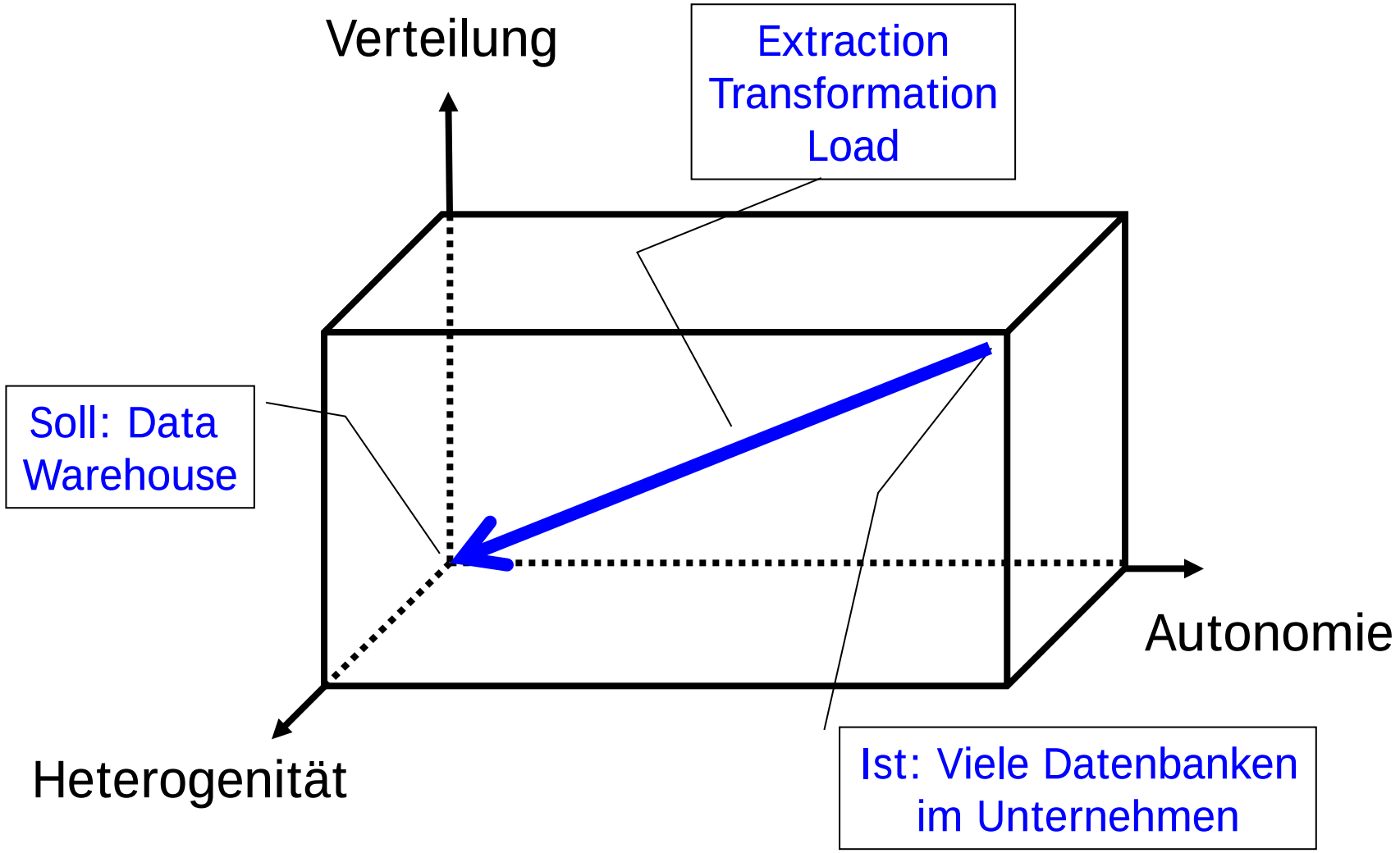
Informationsintegration

Data Warehouse
(= Materialisierte Integration)

Ulf Leser

Date Warehouses



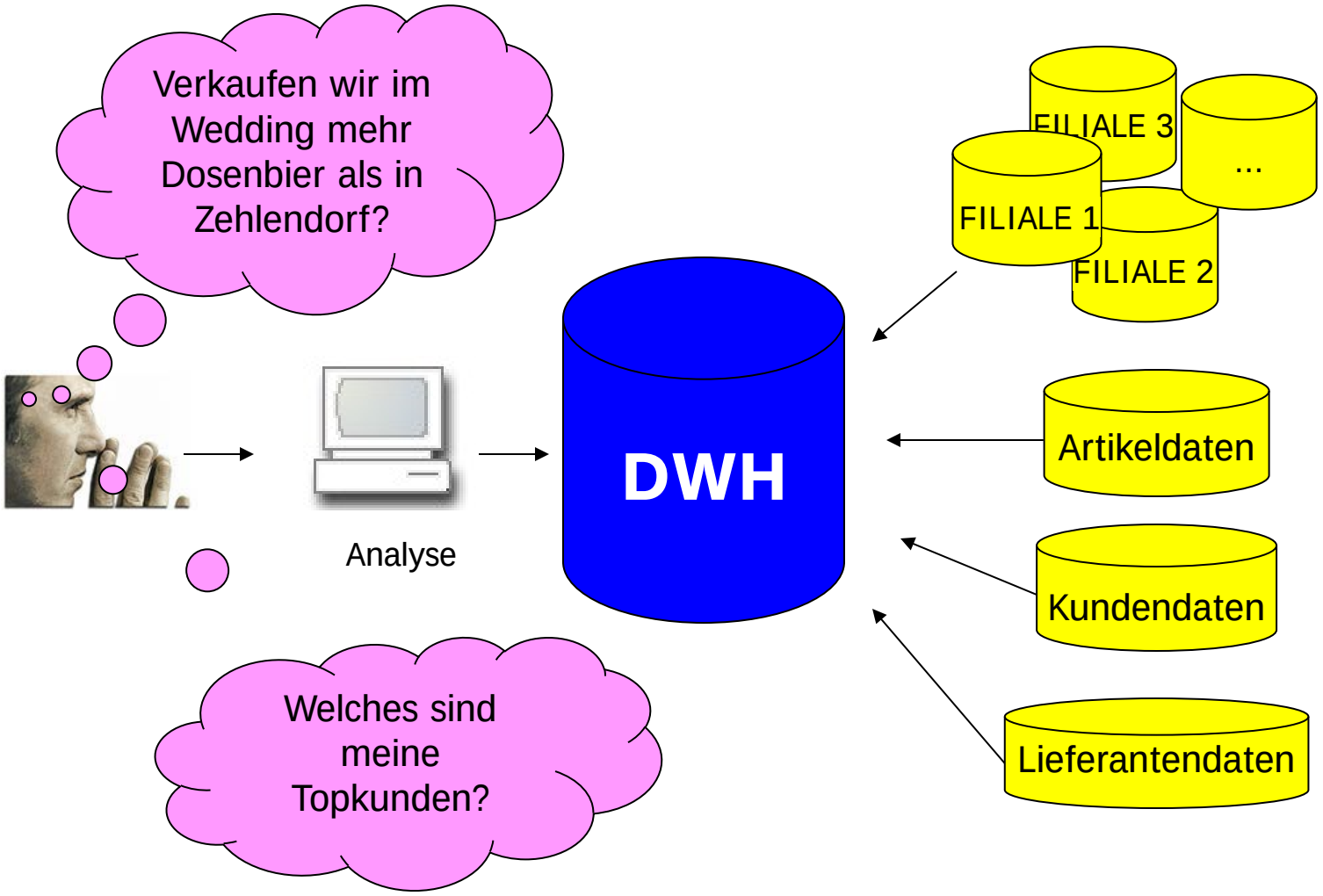
Inhalt dieser Vorlesung

- Data Warehouses
  - OLAP versus OLTP
  - Multidimensionale Modellierung
  - ETL Prozess
- Beispiel: Columba
- 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

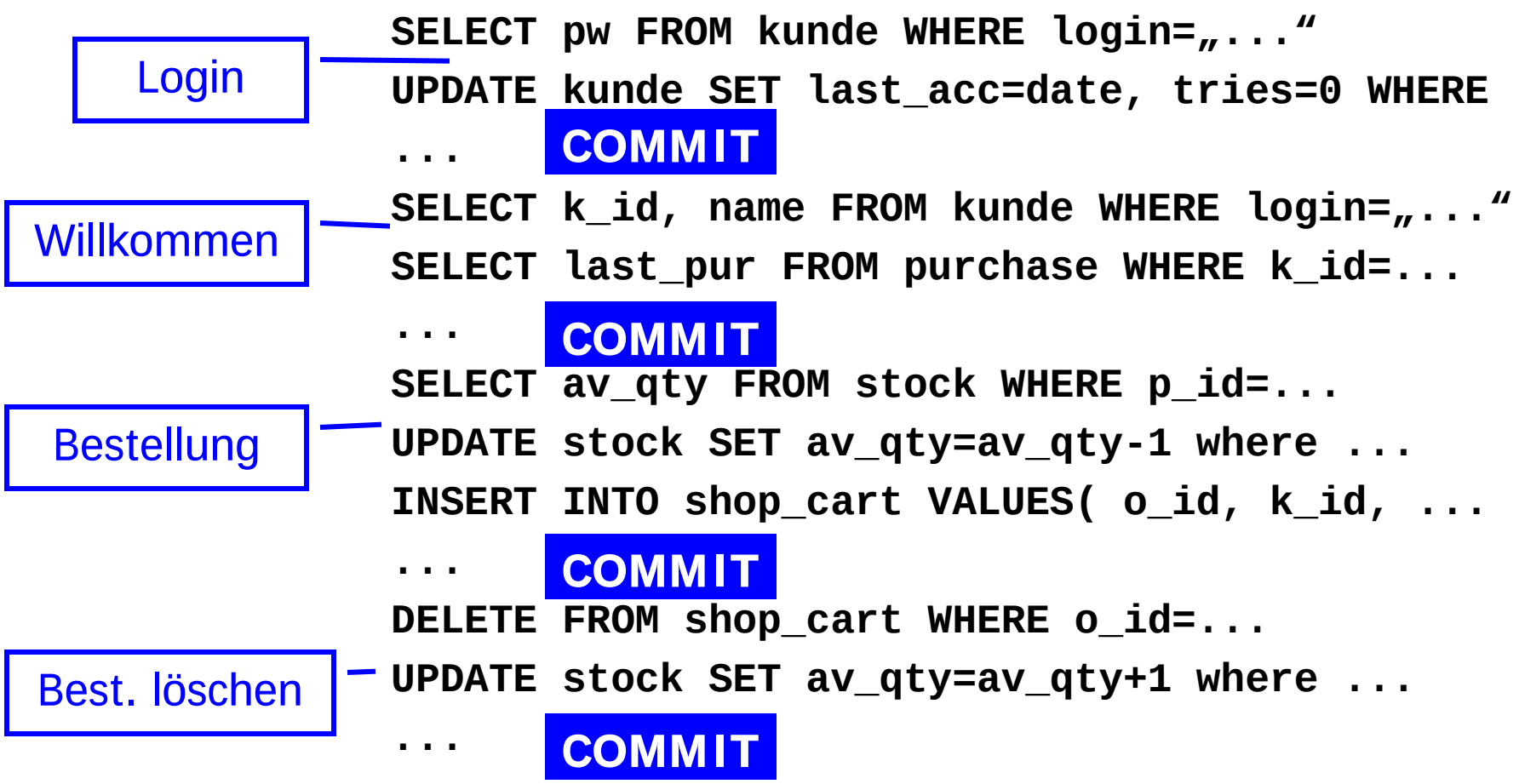
Handels - DWH


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?

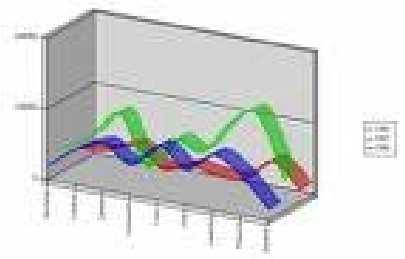
OLTP Beispiel

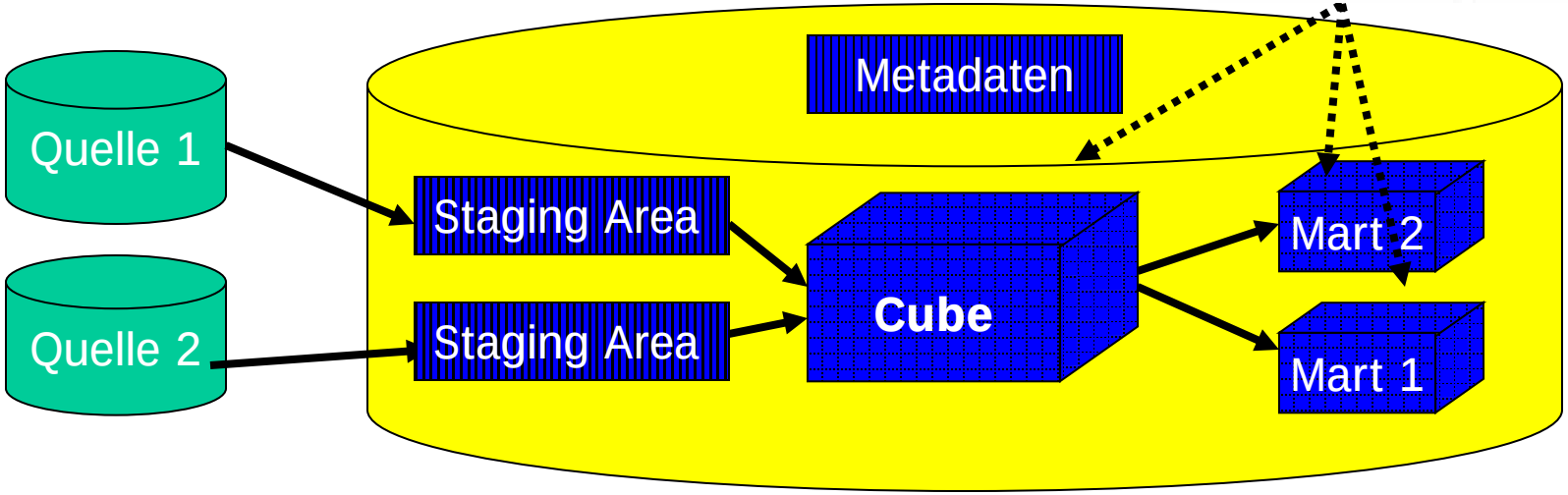


OLAP versus OLTP

	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

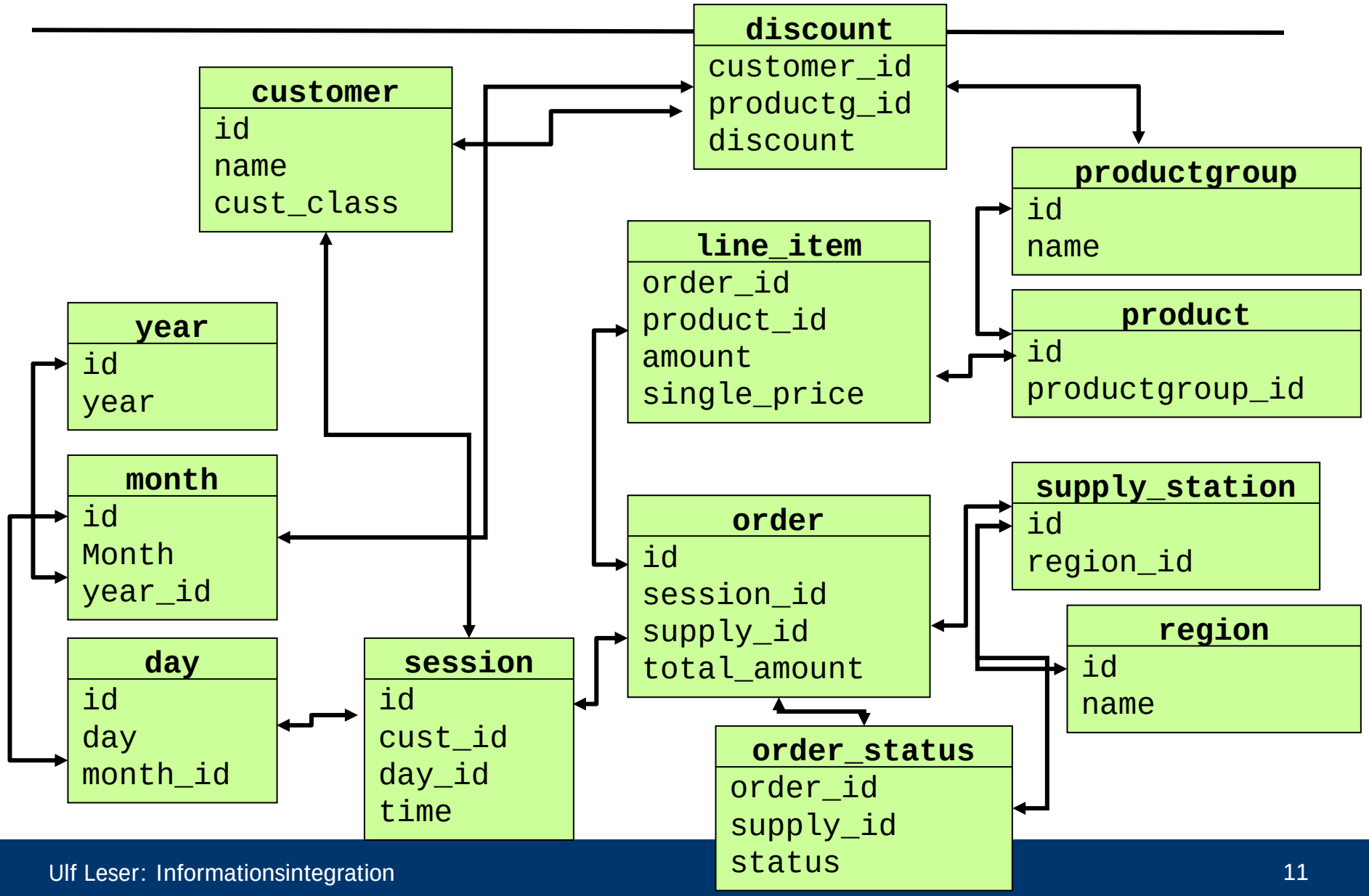




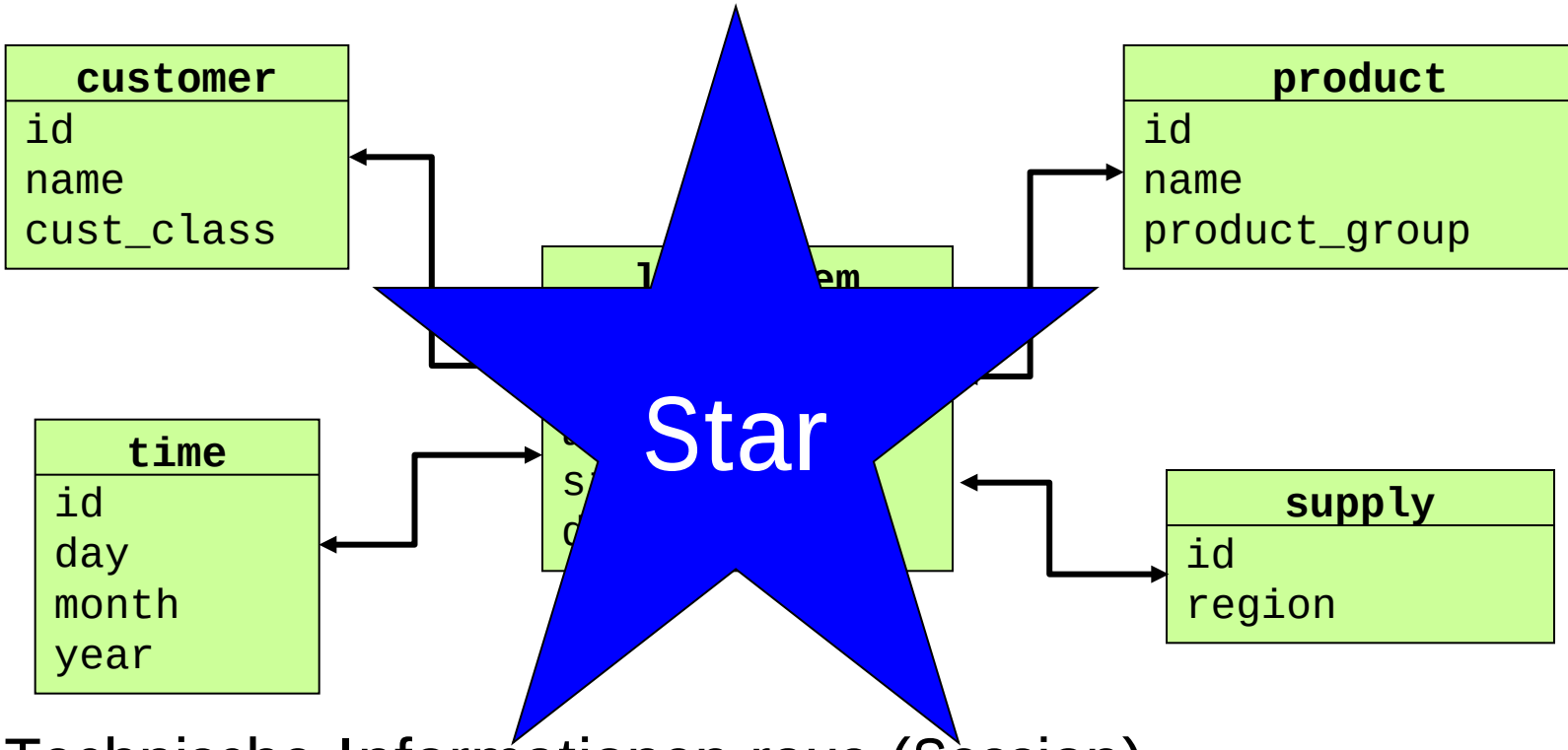
Inhalt dieser Vorlesung

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

Normalisiertes Schema

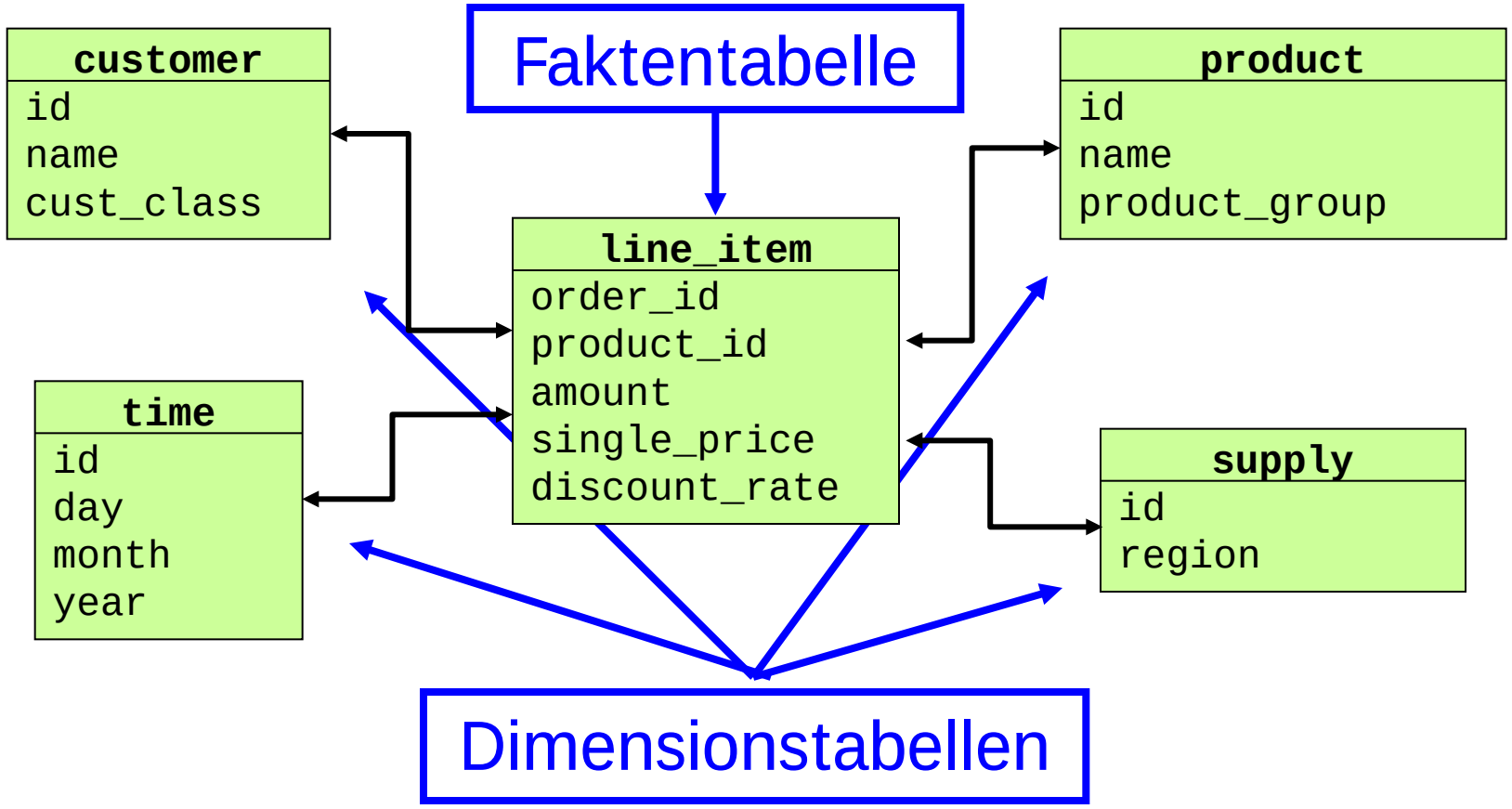


Multidimensionales Schema

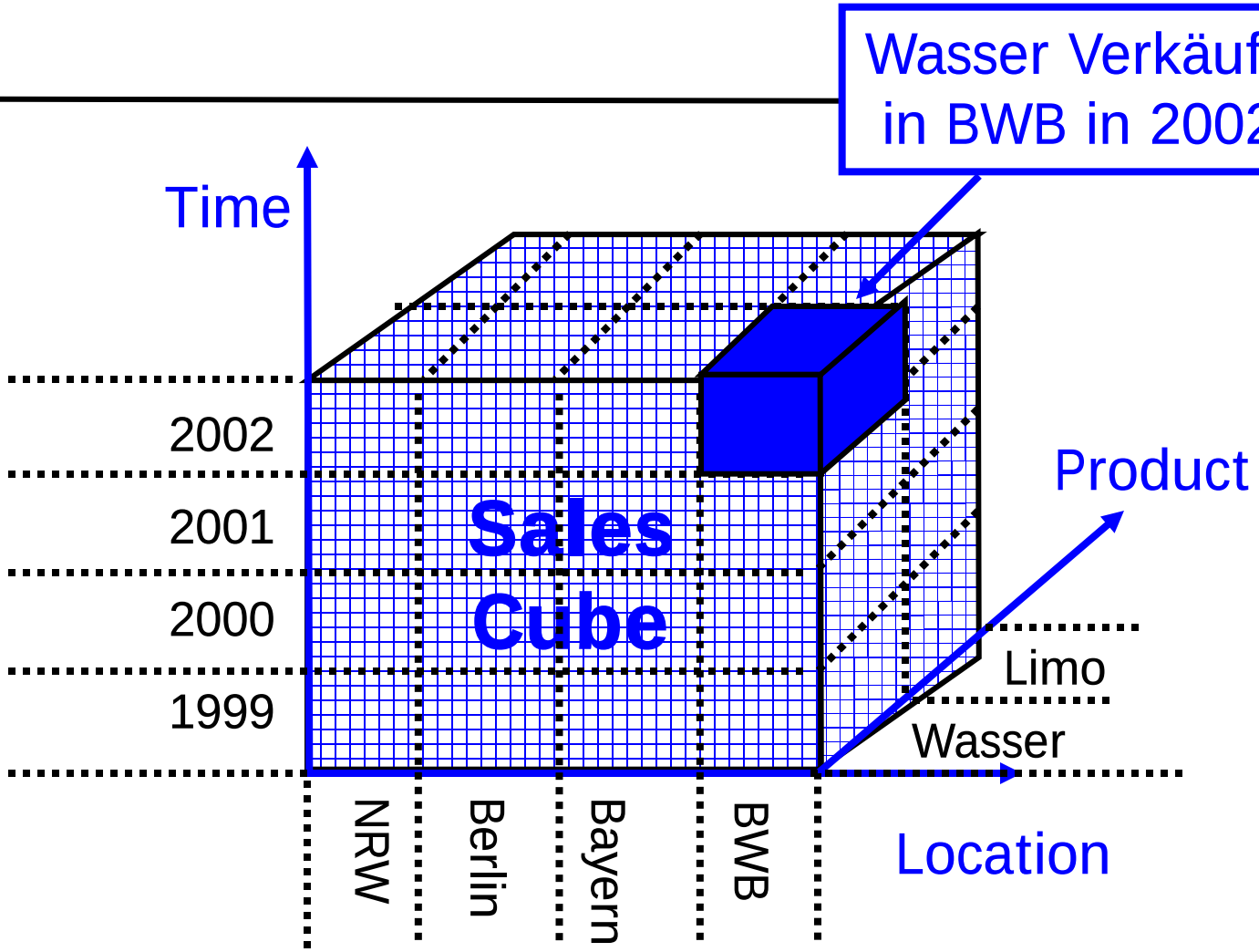


- Technische Informationen raus (Session)
- Nur abgeschlossene Bestellungen aufnehmen (Orderstatus)
- Zusammenfassen (discount_rate)
- Denormalisieren

Dimensionen und Fakten

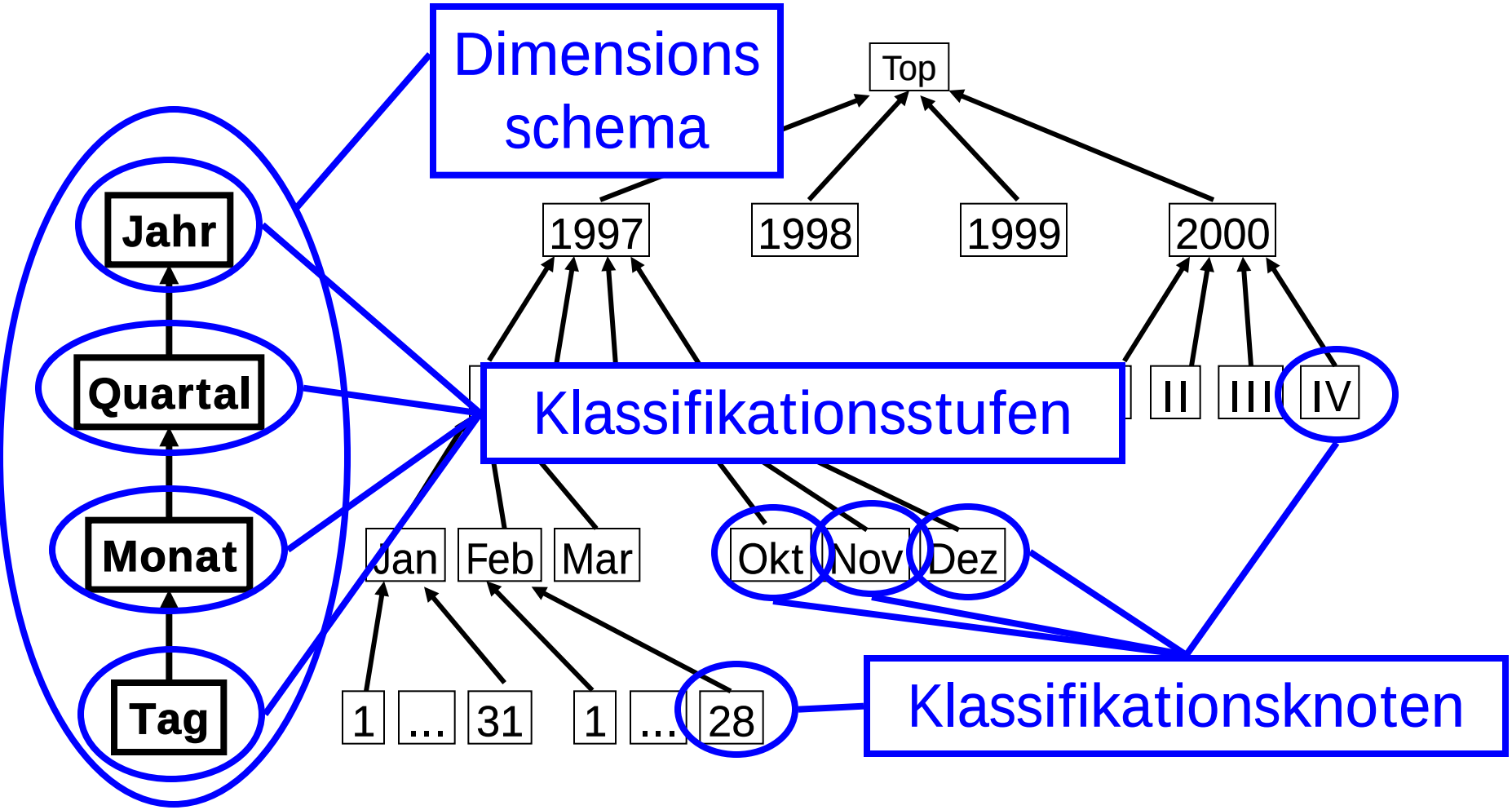


Cube



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

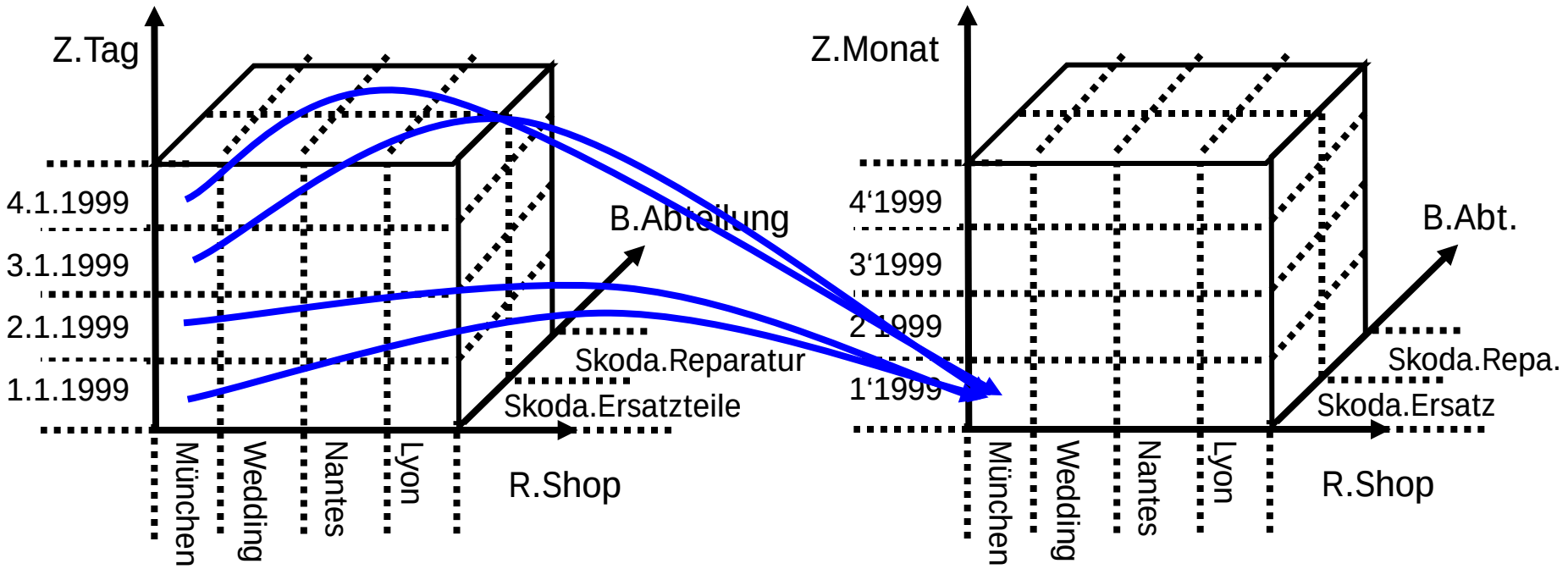
Hierarchische Dimensionsschemata



OLAP Operationen

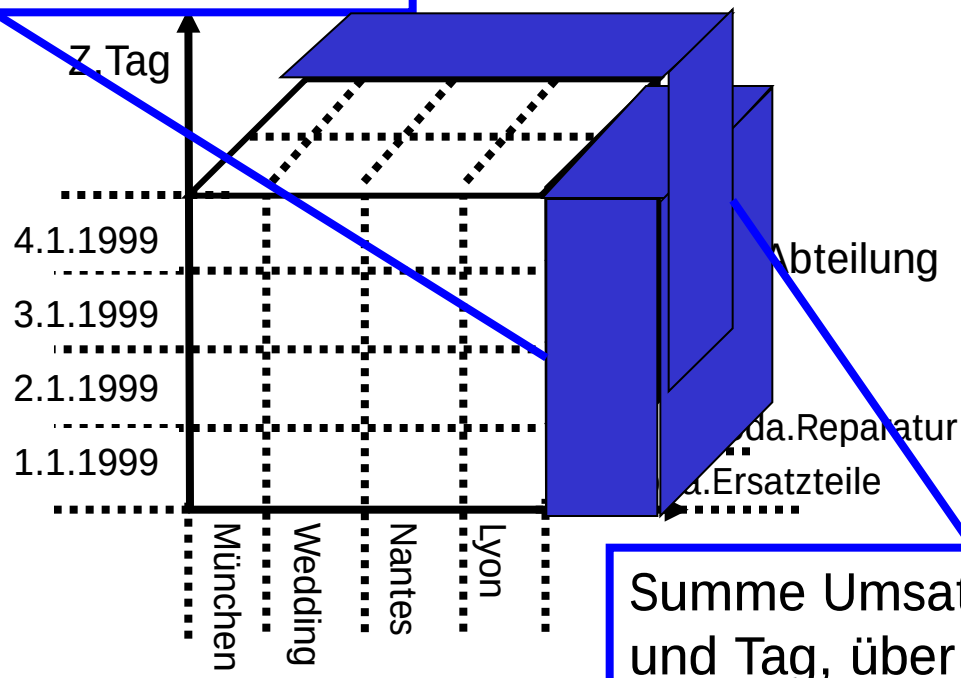
- Multidimensionale Datenmodellierung (MDDM)
- MDDM spricht die Sprache des Analysten
  - Operationen auf dem MDDM sollten die Analysevorgänge abbilden bzw. unterstützen
  - Ziel: Schnelle und intuitive Manipulation der Daten
- Notwendig
  - Definition der Operationen
  - Spezielle Anfragesprache (SQL-OLAP,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

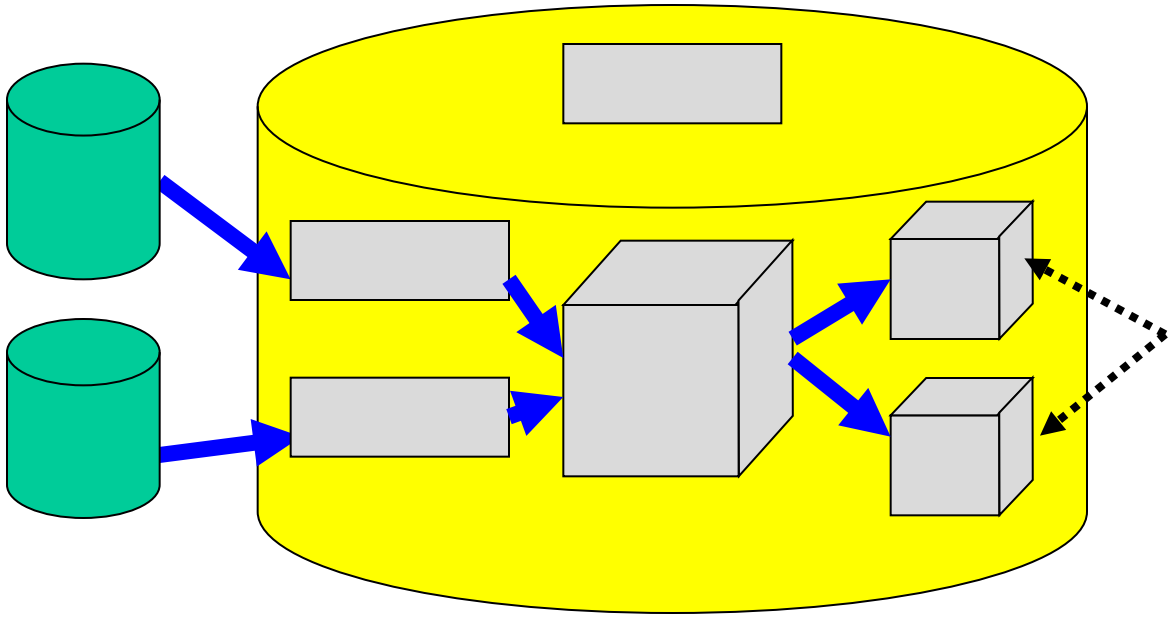


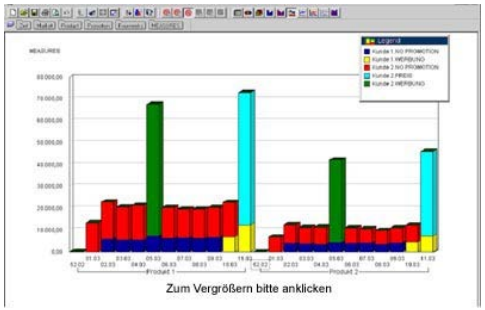
Summe Umsatz pro Shop und Tag, über alle Abteilungen

Inhalt dieser Vorlesung

- Data Warehouses
  - OLAP versus OLTP
  - Multidimensionale Modellierung
  - ETL Prozess
- Beispiel: Columba
- 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

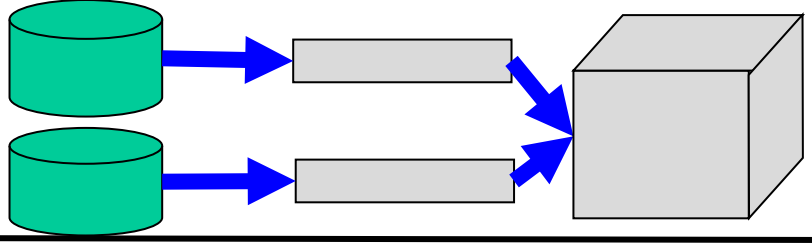
Art der Daten

- 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
  - 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

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

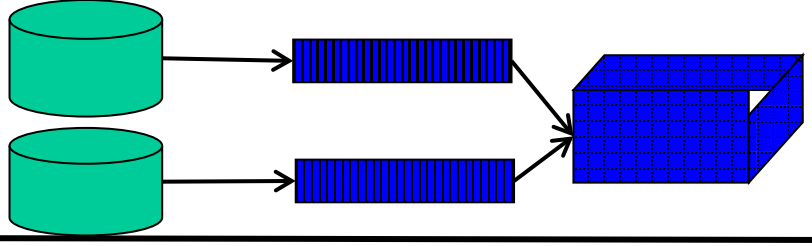
```
while (read_line)
  parse_line
  if (f[10]=2 & f[12]>0)
    write(file, f[1], string(f[4]), f[6]+f[7],...
  ...
bulk_upload( file)
```

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

Inhalt dieser Vorlesung

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

Columba

- Interdisciplinary project
  - Humboldt-Universität zu Berlin
  - Charité – Universitätsmedizin Berlin
  - Zuse-Institut Berlin
  - Technische Fachhochschule Berlin
- Integrated database
  - Data warehouse approach
  - Centered around PDB entries
  - 16 different data sources

Proteinstrukturen

Describing a Protein

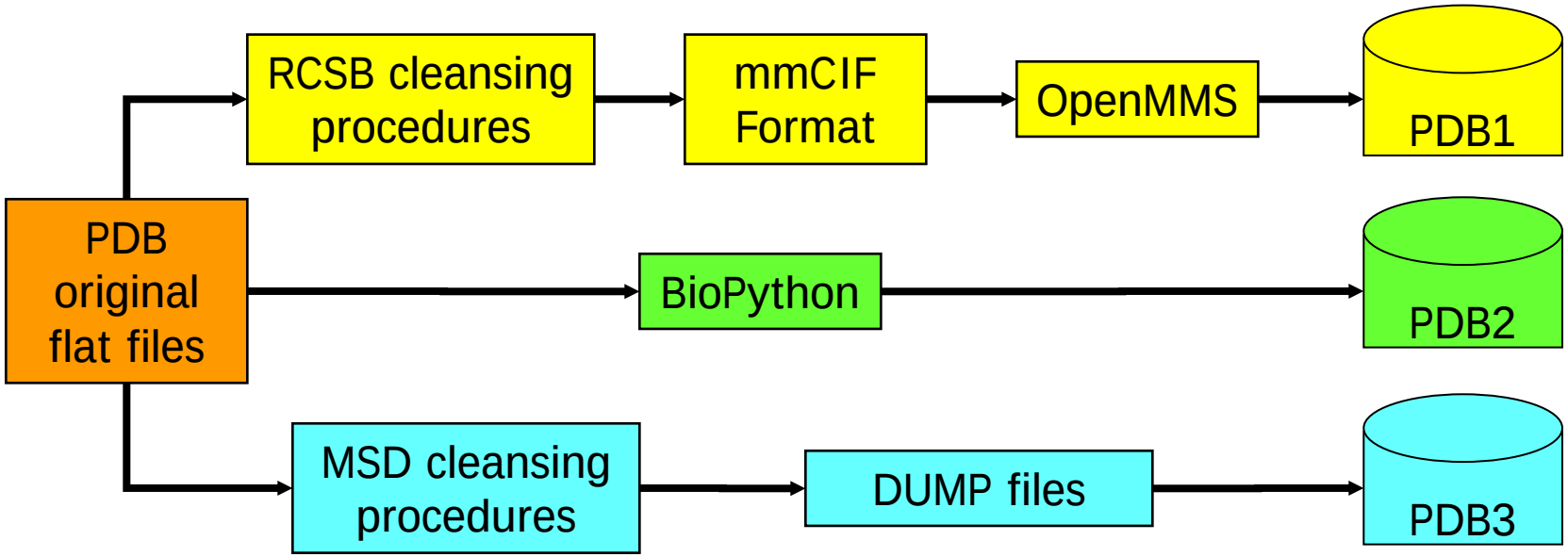
- Sequence (primary, secondary, ...)
- Motifs and domains
- Homologues in other species
- Members of the same protein family
- Function
- Association to diseases
- Mutations
- Isoforms
- Binding agents (ligands etc.)
- Interaction and pathways
- ...

Data sources

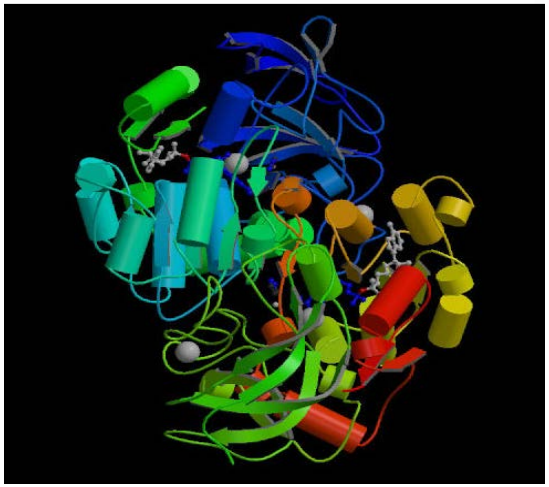
- PDB entries are the center of interest
  - Metadata/annotation, not on coordinates
  - Three potential sources
- Protein classification
  - CATH, SCOP (Structure), Systers
- Pathway / Enzyme
  - KEGG, Böhringer Map, Enzyme, Brenda
- Sequence
  - Swiss-Prot, Picses, InterPro
- Other
  - GeneOntology, NCBI Taxonomy, HSSP

Problem: Inconsistencies

- PDB has major problems with data quality
  - „Legacy“ data – format revisions
  - Misspellings, no controlled vocabularies, omitted updates, ...
- Three instances



Examples



No of entries per year, OpenMMS	
Year	**No**
1900	**35**
1909	**5**
1971	**1**
1972	**2**
...	**…**
1.4.2003	**7737**
…	**…**
2003	**1030**

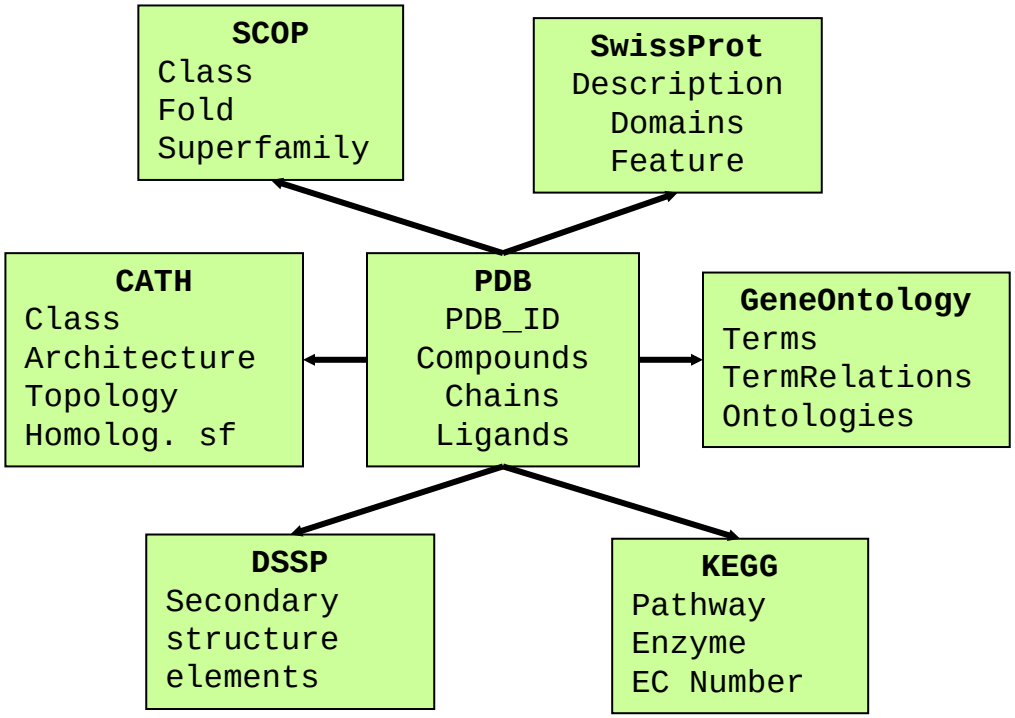
Examples

Mismatches in attribute values between OpenMMS and PDB – BioPython (20316 structures total)						
Name	**Year**	**Date**	**Method**	**Res**	**R_Val**	**R_Free**
20301	**75**	**121**	**3709**	**9165**	**20304**	**20306**

- Problem
  - Find inconsistencies quickly
  - Find pattern in inconsistencies
  - Help in resolving inconsistencies
  - Note: These inconsistencies are close to “resolvable errors”; there are no different experimentations or analysis results

Multidimensional Integration

- Sources are dimensions for PDB entries
- There is no “data” integration



Advantages

- Clear data provenance
- Simplified maintenance
- Users recognize their sources
  - Quality, trust, ...
- Intuitive query concept

Disadvantage

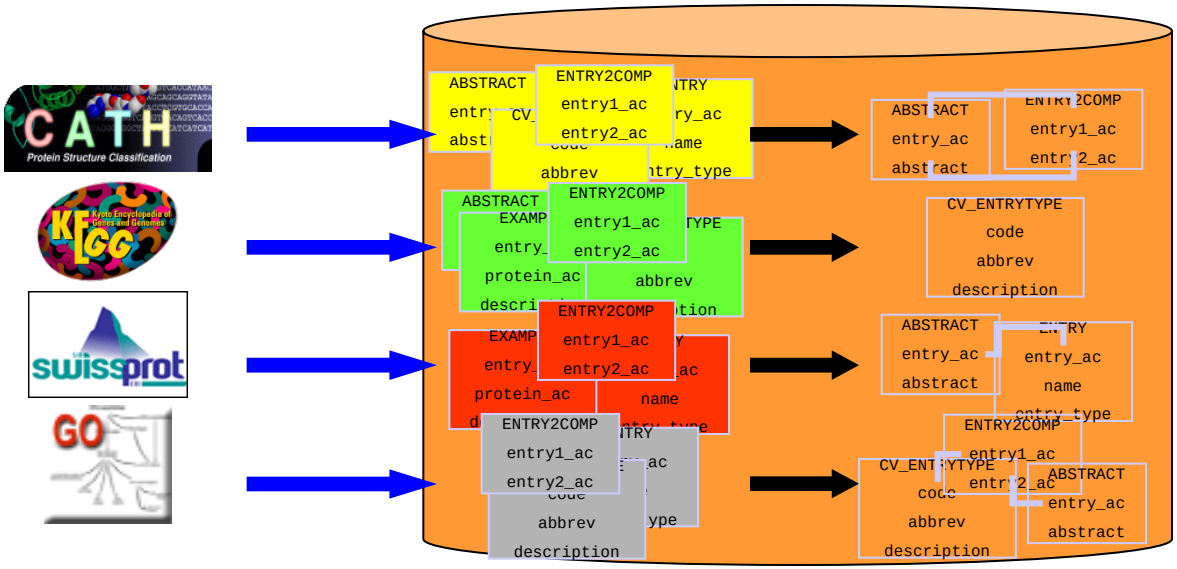
- No semantic integration
- Redundancy, duplicates, data conflicts

Widersprüche

- „Alle Enzyme, die ein ‚Tim Barrel‘ Motif enthalten“
  - Funktionstragende Substruktur, die in vielen Proteinen vorkommt
- Mögliche Quellen
  - CATH : Hierarchische Klassifikation von Substrukturen
  - SCOP: Dito (aber anders hergestellt und andere Definitionen von „Substruktur“)
  - Volltext
- Ergebnisse
  - Volltextsuche: 95 Proteine
  - Nur CATH: 79 Proteine
  - Nur SCOP: 90 Proteine
  - Jede Menge enthält Proteine, die in keiner der anderen enthalten sind

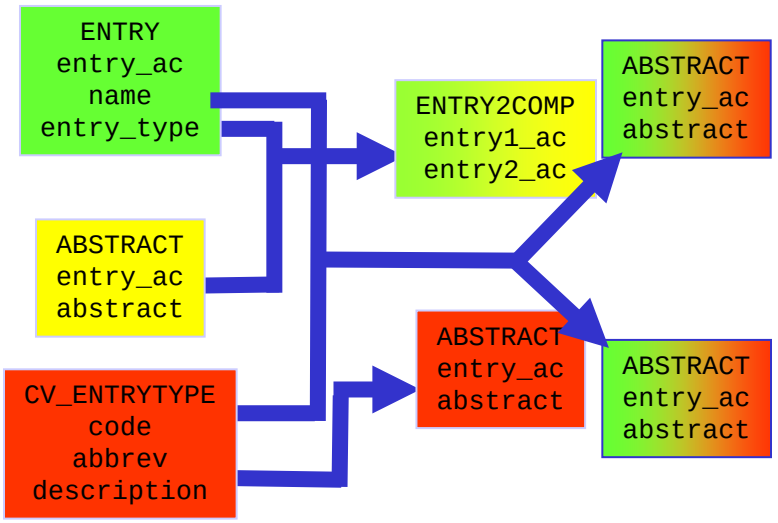
Step 1: Source integration

- Start: data source in original format
- Usage of existing tools (BioSQL, OpenMMS, ...)
- Parsing into native schema

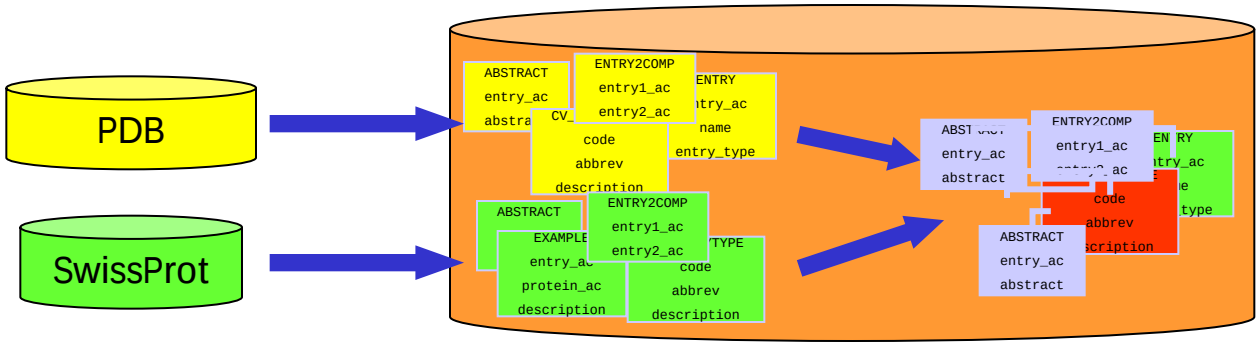


Step 2: Selection & Transformation

- All data exists in two schemas
  - Source schema:
    filled by external tools
  - Target schema:
    designed for Columba‘s needs
- All target schemas together form “the” Columba schema
  - Full query support,
    maximal performance
- Transformation through scripts



Advantages of two-stage integration



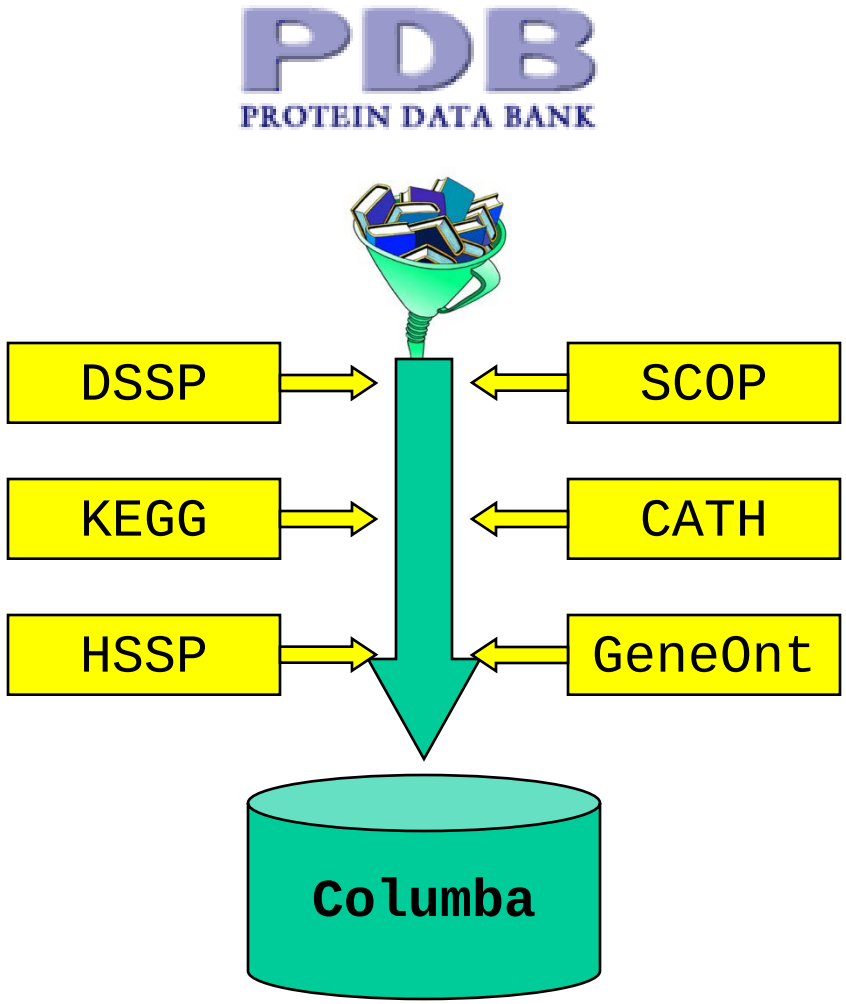
- Source schemas remain in realm of tool providers
- Clear separation of parsing & transformation – enhanced maintainability
- Original data always available for “drill-down”
- Target schema with full query capabilities, maximal performance

Materialized Integration is Never Finished

- Keeping data up-to-date is a challenge
- New PDB entries
  - Annotation pipeline
- New content of other data sources
  - Complete rebuild of target schema ☹
- New data sources
  - Parsing, transformations, pipeline interface

Keeping Columba Up-To-Date

- Annotation pipeline
  - New PDB entries are presented to source agents
  - Source-specific modules perform annotation process
    - Simple: Adjusting some keys
    - Complex: Resolve chains
    - External: Call program
- Inherently incremental
- Modules implement simple interface



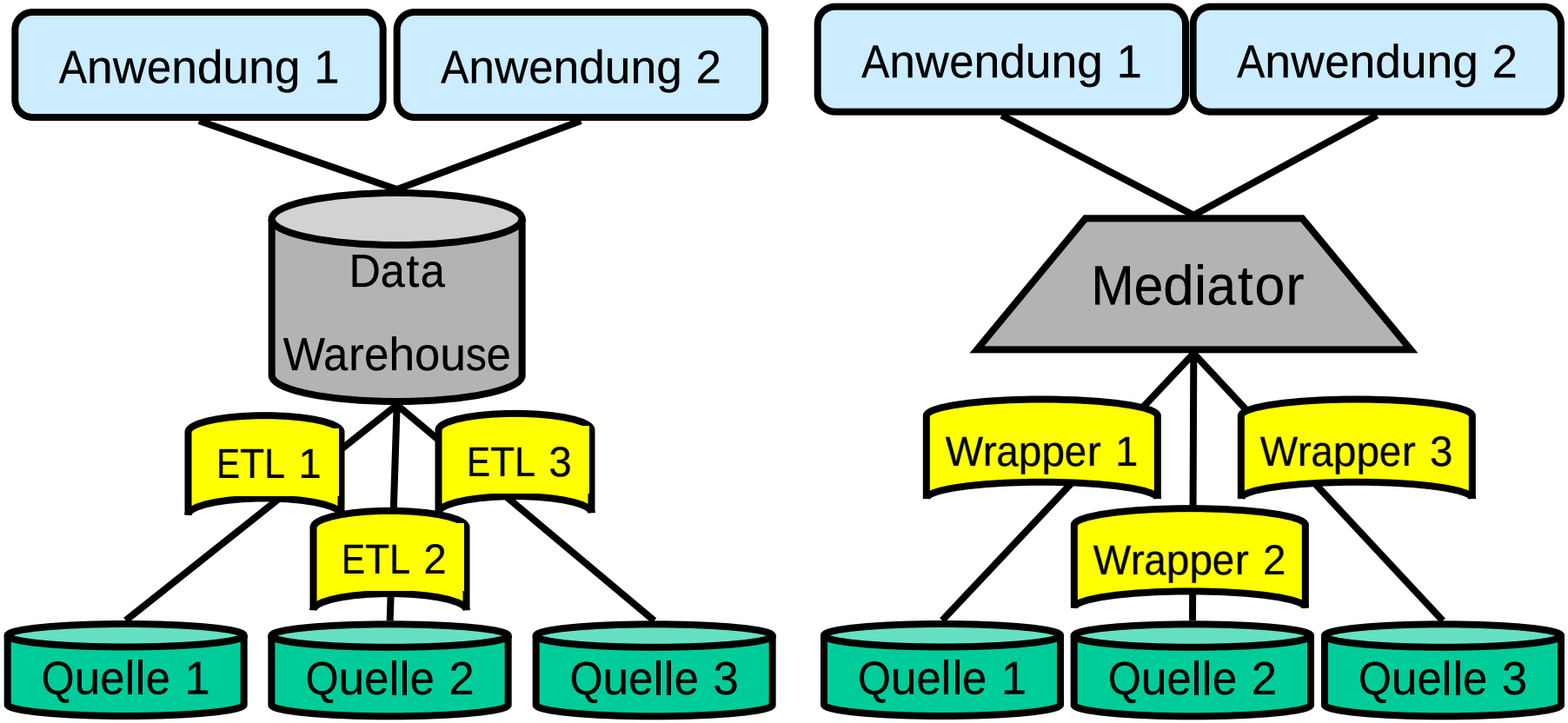
Einordnung

- Materialisierte Integration
- Heterogenität überwinden
  - Datenmodelheterogenität durch Parser
  - Strukturelle / semantische Het. durch SQL/Transformationsskripte
- Logische Verteilung bleibt im integrierten Schema erhalten
  - Erhöhte Wartbarkeit
  - Fehlende Autorität zur semantischen Integration
  - Vertrautheit der Benutzer mit ihren Lieblingsquellen
- Quellen existieren in zwei Schemas
  - Originaldaten bleiben erhalten (Kontrolle und Drill-Down)
  - Zielschema ist auf Columba abgestimmt

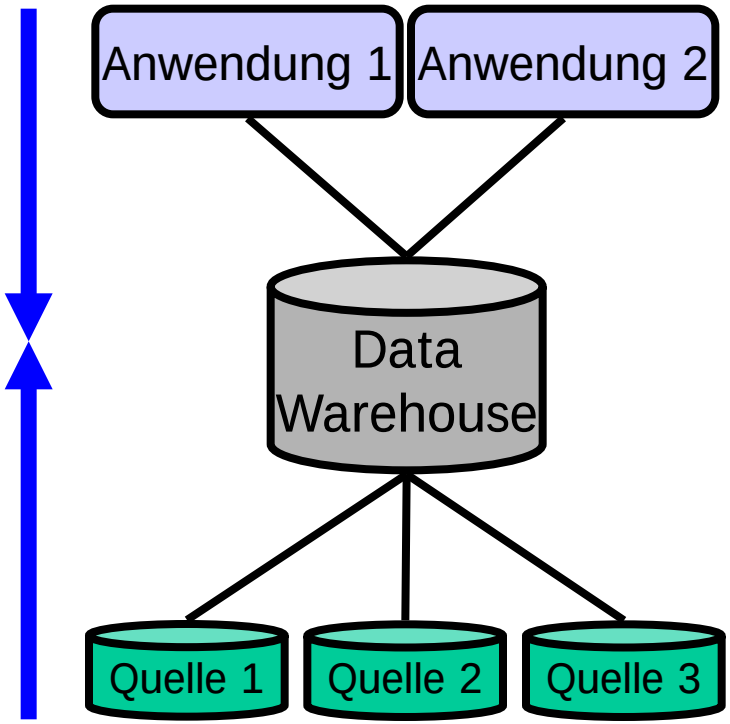
Inhalt dieser Vorlesung

- Data Warehouses
- Beispiel: Columba
- Vergleich: Materialisierte versus virtuelle Integration

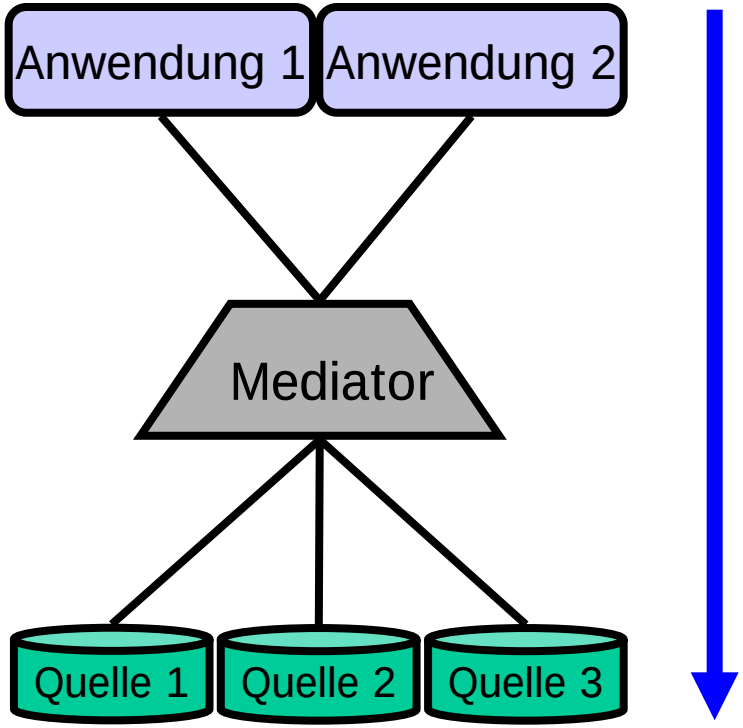
Data Warehouse vs. Mediator



Datenfluss



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