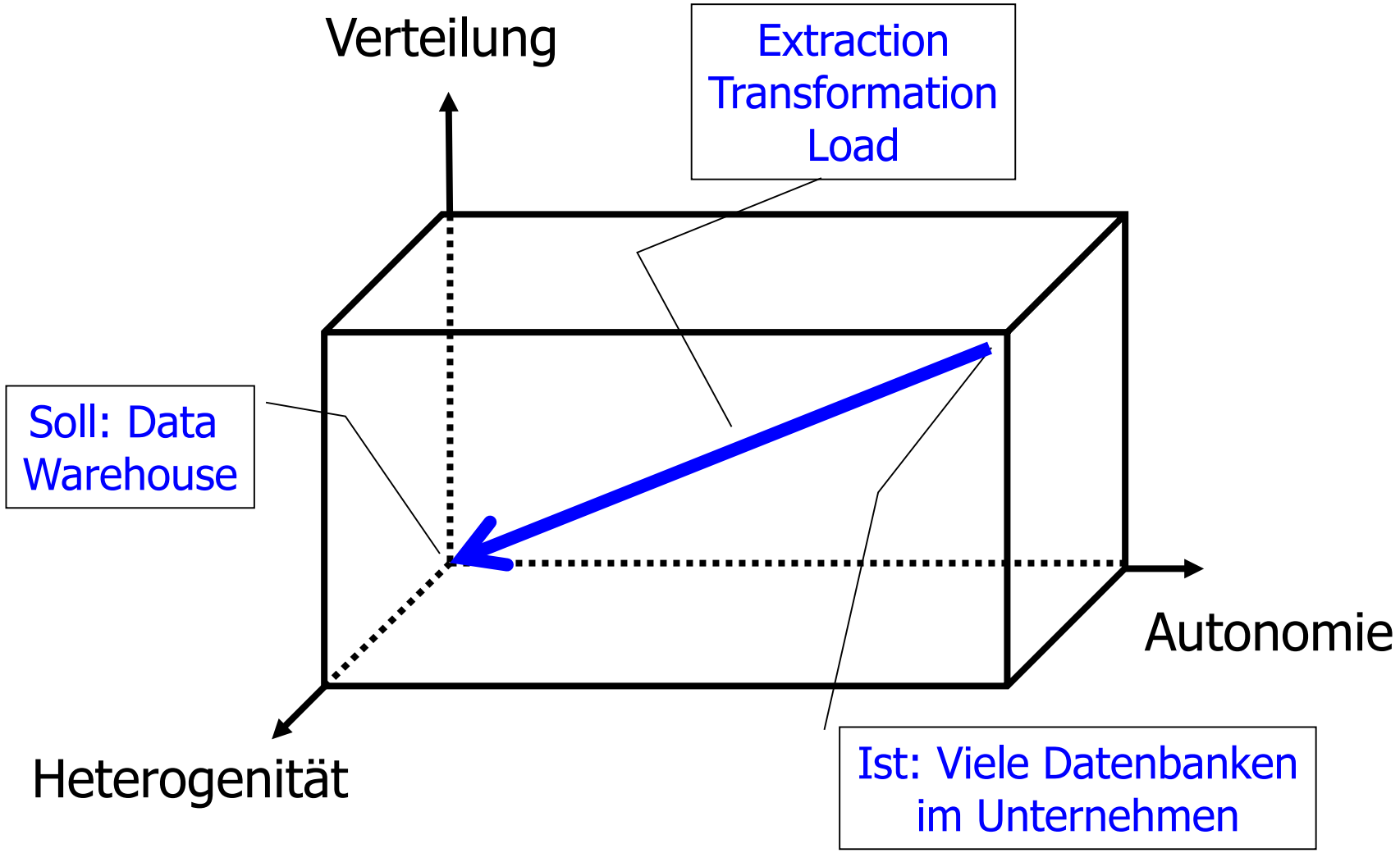
Informationsintegration

Data Warehouse
(= Materialisierte Integration)

Ulf Leser

Date Warehouses



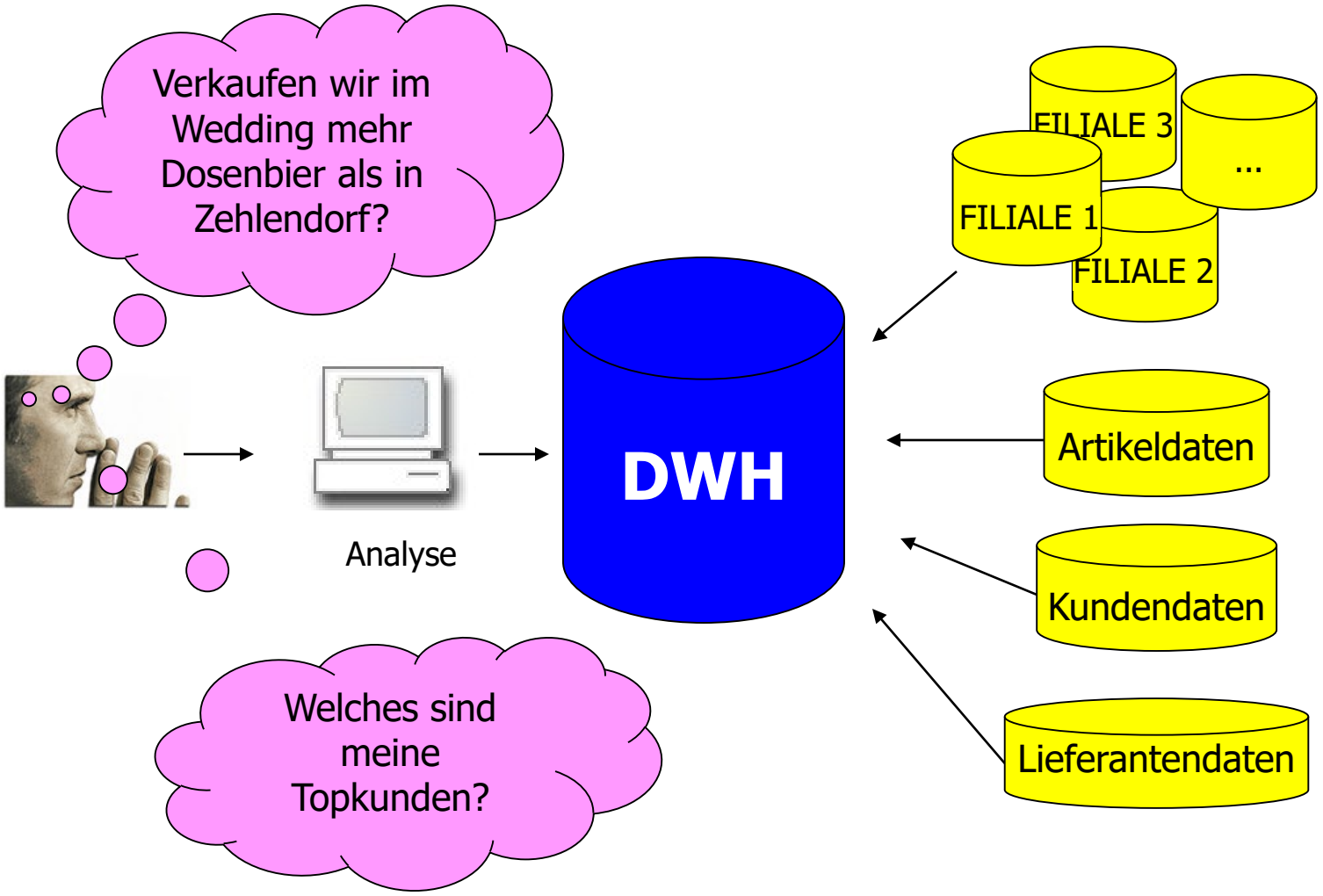
Inhalt dieser Vorlesung

- Data Warehouses
  - OLAP versus OLTP
  - Multidimensionale Modellierung
  - ETL Prozess
- Vergleich: Materialisiert / virtuell

Beispielszenario

- Ein beliebiges Handelshaus
- Physikalische Datenverteilung
  - Viele Niederlassungen (bis zu mehrere tausend)
  - Noch mehr Registerkassen
- Aber: Zentrale Planung, Beschaffung, Verteilung
  - Was wird wo und wie oft verkauft?
  - Was muss wann wohin geliefert werden?
- ... nur möglich, wenn
  - Zentrale Übersicht über Umsätze
  - Integration mit Lieferanten / Produktdaten

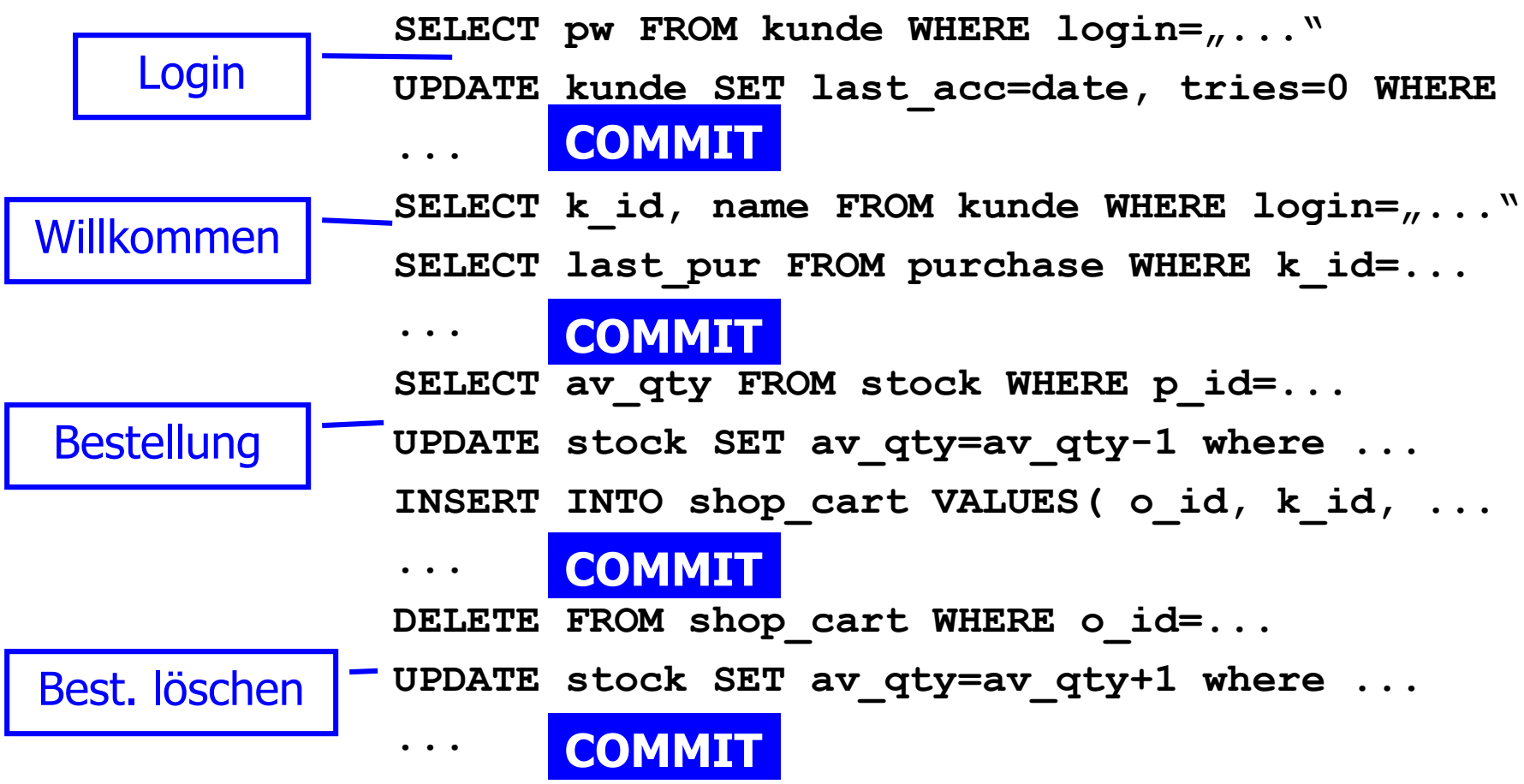
Handels - DWH


Verkaufen wir im Wedding mehr Dosenbier als in Zehlendorf?
Analyse
DWH
FILIALE 3
...
FILIALE 1
FILIALE 2
Artikeldaten
Kundendaten
Lieferantendaten
Welches sind meine Topkunden?

OLAP Beispiel

- Welche Produkte hatten im letzten Jahr im Bereich Bamberg einen Umsatzrückgang um mehr als 10%?
  - Welche Produktgruppen sind davon betroffen?
  - Welche Lieferanten haben diese Produkte?
- Welche Kunden haben über die letzten 5 Jahre eine Bestellung über 50 Euro innerhalb von 4 Wochen nach einem persönlichen Anschreiben aufgegeben?
  - Wie hoch waren die Bestellungen im Schnitt?
  - Wie hoch waren die Bestellungen im Vergleich zu den durchschnittlich. Bestellungen des jeweiligen Kunden in einem vergleichbaren Zeitraum?
  - Lohnen sich Mailing-Aktionen?

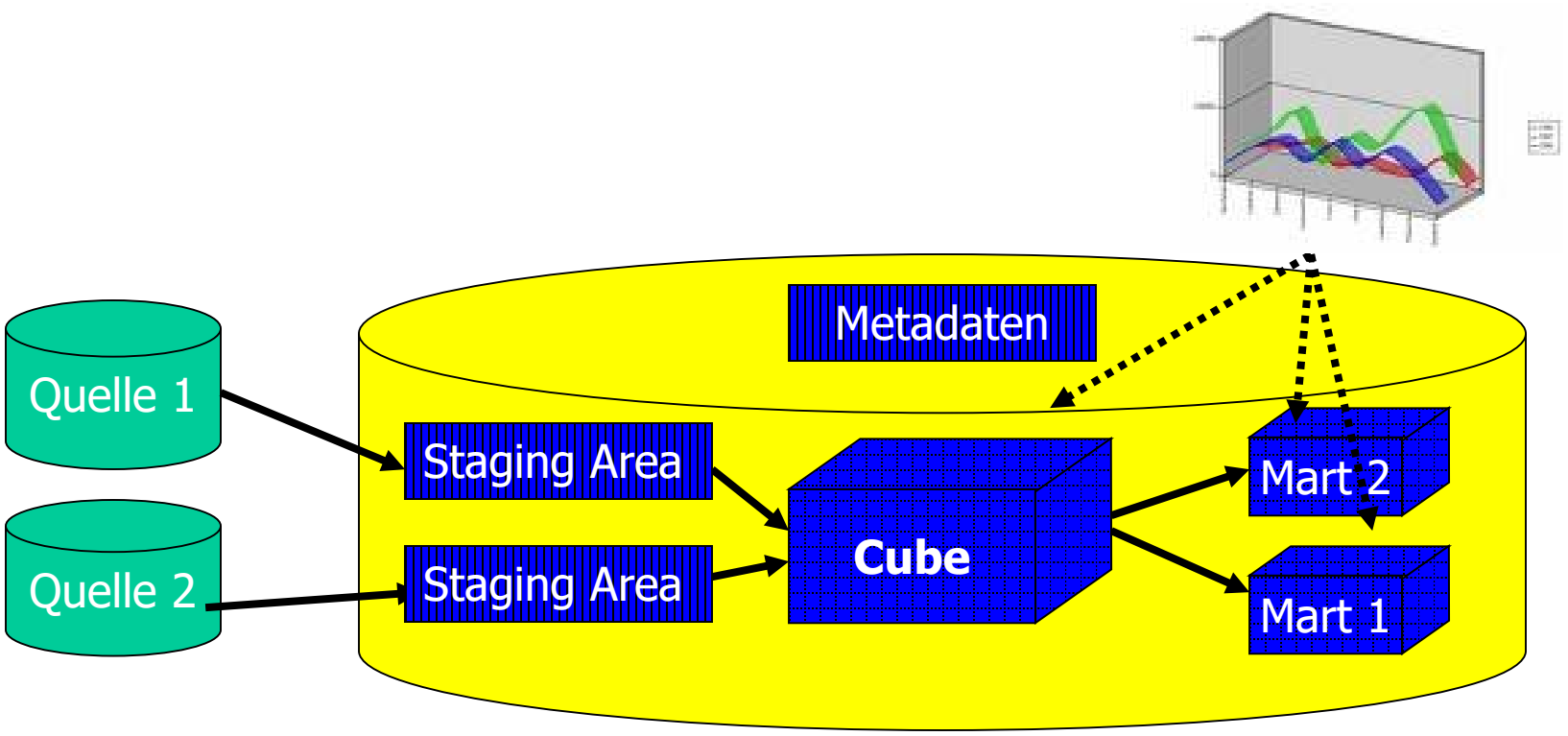
OLTP Beispiel



OLAP versus OLTP

	OLTP	**OLAP**
Typische Operationen	Insert, Update, Delete, Select	Select Bulk-Inserts
Transaktionen	Viele und kurz	Lesetransaktionen
Typische Anfragen	Einfache Queries, Primärschlüsselzugriff, Schnelle Abfolgen von Selects/inserts/updates/deletes	Komplexe Queries: Aggregate, Gruppierung, Subselects, etc. Bereichsanfragen über mehrere Attribute
Daten pro Operation	Wenige Tupel	Mega-/ Gigabyte
Datenmenge in DB	Gigabyte	Terabyte
Eigenschaften der Daten	Rohdaten, häufige Änderungen	Abgeleitete Daten, historisch & stabil
Erwartete Antwortzeiten	Echtzeit bis wenige Sek.	Minuten
Modellierung	Anwendungsorientiert	Themenorientiert
Typische Benutzer	Sachbearbeiter	Management

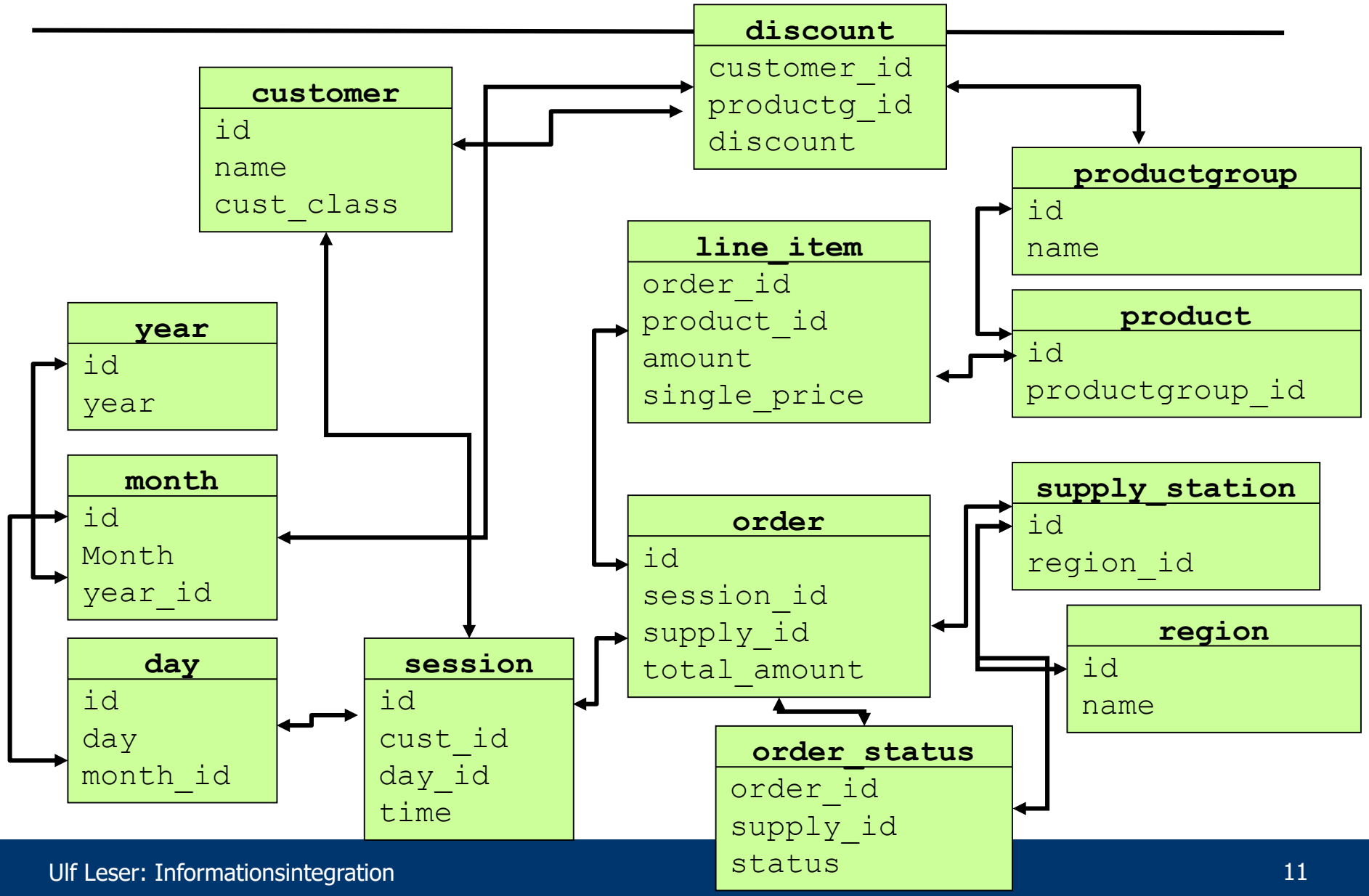
Komponenten eines DWH



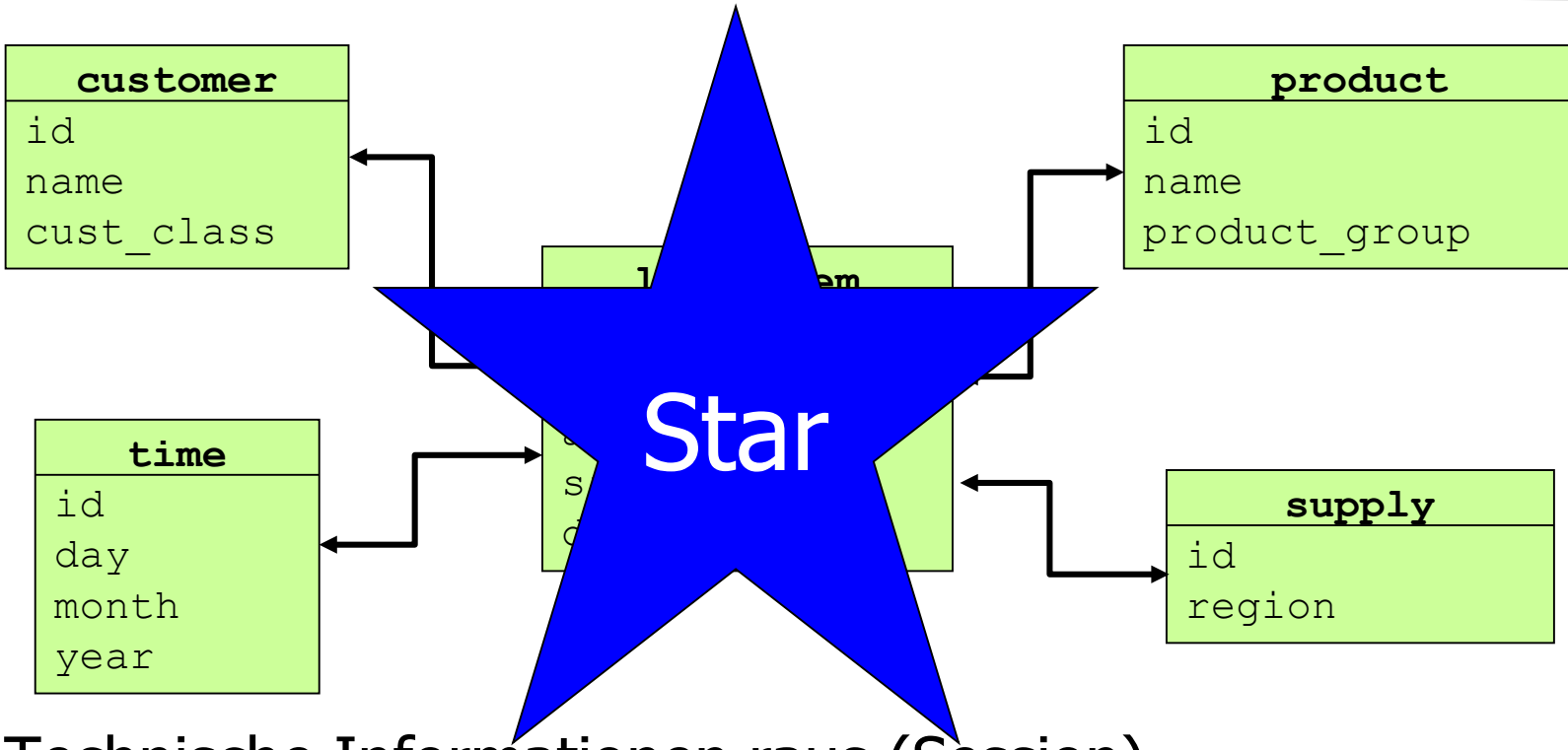
Inhalt dieser Vorlesung

- Data Warehouses
  - OLAP versus OLTP
  - Multidimensionale Modellierung
  - ETL Prozess
- Vergleich: Materialisiert / virtuell

Normalisiertes Schema

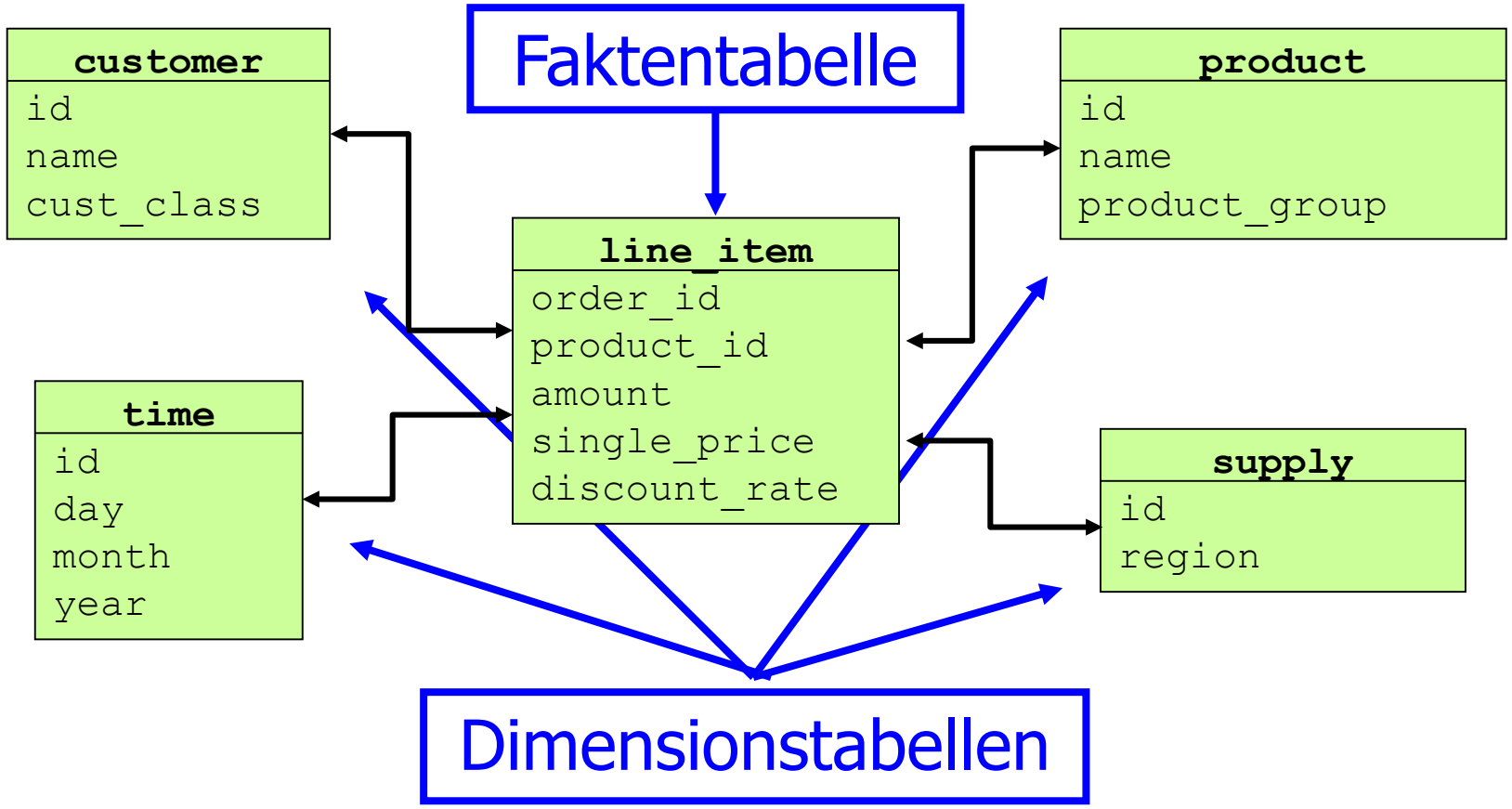


Multidimensionales Schema

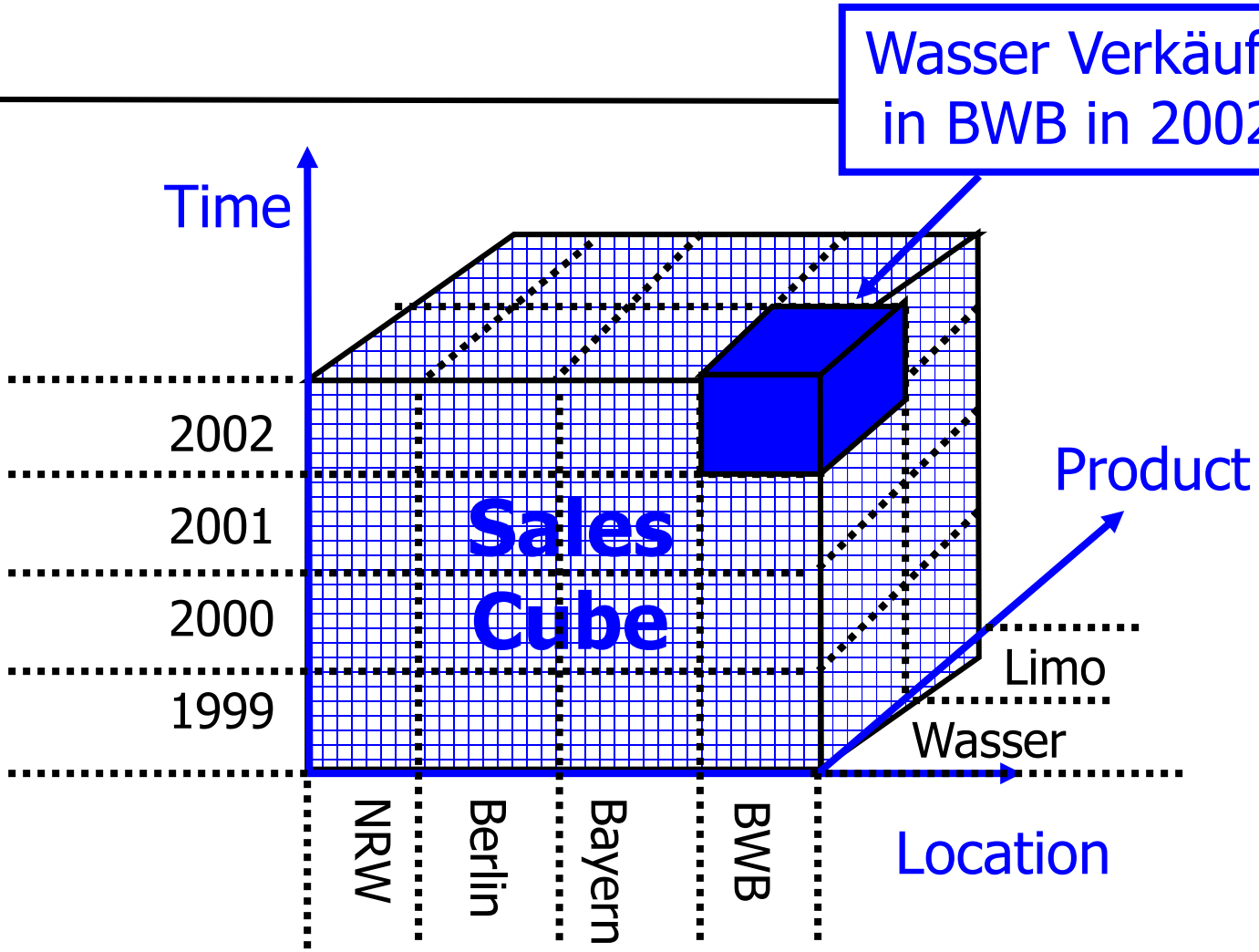


- Technische Informationen raus (Session)
- Nur abgeschlossene Bestellungen aufnehmen (Orderstatus)
- Zusammenfassen (discount_rate)
- Denormalisieren

Dimensionen und Fakten

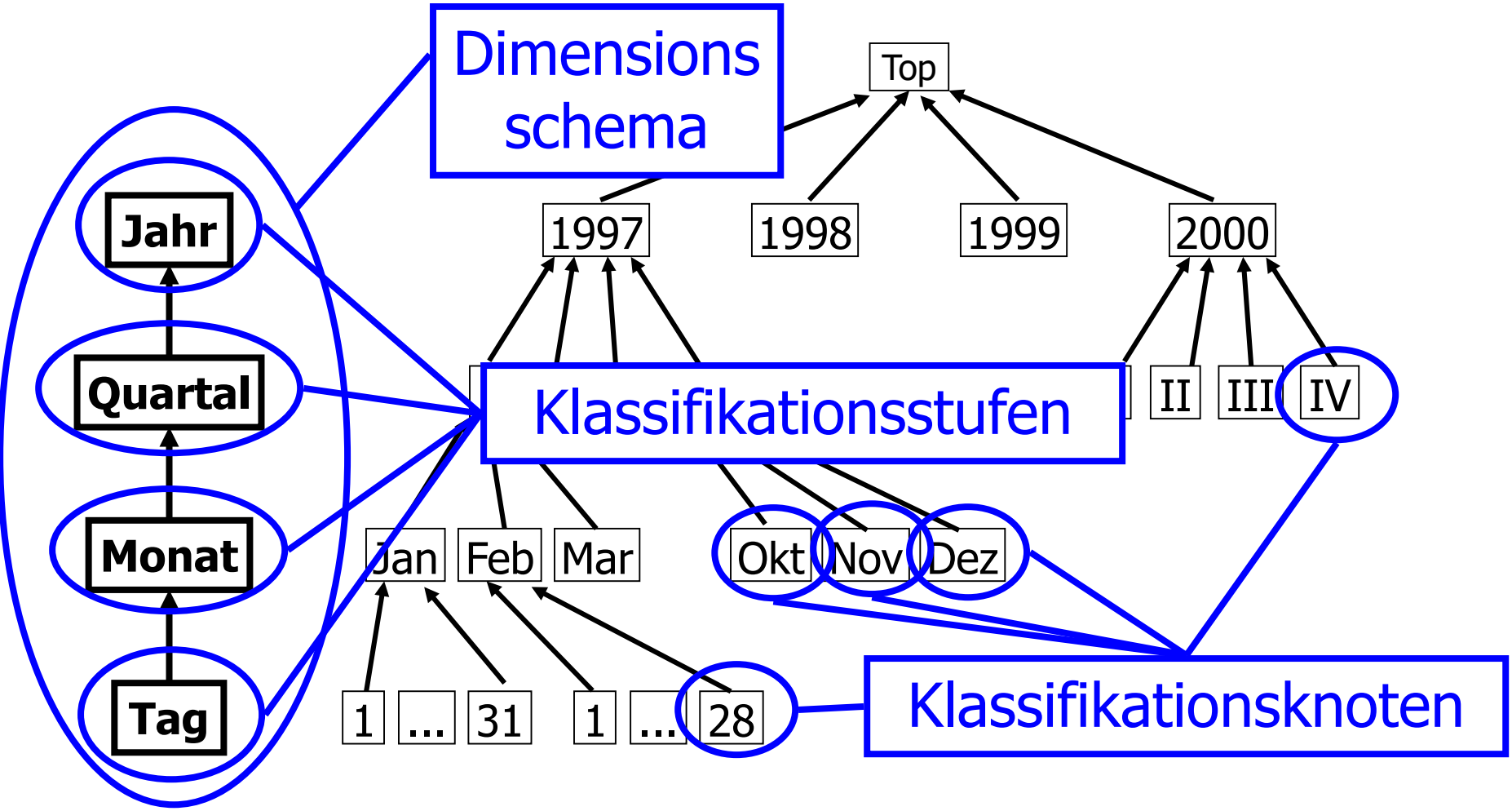


Cube



Cube -> Hypercube: Bon / Lieferant / Kunde / ...

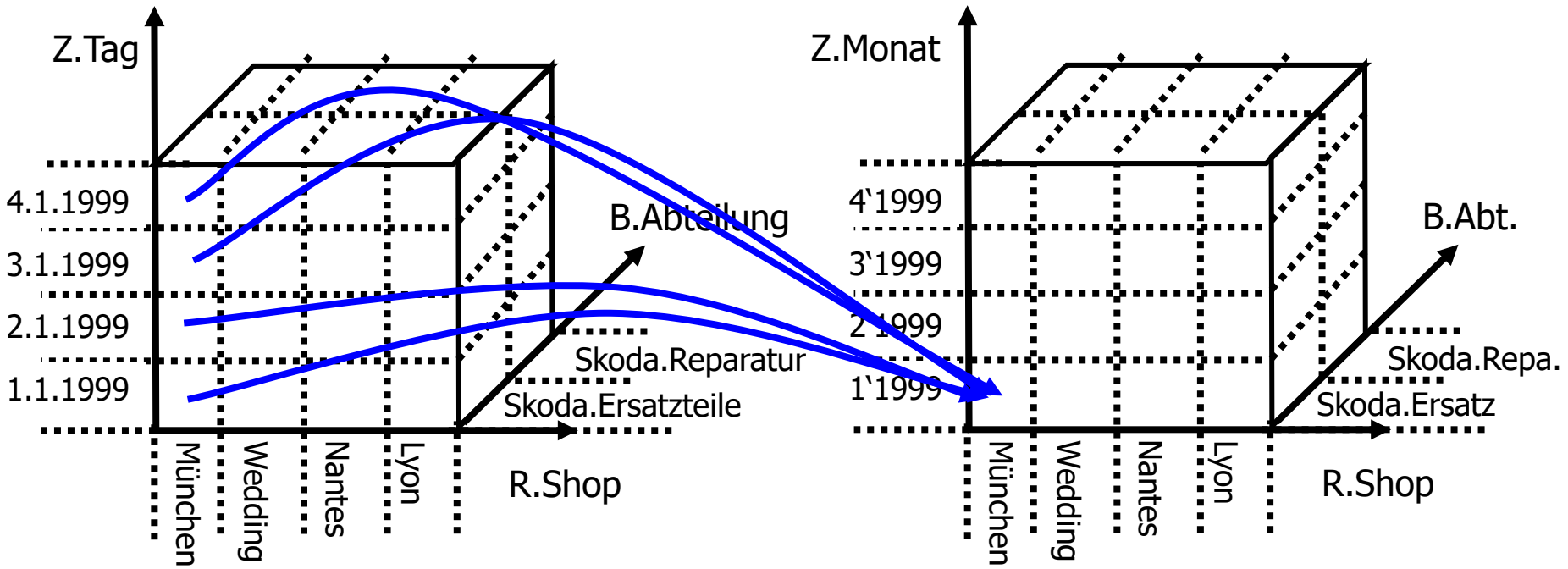
Hierarchische Dimensionsschemata



OLAP Operationen

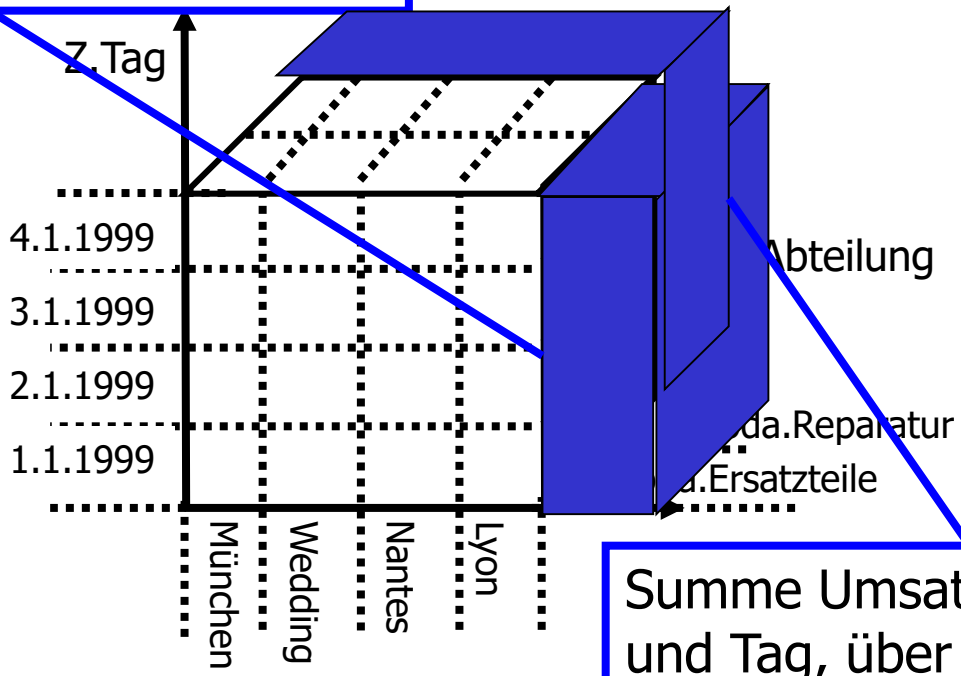
- Multidimensionale Datenmodellierung (MDDM)
- MDDM spricht die Sprache des Analysten
  - Operationen auf dem MDDM sollten die Analysevorgänge abbilden bzw. unterstützen
  - Ziel: Schnelle und intuitive Manipulation der Daten
- Notwendig
  - Definition der Operationen
  - Spezielle Anfragesprache (SQL-OLAP, MDX)
  - Unterstützung durch graphische Werkzeuge
  - Umsetzung in ausführbare Operationen

Beispiel Aggregation (Roll-Up)



Aggregation bis TOP

Summe Umsatz pro Tag und Abteilungen über alle Shops

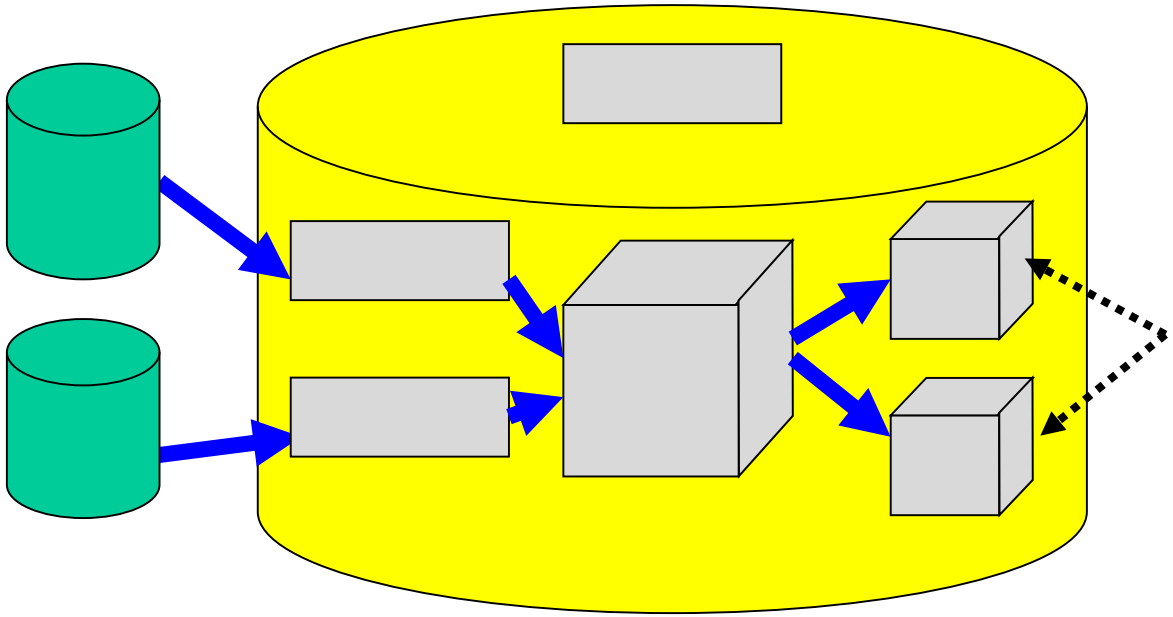


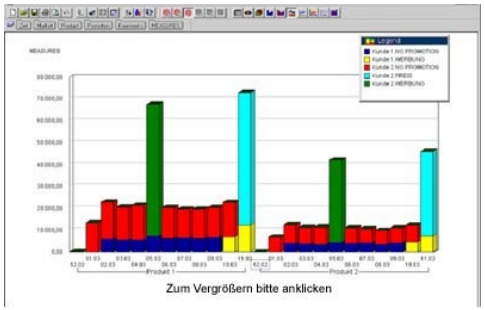
Summe Umsatz pro Shop und Tag, über alle Abteilungen

Inhalt dieser Vorlesung

- Data Warehouses
  - OLAP versus OLTP
  - Multidimensionale Modellierung
  - ETL Prozess
- Vergleich: Materialisiert / virtuell

ETL





- Extraction
- Transformation
- Load

Zeitpunkt der Extraktion

- Synchrone Extraktion: Quelle propagiert jede Änderung
- Asynchrone Extraktion
  - Periodisch
    - Push: Quellen erzeugen regelmäßig Extrakte (Reports)
    - Pull: DWH fragt regelmäßig Datenbestand ab
  - Ereignisgesteuert
    - Push: Quelle informiert z.B. alle X Änderungen
    - Pull: DWH erfragt Änderungen bei bestimmten Ereignissen
      - Jahresabschluss, Quartalsabschluss, …
  - Anfragegesteuert
    - Push: -
    - Pull: DWH erfragt Änderungen bei jedem Zugriff auf die (noch nicht vorhandenen oder potentiell veralteten) Daten

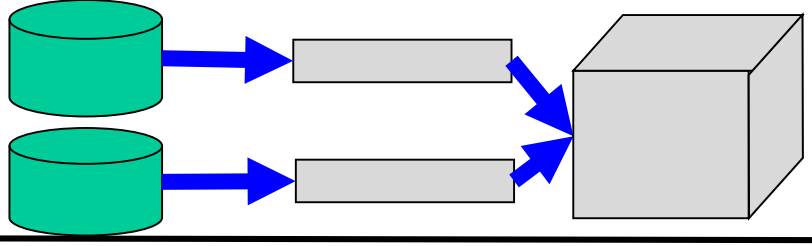
Art der Daten

- Snapshot: Quelle liefert Bestand zu festem Zeitpunkt
  - Lieferantenkatalog, Preisliste, etc. („Stock"-Fakten)
  - Import erfordert Erkennen von Änderungen
    - Differential Snapshot-Problem
    - Alternative: Komplettes Überschreiben (aber IC!)
  - Historie aller Änderungen kann nicht exakt abgebildet werden
- Log: Quelle liefert jede Änderung seit letztem Export
  - Transaktionslogs oder anwendungsgesteuertes Logging
  - Kann direkt importiert werden
    - DWH speichert ggf. Ereignis und seine Stornierung
- Nettolog: Quelle liefert das Delta zum letzten Export
  - Kann direkt importiert werden
  - Eventuell keine komplette Historie möglich

Datenversorgung

Quelle …		Technik	Aktualität DWH	Belastung DWH	Belastung Quellen
… erstellt periodisch Exportfiles		Reports, Snapshots	Je nach Frequenz	Eher niedrig	Eher niedrig
… pushed jede Änderung		Trigger, Changelogs, Replikation	Hoch	Hoch	Hoch
… erstellt Extrakte auf Anfrage (Poll)	**Vor Benutzung**	Sehr schwierig	Hoch	Hoch	Ja nach Anfragen
	Anwendungs-gesteuert	Je nachdem	Je nach Frequenz	Je nach Frequenz	Je nach Frequenz

ETL - Transformation



- Von den Quellen zur Staging Area
  - Extraktion von Daten aus den Quellen
  - Erstellen / Erkennen von differentiellen Updates
  - Erstellen von LOAD Files
- Von der Staging Area zur Basisdatenbank
  - Data Cleaning und Tagging
  - Duplikaterkennung
  - Erstellung integrierter Datenbestände

Transformationen beim Laden der Daten

- Extraktion der Daten aus Quelle mit einfachen Filtern
- Erstellen eines LOAD Files mit einfachen Konvertierungen
- Transformation meist zeilenorientiert
- Vorbereitung für den DB-spezifischen BULK-LOADER

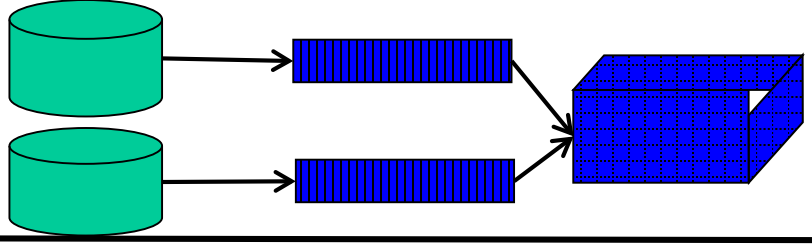
```
while (read_line)
  parse_line
  if (f[10]=2 & f[12]>0)
    write(file, f[1], string(f[4]), f[6]+f[7],...
  ...
bulk_upload( file)
```

Transformationen in Staging Area

- Effiziente mengenorientierte Operationen (SQL) möglich
- Vergleiche über Zeilen hinaus möglich
- Vergleiche mit Daten in Basisdatenbank möglich
- Typisch: Tagging von Datensätzen durch Prüf-Regeln

```
UPDATE sales SET price=price/MWST;
UPDATE sales SET cust_name=
  (SELECT cust_name FROM customer WHERE id=cust_id);
...
UPDATE sales SET flag1=FALSE WHERE cust_name IS NULL;
...
INSERT INTO DWH
  SELECT * FROM sales WHERE f1=TRUE & f2=TRUE & ...
```

ETL - Transformation



- Von den Quellen zur Staging Area
  - Extraction von Daten aus den Quellen
  - Erstellen / Erkennen von differentiellen Updates
  - Erstellen von LOAD Files
- Von der Staging Area zur Basisdatenbank
  - Data Cleaning und Tagging
  - Duplikaterkennung
  - Erstellung integrierter Datenbestände

Wrapper

- Quellspezifisch
- Datenmodell / (syntaktische) Heterogenität

Mediator

- Quellübergreifend
- Strukturelle / semantische Heterogenität

Aktuelles Beispiel

3313	Institut für Informatik	Ausgaben 2018	Ausgaben 2019	Ausgaben 2020	2
	Drittmittelausgaben (in Euro)				
	Institut				
331300	Drittmittel (dezentral verwaltet; TGR 94)				
331310	**Theoretische Informatik**				
331311	Automaten- und Systemtheorie				
331313	Algorithmen und Komplexität II				
331314	Modellierung und Analyse komplexer Systeme				
	Praktische Informatik				
331321	Systemanalyse				
331322	Systemarchitektur				
331323	Softwaretechnik und Theorie der Programmierung I				
331325	Datenbanken und Informationssysteme				
331327	Process-Driven Architectures (J)				
331328	Informatik in Bildung und Gesellschaft				
331329	Kognitive Robotik (J)				
331331	Maschinelles Lernen (J)				
331332	Softwaretechnik				
331333	Modellgetriebene Softwareentwicklung (J)				
	Technische Informatik				
331341	Rechnerorganisation und Kommunikation				
331352	Softwaretechnik (S)				
331353	Wissensmanagement in der Bioinformatik (S)				
331354	Computer Vision (S)				
331355	Visual Computing (S)				
331357	Drahtlose Breitbandkommunikationssystem (S)				
33138801	NW/1: Prozessorientierung in Ereignisgetriebenen Systemen				
33138802	NW/1: Statistisches Lernen				
33138803	NW/2: Datenreduktion				
33138804					
331397	Seniorprofessor(inn)en				
	** Rückzahlung von Mitteln, die im Vorjahr verausgabt wurden.				

Aktuelles Beispiel

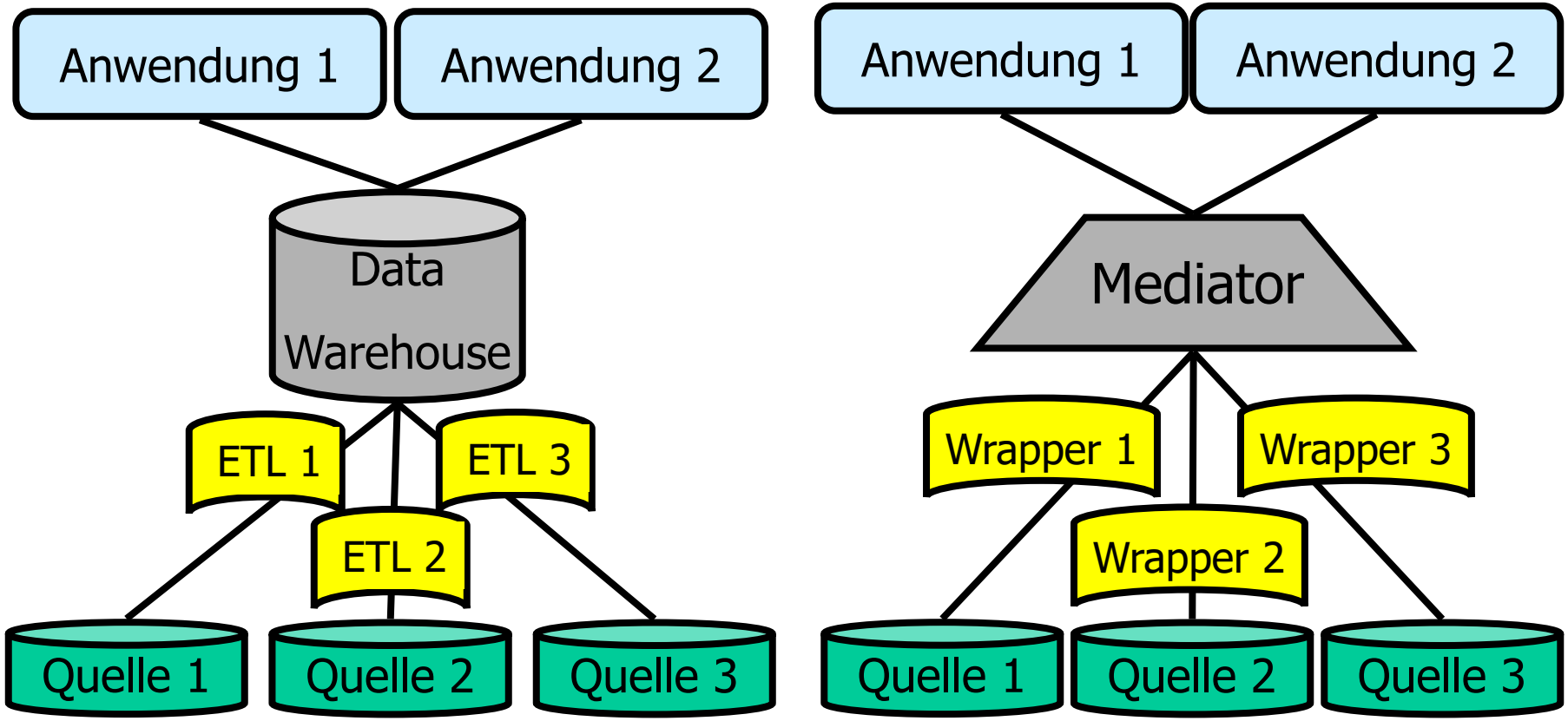
3313	**Institut für Informatik**
	Drittmittelausgaben (in Euro)
	Institut
331300	Drittmittel (dezentral verwaltet; TGR 94)
331310	**Theoretische Informatik**
331311	Automaten- und Systemtheorie
331313	Algorithmen und Komplexität II
331314	Modellierung und Analyse komplexer Systeme
	Praktische Informatik
331321	Systemanalyse
331322	Systemarchitektur
331323	Softwaretechnik und Theorie der Programmierung I
331325	Datenbanken und Informationssysteme
331327	Process-Driven Architectures (J)
331328	Informatik in Bildung und Gesellschaft
331329	Kognitive Robotik (J)
331331	Maschinelles Lernen (J)
331332	Softwaretechnik
331333	Modellgetriebene Softwareentwicklung (J)
	Technische Informatik
331341	Rechnerorganisation und Kommunikation
331352	Softwaretechnik (S)
331353	Wissensmanagement in der Bioinformatik (S)
331354	Computer Vision (S)
331355	Visual Computing (S)
331357	Drahtlose Breitbandkommunikationssystem (S)
33138801	NW/1: Prozessorientierung in Ereignisgetriebenen Systemen
33138802	NW/1: Statistisches Lernen
33138803	NW/2: Datenreduktion
33138804	
331397	Seniorprofessor(inn)en
	** Rückzahlung von Mitteln, die im Vorjahr verausgabt wurden.

- Es gibt
  - unterschiedliche Datenbasen gibt
    - Welche Professuren gibt es im IfI?
  - unterschiedliche Synchronisationszeit-punkte
    - Wer ist noch da?
  - unterschiedliche Bezugsrahmen
    - Der Prof ist zwar nicht mehr richtig da und seine Stelle schon neu besetzt, aber auf den alten Namen laufen noch Projekte
  - unterschiedliche Reportsysteme, die zueinander inkompatibel sind
  - keine klaren Ansprechpartner*innen bei Nachfragen
  - Probleme werden nicht gelöst, sondern oberflächlich repariert
  - Zu wenig Kontext
    - Prof- Namen, Projektnamen, …

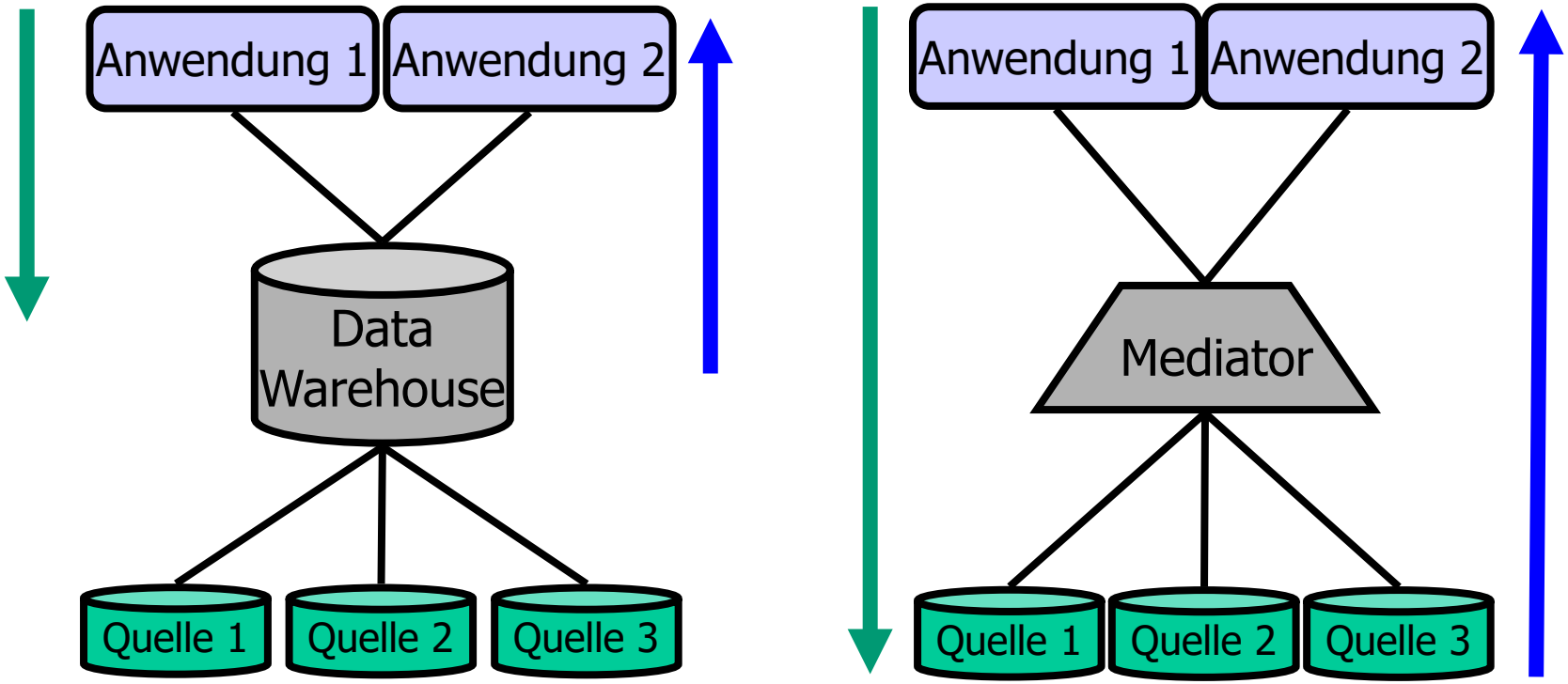
Inhalt dieser Vorlesung

- Data Warehouses
  - OLAP versus OLTP
  - Multidimensionale Modellierung
  - Relationale Implementierung
  - ETL Prozess
- Vergleich: Materialisiert / virtuell

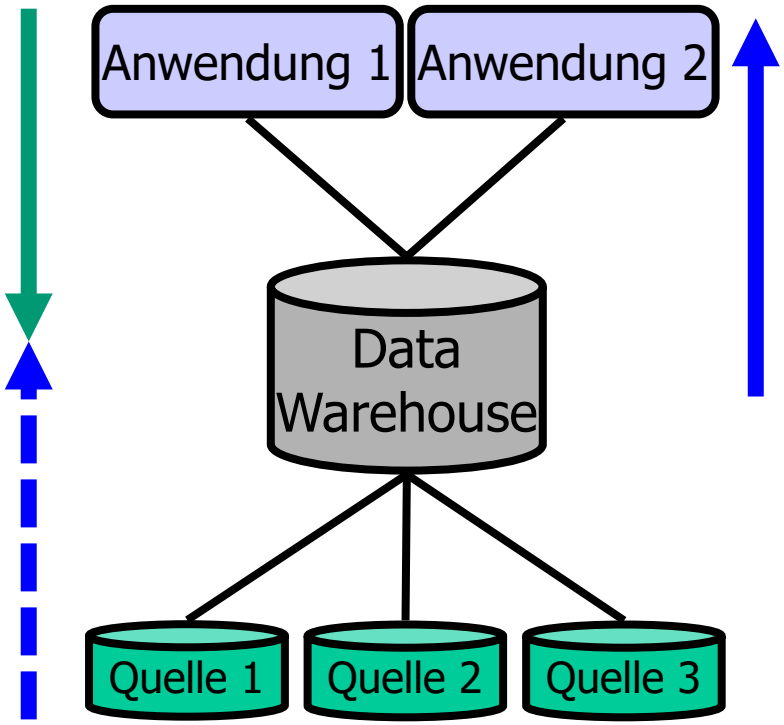
Data Warehouse vs. Mediator



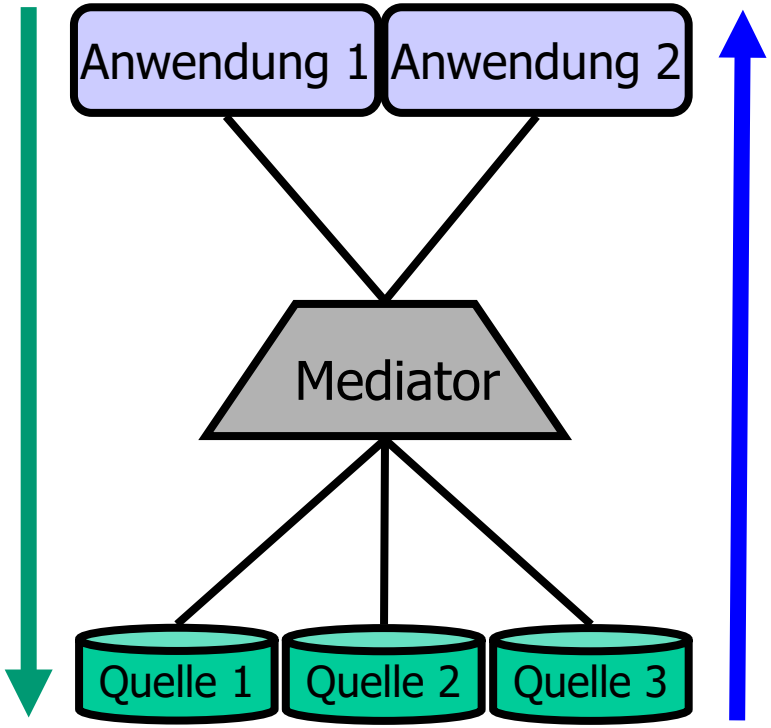
Datenfluss


Anwendung 1
Anwendung 2
Data Warehouse
Quelle 1
Quelle 2
Quelle 3
Anwendung 1
Anwendung 2
Mediator
Quelle 1
Quelle 2
Quelle 3

Load ist asynchron



Push / Pull



Nur Pull

Vor- und Nachteile

	Materialisiert	Virtuell
Aktualität	-	+
Antwortzeit	+	-
Komplexität	O	--
Anfragemächtigkeit	+	--
Read/Write	+/+	+/-
Ressourcenbedarf im Int. System	Zentral	Verteilt
Datenreinigung	+	-
Online-Zugriff auf Quellen notwendig	Nein	Ja
Download von Quellen notwendig	Ja	Nein