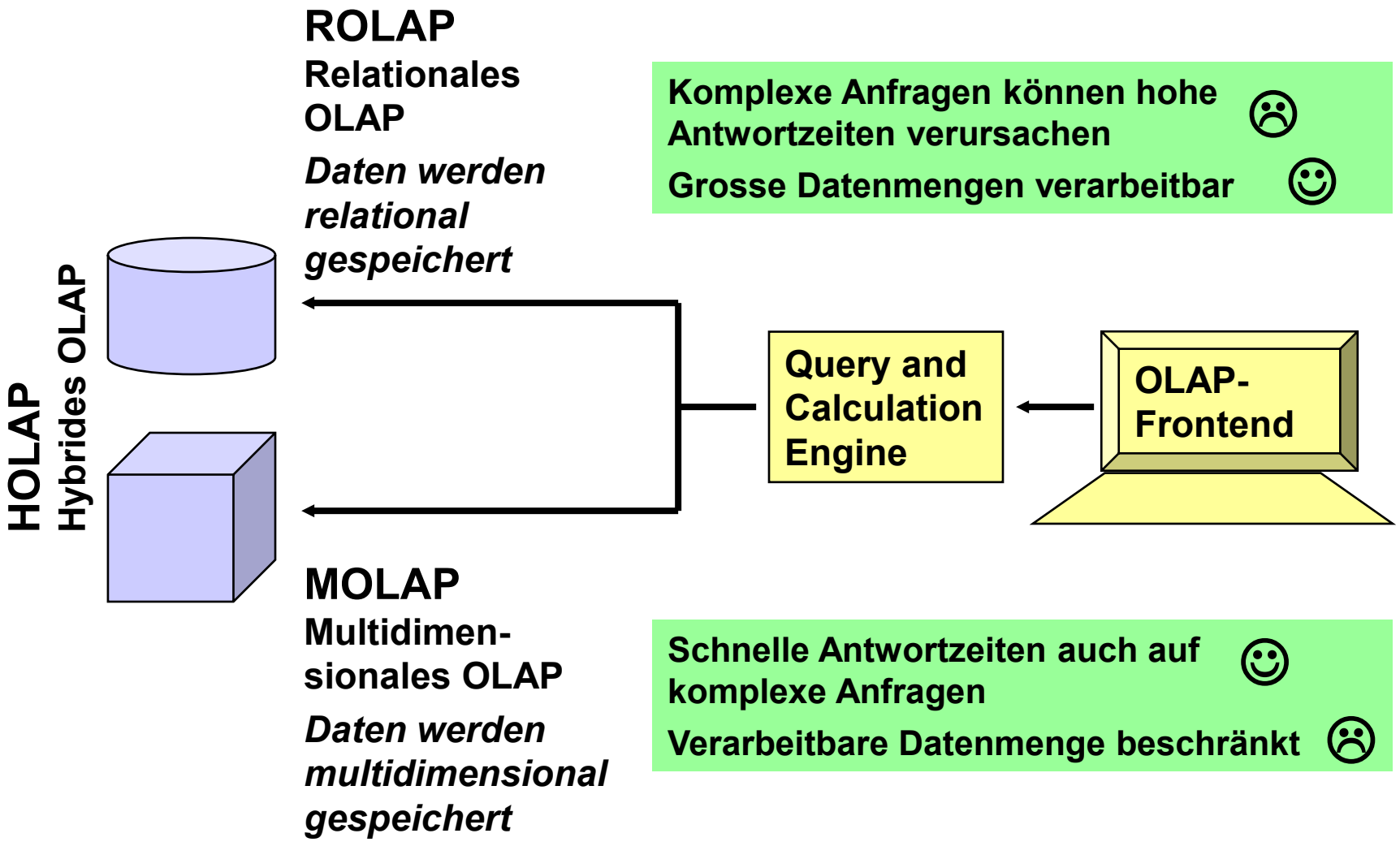
ROLAP - Star-Schema

- erstellen von Fakten- und Dimensionstabellen
- Faktentabelle mit Schlüsseln für Dimensionstabellen
- in Dimensionstabellen stehen relevante Daten
- Redundanz
  - Alternative wäre Snowflake-Schema
  - Dimensionsdaten relativ stabil

Architekturkonzept MOLAP

- Speicherung erfolgt in multidimensionalen Speicher-Arrays
- Ordnung der Dimensionen zur Adressierung der Würfelzellen notwendig
- Klassifikationshierarchien und Aggregation (Echtzeit oder Vorberechnung?)
- optional: Attribute
- Behandlung mehrerer Kenngrößen?
- Single-Cube-Ansatz (Datenbestand in einem Würfel) vs. Multicube-Systeme (mehrere kleinere Würfel)
- Bewertung des Ansatzes:
  - Begrenzte Skalierbarkeit bei Dünnbesetztheit
  - Verbesserung durch Nutzung von Indexierungstechniken

Architekturkonzepte



Unterschiede OLTP/OLAP

Transaktionsorientierte Systeme ***Operative Systeme***	**Auswertungsorientierte Systeme**
OLTP **(Online Transaction Processing)**	**OLAP** **(Online Analytical Processing)**
Häufige, einfache Anfragen	Weniger häufige, komplexe Anfragen
Kleine Datenmengen je Anfrage	Grosse Datenmengen je Anfrage
Operieren hauptsächlich auf aktuellen Daten	Operieren auf aktuellen und historischen Daten
Schneller Update wichtig	Schnelle Kalkulation wichtig
→ Datenbanksystem kann nicht gleichzeitig für OLTP- und für OLAP-Anwendungen optimiert werden	
Paralleles Ausführung von OLAP-Anfragen auf operationalen Datenbeständen könnte Leistungsfähigkeit der OLTP-Anwendungen beeinträchtigen	

OLAP Funktionalität

- Drill Down
  - erhöhen des Detaillierungsgrades, d.h. Navigation von den verdichteten Daten zu den detaillierten
- Roll Up
  - invers zu Drill Down
  - Aggregration entlang des Konsolidierungspfades
- Pivotieren / Rotieren
  - Betrachten aus unterschiedlichen Perspektiven (vertauschen der Dimensionen um seine Achsen)
- Slice & Dice
  - Einschränken des Analyseblickwinkels (Erzeugung von Scheiben oder Teilwürfeln)

OLAP Funktionen

Die multidimensionalen **Daten können am Bildschirm flexibel präsentiert werden**.

Umsatz, Jun				
	SA	Th	S	Alle Regionen
Kosmetik	200	100	50	350
Elektro	120	110	30	260
Alle Produkte	320	210	80	610

Drill-down

Roll-up

Umsatz, Jun				
	CH	D	A	Alle Regionen
Kosmetik	200	100	50	350
DVD	100	75	20	180
CD	20	35	10	80
Elektro	120	110	30	260
Alle Produkte	320	210	80	610

OLAP-Anbieter und -Produkte

1. SSAS
2. Power BI
3. Oracle (Express)
4. Cognos (PowerPlay)
5. MicroStrategy (MicroStrategy)
6. Microsoft (OLAP-Server)
7. Business Objects (Business Objects)
8. Tableau

Gartner Report 2023

. . .

alfatraining

Integration von OLAP und Data Mining und anderen Methoden der entscheidungsorientierten Informationsverarbeitung

Stärkere Beteiligung der **akademischen Welt** an der OLAP-Weiterentwicklung

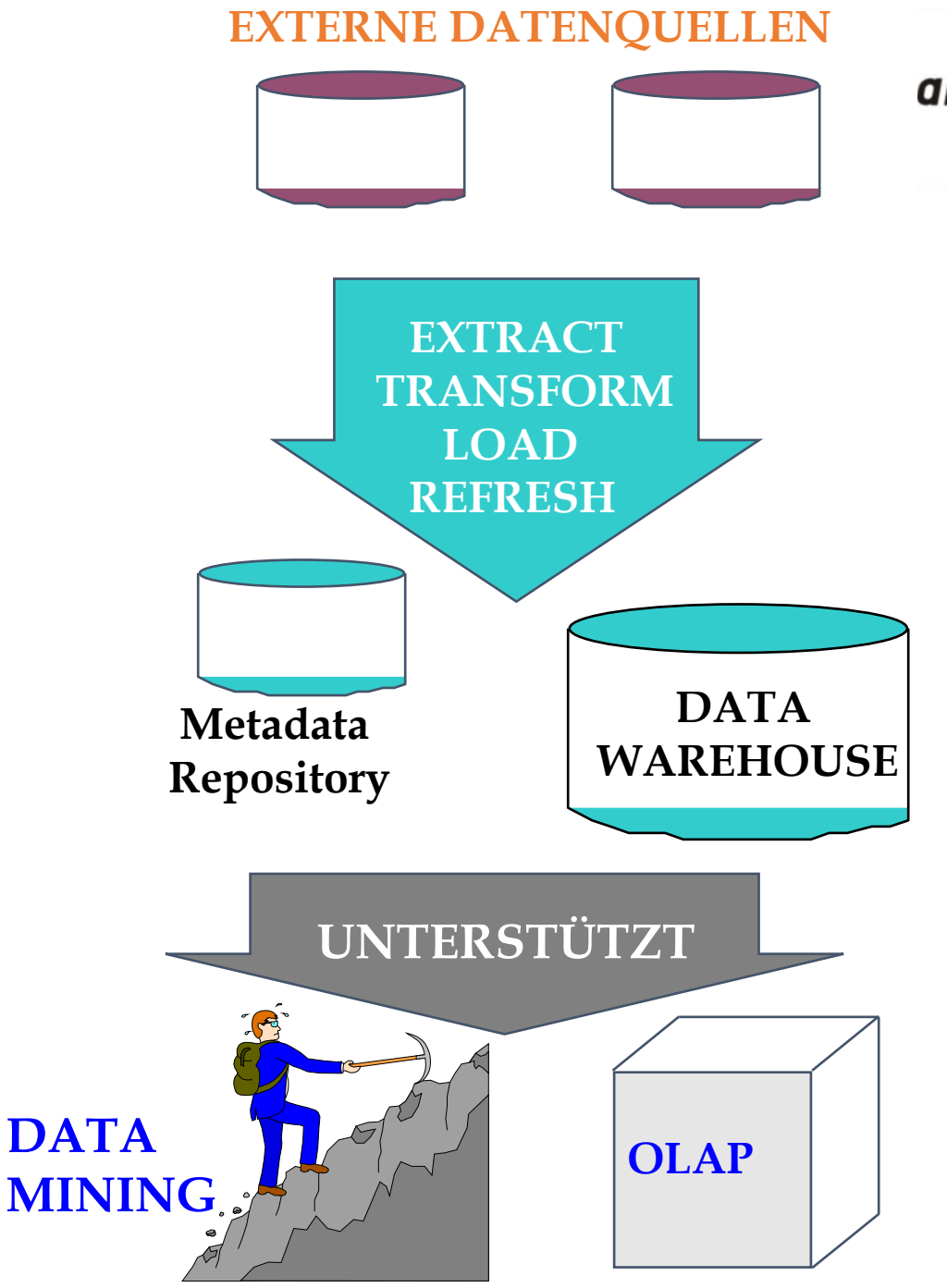
Weiterentwicklung und rasche Verbreitung von **Web-OLAP**

Auf spezifische **vertikale oder horizontale Märkte** ausgerichtete OLAP-Applikationen

Weiterentwicklung der **technischen Konzepte** (z.B. optimale Verteilung von Speicherung und Kalkulation, verbesserte Metadatenverwaltung, ...)

Data Warehousing

- *Data Warehouse* Integrierter Datenbestand, der sich über lange Zeit-perioden erstreckt, oft mit zusätzlicher Information angereichert
- Mehrere Gigabytes bis Terabytes
- Interaktive Antwortzeiten für komplexe Anfragen erwartet; ad-hoc Updates nicht üblich



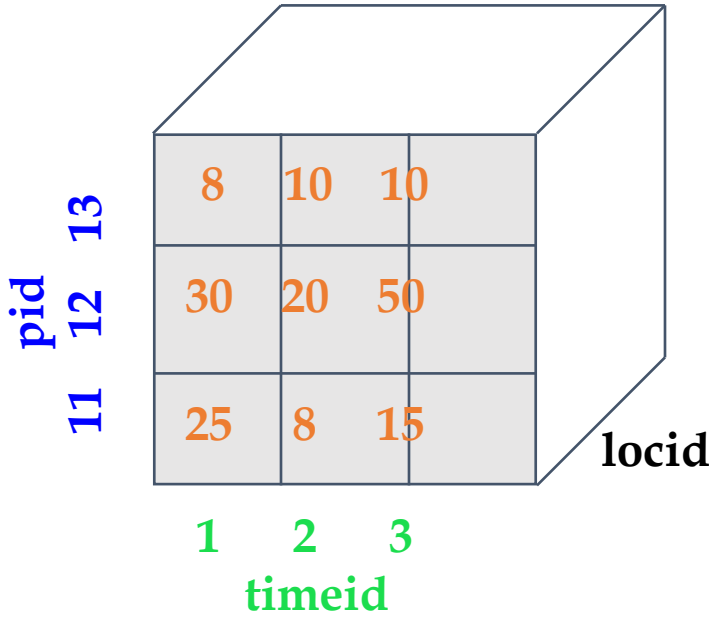
Aufgaben beim Warehousing

- Semantische Integration: Beim Bezug von Daten aus unter-schiedlichen Quellen, sind alle Arten von Heterogenitäten zu beseitigen, z.B.
  - Verschiedene Währungen und Maßeinheiten
  - Unterschiede in den Schemas
  - Verschiedene Wertebereiche
- Heterogene Quellen: Zugriff auf Daten in unterschiedlichsten Formaten und Repositories
  - Möglichkeiten der Replikation ausnutzen
- Load, Refresh, Purge:
  - Daten müssen ins Warehouse geladen werden (Load)
  - Daten müssen periodisch aktualisiert werden (Refresh)
  - Veraltete Daten müssen entfernt werden (Purge)
- Metadata-Management: Verwaltung der Informationen über Daten im Warehouse (Quellen, Ladezeit, Konsistenz-anforderungen etc.)

Multidimensionales Daten Model

- Sammlung von numerischen Größen, die von einer Menge von Dimensionen abhängen.
  - Z.B. Größe **Verkauf**, mit 3 Dimensionen:
    - **Produkt** (Schlüssel: pid)
    - **Ort** (locid)
    - **Zeit** (timeid).

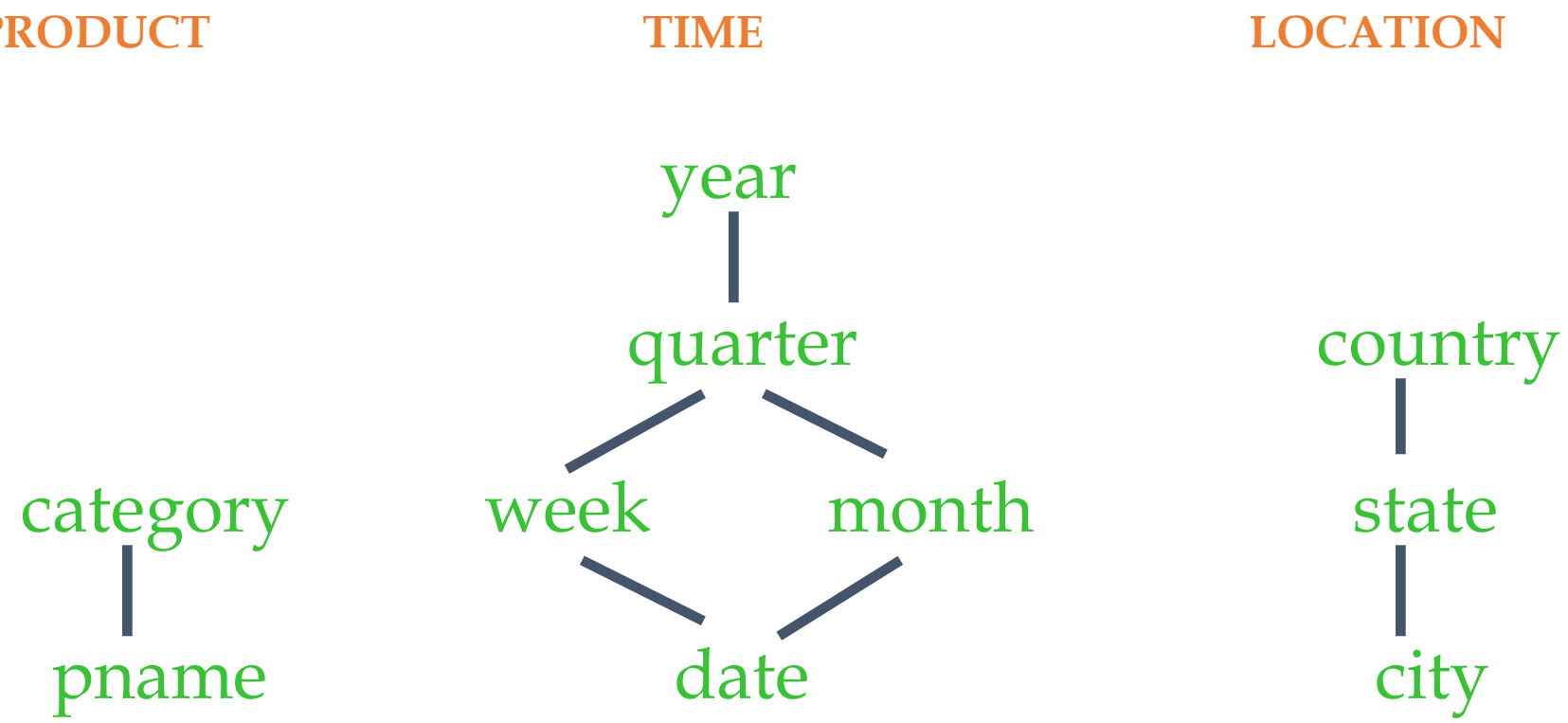
Beispiel mit Slice locid=1



pid	timeid	locid	sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35

Hierarchien in Dimensionen

- In jeder Dimension kann die Menge der Werte in Hierarchien organisiert sein



MOLAP vs. ROLAP

- MOLAP
  Physische Speicherung multidimensionaler Daten in einem (disk-residenten, persistenten) Array gespeichert
- ROLAP
  Physische Speicherung multidimensionaler Daten in Relationen
- Fakten-Tabelle
  Hauptrelation, die Dimensionen mit einer Größe verbindet
  Beispiel:
  Sales (pid, timeid, locid, sales)
- Dimensionen-Tabelle
  Assoziiert mit einer Dimension, enthält zusätzliche Attribute
  Beispiel:
  Products (pid, pname, category, price)
  Locations (locid, city, state, country)
  Times (timeid, date, week, month, quarter, year, holiday_flag)

Fakten-Tabellen sind viel breiter als Dimensionen-Tabellen und größer

Data Engineer

Grundlagen DWH
Umgang und Verarbeitung allen Arten von Daten

Überblick

- Historie
- Funktionen
- Architektur
- Data Warehouse
- OLAP
- Data Mining

Historie

- Wurzeln
  - 60er Jahre: Executive Information Systems (EIS)
    - qualitative Informationsversorgung von Entscheidern
    - kleine, verdichtete Extrakte der operativen Datenbestände
    - Aufbereitung in Form statischer Berichte
    - Mainframe
  - 80er Jahre: Management Information Systems (MIS)
    - meist statische Berichtsgeneratoren
    - Einführung von Hierarchieebenen für Auswertung von Kennzahlen (Roll-Up, Drill-Down)
    - Client-Server-Architekturen, GUI (Windows, Apple)

Historie (Forts.)

- 1992: Einführung des Data-Warehouse-Konzeptes durch W.H. Inmon
  - redundante Haltung von Daten, losgelöst von Quellsystemen
  - Beschränkung der Daten auf Analysezweck
- 1993: Definition des Begriffs OLAP durch E.F. Codd
  - Dynamische, multidimensionale Analyse
- Weitere Einflussgebiete
  - Verbreitung geschäftsprozessorientierter Transaktionssysteme (SAP R/3) → Bereitstellung von entscheidungsrelevanten Informationen
  - Data Mining
  - WWW (Web-enabled Data Warehouse etc.)

Funktionen

- periodische und standardisierte Berichte
- Verfügbarkeit auf allen Managementebenen
- verdichtete, zentralisierte Informationen über alle Geschäftsaktivitäten
- interaktive Beschaffung von entscheidungs-relevanten Daten, die den Ist-Zustand des Unternehmens beschreiben
- größtmögliche Interaktivität
- Darstellung von Kennzahlen / Visualisierung / Erkennen von Trends
- regelmäßige und ad-hoc Berichte

Funktionen (Forts.)

- Unterstützung des Managers im Sinne einer Assistenz
- Management von Modellen und Methoden
- Datenbankmanagement
- konzentriert auf fachliche Teilprobleme
- eingebettet in komplexe Informationssysteme (z.B. ERP-Systeme, SAP BW)
- als Decision Support System
  - in den frühen Phasen von Entscheidungsprozessen
  - strategische Funktionen

Data Warehouse

Data Warehouse Überblick

- Begriff
- Anwendungen
- Definition und Abgrenzung
- Architekturmodell
  - Komponenten
- Phasen des Data Warehousing
  - ETL
  - Datenkonflikte

Was ist Data Warehousing?

- **Data Warehouse**:
  Sammlung von Technologien zur Unterstützung von Entscheidungsprozessen
- **Herausforderung an Datenbanktechnologien**
  - Datenvolumen (effiziente Speicherung und Verwaltung, Anfragebearbeitung)
  - Datenmodellierung (Zeitbezug, mehrere Dimensionen)
  - Integration heterogener Datenbestände

Anwendungen

- **Betriebswirtschaftliche Anwendungen**
  - Informationsbereitstellung
  - Analyse
  - Planung
  - Kampagnenmanagement
- **Wissenschaftliche Anwendungen**
  - Statistical und Scientific Databases
- **Technische Anwendungen**
  - Öffentlicher Bereich: DW mit Umwelt- oder geographischen Daten (z.B. Wasseranalysen)

Definition Data Warehouse

- **Begriff**

  „A Data Warehouse is a subject-oriented, integrated, non-volatile, and time variant collection of data in support of managements decisions."

  (W.H. Inmon 1996)

- **Charakteristika**

1. Themenorientierung (subject-oriented):

- Zweck des Systems ist nicht Erfüllung einer Aufgabe (z.B. Verwaltung), sondern Modellierung eines spezifischen Anwendungsziels

2. Integrierte Datenbasis (integrated):

- Verarbeitung von Daten aus mehreren verschiedenen Datenquellen (intern und extern) in einheitlicher konsistenter Sicht

3. Nicht-flüchtige Datenbasis (non-volatile):

- stabile, persistente Datenbasis
- Daten im DW werden nicht mehr entfernt oder geändert (Beständigkeit)

4. Historische Daten (time-variant):

- Speicherung der Daten zeitraumbezogen
- Vergleich der Daten über Zeit möglich (Zeitreihenanalyse)

Trennung operativer und analytischer Systeme

- **Klassische operative Informationssysteme (OLTP)**
  - Erfassung und Verwaltung von Daten
  - Verarbeitung unter Verantwortung der jeweiligen Abteilung
  - Transaktionale Verarbeitung: kurze Lese-/ Schreibzugriffe auf wenige Datensätze
- **Data Warehouse**
  - Analyse im Mittelpunkt
  - lange Lesetransaktionen auf vielen Datensätzen
  - Integration, Konsolidierung und Aggregation der Daten
- **Gründe**
  - Antwortzeitverhalten
  - Verfügbarkeit, Integrationsproblematik
  - Vereinheitlichung des Datenformats
  - Gewährleistung der Datenqualität

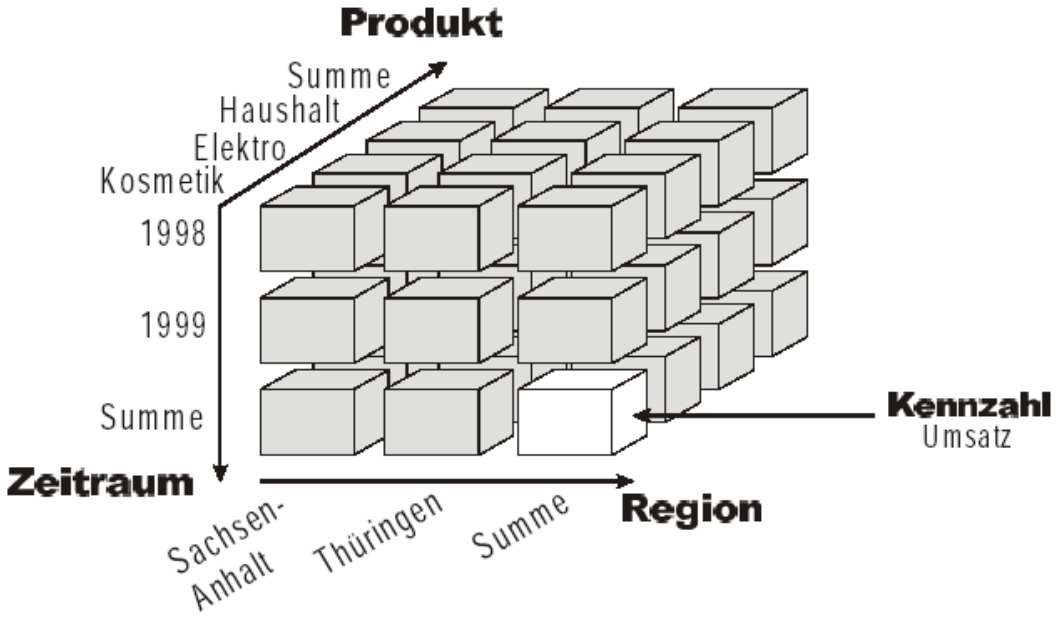
Beispiel einer Anfrage

„Welche Umsätze sind in den Jahren 1998 und 1999 in den Abteilungen Kosmetik, Elektro und Haushaltswaren in den Bundesländern Sachsen-Anhalt und Thüringen angefallen?“

Umsatz		Kosmetik	Elektro	Haushalt	SUMME
1998	Sachsen-Anhalt	45	123	17	185
	Thüringen	43	131	21	195
	SUMME	88	254	38	380
1999	Sachsen-Anhalt	47	131	19	197
	Thüringen	40	136	20	196
	SUMME	87	267	39	393
SUMME		175	521	77	773

Multidimensionales Datenmodell

- **Datenmodell zur Unterstützung der Analyse**
  - Fakten und Dimensionen
  - Klassifikationsschema
  - Würfel
  - Operationen



- **Notationen zur konzeptuellen Modellierung**
- **Relationale Umsetzung**
  - Star-, Snowflake-Schema
- **Multidimensionale Speicherung**

Fallbeispiel Wal-Mart

- Marktführer im amerikanischen Einzelhandel
- Weltgrößtes Data Warehouse mit ca. 0.5 PB (2006): 100 Mio Kunden, Milliarden Einkäufe pro Woche

Wal-Mart Data Center in MacDonald County

Fallbeispiel Wal-Mart: Orange Juice

- How much orange juice did we sell last year, last month, last week in store X?
- Comparing sales data of orange juice in various stores?
- What internal factors (position in store, advertising campaigns...) influence orange juice sales?
- What external factors (weather...) influence orange juice sales?
- Who bought orange juice last year, last month, last week?
- And most important: How much orange juice are we going to sell next week, next month, next year?

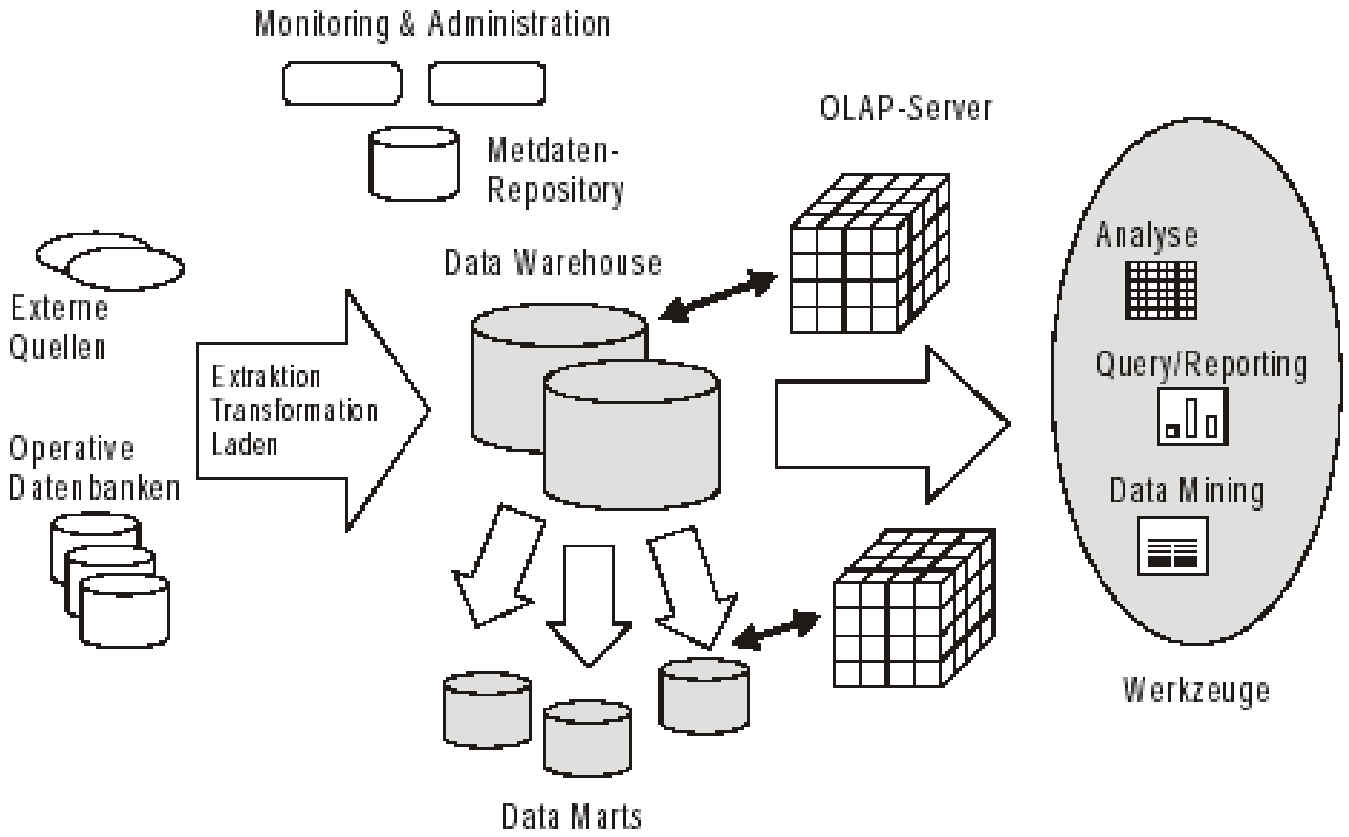
Other business questions include:

- What is the suppliers price of orange juice last year, this year, next year?
- How can we help suppliers to reduce their cost?
- What are the shipping/stocking costs of orange juice to/in store X?
- How can suppliers help us reduce those cost?

Data Warehouse Anforderungen

- **Unabhängigkeit zwischen Datenquellen und Analysesystemen** (bzgl. Verfügbarkeit, Belastung, laufender Änderungen)
- **Dauerhafte Bereitstellung integrierter und abgeleiteter Daten** (Persistenz)
- **Mehrfachverwendbarkeit der bereitgestellten Daten**
- **Möglichkeit der Durchführung prinzipiell beliebiger Auswertungen**
- **Unterstützung individueller Sichten** (z.B. bzgl. Zeithorizont, Struktur)
- **Erweiterbarkeit** (z.B. Integration neuer Quellen)
- **Automatisierung der Abläufe**
- **Eindeutigkeit über Datenstrukturen, Zugriffsberechtigungen und Prozesse**
- **Ausrichtung am Zweck: Analyse der Daten**

Data Warehouse Architekturmodell



Manager & Datenquellen

- **Data-Warehouse-Manager**
  - Zentrale Komponente eines DW-Systems
  - Initiierung, Steuerung der einzelnen Prozesse (Ablaufsteuerung)
  - Überwachung + Koordination
  - Fehlerhandling
  - Zugriff auf Metadaten aus dem Repository
- **Datenquellen**
  - Gehören nicht zum DWH
  - Klassifikation nach Herkunft, Zeit, Nutzungsebene
  - Auswahlkriterien: Zweck, Qualität, Verfügbarkeit, Preis
  - Qualitätsforderungen: Konsistenz , Korrektheit, Vollständigkeit, Genauigkeit und Granularität, Zuverlässigkeit und Glaub-würdigkeit, Verständlichkeit, Verwendbarkeit und Relevanz

Monitore & Arbeitsbereich

- **Monitore**
  - Entdeckung von Datenmanipulationen in einer Datenquelle
  - Strategien:
    Trigger-basiert, replikationsbasiert, Log-basiert, zeitstempelbasiert, Snapshot-basiert
- **Arbeitsbereich**
  - Zentrale Datenhaltungskomponente des Daten-beschaffungsbereichs (staging area)
  - Temporärer Zwischenspeicher zur Integration
  - Ausführungsort der Transformationen

  → Keine Beeinflussung der Quellen oder des DW
  Keine Übernahme fehlerbehafteter Daten

Extraktions-, Transformations- und Ladekomponente

- **Extraktionskomponente**
  - Übertragung von Daten aus Quellen in den Arbeitsbereich
  - abhängig von Monitoring-Strategie
  - Nutzung von Standardschnittstellen
  - Ausnahmebehandlung zur Fortsetzung im Fehlerfall
- **Transformationskomponente**
  - Vorbereitung und Anpassung der Daten für das Laden
  - Überführung aller Daten in ein einheitliches Format
  - Data Cleaning, Data Scrubbing, Data Auditing
- **Ladekomponente**
  - Übertragung der bereinigten und aufbereiteten (z.B. aggregierten) Daten in das DWH
  - Nutzung spezieller Ladewerkzeuge (z.B. SQL*Loader von Oracle)
  - Historisierung: Änderung in Quellen dürfen DWH-Daten nicht überschreiben, stattdessen zusätzliches Abspeichern
  - Online/Offline Ladevorgang

Data Warehouse & Data Marts

- **Data Warehouse**
  - Datenbank für Analysezwecke; orientiert sich in Struktur an Analysebedürfnissen
  - Basis: DBMS
  - Unterstützung des Ladeprozesses
  - Unterstützung des Analyseprozesses
- **Data Marts**
  - Bereitstellung einer inhaltlich beschränkten Sicht auf das DW (z.B. für Abteilung)
  - Gründe: Eigenständigkeit, Datenschutz, Lastverteilung, Datenvolumen, etc.
  - Abhängige Data Marts / Unabhängige Data Marts

Repository & Metadaten-Manager

- **Repository**
  - Speicherung der Metadaten des DWH-Systems
- **Metadaten**
  - Informationen, die Aufbau, Wartung und Administration des DW-Systemsvereinfachen und Informationsgewinnung ermöglichen
  - Beispiele: Datenbankschemata, Zugriffsrechte, Prozessinformationen (Verarbeitungsschritte und Parameter), etc.
- **Metadaten-Manager**
  - Steuerung der Metadatenverwaltung
  - Zugriff, Anfrage, Navigation
  - Versions- und Konfigurationsverwaltung

Phasen des Data Warehousing

- **Phasen**
  1. Überwachung der Quellen auf Änderungen durch Monitore
  2. Kopieren der relevanten Daten mittels Extraktion in temporären Arbeitsbereich
  3. Transformation der Daten im Arbeitsbereich (Bereinigung, Integration)
  4. Laden der Daten in das Data Warehouse
  5. Analyse: Operationen auf Daten des DWH
- **ETL-Prozeß**
  1. Extraktion: Selektion eines Ausschnitts der Daten aus den Quellen und Bereitstellung für Transformation
  2. Transformation: Anpassung der Daten an vorgegebene Schema- und Qualitätsanforderungen
  3. Laden: physisches Einbringen der Daten aus dem Arbeitsbereich (staging area) in das Data Warehouse

Datenkonflikte

- **Probleme**

1. heterogene Bezeichungen, Formate etc.
   → Beispiel
2. inkorrekte Einträge:
   - Tippfehler bei Eingabe von Werten
   - falsche Einträge aufgrund von Programmierfehlern in einzelnen Anwendungsprogrammen → i.d.R. nicht automatisch behebbar !!!
3. veraltete Einträge:
   - durch unterschiedliche Aktualisierungszeitpunkte
   - „vergessene“ Aktualisierungen in einzelnen Quellen

- **Behebung**

   - explizite Werteabbildung
   - Einführung von Ähnlichkeitsmaßen
   - Bevorzugung der Werte aus einer lokalen Quelle
   - Verwendung von Hintergrundwissen
     → Einsatz wissensbasierter Verfahren

Name	Geb.Jahr	Beruf
Peter Meier	1962	Dipl.-Inform.
Ingo Schmitt	1928	Dichter
...	...	...

Name	Geb.Jahr	Beruf
Meier, Peter	62	Informatiker
Schmitt, Ingo	28	Lyriker
...	...	...

Data Cleaning, Data Scrubbing, Data Auditing

- **Data Cleaning**
  - Korrektur inkorrekter, inkonsistenter oder unvollständiger Daten
  - Techniken:
    - Domänenspezifische Bereinigung
    - Domänenunabhängige Bereinigung
    - Regelbasierte Bereinigung
    - Konvertierungs- und Normalisierungsfunktionen
- **Data Scrubbing**
  - Ausnutzung von domänenspezifischen Wissen (z.B. Geschäftsregeln) zum Erkennen von Verunreinigungen
  - Beispiel: Erkennen von Redundanzen
- **Data Auditing**
  - Anwendung von Data-Mining-Verfahren zum Aufdecken von Regeln
  - Aufspüren von Abweichungen

Daten- und Informationsqualität

Management der Informationsqualität

- Keine verbindlichen Standards oder Vorgaben für Informationsqualität
- Allgemeine Definition von Qualität gemäß der ISO-Norm zu Qualitätsmanagement
  - aus der Sicht des Kunden eines Produkts
  - durch gesetzliche Vorgaben
- Qualität intuitiv charakterisiert durch „**Fitness for use**“ (Wang 1998), d.h. Eignung der Information für jeweiligen Einsatzzweck bestimmt deren Qualität
- Zahlreiche Ansätze und Modelle zur Beschreibung der Info-Qualität in verschiedenen Dimensionen
- Grundlage: Datenqualität

Datenqualität in der Praxis

- totale Kosten von schlechter Datenqualität liegen in Größenordnung zwischen 8% und 12% des Gesamtumsatzes
- Ca. 15-20% der Datenwerte einer typischen Kunden-Datenbank sind falsch
- schlechte Auswirkungen auf Geschäftsprozesse eines Unternehmens vorprogrammiert
- Kundenbeschwerden aufgrund z.B. falscher Rechnungen führt zu Vertrauensverlust
- erwarteter Nutzen eines DWH wird nicht erreicht
- falsche Zielgruppen bei Werbemaßnahmen →Kundenpotenzial wird nicht genutzt
- Cross-Selling-Möglichkeiten werden falsch erkannt oder nicht erkannt

⇒ Großer Imageverlust

Aspekte der Datenqualität

- Datenqualität ist ein mehrdimensionales Maß
- Verschiedene Aspekte, die miteinander konkurrieren (erfordert Kompromisse)
  - Genauigkeit
  - Vollständigkeit
  - Zeitbezogene Aspekte
  - Konsistenz
- Beispiel

ID	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

Datenqualität: Genauigkeit

- Abstand zwischen dem tatsächlichen Wert w und dem als exakt geltenden Wert w‘
- Unterteilung in zwei Arten:
  - syntaktische Genauigkeit: Kosten der Konvertierung eines Strings s in einen String s‘
  - semantische Genauigkeit: w ist syntaktisch korrekt aber dennoch von w‘ verschieden
- syntaktische Fehler sind leichter zu finden als semantische
- semantische Fehler korrigieren durch Vergleich mit einem äquivalentem Datensatz einer anderen Quelle
- führt aber zu neuem Problem (*record matching*):
  Wann sind zwei Datensätze gleich?

  J.E. Miller **vs.** John Edward Miller
  - Identifizierung: Verschiedene Bezeichner in verschiedenen Quellen
  - Entscheidung: Repräsentieren beide Datensätze das Gleiche?

Datenqualität: Genauigkeit (Forts.)

- Genauigkeit nicht nur für Werte interessant, auch für Attribute (*column accuracy*), die Relation oder die gesamte DB
- dazu muss man auch die Redundanz betrachten
- Redundanz wird vor allem in nicht relationaler Datenspeicherung zu großem Problem
- Doppelt verschickte Briefe schaden nicht nur der Portokasse eines Unternehmens !

- Bestimmen der Genauigkeit einer DB

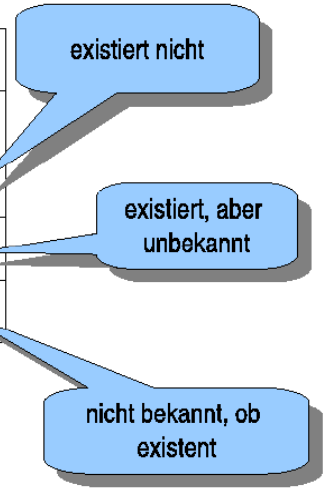
  Meist durch ein Verhältnis:

  # korrekter Spalten
  # Spalten

Datenqualität: Vollständigkeit

- Definition Vollständigkeit
  - abgeleitet vom Ausdruck „vollen Bestand haben“
  - Wenn sämtliche zu etwas gehörenden Teile vorhanden sind
- Behandlung im Relationenmodell: NULL-Werte
  - in Modell mit NULL-Werten muss deren Bedeutung interpretiert werden
  - 4 Arten: Wert-, Tupel-, Attribut-, Relations-vollständigkeit

ID	Name	Surname	Bithdate	Email
1	John	Smith	17.03.74	smith@home.net
2	Edward	Monroe	02.03.67	NULL
3	Anthony	White	01.01.36	NULL
4	Marianne	Collins	20.11.55	NULL



Datenqualität: Zeitbezogen

- Daten können im Laufe der Zeit variieren (temporale Daten)
- drei Kriterien zeitbezogener Daten:
  - Aktualität (*Currency*)
  - Änderungsfrequenz (*Volatility*)
  - Rechtzeitigkeit (*Timeliness*)
- korrekt heißt also sicherlich aktuell aber der Zeitpunkt des Gebrauchs der Daten muss berücksichtigt werden!

Datenqualität: Konsistenz

- aufdecken von Verletzungen semantischer Regeln
- semantische Regeln sind z.B. Integritätsbedingungen
- es gibt Intra- und Inter-Relations-Integritäts-Bedingungen
- schon geraume Zeit Gegenstand der Forschung
- Tools verfügbar
- Konsistenzregeln auch definierbar auf nicht-relationalen Daten
- dort gibt es auch entsprechende Möglichkeiten, Konsistenzüberprüfungen zu machen (*edit-imputation* Ansatz)

Datenqualitäts-Tools

- Vielzahl von kommerziellen und nichtkommerziellen Tools verfügbar
- die allgemeinen Anforderungen lassen erkennen : Es gibt kein „All-in-One-Tool”
- Tools lassen sich in Kategorien einordnen
- Eliminierung von Datenfehlern wird als *data cleaning* oder auch *data cleansing* bezeichnet
- Ziel : Erhöhung der Datenqualität (schwerer Weg)
- aktuelle Technologien lösen dieses Problem auf verschiedene Arten:
  - ad-hoc Programme in C / Java oder PL/SQL (in Oracle)
  - RDBMS Mechanismen die Integritätsbedingungen garantieren
  - Datentransformationsskripte, die Datenqualitätstools nutzen
- proprietäre RDBMS-Tools machen es Datenqualitätsprogrammen schwer
- großer Markt für Tools, die es ermöglichen Daten zu transformieren um DWHs zu bilden (ETL-Tools)

Funktionen von DQ-Tools

- Heterogene Datenquellen
- Steuerung der Extraktion
- Möglichkeiten des Ladens von Daten ins Zielsystem
- Schrittweise Updates (nicht immer wieder from scratch)
- GUI
- Metadatenverzeichnis
- Performance Funktion
- Versionierung
- Funktionsbibliothek
- integrierte Programmiersprache
- Debugging und Tracing
- Ausnahmebehandlung der Datensätze bei Fehlschlagen der Transformation

DQ-Tools: Kategorien

1. ***Analyse*** – zur Regelfestlegung und Sicherstellung, dass die Daten nicht die Anwendungsdomänen-Constraints verletzen
2. ***Data Profiling*** – anwendungsspezifische Datenqualitätsaspekte bestimmen
3. ***Transformation*** – Operationen die Quelldaten in Zielsystem integrieren
4. ***Säuberung*** – Entdecken, Löschen oder Korrigieren von schmutzigen Daten (inkorrekt, veraltet, redundant,inkonsistent, falsch formatiert)
5. ***Duplikate löschen*** – Erkennen und Löschen von Duplikaten
6. ***Erweiterung*** – Zusatzinformationen aus internen oder externen Quellen um Qualität der Eingangsdaten zu erhöhen

Datenqualität - Fazit

- Messen von Datenqualität ist sehr komplex
- Zahlreiche Tools vorhanden, die sich darauf spezialisiert haben
- Qualitätsdimensionen müssen in anwendungs-spezifischem Kontext evtl. erweitert werden
- Bis jetzt kein Standard verfügbar, aber auf gutem Weg