

Kurs  
Einführung  
In das Thema  
Data Ware House  
&  
Business Intelligence

*Organisatorisches*

# Warm werden: Vorstellungsrunde

- Name: Joachim Brock
- Baujahr 1965, Verheiratet und zwei Töchter
- Ausbildung im Handwerk
- Studium: Kommunikationsinformatik
- Kerngebiete: Datenbanken, Business Intelligence, Schnittstellen und Software Entwicklung
- Werdegang: Softwarehaus, Eigene GmbH, Big-Date&KI, Expertensysteme, aktuell: Head of SWE bei AUVESY-MDT GmbH in Landau
- Erwartung: Interessierte lernbereiter Kurs mit engagierter Zusammenarbeit



# Warm werden: Vorstellungsrunde

- Vorstellungsrunde
  - Name
  - Werdegang
  - Erwartungshaltung an den Kurs

# Warm werden: Gemeinsame Regeln

- Regeln der Vorlesung:
  - 1) Der aktuelle Redner darf ausreden
  - 2) Fragen sind von allen erwünscht und nicht nur von wenigen
    - Manche Blöcke muss ich am Stück erklären. Keine Panik, Zeit zum Fragen kommt am Ende des Blocks; Dies Blöcke benenne ich extra
  - 3) Pausen bitte einfordern, ich vergesse manchmal die Zeit
  - 4) Bei Online: Wenn die Technik mit macht sollten wir die Kameras einsetzen

# Warm werden: Orga

- Organisation der Vorlesung
  - Theorie
  - Beispiele aus der Praxis
  - Aufgaben und Übungen
    - Vorstellung durch Teilnehmer
  - Wiederholung der vorherigen Vorlesung im Schnelldurchgang
  - Prüfung KW 51/2023
    - **Fragen vorab oder noch offene Punkte?**

# Kurs Einführung In das Thema Data Ware House & Business Intelligence

*Capitel 1: Erste Grundlagen*

# Business Intelligence

## Was umfasst BI?

- OLAP (Online Analytical Processing)
  - Umfasst auch das DWH samt Datenbanken
  - Weitere Themen sind Verteilung, Datenbeschaffung, Daten Vorbereitungen
- Analyse
  - Auswertungen, Statistiken, zyklisch & adhoc, Entscheidungsvorlagen
- Data Mining
  - Korrelationen, Kausalitäten, Wissenbasiertes Lernen und Prognosen
- Projektorganisation
  - Planung, Aufbau, Pflege und Betrieb von BI-Systemen

# OLAP

## 12 Regeln nach Edward F. Codd aus dem Jahr 1993

- 1) Multidimensionale Sicht auf die Daten
- 2) Transparenz (Trennung von UI und Architektur)
- 3) Zugriffsmöglichkeiten (Daten aus Operativen Systemen)
- 4) Konsistente Leistungsfähigkeit der Berichterstattung
- 5) Client-Server-Architektur mit Lasterverteilung
- 6) Generische Dimensionalität mit einheitlicher Dimensionierung
- 7) Dynamische Handhabung dünn besetzter Matrizen
- 8) Mehrbenutzerunterstützung
- 9) Einheitliche dimensionsübergreifende Operationen
- 10) Intuitive Datenanalyse
- 11) Flexibles Berichtswesen
- 12) Unbegrenzte Anzahl von Dimensionen und Konsolidierungsebenen

Quelle: <https://www.hdm-stuttgart.de/~riekert/lehre/db-kelz/chap6.htm>

# OLAP

## FASMI-Regeln nach Pendse und Creeth

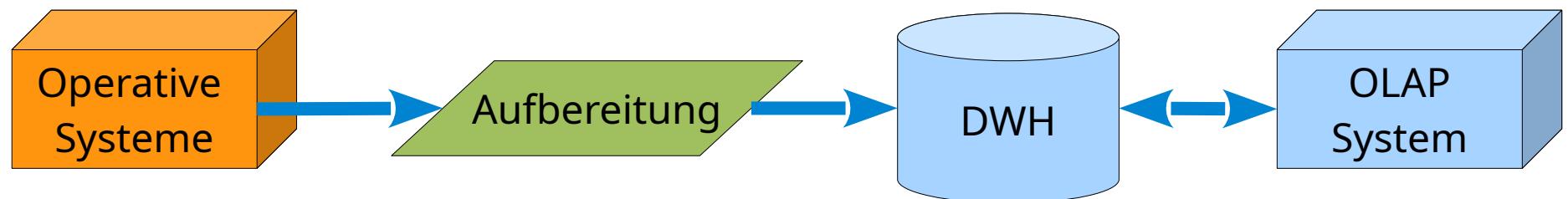
- 1) Fast** = Schnelle Abfragen mit durchschnittlich 5 s bis max. 20 s
- 2) Analysis** = Einfache Analyse der Daten ermöglichen möglichst ohne Programmieraufwand
- 3) Shared** = Mehrbenutzerbetrieb ermöglichen mit entsprechenden Schutzmaßnahmen
- 4) Multidimensional** = Struktur der Daten ermöglichen beliebige Dimensionshierarchien
- 5) Information** = Die Daten dürfen nicht durch das Systems in ihrer Transparenz beschränkt werden

Quelle: <https://www.datenbanken-verstehen.de/business-intelligence/business-intelligence-grundlagen/anforderungen-business-intelligence/fasmi-regeln-pendse-creeth/>

# OLAP DWH-Architektur

## DWH Systeme samt Aufbau und Betrieb

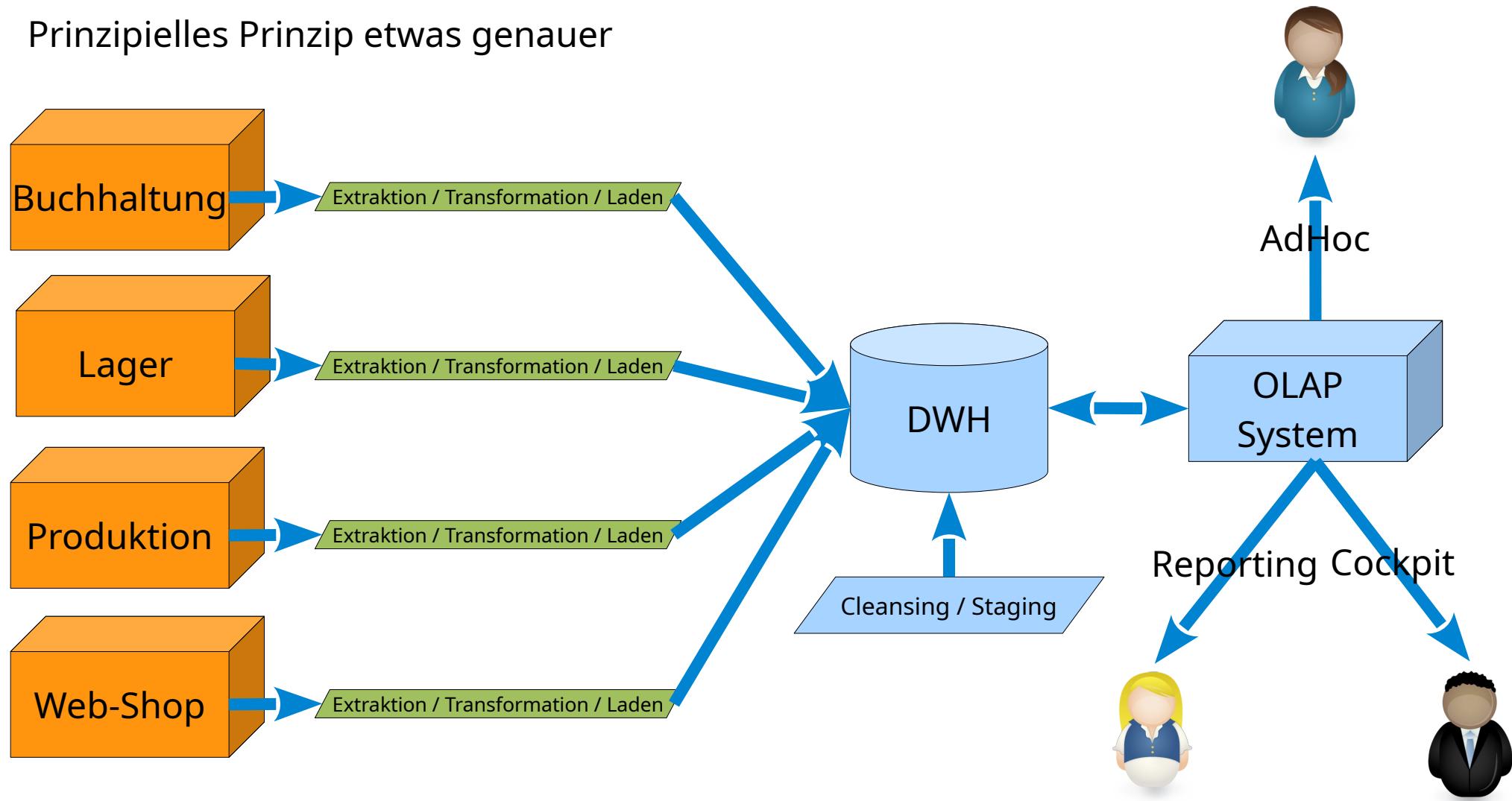
Prinzipielles Prinzip



# OLAP DWH-Architektur

## DWH Systeme samt Aufbau und Betrieb

Prinzipielles Prinzip etwas genauer



# OLAP

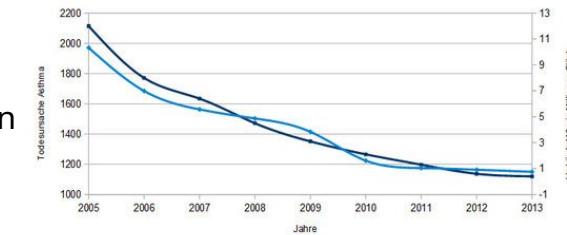
## OLAP Anwendung

- Konzentration unterschiedliche Datenquellen
  - Datenreihen, Produktionsanlagen, Wirtschaftssysteme, Statistiken, ...
- Ermöglicht globale Sicht auf Daten
  - Konzentration und Aggregation von unübersichtlichen Datenmengen
- Ermitteln von Korrelationen und Kausalität
  - Bitte nicht verwechseln!
  - Beispiel: Selbstmordrate und Nasa-Investitionen
- Beobachtung von Daten-Entwicklungen, auch von temporalen
  - Bis hin zu echzeitbeachtungen in Monitoring-Systemen
- Entscheidungsunterlage bieten
  - Analyse-Ergebnisse als Basis für Entscheidungen
  - Beispiel: Vorhersage Papierdicke zur Produktionssteuerung

# OLAP

## OLAP Risiken

- Unvollständige Dimensionen
  - Korrelationen hängen an Dimensionen die nicht enthalten sind
- Fehlerhafte Daten
  - Verfälschung von Analyseergebnissen
- Große Lücken in den Faktendaten
  - Lücken in den Daten reduzieren Aussagekraft
- Dupletten in den Daten oder fehlende single Point of Truth
  - Widersprechende oder falsch verstärkende uneindeutige Wirkungen
- Datenschutzverletzungen
  - Einfache personenbezogene Daten
  - Besonders geschützte personenbezogene Daten, z.B. Medizin
- Scheinkorrelationen ohne Kausalitäten
  - Beispiel: MC<=>Asthma, Quelle: Marc Oliver Schmidt, Scheinkorrelation

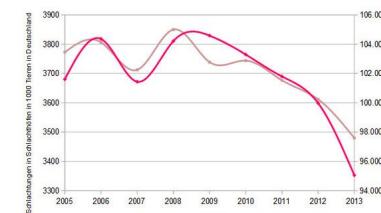




# OLAP

## Verständnisfragen

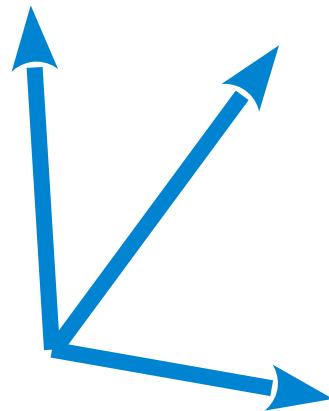
- Warum werden operative Daten von Auswertungsdaten getrennt?
- Welche Bereiche in einem Unternehmen haben Interesse an OLAP?  
Bitte nennen Sie min. 4 Bereiche und deren Nutzen.
- Zwei Vertriebsstellen mit überschneidenden Rechnungskreise benötigen ein gemeinsames OLAP, was empfehlen Sie als ersten Schritt?
- Ein Pharmakonzern bietet Ihnen hohe Summen für Ihre medizinischen Labordaten. Was ist zu beachten?
- In den Jahren 2005 bis 2013 korrelieren die Schlachtungen in deutschen Schlachthöfen und die Sitzplatzkapazität der österreichischen Kinos.  
Was leiten Sie daraus ab?



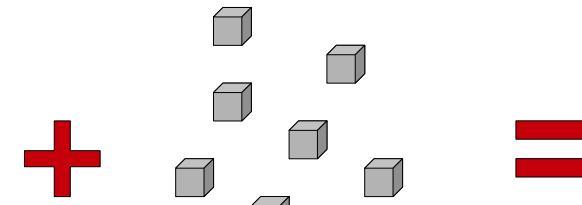
# OLAP DWH-Architektur

## DWH grundsätzliche logische Architektur

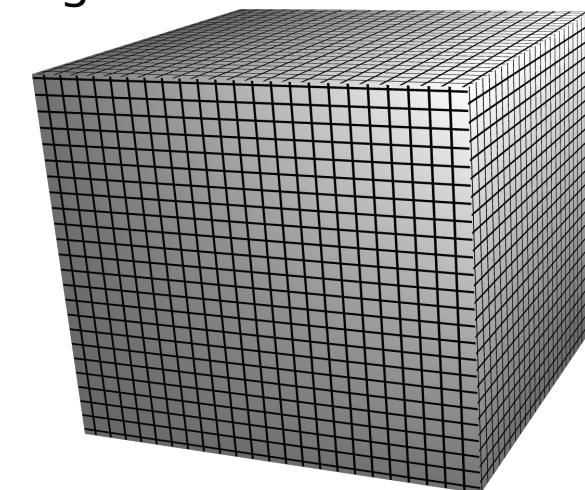
- Mehrdimensionale logische Datenarchitektur
- Fakten werden in Form einer Matrize gehalten
- Gruppierung der Stammdaten bestimmen die Dimensionen
- Symbolische Ähnlichkeit mit einem Würfel
- Normalisierung zugunsten Performance vernachlässigt



Dimensionen



Fakten



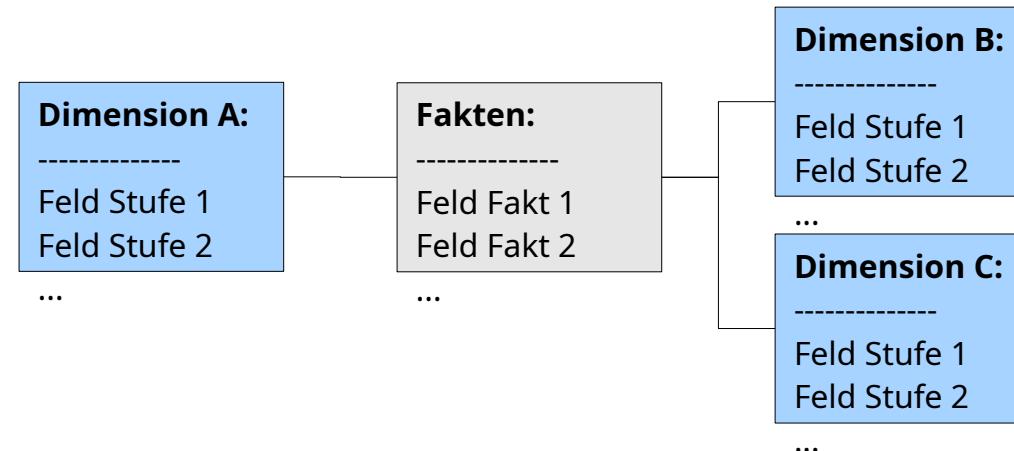
Data Warehouse Würfel

# OLAP DWH-Architektur

## DWH logische Architektur der Dimensionen

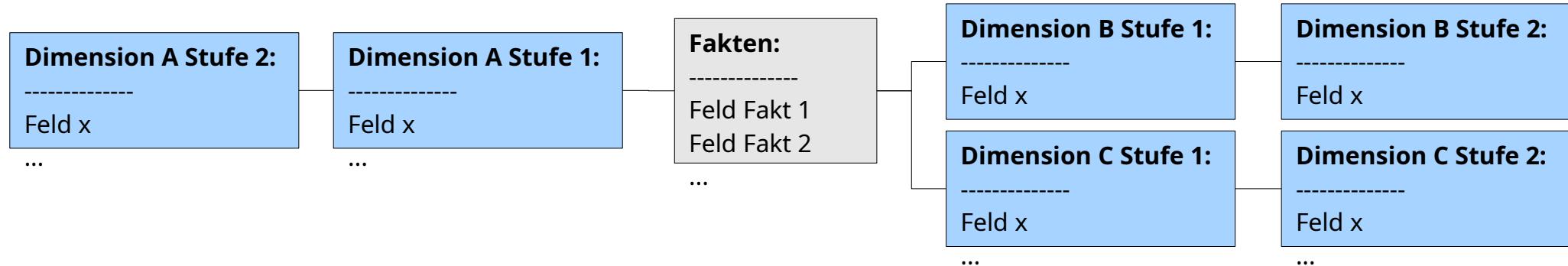
### → Star Schema

- Jede Dimension wird in einer Tabelle/Objekt zusammengefasst



### → Snowflake Schema

- Jede Hierarchie einer Dimension wird in einer Tabelle/Objekt gehalten



# OLAP DWH-Architektur

## DWH logische Architektur der Dimensionen

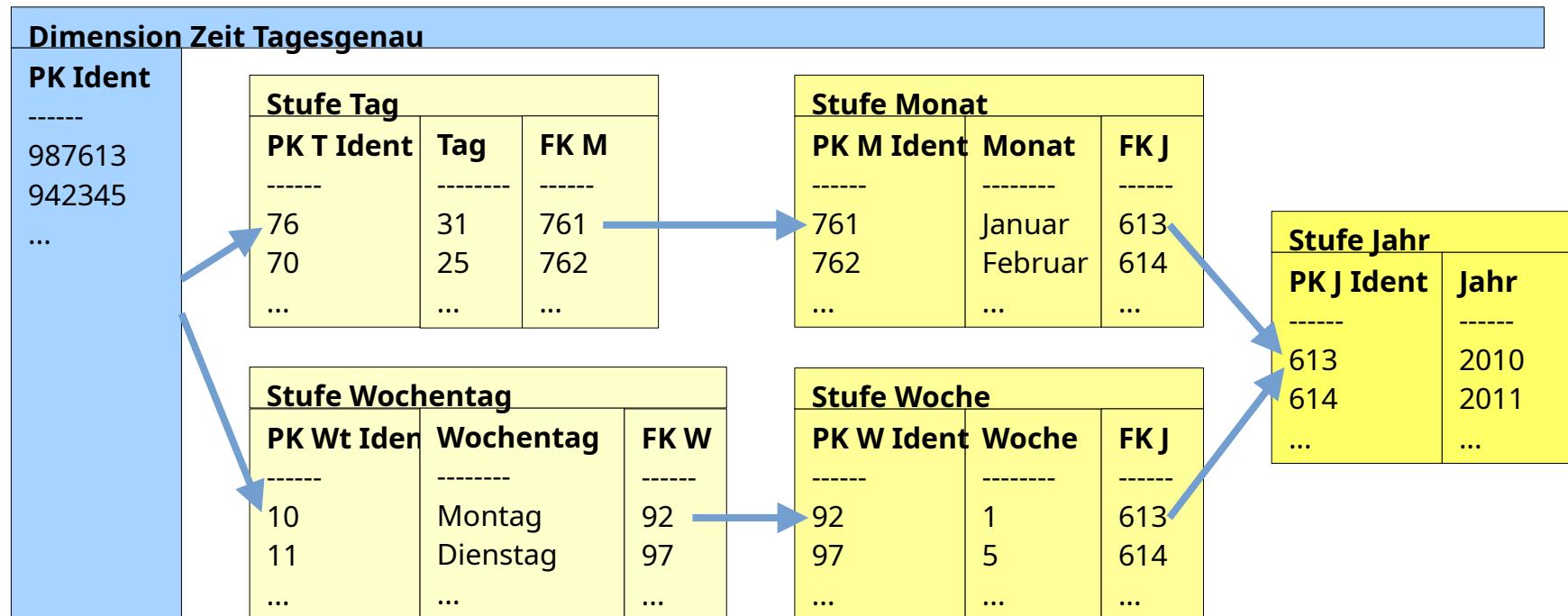
- Beispiel Star Schema für Zeitreihen oder Temporalisierung
  - Pro Tag einen Eintrag
  - Mehrere Pfade der Hierarchie gleichzeitig möglich
  - Einen künstlichen Primärschlüssel ist empfehlenswert

Dimension Zeit Tagesgenau					
PK Ident	Jahr	Monat	Woche	Tag	Wochentag
-----	-----	-----	-----	-----	-----
987613	2010	Januar	1	31	Montag
942345	2011	Februar	5	25	Dienstag
...	...	...	...	...	...

# OLAP DWH-Architektur

## DWH logische Architektur der Dimensionen

- Ein Beispiel Snowflake Schema für Zeitreihen oder Temporalisierung
  - Pro Tag mehrere gleichzeitige Pfade möglich
  - Jeweilige künstlichen Primärschlüssel ist empfehlenswert
  - Traversierung durch die Hierarchiestufen notwendig



# OLAP DWH-Architektur

## DWH logische Architektur der Dimensionen

Stufe Tag		
PK T Ident	Tag	FK Monat
---	---	---
73	1	761
75	1	792
76	31	901
70	25	901
...	...	...

Stufe Monat		
PK M Ident	Monat	FK Jahr
---	---	---
761	Januar	613
762	Februar	613
792	Januar	614
901	Januar	615
...	...	...

Stufe Jahr	
PK J Ident	Jahr
---	---
613	2020
614	2021
...	...

Fakten Vertrieb			
PK ID	FK T	FK N	Umsatz
5001	73	1001	10.333,05 €
5002	75	1001	42.123,44 €
5003	73	1003	72.042,42 €
5042	70	1042	1.099,01 €
...	...	...	...

- Beispiel Snowflake Schema Zeitreihen und Regionen
- Fakten referenzieren nur auf die niedrigste Granularität
  - Redundanzen in den Stufen ggf. notwendigen

Stufe Niederlassung		
PK N Ident	Name	FK S
---	---	---
1001	Zentrale	761
1002	Notre Dame	792
1003	Limmat	901
1042	Oerlikon	901
...	...	...

Stufe Stadt		
PK S Ident	Stadt	FK Land
---	---	---
761	Mannheim	49
792	Paris	33
712	Luzern	41
901	Zürich	41
...	...	...

Stufe Land	
PK L Ident	ISO2
---	---
49	DE
41	SW
33	FR
...	...

Umsatz 42.123,44 €  
am 1. Januar 2021  
in der deutschen Zentral in  
Mannheim



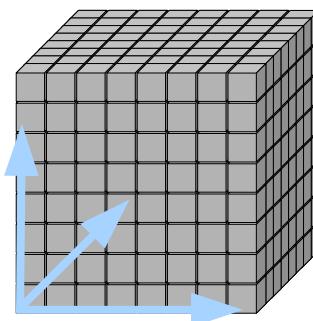
# OLAP DWH-Architektur

## Verständnisfragen

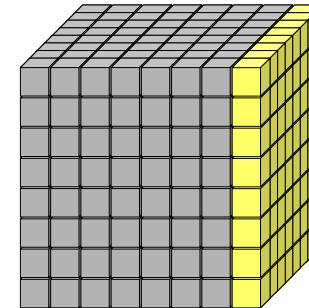
- Warum wird im DWH gerne gegen die Normalformen verstossen?
- Nennen Sie min. zwei weitere Dimensionen, bei denen ein Verstoß gegen Normalform sinnvoll ist.
- Sie sollen ein DWH entwerfen mit extrem schnellen Zugriffzeiten. Welches Modell (Star oder Snowflake) wählen Sie?
- Bauen Sie ein Starschema für den Vertrieb mit folgenden Inhalten: Verkaufzeitstempel, Menge, Einzelpreis, MwSt, Produktnname, Produktgruppe, Filiale, Ort, PLZ.
- Schreiben Sie eine SQL Abfrage, welche die Verkaufssumme der jeweiligen Produktgruppen in den Filialen zwischen 12:00 Uhr und 13:00 Uhr am 18.08.2022 ausgibt.

# OLAP DWH-Datenabfragen

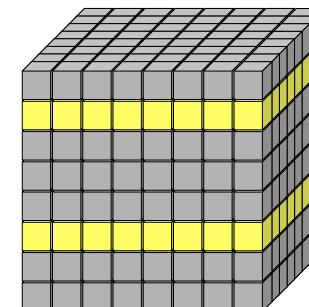
## DWH logisches Bewegen in den Daten



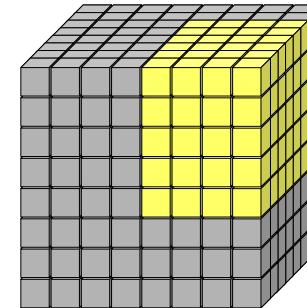
Dimensionen



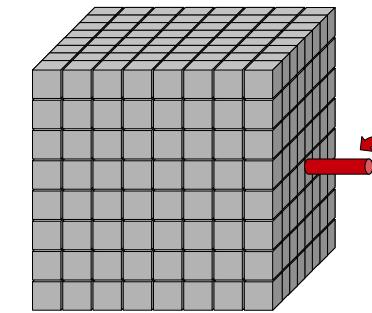
Slice = Nur eine  
Scheibe  
betrachten



Dice = Nur einen  
Subwürfel  
betrachten



Drill-Down oder Drill-  
Thru und  
Drill-Up = Aggregation  
in den Hierarchie tiefer  
und höher einsehen



Pivot oder Rotate =  
Drehen der Daten,  
um weitere  
Dimensionen sichtbar  
zu machen

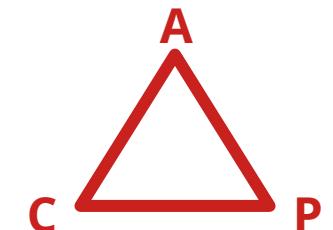
# OLAP DWH-Architektur

## DWH Transaktionsmodell versus operative Datenbanken

- Operative Datenbanken meist im ACID/AKID Modell
  - AKID = Atomar, Konsistent, Isoliert und Dauerhaft
- DWH oft andere Transaktionsmodelle
  - CAP/KVA-Theorem = konsistent, verfügbar und ausfallsicher
  - NoSQL und Eventual consistency = schlussendlich Konsistenz

### → Nachteile

- Massendatentauglich
- Verteilbare Last
- Hohe Verfügbarkeit
- Konsistenzmanagement
- Schreib-Lese-Konflikte
- Multiversionen





# OLAP Navigation und Transaktionsmodell

## Verständnisfragen

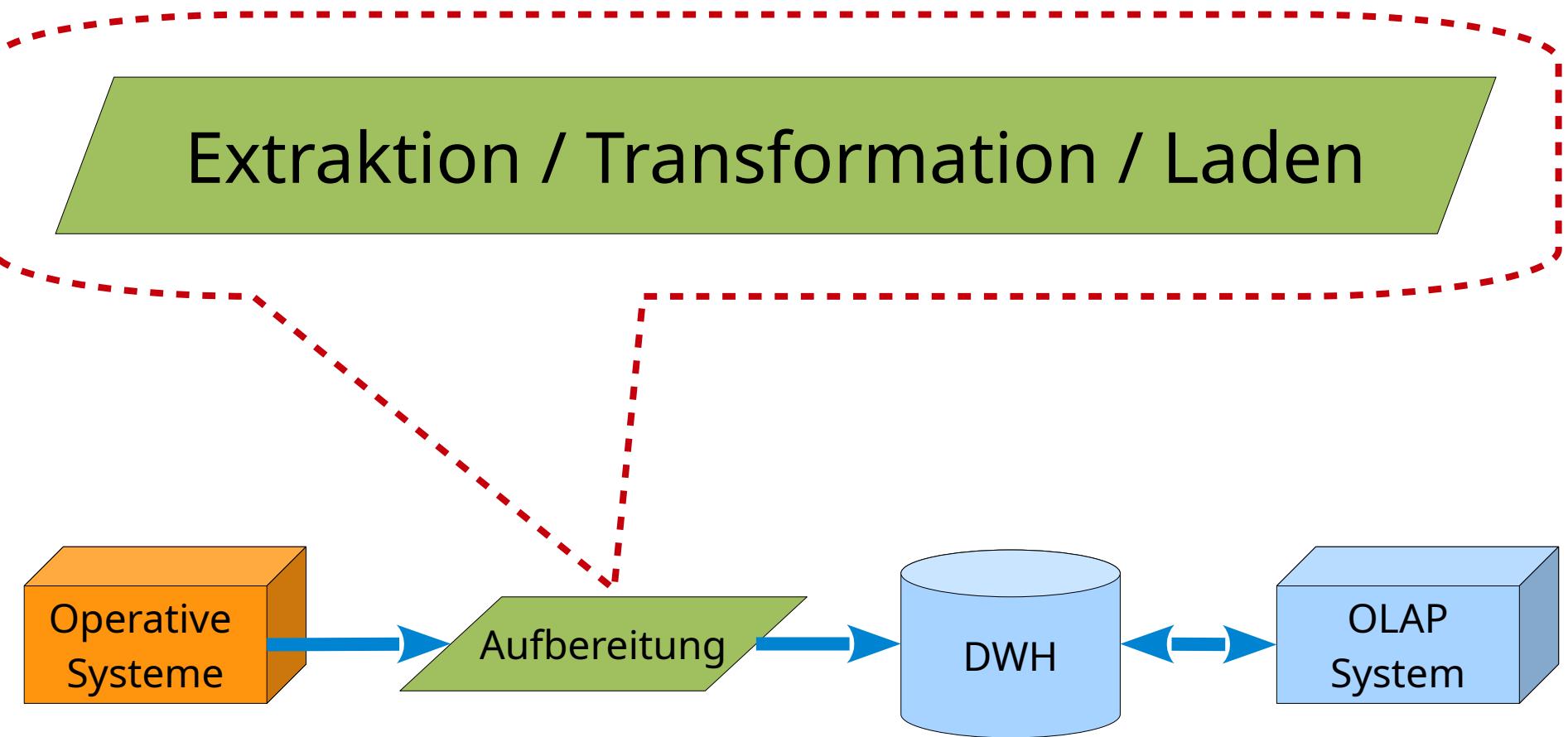
- Mit welchen SQL Schlüsselwörtern können händisch Slice und Dice umgesetzt werden?
- Zu welcher DWH Navigation passt das SQL Schlüsselwort "UNION"?
- Wieso kann in OLAP Systemen von dem ACID Modell abgewichen werden?
- Zeigen Sie ein Szenario, in dem es zu Multiversionen in einem OLAP System kommen kann.
- Welche Auswirkungen können Konsistenzprobleme auf die Korrelationsanalyse haben?

Kurs  
Einführung  
In das Thema  
Data Ware House  
&  
Business Intelligence

*Kapitel 2: Datenprozesse*

# Datenprozesse

## Was macht ETL?



# Datenprozesse

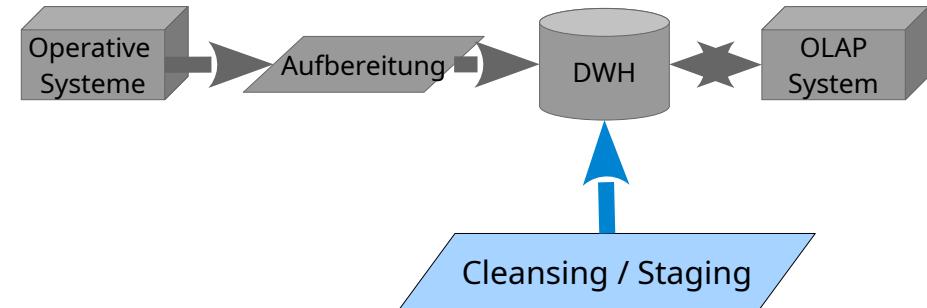
## Was macht ETL?

Extraktion / Transformation / Laden

- Extraktion
  - Heterogene Welt an BI anbinden
  - Unterschiedliche Quellsysteme, d.h. Datenquellen und Schnittstellen
  - Extraktion nach Zyklus, Ereignis oder Bereitstellung durch Quelle
  - Deltaextraktion
- Transformation
  - Vereinheitlichen von Datenformate, Bezugspunkte und Datenstrukturen
  - Mergen von Daten
  - Entfernen von Dubletten
  - Erhöhen der Datenqualität
- Loading
  - Laden der Daten in das DWH-System
  - Pufferung in einem Staging-Bereich der Daten möglich

# Datenprozesse

## Erweiterung des ETL-Prozesses



- Staging-Area
  - Pufferbereich um Daten
  - Schnellstmögliche Laden der Daten in den Bereich
  - Nachverarbeitung möglich
  - Datenverteilung aus dem Stagingarea in die Fakten und Dimensionen
- Cleansing Prozess
  - Verknüpfung der Staging Daten mit den Dimension
  - Temporale Bereinigung von Daten
  - Reduzierung von Datenlücken in den Faktentabellen
  - Zusätzlicher Ladeprozess der Daten in Fakten und Dimensionen
  - Weitere Erhöhung der Datenqualität

# Datenprozesse

## Keine Starre Kombination der ETL-Prozesses

- E T L
  - Traditionelle Vorgehenweise
- E L T
  - Transformation finden im DWH statt
  - Beliebt in Big Data Bereichen
  - Meist umfangreicher Cleansing Prozess oder Verzicht auf Cleansing
- T E L
  - Die Daten werden in der Quelle aufbereiten
  - Seltene Form da höhere Bindung an das Quellsystem

# Datenprozesse

## Tools für den ETL-Prozesses

- Classische ETL-Tools
  - Aufbau nach dem Filter&Pipe Muster
  - Grafischer Editor für das Erstellen des Ablaufs
  - Typische Vertreter: SAS, DWH Builder (OWB), Pentaho DI (NEC), Talend, ...
  - Verteilte Ausführung oft integriert
- Aufbau nach dem Map Reduce Princip
- Eigenentwicklung leider auch oft anzufinden, z.B. reine SQL, Python Entwicklung



# ETL

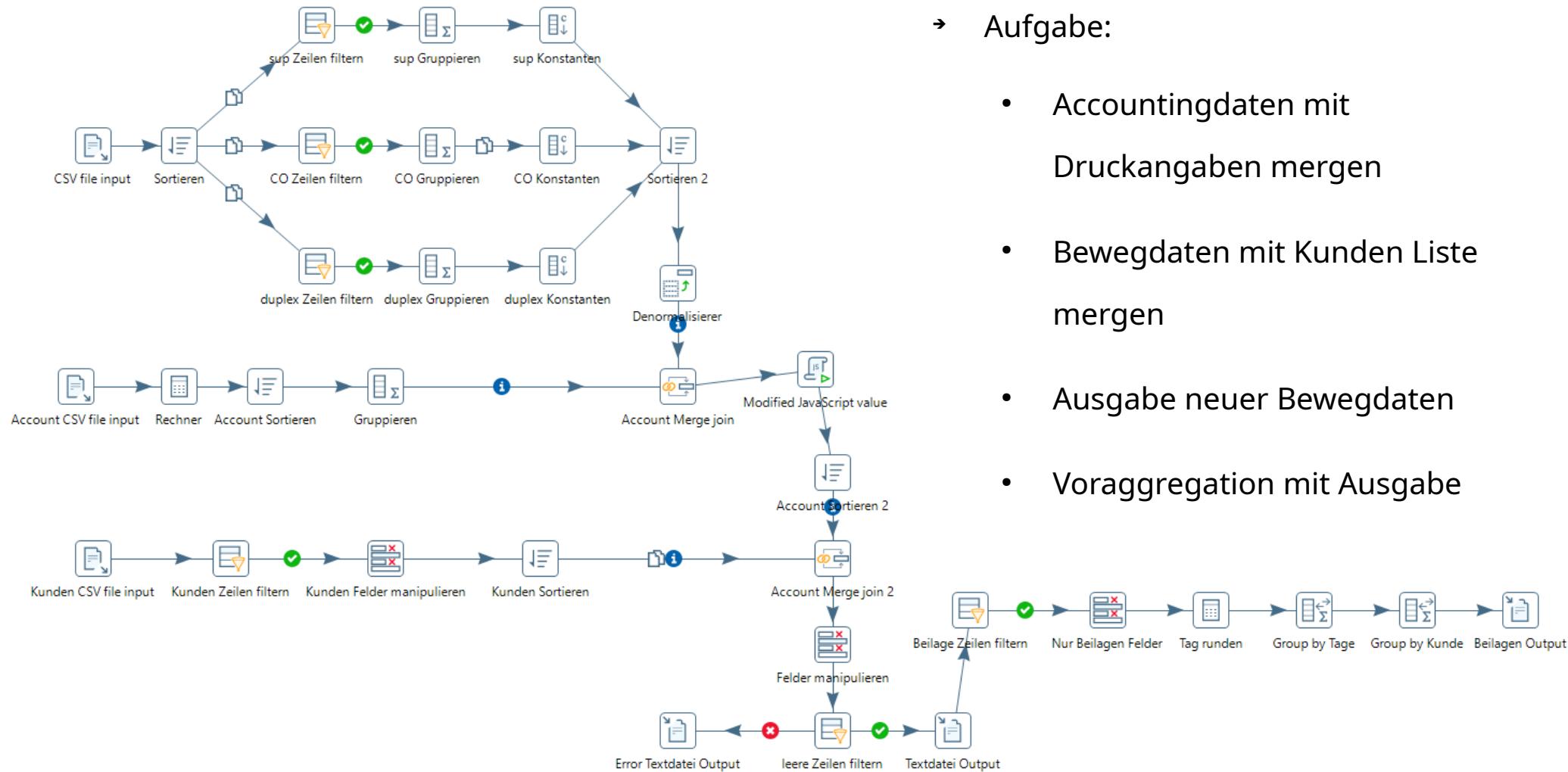
## Verständnisfragen

- Nennen Sie min. 5 Hersteller von ETL Tools und für welchen Einsatzzweck diese besonders geeignet sind.
- Welche Datenbereinigungen sind aus einer Produktionsmaschine denkbar deren Produktionslogs in ein DWH transformiert werden sollen?
- Nennen Sie 8 Problematische Datenquellen für DWH die schwer zu transformieren sind.
- Welchen maximalen Laufzeit darf ein ETL-Prozess haben?  
(Bitte keine Angabe in s, min., h)

# Datenprozesse

## Beispiel für den ETL-Prozess

Pentaho Data Integration

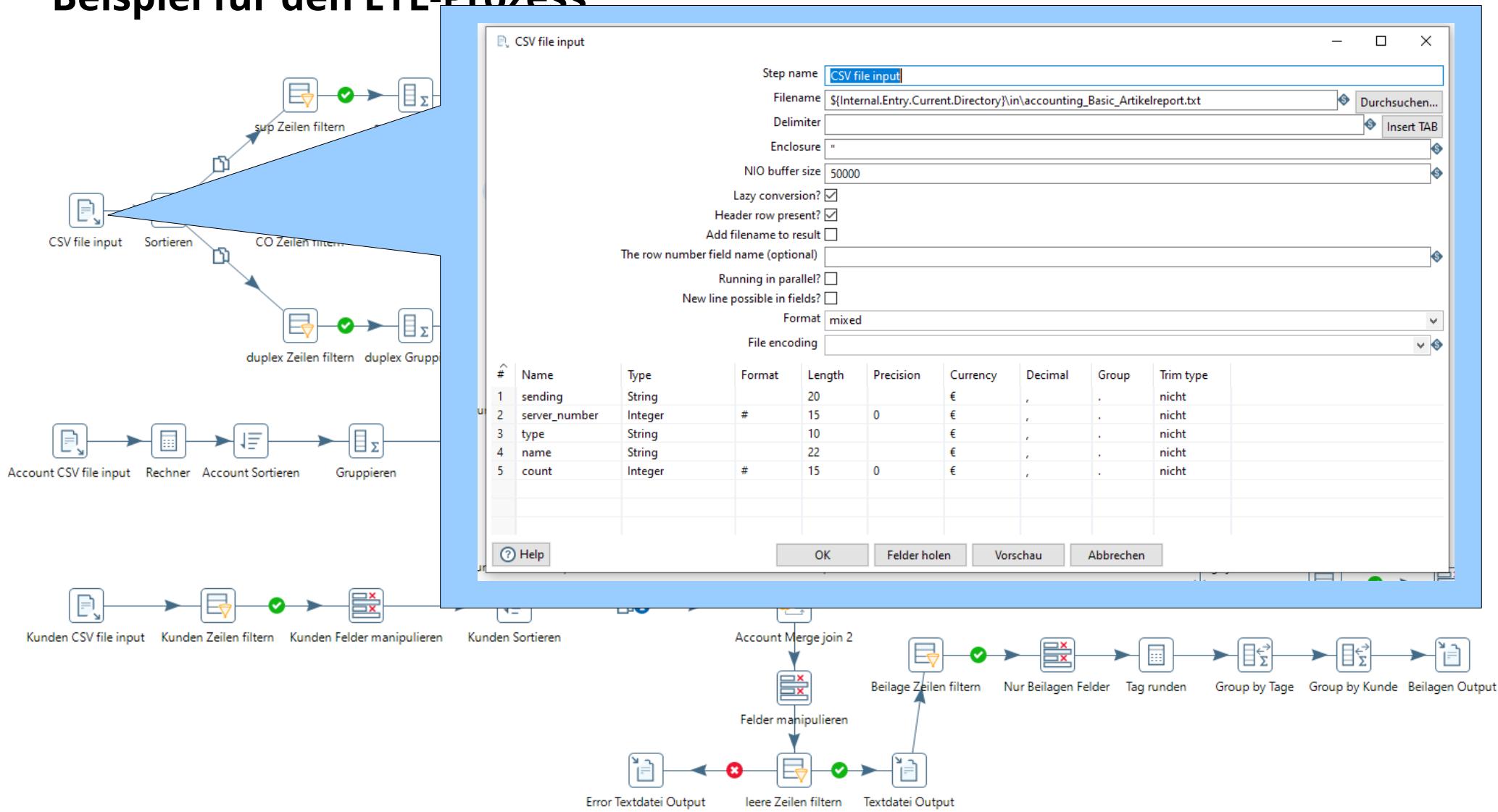


→ Aufgabe:

- Accountingdaten mit Druckangaben mergen
- Bewegdaten mit Kunden Liste mergen
- Ausgabe neuer Bewegdaten
- Voraggregation mit Ausgabe

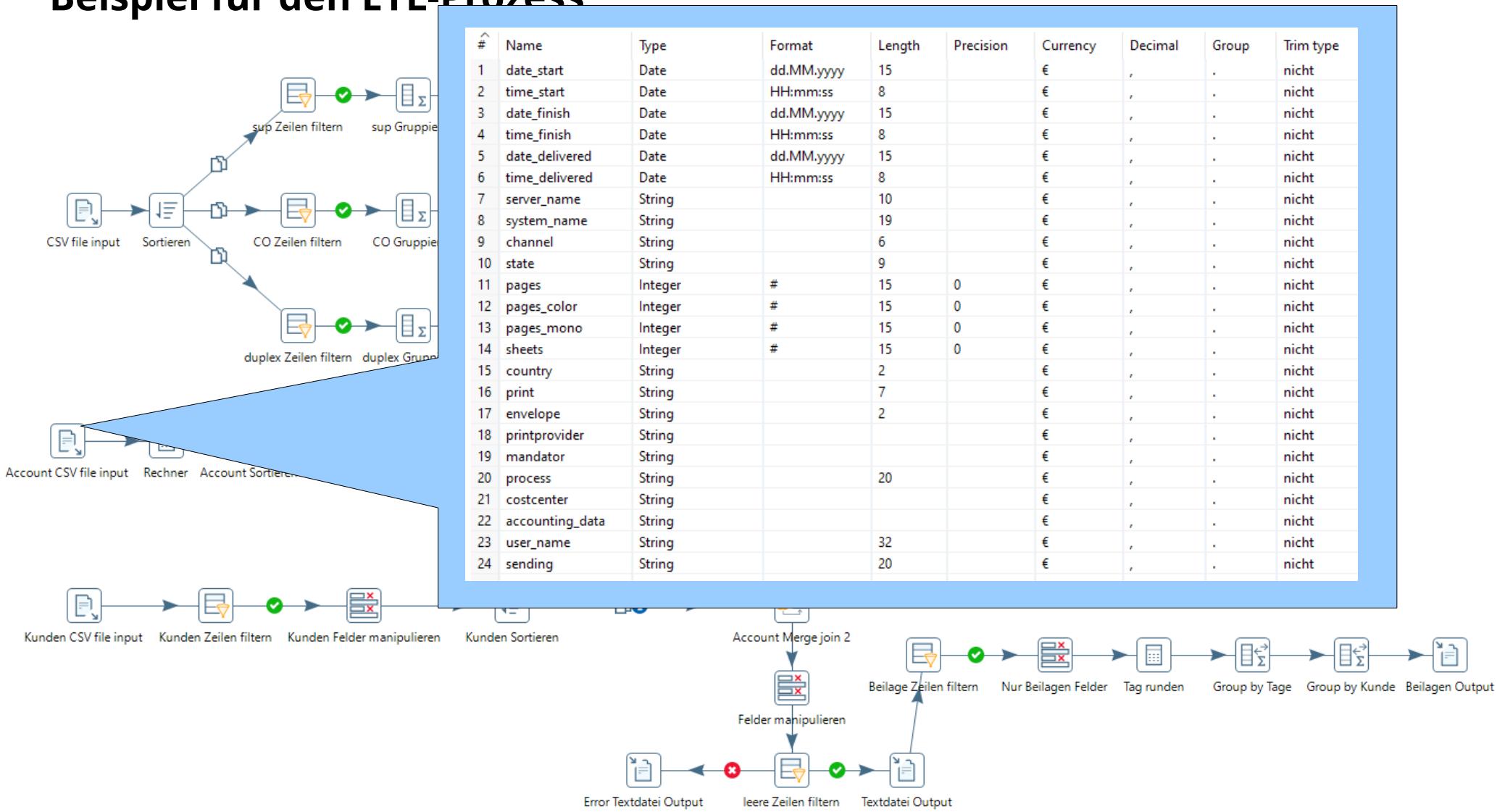
# Datenprozesse

## Beispiel für den ETL-Prozess



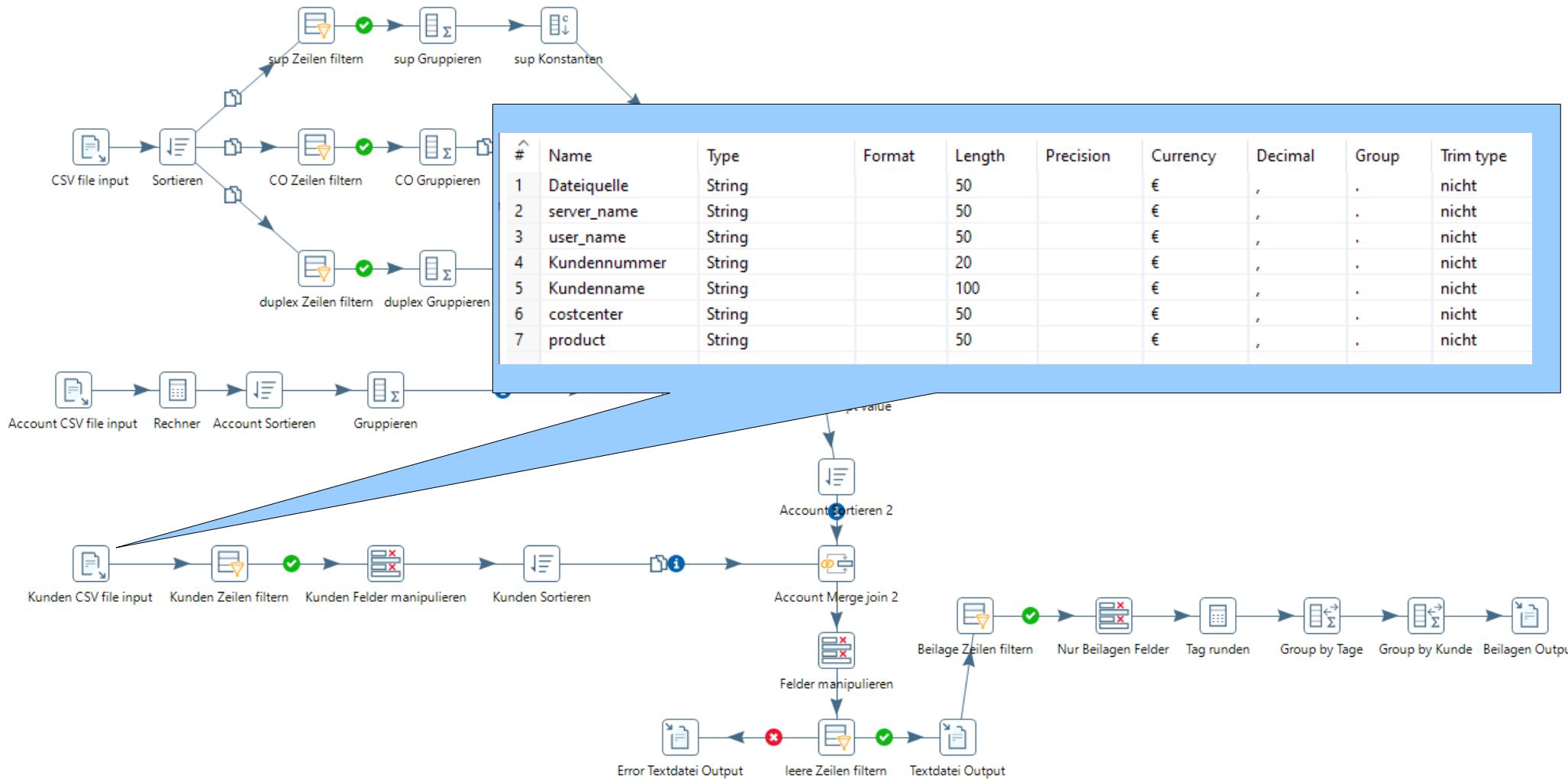
# Datenprozesse

## Beispiel für den ETL-Prozess



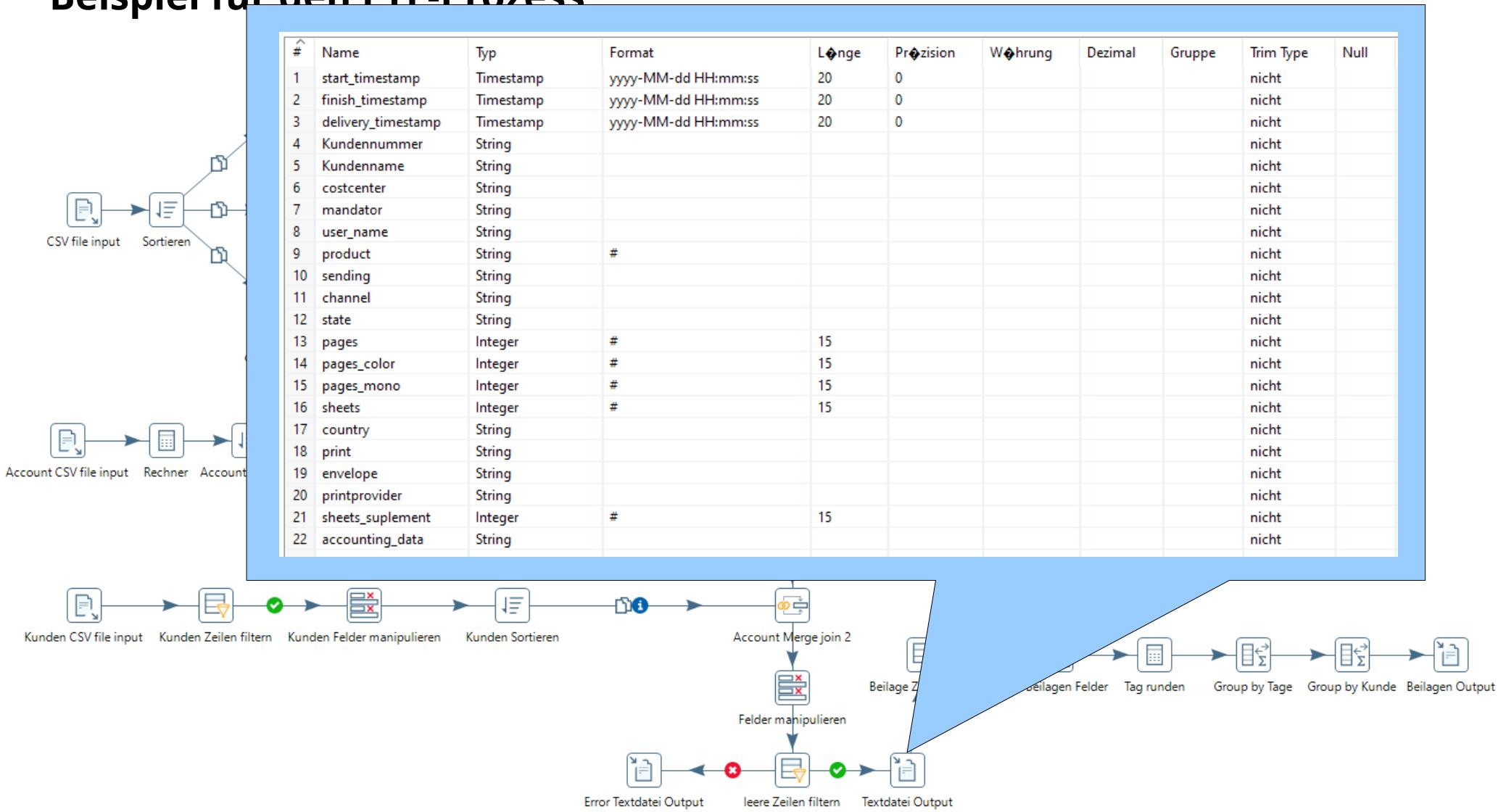
# Datenprozesse

## Beispiel für den ETL-Prozess



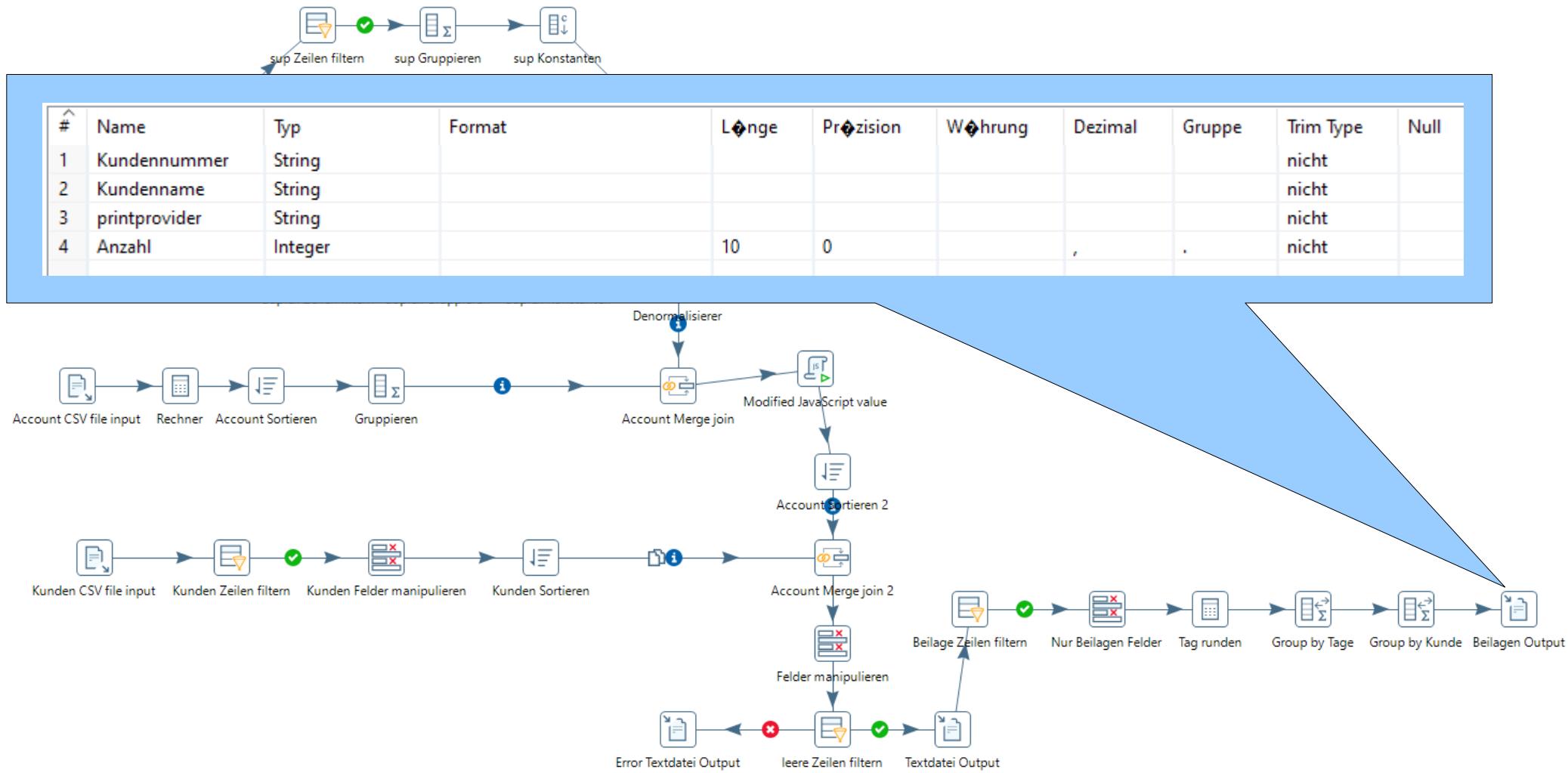
# Datenprozesse

## Beispiel für den FTI -Prozess



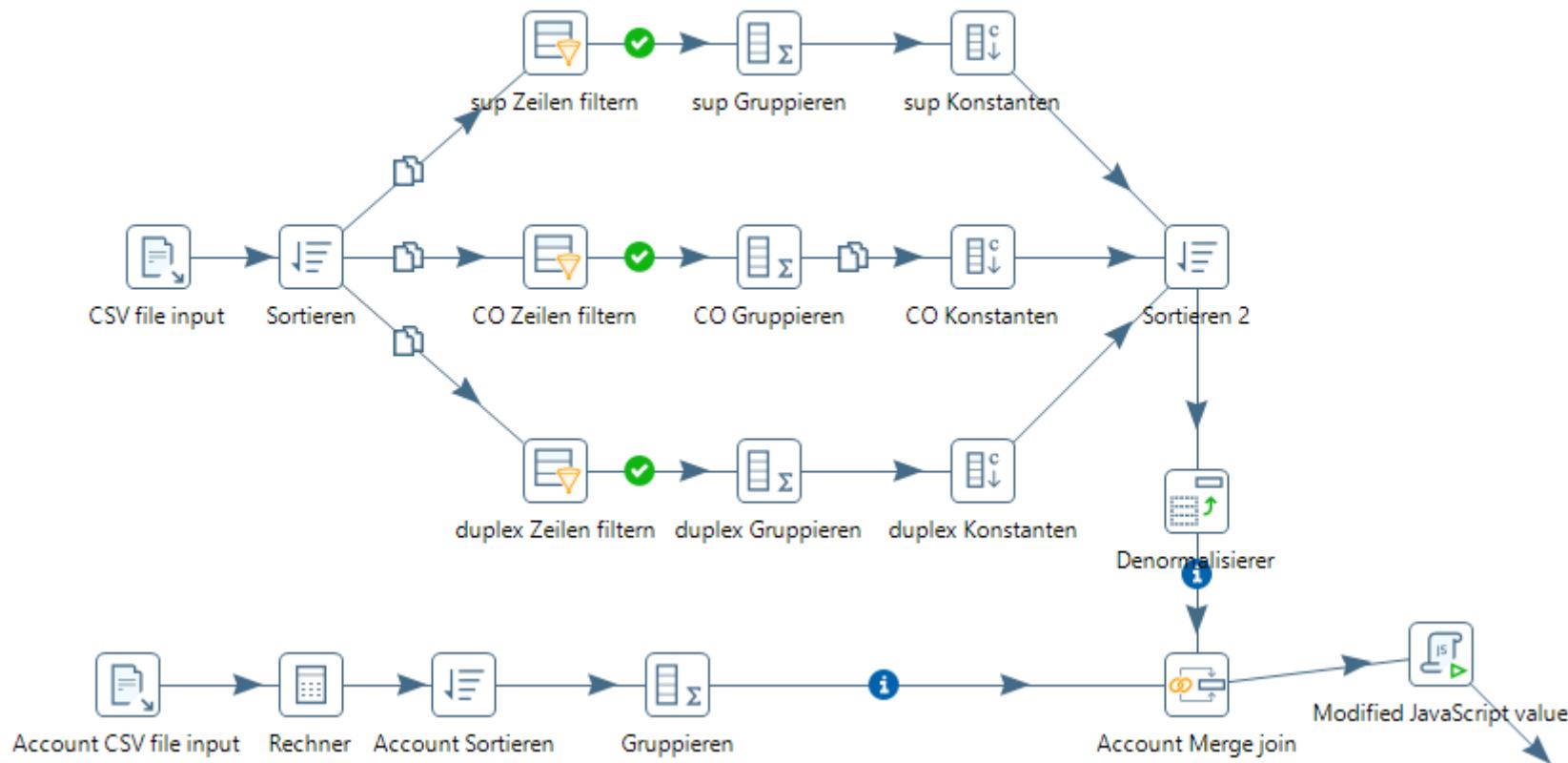
# Datenprozesse

## Beispiel für den ETL-Prozess



# Datenprozesse

## Beispiel für den ETL-Prozess



# Datenprozesse

## Beispiel für den ETL-Prozess

```
JavaScript:
Script 1 X
var lsupplement_count = count_supplement.getInteger();
var lduplex_count = count_duplex.getInteger();
var lcolor_count = count_color.getInteger();
var lcolor_pages = pages_color.getInteger();
var lsheets = sheets.getInteger();
var lenvelope = envelope.getString();
var lprint = print.getString();
var lpages = pages.getInteger();
var lChannel = channel.getString();

if(sending_1)
{
    if(lsupplement_count && lsupplement_count>0) lsheets += lsupplement_count;
    else lsupplement_count = 0;
    if (lduplex_count > 0) lprint = "duplex";
    if (lcolor_count > lcolor_pages) lcolor_pages = lcolor_count;
// Au farbe setzen
}

if (lprint == "duplex")
{
    lpages = lsheets + lsheets;
    if(lcolor_pages && lcolor_pages > 0) lcolor_pages = lpares;
}

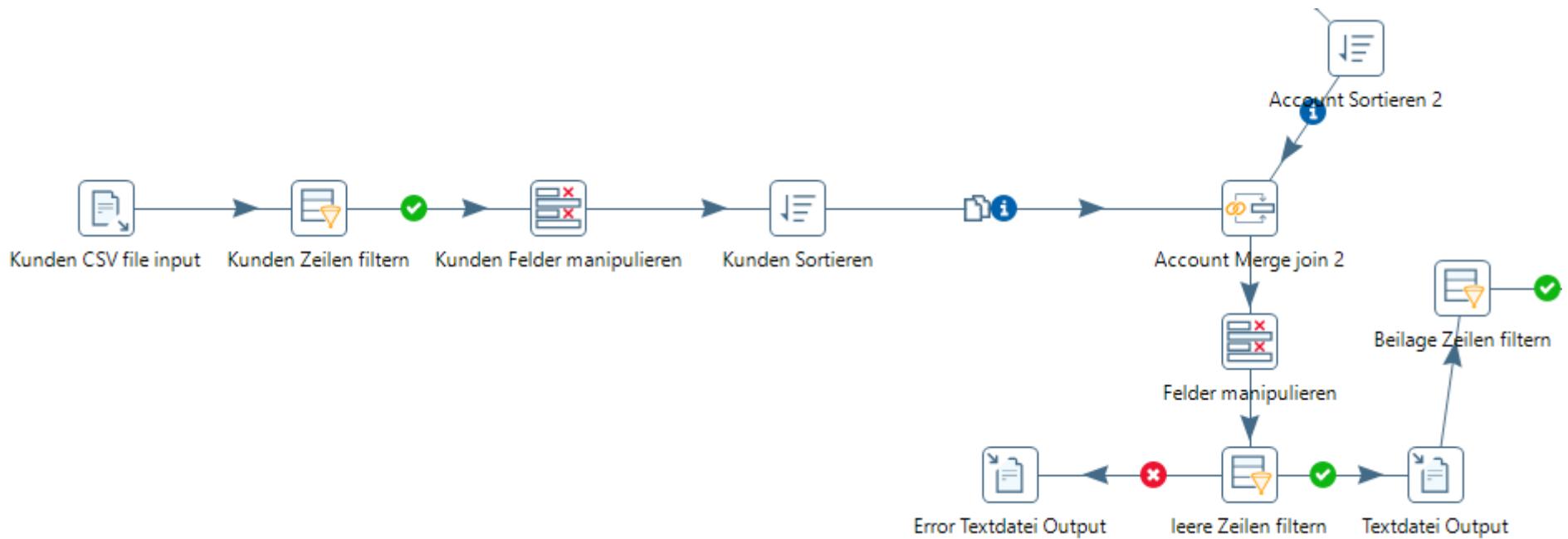
if(!lpages || lpares < lcolor_pages) lpares = lcolor_pages + pages_mono.getInteger();
// if (lsheets > lpares) lpares = lsheets;
if (!lenvelope) lenvelope = "DL";
if (lsheets > 9 && lenvelope == "DL") lenvelope = "C4";
if (lenvelope == "SWT01EDL") lenvelope = "DL";
if (lChannel == "pzu") lenvelope = "PZA";

count_supplement.setValue(lsupplement_count);
sheets.setValue(lsheet);
envelope.setValue(lenvelope);
print.setValue(lprint);
pages_color.setValue(lcolor_pages);
pages_mono.setValue(lpares - lcolor_pages);
pages.setValue(lpares);
```

The diagram illustrates the data flow for the ETL process. It starts with three 'Konstanten' (constant) inputs, each represented by a box with a downward arrow and labeled 'c'. These inputs feed into a 'Sortieren 2' (Sort 2) node, which is a square with two horizontal arrows pointing downwards. The output of 'Sortieren 2' feeds into a 'Denormalisierer' (Denormalizer) node, which is a square with a green checkmark and a small icon. The 'Denormalisierer' node has a single output arrow that points to a 'Modified JavaScript value' target, represented by a box with a green play button icon. A large blue triangle is positioned at the bottom right of the diagram.

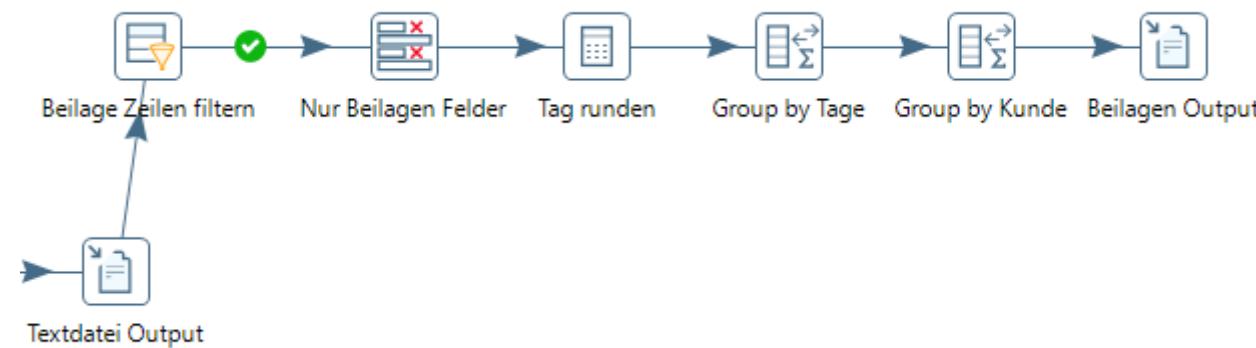
# Datenprozesse

## Beispiel für den ETL-Prozess



# Datenprozesse

## Beispiel für den ETL-Prozess





# Der Kern des DWH-Systems

## Verständnisaufgabe

- Empfohlene Toolauswahl: Pentaho Data Integration, Jupyter oder Google BigQuery
- Laden Sie die Daten von der Seite: "<https://www.kaggle.com/carrie1/ecommerce-data>". Wandeln Sie die Daten per ETL-Tool so um, dass alle Umsätze außerhalb Großbritanien pro Produkt und Tag zu einem Kurs  $1 \text{ £} = 1,1865 \text{ €}$  sortiert nach Menge als CSV gespeichert werden.
- Laden Sie die Daten von der Seite: "<https://www.kaggle.com/deepcontractor/dark-netflix-character-relationship>". Wandeln Sie Daten per ETL-Tool so um, dass sie mittels Graphviz alle Beziehungen (Eltern, Kindern, Partner) als gerichteten Graph dargestellt werden.



# System Sicht auf DWH

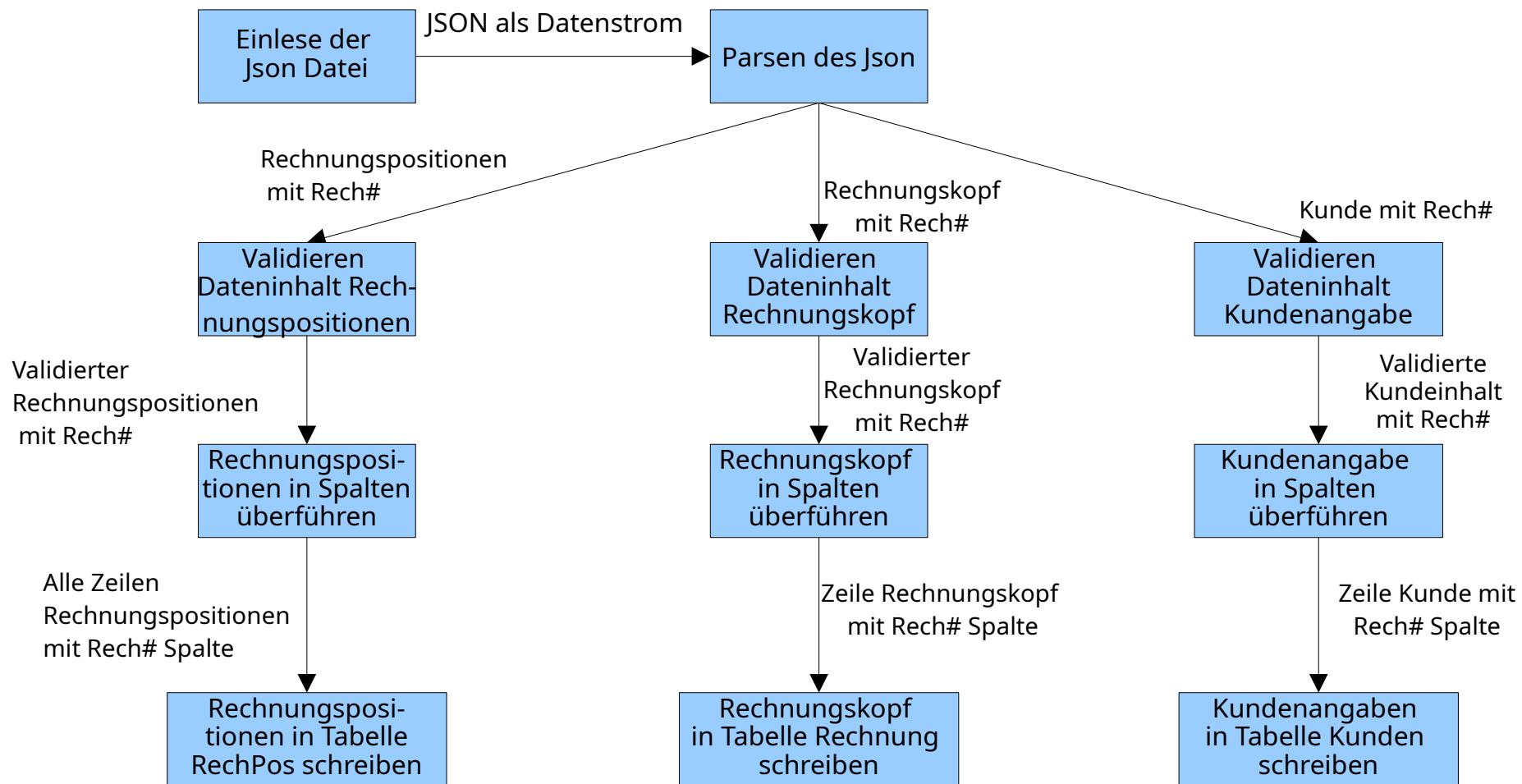
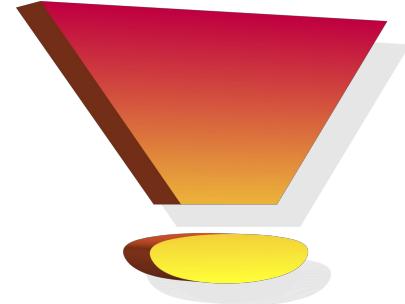
## ETL Prozess

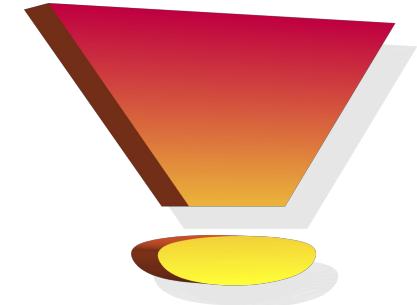
- Bauen Sie ein ETL-Programm auf, in dem Rechnungs-Informationen in Tabellenform für einen Staging Bereich umgewandelt werden. Bitte beachten Sie, dass die Daten aus JSON-Rechnungs-Dateien ermittelt werden müssen. Wählen Sie Ihre Schritte so, dass die Arbeitsweise und der Datenaustausch zwischen den Schritten nachvollzogen werden kann.

```
{ "Rechnung": 733601475250000001,  
  "Kunde": {"Erwin Müller GmbH", "Haptstrasse 42", "60322 Frankfurt Am Main", "Ustid":"0815"},  
  "Lieferer": {"Soda AG", "Brunnenweg 42", "60322 Frankfurt Am Main", "Ustid": "4711"},  
  "Lieferung": [  
    {"Position": 1, "Produkt": "Soda", "Produkt#": "221144", "Menge": "20 t",  
     "Epreis": "3001,00 Euro", "Gpreis": "60020,00 Euro"},  
    {"Position": 2, "Produkt": "Pottasche", "Produkt#": "9976123", "Menge": "5 t",  
     "Epreis": "400,00 Euro", "Gpreis": "2000,00 Euro"},  
    {"Position": 3, "Produkt": "Ethanol", "Produkt#": "78658943", "Menge": "3,5 t",  
     "Epreis": "1200,00 Euro", "Gpreis": "4200,00 Euro"}  
  ],  
  "GpreisNetto": "66220,00 Euro", "MwSt": "19 %", "MwSt_Betrag": "12581,80 Euro", "Gpreis": "78801,80 Euro",  
  "Ustid": "4711", "Lieferart": "frei Lager", "Zahlund_bis": "2021-12-31"  
}
```

# System Sicht auf DWH

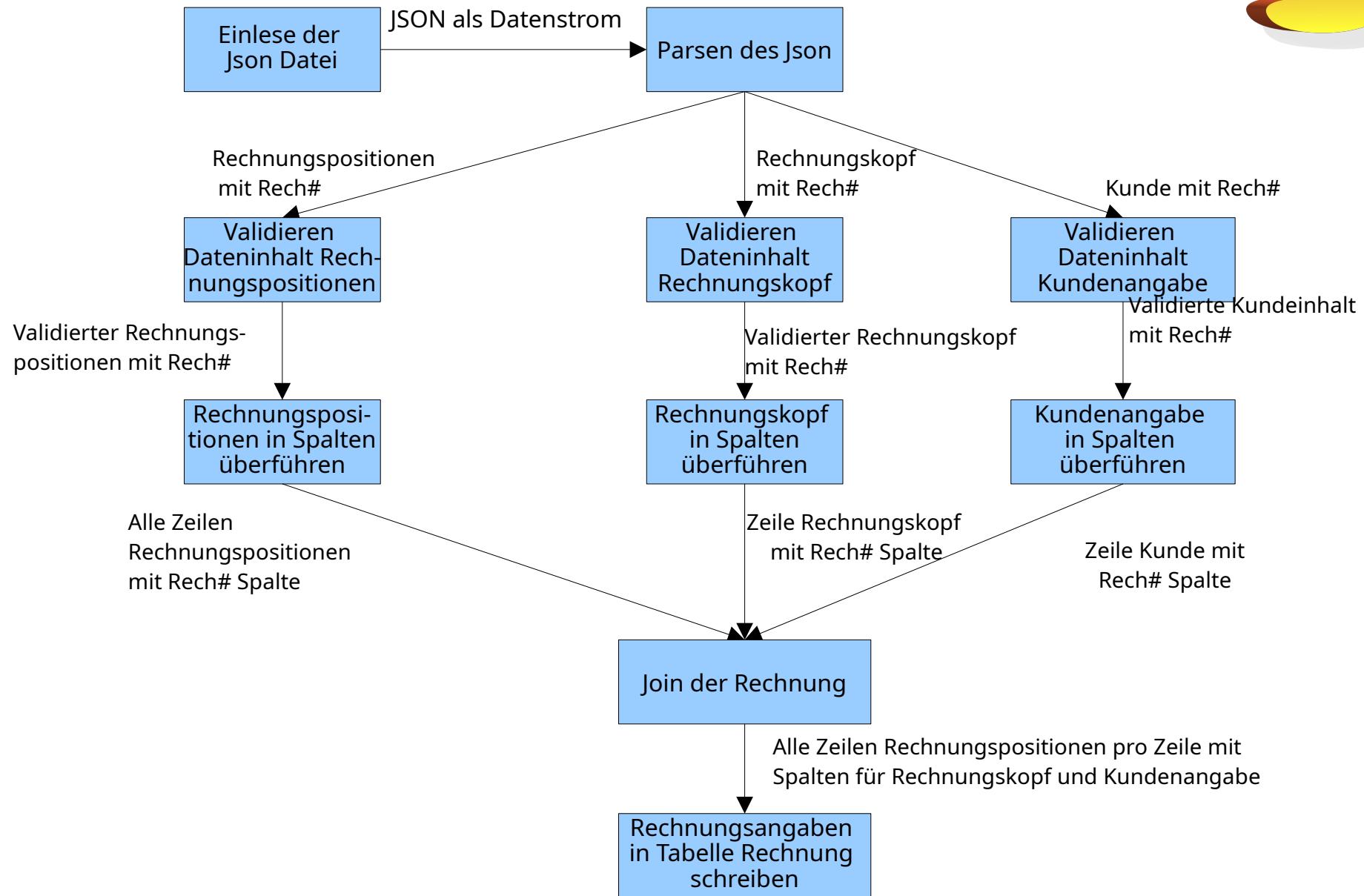
## ETL Prozess Lösung jede Liste in eine Tabelle





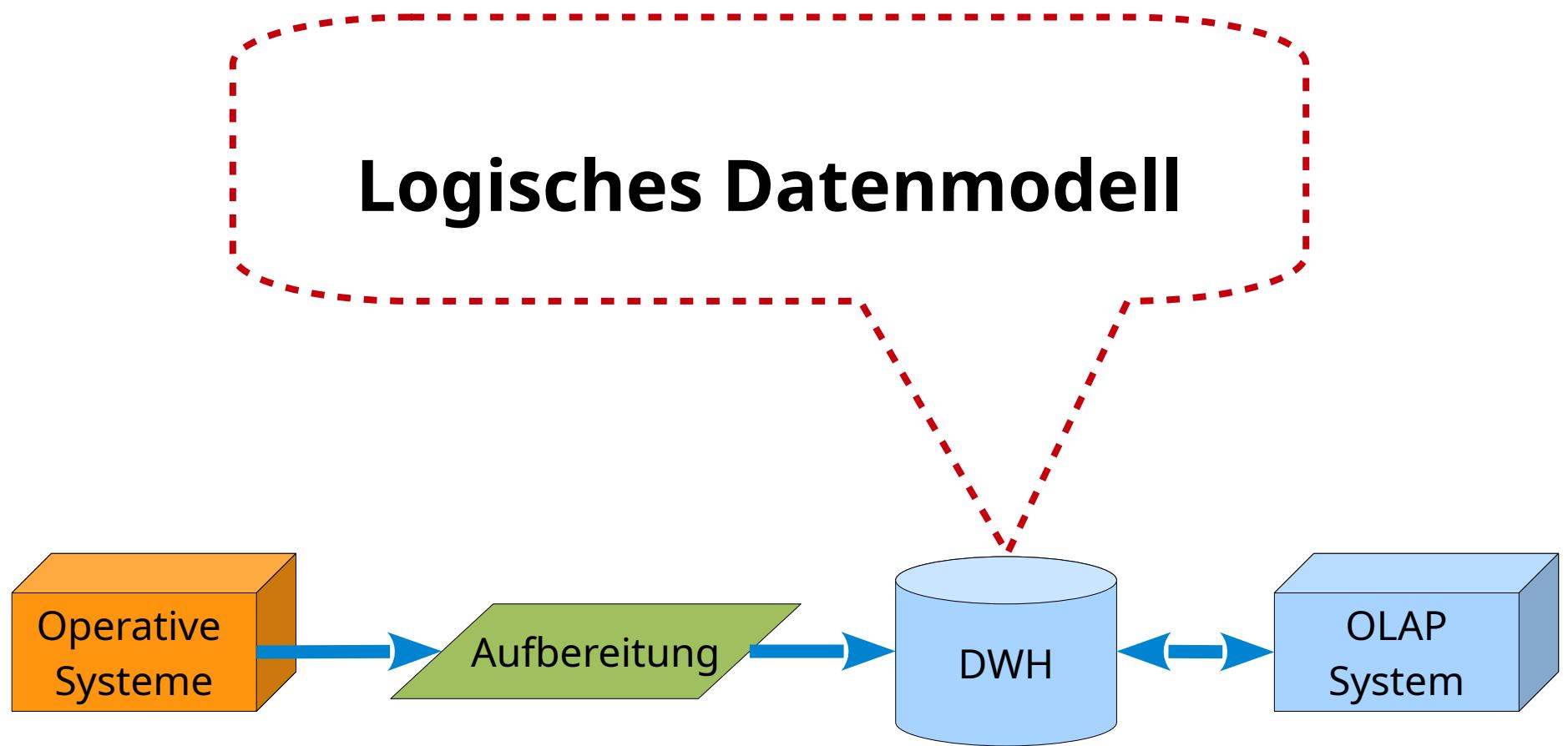
# System Sicht auf DWH

## ETL Prozess Lösung um eine Tabelle zu füllen



# Der Kern des DWH-Systems

## Möglichkeiten der logischen Datenmodellierung



# Der Kern des DWH-Systems

## Bestandteile des logischen Datenmodells

- Detaildaten
  - Datenfakten und die Dimensionen
- Aggregatsdaten
  - Multidimensionale Aggregate der Detaildaten
  - Oft optimiert auf wichtigsten Permutationen zu aggregieren
  - Besondere Behandlung von Lücken und dünnbesetzter Tabellen (sparse matrix)
- Metadaten
  - Daten über die Daten
  - Struktur und Inhalsauskünfte zu den Detaildaten oder Aggregatsdaten

# Der Kern des DWH-Systems

## **Bestandteile des logischen Datenmodells, Probleme Aggregatsdaten**

- Einfache Aggregation aller Permutationen
  - Nachteil ist extrem große Datenmenge
- Aggregation aller besetzten Fakten
  - Nachteil ist die vollständige Faktenanalyse
- Aggregation aller nicht dünn besetzter Fakten
  - Nachteil wichtige dünn besetzte Fakten sind nicht voraggregiert
- Mitlernende Aggregation
  - Nachteil: Optimum wird erst nach Füllen und intensiver Nutzung erreicht

# Der Kern des DWH-Systems

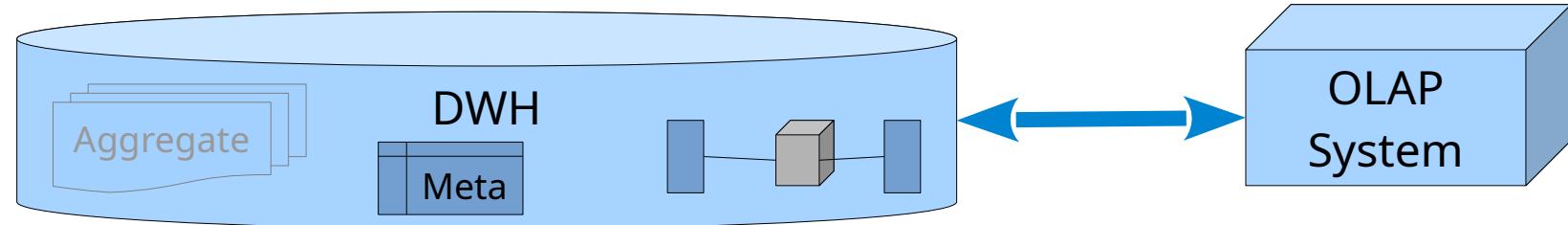
## Möglichkeiten der logischen Datenmodellierung

- ROLAP
  - Basis ist eine relationale Datenbank
- MOLAP
  - Basis ist eine multidimensionale Datenbank
- HOLAP
  - Basis ist ein Hybrid aus ROLAP und MOLAP
- DOLAP
  - Basis ist eine Desktop Verarbeitung der Daten
- Data Lake
  - Quelldaten werden quasi unverändert gesammelt abgelegt

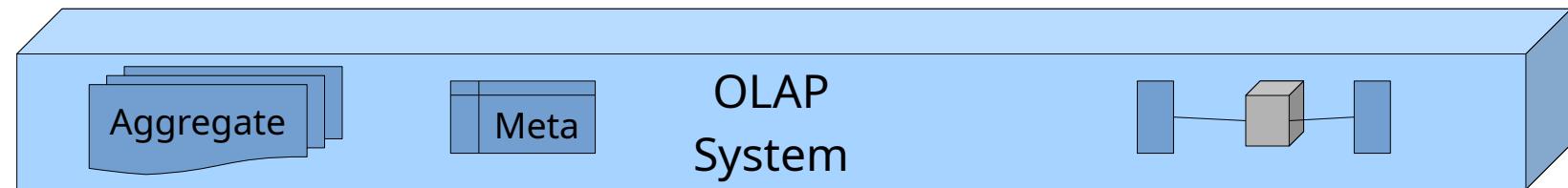
# Der Kern des DWH-Systems

## Möglichkeiten der logischen Datenmodellierung

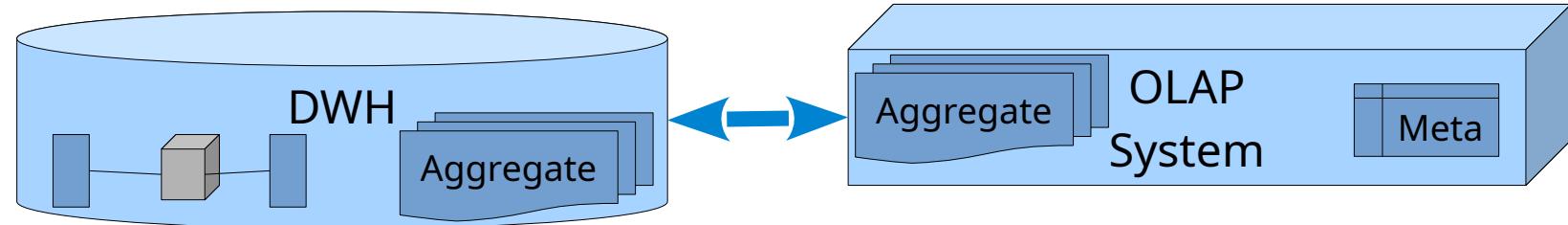
**ROLAP**



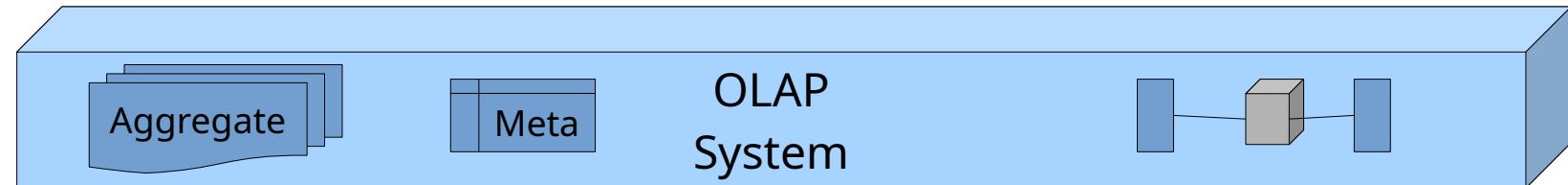
**MOLAP**



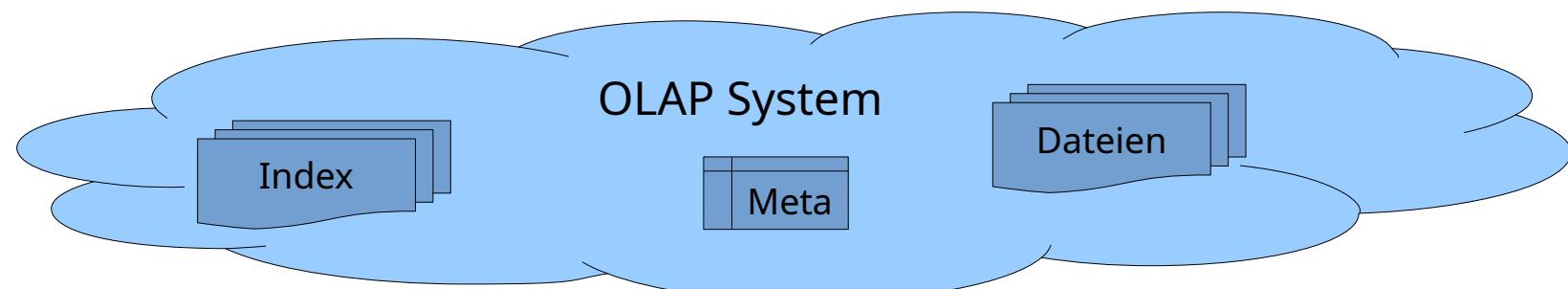
**HOLAP**



**DOLAP**



**Data Lake**



# Der Kern des DWH-Systems

## Datenmodellierung mit ROLAP

- Basis ist eine relationale Datenbank
- Detaildaten und Metadaten liegen in der Datenbank
- Aggregatsdaten werden meist nicht aufgebaut
- Geeignet für große strukturierbare Datenmengen
- Der Datenzugriff erfolgt über übliche Methoden z.B. SQL
- Speicherkapazität entspricht ca. den operativen Systemen
  - Sofern keine Aggregate multidimensionaler Fakten vorgehalten werden
- Zugriffsgeschwindigkeit entspricht vergleichbaren operativen Systemen
- Ladevorgänge erfolgen im Insert/Update Schritten

# Der Kern des DWH-Systems

## Datenmodellierung mit MOLAP

- Daten werden in proprietären multidimensionalen Datenstrukturen gehalten
- Detaildaten, Aggregatsdaten und Metadaten sind Teil der proprietären Struktur
- Geeignet für kleine und mittlere multidimensionalen Datenmengen
- Der Datenzugriff erfolgt meist über proprietäres OLAP Protokoll
- Speicherkapazität ist sehr hoch
  - Optimierung durch multidimensionale Aggregation
- Zugriffsgeschwindigkeit sehr schnell
- Ladevorgänge sehr langsam
  - Beim Laden werden multidimensionale Aggregate erstellt

# Der Kern des DWH-Systems

## Datenmodellierung mit HOLAP

- Daten werden sowohl in proprietären multidimensionalen Datenstrukturen als auch in einer relationalen Datenbank gehalten
- Detaildaten liegen in einer relationalen Datenbank
- Aggregatsdaten und Metadaten sind Teil der proprietären Struktur
- Geeignet für große multidimensionale Datenmengen
- Der Datenzugriff erfolgt meist über proprietäres OLAP Protokoll
- Speicherkapazität ist größer als in operativen Systemen
  - Aggregate werden zusätzlich zu Dimensionen und Fakten gehalten
- Zugriffsgeschwindigkeit kann sehr schnell sein, sofern vorausgesehen Aggregate abgefragt werden
- Ladevorgänge sind langsamer
  - Beim Laden werden multidimensionale Aggregate erstellt

# Der Kern des DWH-Systems

## Datenmodellierung mit DOLAP

- Daten werden aus den Quellen direkt in eine Desktop Anwendung geladen
- Detaildaten, Aggregatsdaten und Metadaten sind Teil der proprietären Hauptspeicherstruktur
- Geeignet für kleine Datenmengen
- Der Datenzugriff erfolgt meist über eine GUI
- Es wird direkt der Arbeitsspeicher des Desktops verwendet
  - Aggregate werden meist auf Anfrage erstellt
- Zugriffsgeschwindigkeit sehr schnell sein, sofern im Speicher ausgeführt
- Ladevorgänge liegt im Dateilade Bereich
- Einschränkungen des Desktop begrenzen den Einsatz von DOLAP

# Der Kern des DWH-Systems

## Datenmodellierung mit einem Data Lake

- Quelldaten werden quasi unverändert in einem "Datensee" abgelegt
- Detaildaten, Indexdaten und Metadaten sind Teil der Dokumentenverwaltung
- Geeignet für riesige unstrukturierte Datenmengen
- Der Datenzugriff erfolgt über Suchmechanismen, z.B. RegEx
- Speicherbedarf liegt über den Quelldaten
  - Indexierungen über die Quelldaten werden zusätzlich gehalten
- Zugriffsgeschwindigkeit im Indexbereich ist sehr schnell und außerhalb nicht ratsam
- Ladevorgänge extrem schnell
  - Beim Laden werden im See abgelegt und Indexdaten ergänzt

# Der Kern des DWH-Systems

## Möglichkeiten der logischen Datenmodellierung mit Data Lake

- Schema freie Datenhaltung im Data Lake
- Zugriffsoptimierung durch Indizierung auf Schlüsselbegriffe
- Filterung ist dabei mehr eine Suche
- Zugriffsheuristik dient als Optimierung zukünftiger Suchen
- Voraggregation von Suchaufgaben Sinnvoll
- Vorteile ist die schnelle Datenintegration
- Geringe oder keine Transformation der neuer Daten
- Geeignet für sehr große Datenmengen und verteilte Ausführung



# Der Kern des DWH-Systems

## Datenmodellierung

- Sie sollen ein OLAP System konzipieren, um die Konzernumsätze eines Handelskonzerns mit 1.000 Filialen Weltweit und einem Warenangebot von 1.000.000 unterschiedlicher Produkten auszuwerten.  
Welches logische Datenmodell wählen Sie und warum?
- Die File-Server eines Dax-Konzerns sollen auf verborgene diskriminierende Inhalte hin permanent durchsucht werden. Welches logische Datenmodell wählen Sie und warum?
- Sie erhalten große Excel Dateien von Umsatzzahlen eines Warenhauses. Daraus sollen sie die Anteile von Warengruppen am jeweiligen Tagesumsatz ermitteln.  
Welches logische Datenmodell wählen Sie und warum?

# Der Kern des DWH-Systems

## Möglichkeiten der logischen Datenmodellierung mit ROLAP

- Abbildung des Datenwürfel in einer relationale Datenbank
  - Star Schema (aus Kapitel 1)
    - Jede Dimension wird in einer Tabelle/Objekt zusammengefasst
  - Snowflake Schema (aus Kapitel 1)
    - Jede Hierarchie einer Dimension wird in einer Tabelle/Objekt gehalten
- Einschränkung von Normalformen aus Performance Gründen
- Achten auf Versionierung von Daten => Temporalisierung



# Der Kern des DWH-Systems

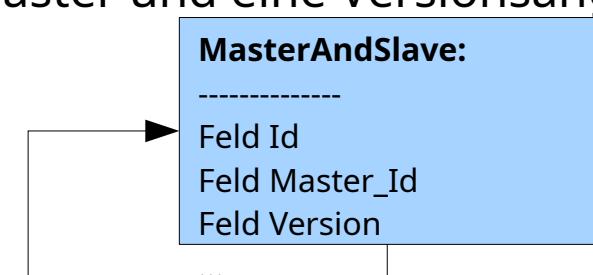
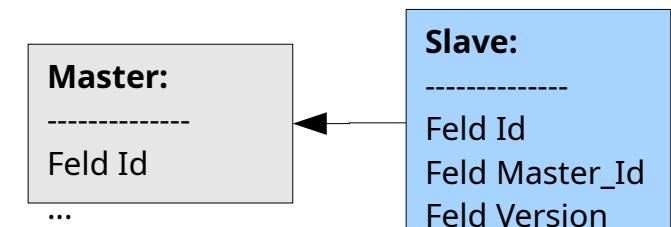
## Logischen Datenmodellierung

- Entwerfen Sie ein Star Schema eines Handelskonzerns im Textil Bereich. Berücksichtigen Sie wenigstens 5 Aggregate zur Unternehmenssteuerung. Begründen Sie Ihre Entscheidung für Dimensionen und Fakten?
- Entwerfen Sie ein Snowflake Dimensions Schema Produkte eines Handelskonzerns im Textil Bereich. Berücksichtigen Sie mindestens 6 Warengruppen Stufen. Erarbeiten Sie eine Strategie, wie sie häufig wechselnde Warengruppen der Produkte Herr werden wollen?
- Entwerfen Sie eine Dimensions-Struktur Zeit, die internationale Kalender berücksichtigt. Dabei sind von Stunden bis Jahre und Gregorianische Kalender, Julianischen Kalender und der Islamischer Kalender zu berücksichtigen.
- Sie erhalten Umsatzzahlen eines weltweiten Konzerns für Ihr OLAP System. Jeder Umsatzeintrag enthält ein Zeitstempel. Wie ermöglichen Sie die Abfragen des Umsatz um 13:00 Uhr eines Tages einer Zeitzone und den Umsatz jeweils um 13:00 Uhr eines Tages der jeweiligen Zeitzone?

# Der Kern des DWH-Systems

## Exkurs: Versionierung von Daten

- Jede Änderung an Daten führt zu einer Kopie der Original Daten
- Versionen können chronologisch nachvollzogen werden
- In relationalen Datenbanken mit Master-Slave Konzept realisiert
- Versionsangabe muss Eindeutigkeit und Chronologie ermöglichen
- Lösung Master-Slave mit zwei Tabellen:
  - Master Tabelle trägt Primärer Schlüssel und unveränderbare Felder
  - Slave Tabelle trägt Fremdschlüssel auf Master, eigenen Primärschlüssel und die Versionsangabe
- Lösung Master-Slave in einer Tabelle
  - Master trägt Primärschlüssel und einen Fremdschlüssel auf sich selbst
  - Slave trägt auch einen Primärschlüssel auf Master und eine Versionsangabe
  - Oft wird aktuelle Version als Master genutzt
    - Nachteil: Zwei Updates



# Der Kern des DWH-Systems

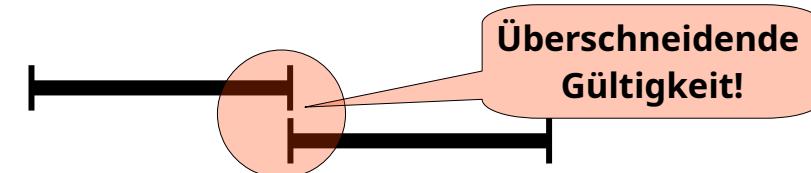
## Exkurs: Temporalisierung von Daten

- Temporalisierung ist eine Spezialform der Versionierung
  - Ohne: Letzter bekannter Wert wird gespeichert
  - Einfach: Gültigkeitszeitraum der Werte werden mit abgelegt
  - 2 D: Zum Gültigkeitszeitraum wird der Kenntniszeitraum der Werte werden mit abgelegt
  - 3 D. Zum Gültigkeitszeitraum und Kenntniszeitraum wird der Aktualisierungszeitraum der Werte werden mit abgelegt
- Bei den Zeitintervallen ist auf die Grenze zu achten
  - Vermeiden von Zeit Lücken oder Überschneidungen
  - Eindeutige und ununterbrochene Zeitintervalle
- Weitere Infos:
  - <https://dspace2.flinders.edu.au/xmlui/bitstream/handle/2328/25829/Roddick%20Survey.pdf?sequence=1&isAllowed=y>
  - <http://www.timeconsult.com/TemporalData/TemporalData.html>
  - <https://martinfowler.com/eaaDev/timeNarrative.html>

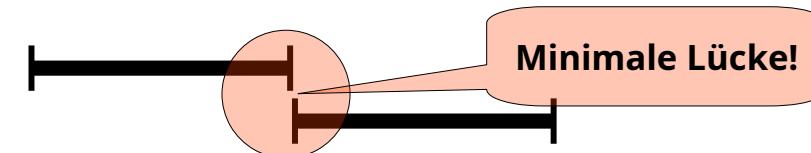
# Der Kern des DWH-Systems

## Exkurs: Temporalisierung von Daten, Best Practice

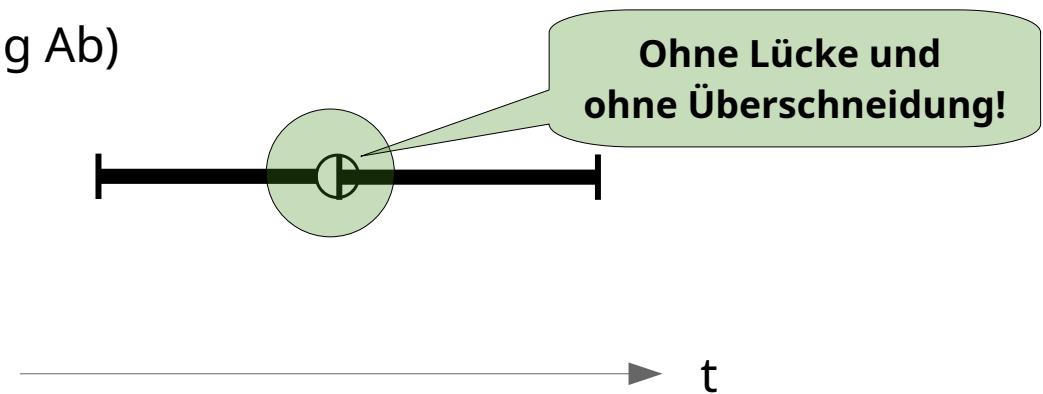
- ✗ Pro Datensatz (Gültig Von, Gültig Bis)



- ✗ Pro Datensatz (Gültig Von, Gültig Bis- $1^{-\infty}$ )



- ✓ Pro Datensatz (Gültig Von, Ungültig Ab)



# Der Kern des DWH-Systems

## Möglichkeiten der logischen Datenmodellierung mit ROLAP

- Versionierung macht in den Dimensionen oft Sinn
- In den Fakten sind neue Versions-Dimensionen wesentlich einfacher
- Versionierung sind bei hierarchischen Daten/Dimensionen sehr aufwendig
  - Es drohen verwaiste Elemente im Baum
  - Der kompletten Baum muss versioniert werden
- Aggregationstabellen, die wichtigste Kennzahlen vor-aggregieren
- Hohe Stufen der Temporealisierung sind Aufwändig
  - Gesetzliche Vorschriften beachten
  - Erstellen von Abfragen erhöhen sich sehr stark
  - Mit Fensterfunktionen werden manche Abfragen vereinfacht



# Der Kern des DWH-Systems

## Versionierung und Temporealisierung

- In einer Versicherung soll ein OLAP aufgebaut werden. Dabei soll analysiert werden, welche Streuung der Zeitverzögerung zwischen Datenstand und Datenmeldung liegen. Welche Form der Temporealisierung empfehlen Sie und warum?
- Ein Textilhandel hat seine Produktgruppen neu sortiert. Dies soll in Ihrem OLAP System nachgepflegt werden und dabei die alte Struktur für Bestandsdaten erhalten bleiben. Begründen Sie Ihr Vorgehen.
- Entwerfen Sie eine Faktentabelle für Produkte mit Umsatz und Tagesbezug. Wie können sie diese Umsätze Tagesgenau Temporealisieren?
- Entwerfen Sie eine Dimension Filiale in der die Filiale-Ident unveränderlich aber der Name -tagesgenau temporealisiert ist. Verbinden Sie Filialen und Fakten aus Aufgabe zuvor miteinander.
- Erstellen Sie eine SQL Abfrage, bei der aus der eben entworfenen Tabelle der Filialname im Jahr 2019 ermittelt werden soll.

# Der Kern des DWH-Systems

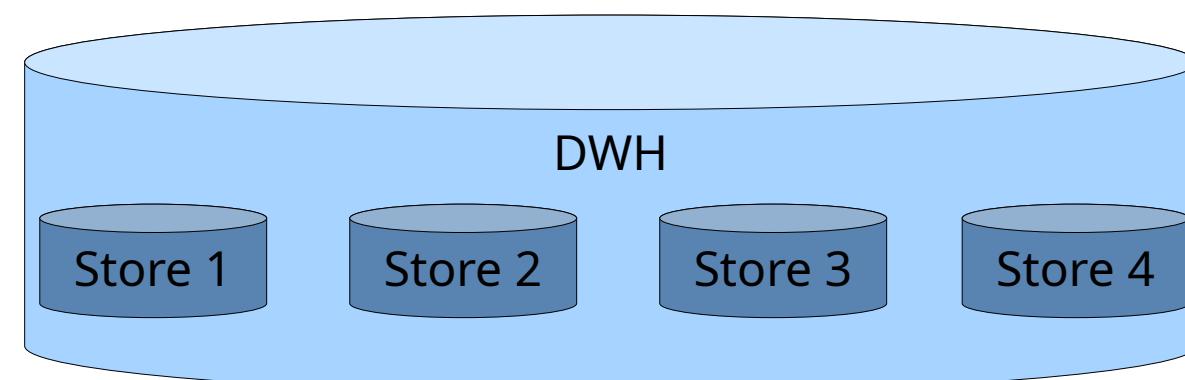
## Datenmodellierung mit großen Datenmengen in ROLAP

- Dimensionen und Fakten sind üblicherweise in einem Datenmodell
- Alternative:
  - Partitionierung des Datenbank Schema bis hin zum Data Mart
  - In externen Dateien halten
  - In externen Datenbanken halten
  - In Web-Service bereitstellen
  - Abbildung von Snowflake-Schema durch Views
  - Abbildung von Slice&Dice durch Views
  - Verteilte Verarbeitung und Datenhaltung

# Der Kern des DWH-Systems

## Möglichkeiten der Partitionierung

- Daten werden über Kriterien über mehrere Speicherorte verteilt
- Reduzierung der Lesezugriffe durch Kriterien Vorauswahl
- Parallel Befüllung unterschiedlicher Speicherbereiche möglich
- Nachteil: Schlechte Kriterienwahl führt zu längeren Suchzeiten
- Mögliche Partitionierung:
  - Faktentabelle nach Zeiträumen (Monat, Quartal, Jahr)
  - Faktentabelle nach Orten (Filialen, Länder, Absatzgebiete)
  - Faktentabelle nach Schreibprozessen bzw. Datenlieferungen
  - Rollende Partitionen mit zyklischem Überschreiben



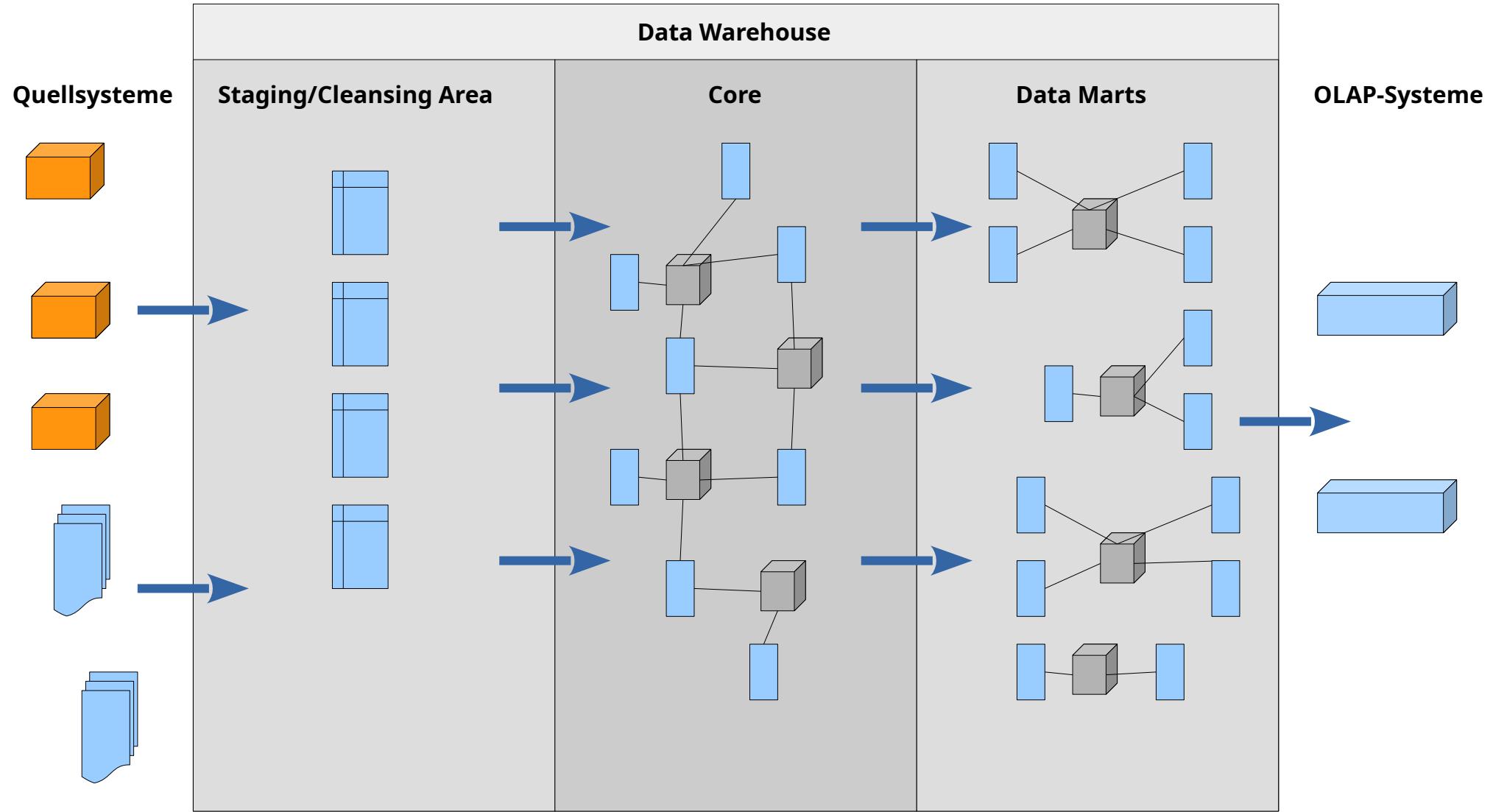
# Der Kern des DWH-Systems

## Datenmodellierung mit Data Marts

- Getrenntes Subsystem, das ein Drittel Anteil von Daten aufnimmt
- Dient der Geschwindigkeitsoptimierung und dem Datenschutz
- Modell Core-Mart und Mart-Core üblich
- Abhängige Data Marts
  - Alle Daten werden zentral im Master zusammengeführt
  - Slaves bekommen einen Drei update als Kopie
  - Slaves pflegen ihre eigene Indices
- Unabhängige Data Marts
  - Slave nehmen ihre quasi Drei Daten direkt auf und sind unabhängig
  - Master fasst ggf. virtuell alle Marts zu einem zusammen
  - Master erstellt ein übergreifenden Index über alle Slave Daten
- Hybride Data Marts
  - Kombination aus beiden Formen mit Datenaustausch in beide Richtungen

# Der Kern des DWH-Systems

## Datenmodellierung mit Data Marts



# Der Kern des DWH-Systems

## Data Marts Performance vs. Zeitfenster

- Data Marts bieten die Möglichkeit große Datenmengen zu verteilen
  - Schneller lokale Daten aber langsamer globale Daten (Zugriff und Integration)
  - Je größer die Abgrenzung desto geeigneter sind Data Marts
  - Globale Analysen werden Aufwändiger
  - Verteilte Verarbeitung kann helfen
- Verteilte Organisationen haben kleine Zeitfenster für Datenintegration
  - Wann ist Systemruhe in den operativen Systemen?
  - Wann können Data Marts synchronisiert werden?
- Optimale Lösung
  - Kriterien nach Bedürfnissen der Anwender ausgeglichen
  - Achten auf Datenmengen, ETL-Prozessoptimierung, DWH Prozessoptimierung

# Der Kern des DWH-Systems

## Datenmodellierung mit Data Marts

- Vorteile von Data Marts
  - Effizienter lokaler Zugriff
  - Kostensparnis gegenüber einem sehr großen zentralen OLAP
  - Verbesserte Datenprozess Performance
  - Dezentrale Datenpflege
  - Verteilter Aufbau
- Nachteil von Data Marts
  - Keine zentrale Datenhaltung
  - Zusätzliche Verwaltung der Einheitlichkeit von Datenstrukturen
  - Lizenz Inflation
  - Globaler Datenzugriff oft langsam

# Der Kern des DWH-Systems

## Datenmodellierung mit externen Dateien

- Einzelne Dimensionen oder Faktenbereiche werden nicht im DWH gehalten
- Speicherung erfolgt in eigenen Dateien
- Beim Zugriff müssen diese gesondert geladen werden
- Vorteil: Möglichkeit Daten DWH klein zu halten und optimierte Zugriffsmethoden zu nutzen
- Nachteil: Zugriff selten schneller als direkt im DWH

# Der Kern des DWH-Systems

## Datenmodellierung mit externen Datenbanken

- Einzelne Dimensionen oder Faktenbereiche werden nicht im DWH gehalten
- Speicherung erfolgt in eigener Datenbank
- Die Datenbanken sind meist mit einem internen Link verbunden
- Beim Zugriff müssen Daten gesondert geladen werden
- Vorteil: Möglichkeit Daten DWH klein zu halten und optimierte Zugriffsmethoden zu nutzen
- Nachteil: Verwaltung von zwei getrennten DB-System die sich technologisch unterscheiden
- Alternative: DWH auf einem Clustern betreiben

# Der Kern des DWH-Systems

## Datenmodellierung mit Web-Services

- Einzelne Dimensionen oder Faktenbereiche werden nicht im DWH gehalten
- Speicherung bzw. Zugriff erfolgt über Web-Services
- Die Web-Service sind meist per Service Link verbunden
- Beim Zugriff müssen Daten gesondert geladen werden
- Vorteil: Möglichkeit Daten DWH klein zu halten und optimierte Zugriffsmethoden zu nutzen
- Nachteil: Verwaltung von zwei getrennter Daten Welten die sich technologisch unterscheiden
- Alternative: DWH auf einem Clustern betreiben

# Der Kern des DWH-Systems

## Datenmodellierung mit Snowflake-Schema Views

- Einzelne Dimensionen im Snowflake-Schema werden über Views zugegriffen
- Zugriff auf Dimension erfolgt über View
- View kann Materialisiert werden
- Beim Zugriff auf Daten wird ein Star-Schema simuliert
- Vorteil: Erleichterung der Abfrage
- Nachteil: Langsamer Snowflake Zugriff ohne Materialisierung
- Nachteil: Zusätzlicher Aufwand die Materialisierung zu erstellen
- Alternative: Direkt ein Star-Schema nutzen

# Der Kern des DWH-Systems

## Datenmodellierung mit Slice&Dice Views

- Einzelne Faktenbereiche werden über Views bereitgestellt
- Zugriff auf den Faktenbereich erfolgt über View
- View kann Materialisiert werden
- Im View können bereits Aggregationen integriert sein
- Vorteil: Erleichterung und Beschleunigung der Abfrage in dem bestimmten Faktenbereich
- Nachteil: Umständlicher Abfrage beim Überschreiten der Faktengrenzen
- Nachteil: Zusätzlicher Aufwand die Materialisierung zu erstellen
- Alternative: Übergang in HOLAP gleitend mit mehr Unterstützung der Aggregation

# Der Kern des DWH-Systems

## Datenmodellierung mit verteilter Verarbeitung und Datenhaltung

- Daten wird über viele ggf. sehr viele Verarbeitungsrechner (Knoten) verteilt
- Zugriffe werden in vielen der Rechner abgearbeitet
- Ergebnisse werden vom angefragten Knoten zusammengeführt
- Schnittstellen zwischen SQL Welt und der verteilten Welt verfügbar
- Vorteil: Ermöglicht schnelle Verarbeitung riesiger Datenmengen (Peta Byte)
- Vorteil: Dynamische Reaktion auf Lastspitzen
- Nachteil: Für strukturierte Daten weniger geeignet
- Nachteil: Komplexe Verwaltung und Synchronisierung der Knoten
- Beispiel: Hadoop (MapReduce-Algorithmus)



# Der Kern des DWH-Systems

## Datenorganisation und Data Marts

- Diskutieren Sie in einer Gruppe von min. 4 Personen, ob es für ein Unternehmen des Mittelstands (bis  $100 \cdot 10^6$  € Umsatz) mit 10 Außenstellen besser ist ein OLAP mit Data Mart oder mit einem Data Lake aufzubauen. Ziel des OLAP ist es, die ERP Daten des Unternehmens auszuwerten. Stellen Sie Ihre Diskussionsergebnisse vor.
- Wie würde das Diskussionsergebnis aussehen, wenn das Unternehmen kein ERP besitzen würde? Stellen Sie auch hier Ihre Diskussionsergebnisse vor.

# Kurs Einführung In das Thema Data Ware House & Business Intelligence

→ Kapitel 3: System Sicht auf DWH

# Betrieb eines Data Ware House

## Logging und Monitoring

- Sichtfenster der Administration
- Überwachung der ETL-Prozesse
  - Fehlerhafte Daten
  - Datenqualität
  - Fehlende Lieferungen
- Überwachung der Datenintegration
  - Integrationsprobleme
  - Fakten Lücken
  - Dead Locks
- Überwachung der Systeme selbst nach RAM, HDD, Prozessor und Netzwerk
- Erkennen von Softwarefehlern
- Proaktiv Systemausfall verhindern
- Erkennen von unerlaubten zugriffen
- Vorsicht: Beim Logging und Monitoring können große Datenmengen entstehen  
=> eigenes DWH für L&M

# Betrieb eines Data Ware House

## Betrieb eines DWH

- Auswahl der Hardware für große Datenmengen mit schnellen Zugriffen
- Mode der 80er: Alles in RAM packen!
  - Wurde in von Spalten-orientierten Datenbanken aufgegriffen
- Bei Rolap-Cluster ist performantes Backbone wichtig
- Hadoop Cluster oft Vernetzung wichtiger als Rechenleistung
- Manche Hersteller bringen optimierte Hardware mit: Oracle Exadata

# Betrieb eines Data Ware House

## Schutz der Daten

- Sicherheit umfasst auch die Schutz der Daten vor fremder Einsicht
  - Vertrauen der Anwender sicherstellen
  - Integrität der Daten erhalten
  - Verfügbarkeit für die Anwender gewährleisten
- Mögliche Maßnahmen
  - Verschlüsseln des Datentransports bei jeder System- oder Prozessgrenze
  - Rollback bei Angriff Detektion ermöglichen
  - Verschlüsseln der Inhalte im DWH
  - Verschlüsseln der Inhalte auf Dateisystem
  - Benutzeridentifikation und Authentifizierung
  - Zugang zu Netzwerk und Hardware

# Betrieb eines Data Ware House

## Benutzeridentifikation und Authentifizierung

- Benutzeridentifikation
  - Nur verschlüsselte Kommunikation
  - Einfache Sicherheit mit User/PW
  - Mehrfache Sicherheit: OTP, Token, Zertifikate,
  - SSO – Systeme & LDAP (auch Active Directory)
- Zugriffskontrolle
  - Lesen und Schreibrollen
  - Beschränkung für einzelne Dimensionen, D-Ebenen, Fakten, Analyse
- Auditieren
  - Protokollieren und Überwachen der Zugriffe

# Betrieb eines Data Ware House

## Datenerhaltung durch Backup

- Backup wegen großer Datenmengen oft Problematisch
- Ein Backup nutzt nur wenn man zurückspielen kann! Testen!
- Backup getrennt und geschützt aufbewahren
- Verteilte Daten durch Redundanz Faktor schützen
- Einfluss des Backup auf Datenprozesse beachten

# Betrieb eines Data Ware House

## Performance Tuning

- Performance Tuning ist Verbesserung der Antwortzeiten
- Maßnahmen sind:
  - 1) Informationsmanagement
  - 2) Datenbank Design
  - 3) Anwendungsumgebung
  - 4) Datenbankzugriff
  - 5) Datenbank Konfiguration
  - 6) Betriebssystem
  - 7) Netzwerk
  - 8) Hardware

# Betrieb eines Data Ware House

## **Performance Tuning, Informationsmanagement**

- Prinzipielle Strukturierung und Konfiguration
- Performance Messung und Engpässe finden
  - Logging der Verarbeitungszeiten
- Subjektive Engpässe angehen
- Organisation des Systems und der zeitlichen und räumlichen Daten Work Flows
  - ETL-Prozesse
  - Backup-Prozesse
  - Reporting
  - Analyse

# Betrieb eines Data Ware House

## **Performance Tuning, Datenbank Design**

- Abwägen zwischen Normalisierung und bewusster Denormalisierung
- Daten-Anomalien aufspüren und beseitigen
- Indexierung der richtigen Spalten
- Abwägen zwischen Insert/Update und Abfrage Optimierung
- Wahl der richtigen Index Typen
- Datenkriterien der Partitionierung
- Verteilungsregeln und Clustering
- Auswirkungen des Füllgrads auf Performance

# Betrieb eines Data Ware House

## **Performance Tuning, Anwendungsumgebung**

- Zentrale oder verteilte Ausführung
- Sperrmechanismen nur bei wirklichem Schreiben
- Backup Mechanismen
- Implementierungslogik
- Einsatz von Ausführungs--Optimierer
- Connection Pooling

# Betrieb eines Data Ware House

## Performance Tuning, Datenbank Konfiguration

- Parallelität der Verarbeitung
  - Dedicated in einem Prozess für alle Aufgaben
  - Multithreaded oder Mehrprozessverarbeitung
  - Parallele DB-Server
- Master-Slave oder Multi-Master
- Speichermanagement, z.B. Pufferbereiche, Erweiterungen
- Verteilung der Datenbasis
  - Replikation: verteiltes lesen
  - Separation: Tabellen, Indexe, Log, ... trennen
  - Partitionierung: Konzept der Verteilen von Tabelleninhalten
  - Caching: Tabellenbereiche auf besondere Hardware legen
- Reorganisation von Tabellen und Indexe
- Defragmentieren von Speicherbereiche

# Betrieb eines Data Ware House

## **Performance Tuning, Betriebssystem**

- Aktualität des Betriebssystems
- Verteilungskonzepte des Filesystems
- Filesystem selbst und seine Blockverwaltung bzw. -größe
- Pufferbereiche für temp. Daten und Caching

# Betrieb eines Data Ware House

## Performance Tuning, Netzwerk

- Latenz und Kapazität
- Monitoring der Last und Latenz
- Separierung der internen Kommunikation
  - Multiplexing
  - Kanaltrennung
  - Kanal
- Art des Netzwerks Aufbau, Protokoll (Iso-Osi bis 3)
  - (3) Vermittlung: ICMP, IP, Ipsec
  - (2) Data-Link: Ethernet, WLAN, FDDI, ArcNet
  - (1) Physisch: Koaxial, Glasfaser (Multi-, Single-Mode), Funk

# Betrieb eines Data Ware House

## Performance Tuning, Hardware

- Hauptspeicherausstattung
- Festplattenkapazität
- Prozessorleistung und Mehrprozessorleistung
- Dauerlastfähigkeit
- Kühlung
- Stromversorgung und USV
- Redundante Hardware



# System Sicht auf DWH

## Betrieb eines Data Ware House

- Im Monitoring eines DWH wird eine permanente Volllast nur eines Prozessors angezeigt. Auf welches Probleme weißt dies hin?
- Sie sollen ein großes DWH einrichten mit vielen unstrukturierten Daten das gleichzeitig Green-IT ist. Wo liegt das Dilemma?
- Recherchieren Sie nach Sicherheitsproblemen in großen Analysesystemen die zu Strafverfahren nach DSGVO führten. Wie hätte sich dies verhindern lassen?
- Sie werden zu einem Kunden zur Beschleunigung eines OLAP-Systems gerufen. Erstellen Sie einen Arbeitsplan wie Sie vorgehen und welche Punkte Sie sich in welcher Reihenfolge annehmen wollen.

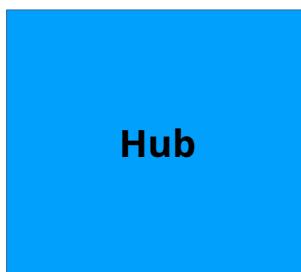
# Kurs Einführung In das Thema Data Ware House & Business Intelligence

*Kapitel 4: Beispiele für OLAP Schemen*

# Beispiele für OLAP Schemen

## Data vault modeling: Die Elemente

- Die Data-Vault-Modellierung wurde in den 1990ern von **Dan Linstedt** entwickelt

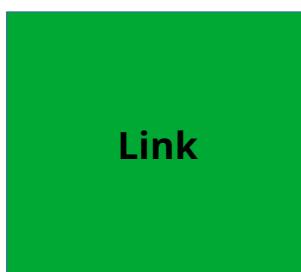


Hubs stehen für die zentrale Informationen die das Geschäftskonzept eindeutig beschreiben.  
Präfix H



Die Satelliten enthalten die Attribute der Hubs und deren Temporalisierung. Ein Hub oder Link kann mehrere Satelliten haben. Ein Satellit darf keine direkte Verbindung zu einem anderen Satellit haben.

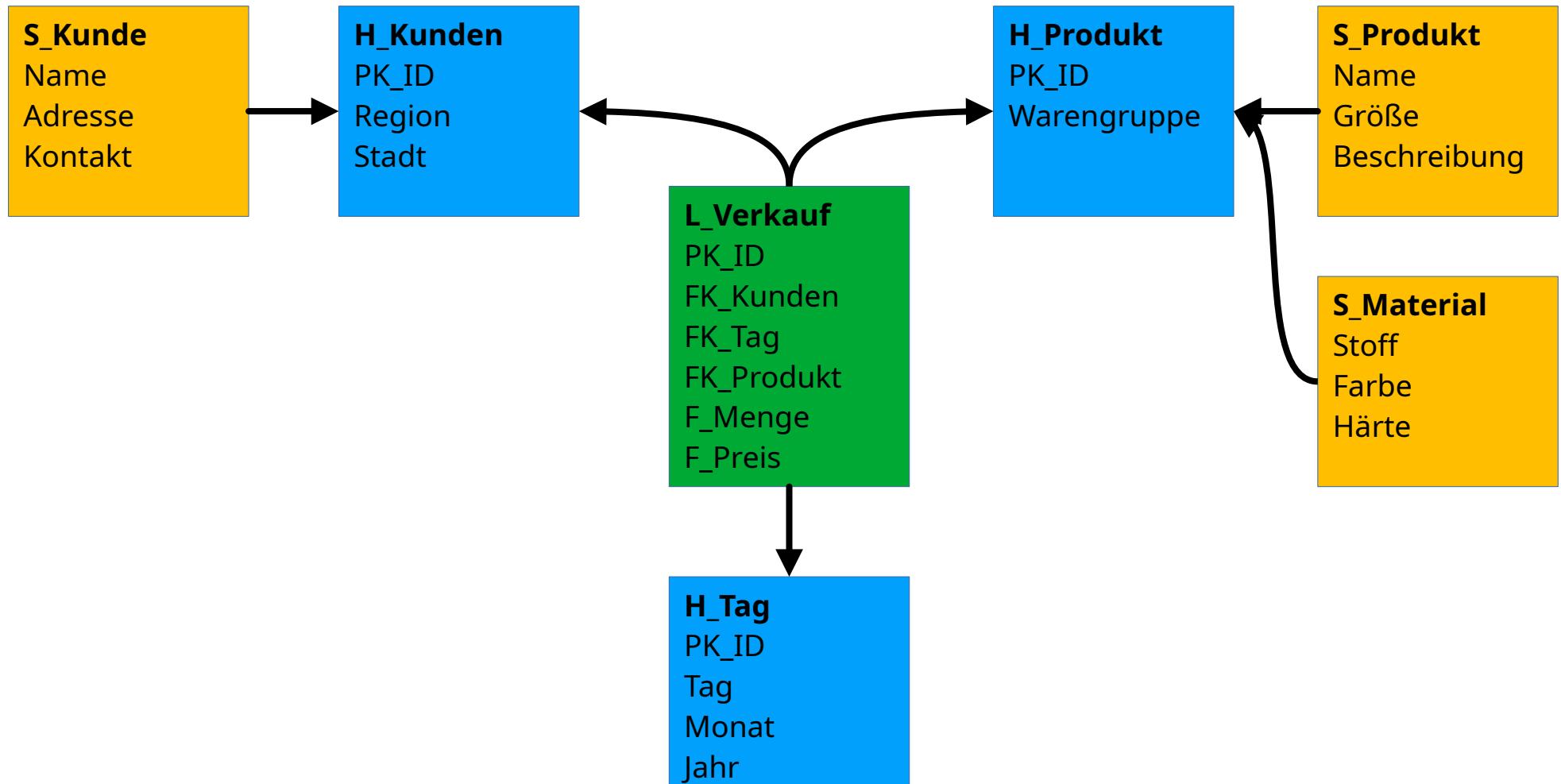
Präfix S



Links stellen die Beziehung zwischen Hub-Entitäten dar.  
Präfix H

# Beispiele für OLAP Schemen

## Data vault modeling: Beispiel Verkaufsdaten



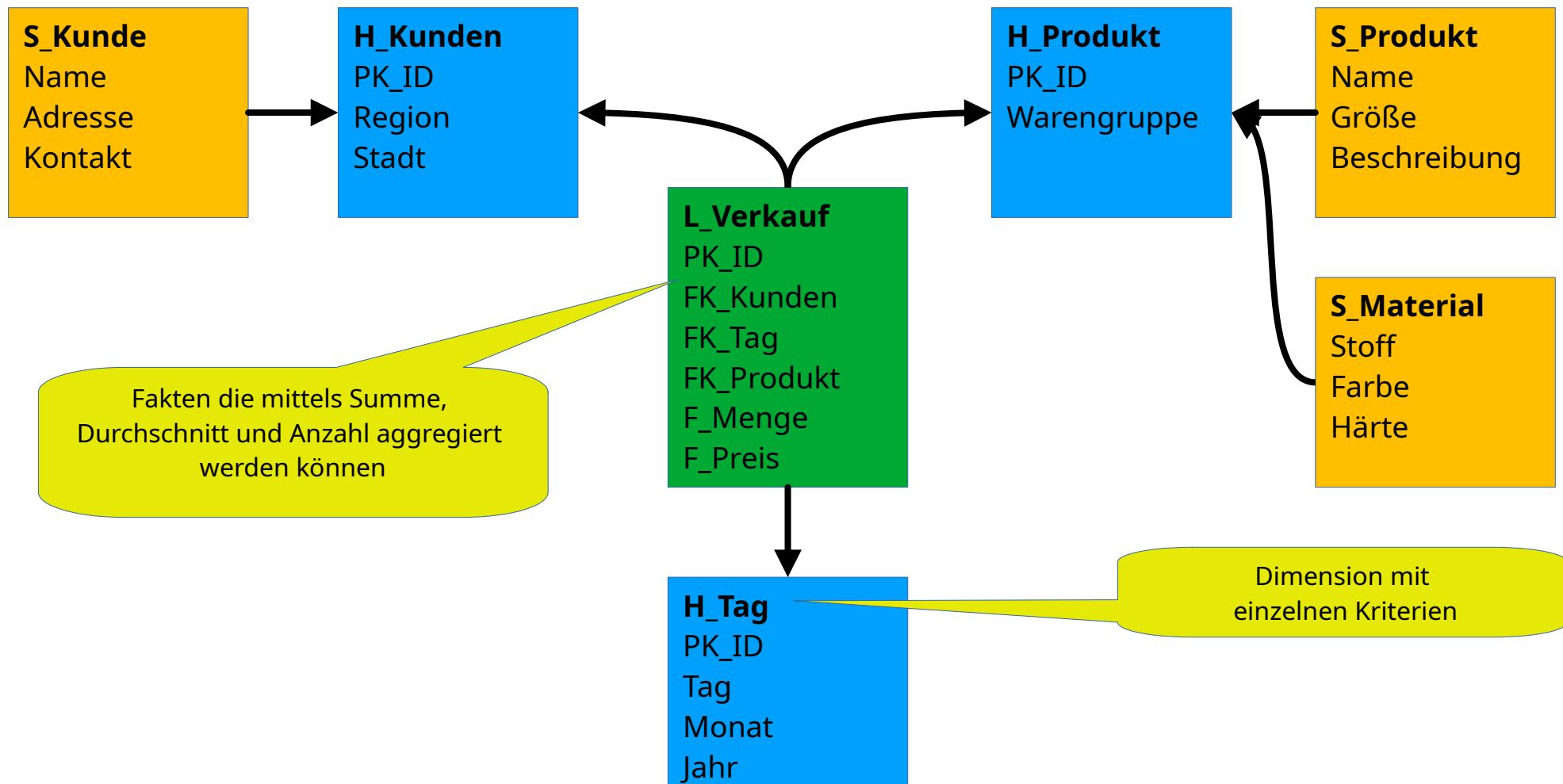
# Beispiele für OLAP Schemen

## Beispiele und Quelle

- **Quelle:** <https://www.goodreads.com/book/show/9123496-the-multidimensional-manager>
- Best Practices von Cognos, heute IBM, für DWH Modelle
- Beispiele zu
  - Finanzen
  - Vertrieb
  - Marketing
  - Materialwirtschaft
  - Produktion
  - Lieferkette
  - Service
  - Personal

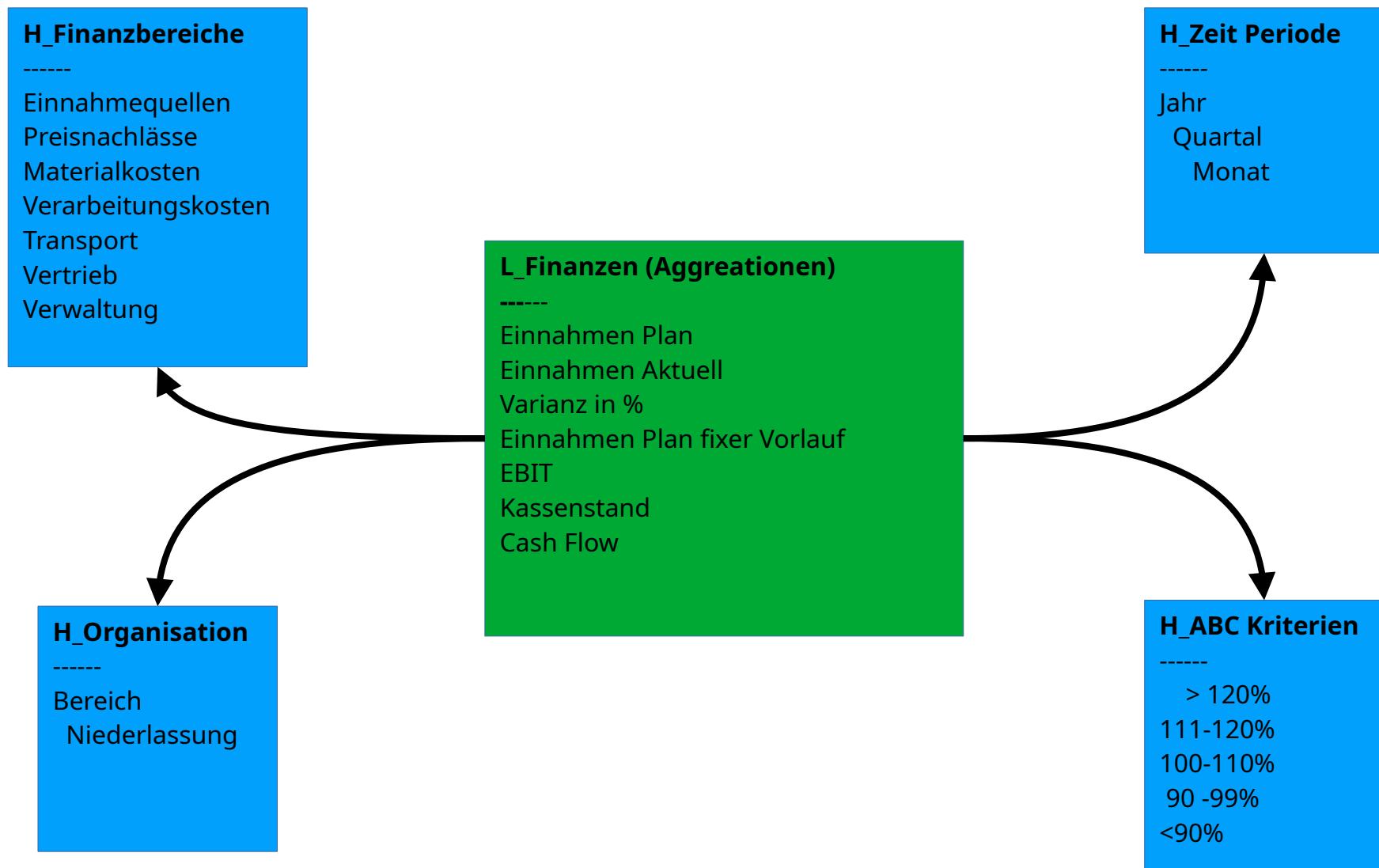
# Beispiele für OLAP Schemen

## Erklärung der Darstellung



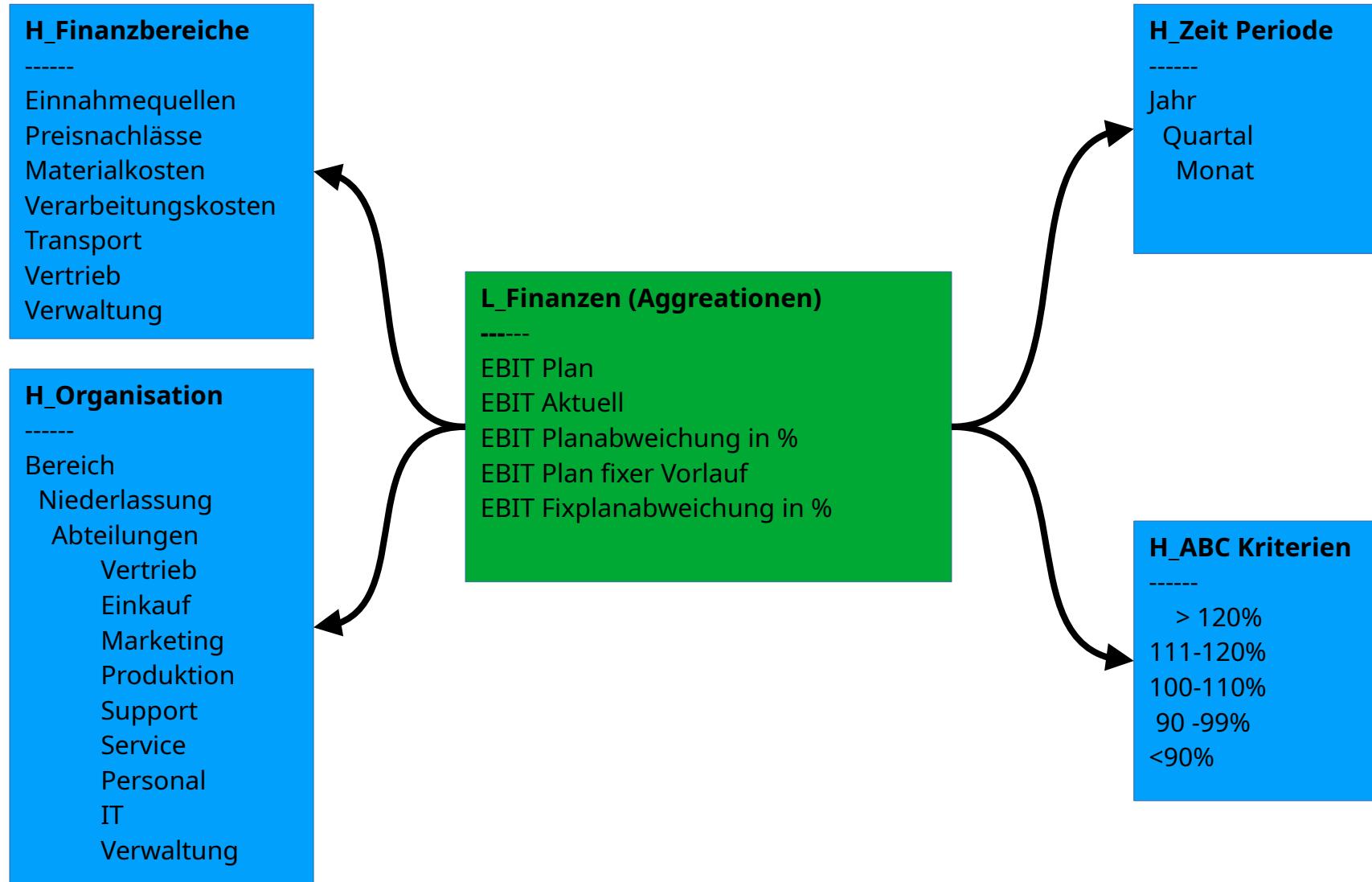
# Beispiele für OLAP Schemen

## Finanzen Einkommensanalyse



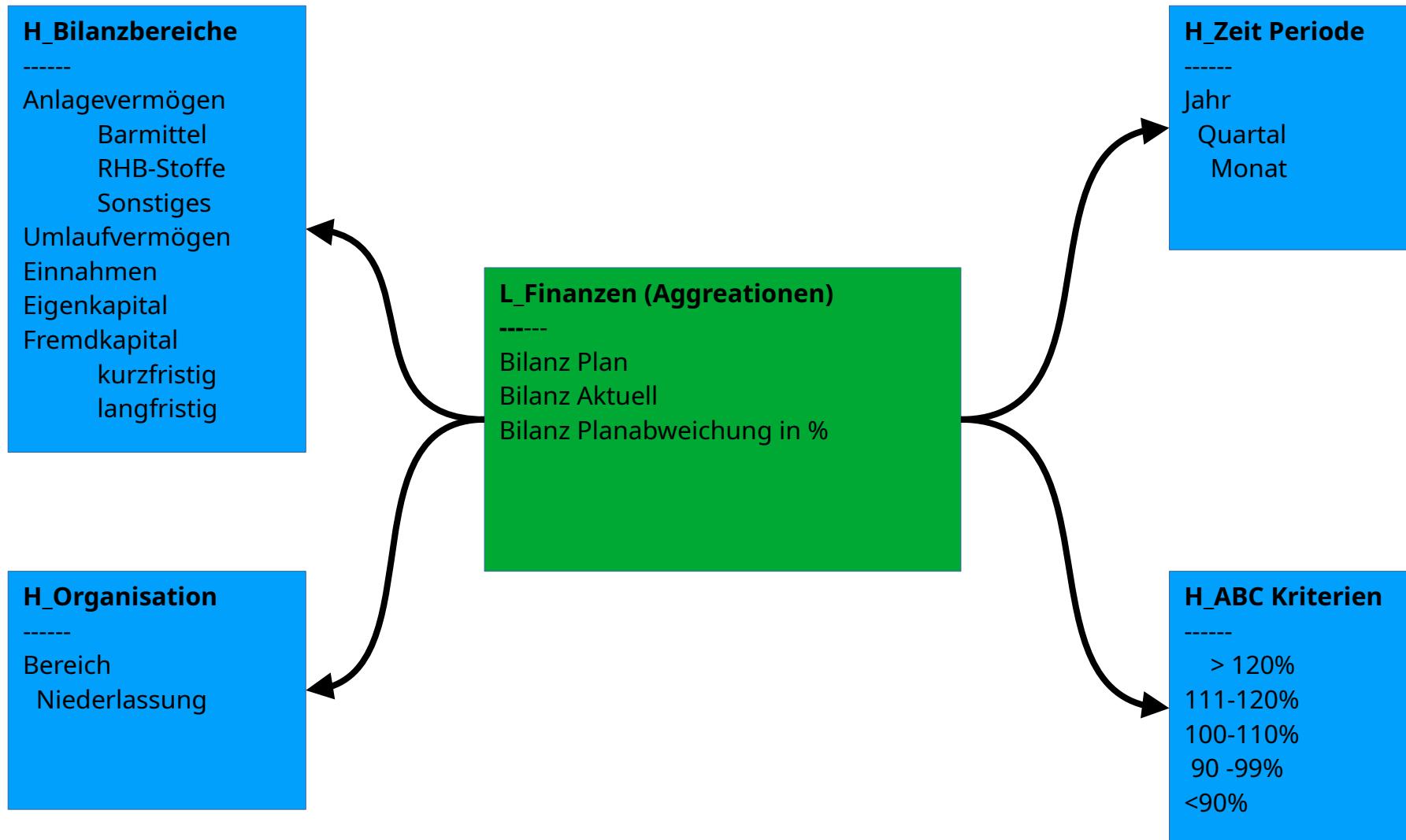
# Beispiele für OLAP Schemen

## Finanzen Gewinnanalyse



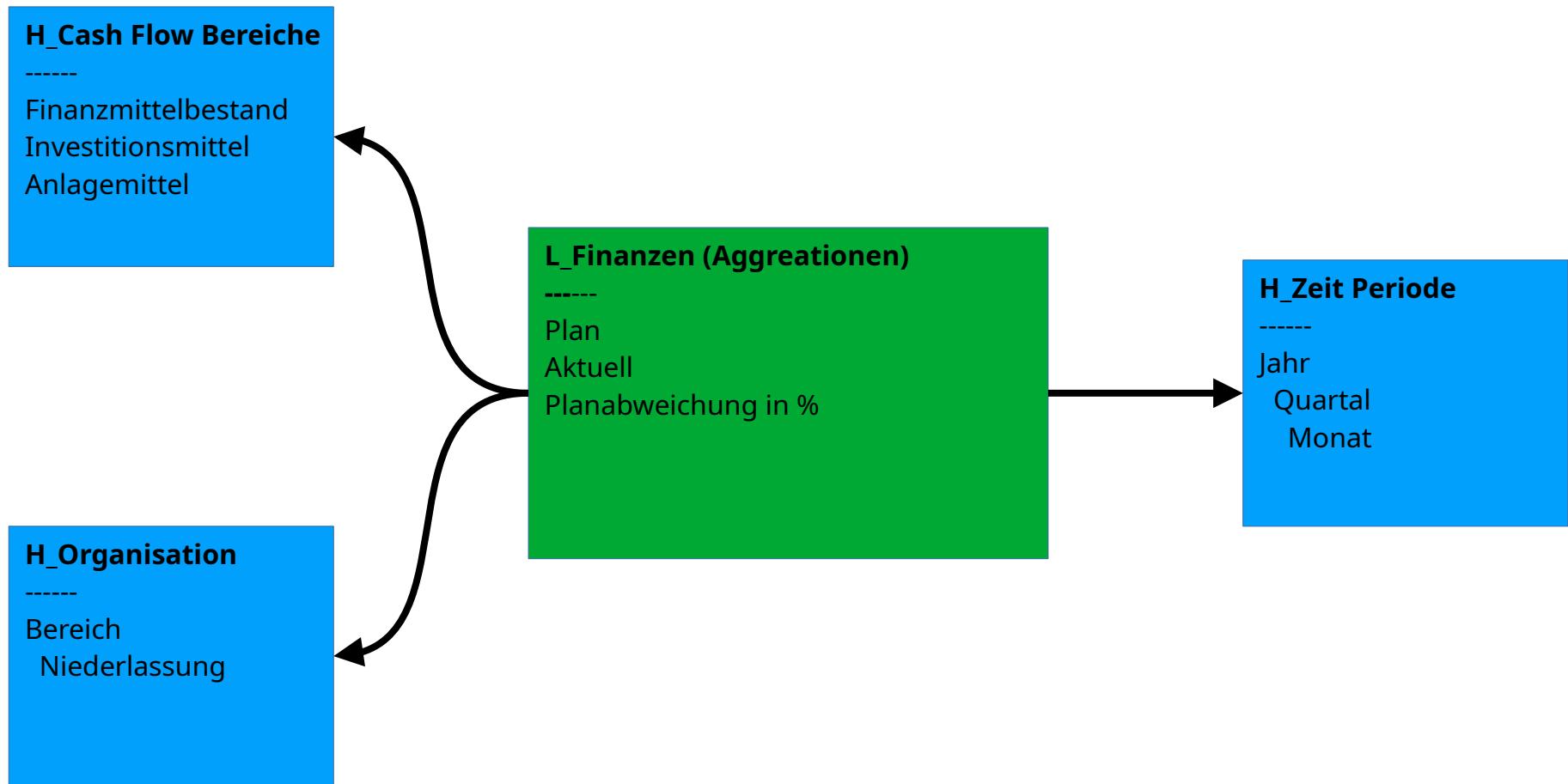
# Beispiele für OLAP Schemen

## Finanzen Bilanzanalyse



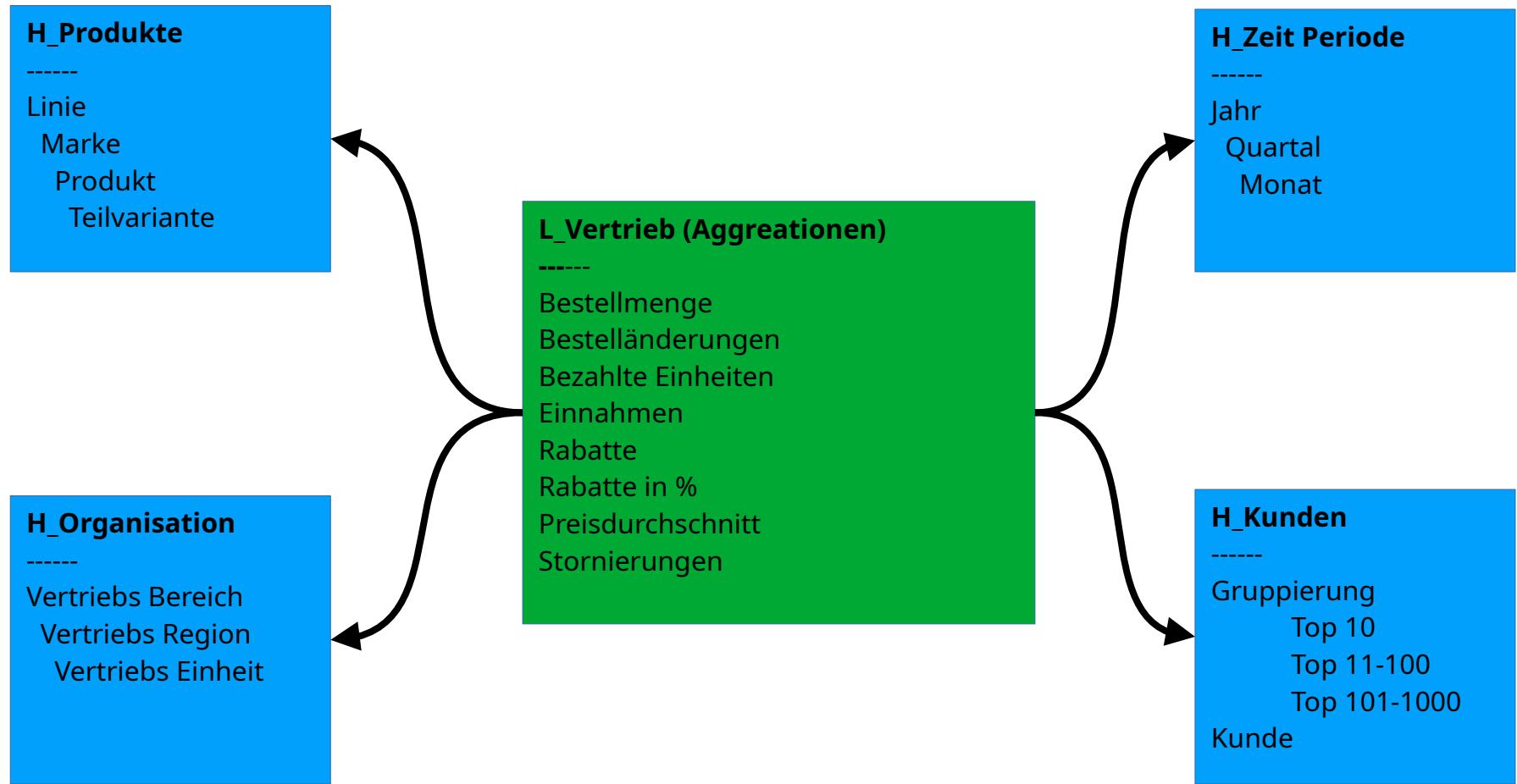
# Beispiele für OLAP Schemen

## Finanzen Cash Flow Analyse



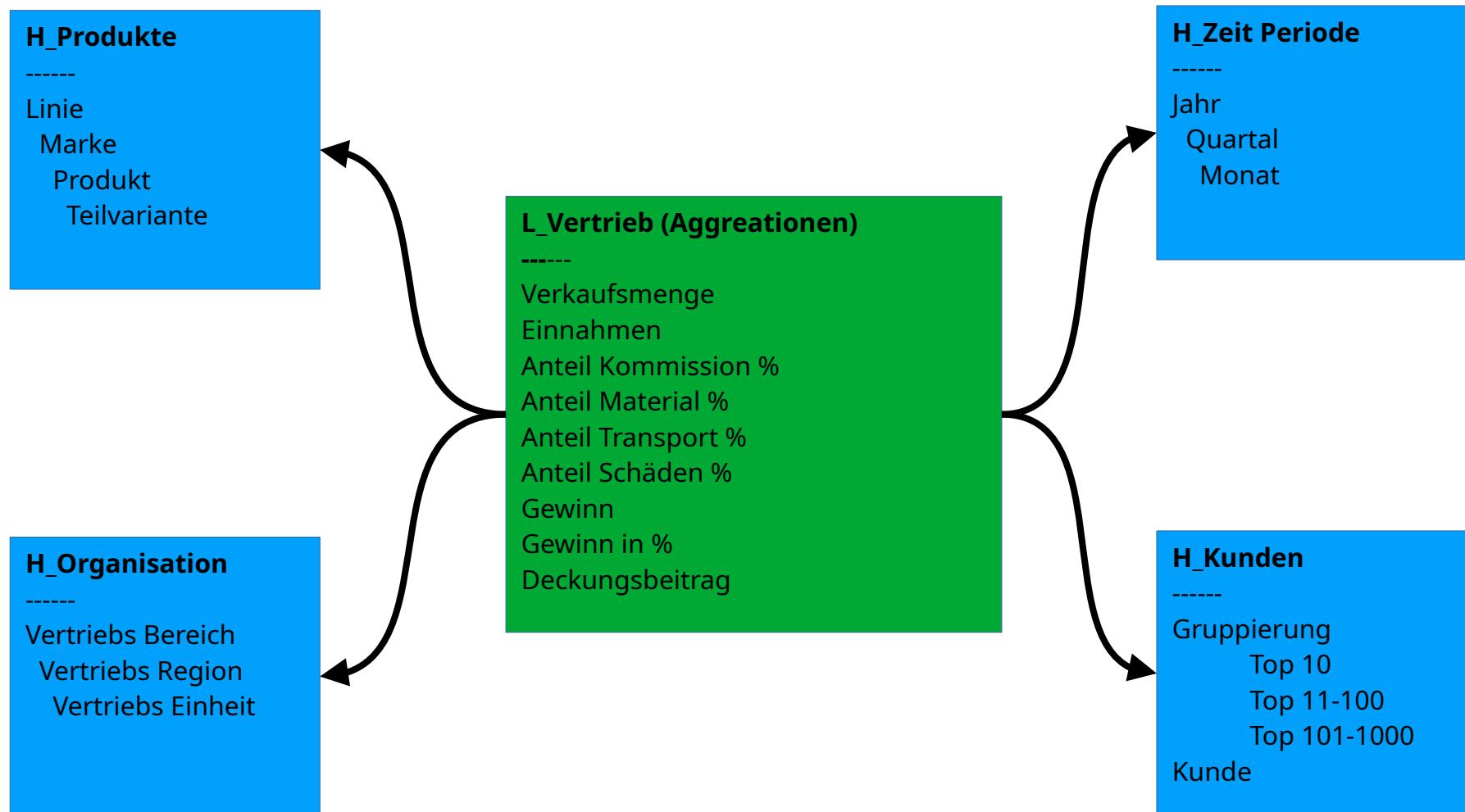
# Beispiele für OLAP Schemen

## Vertriebsanalyse Verkauf



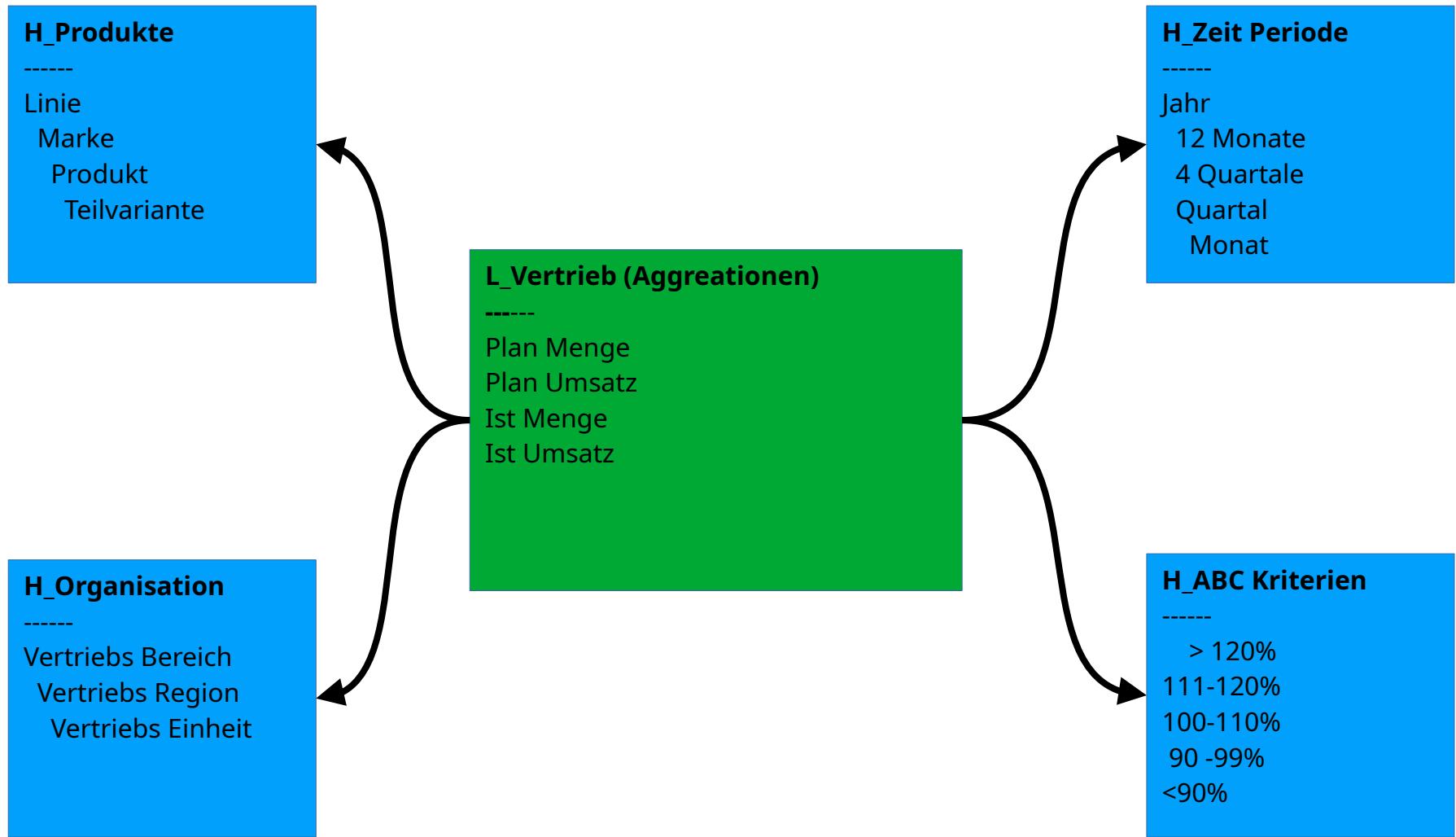
# Beispiele für OLAP Schemen

## Vertriebsanalyse Kunden und Produkt Profitabilität



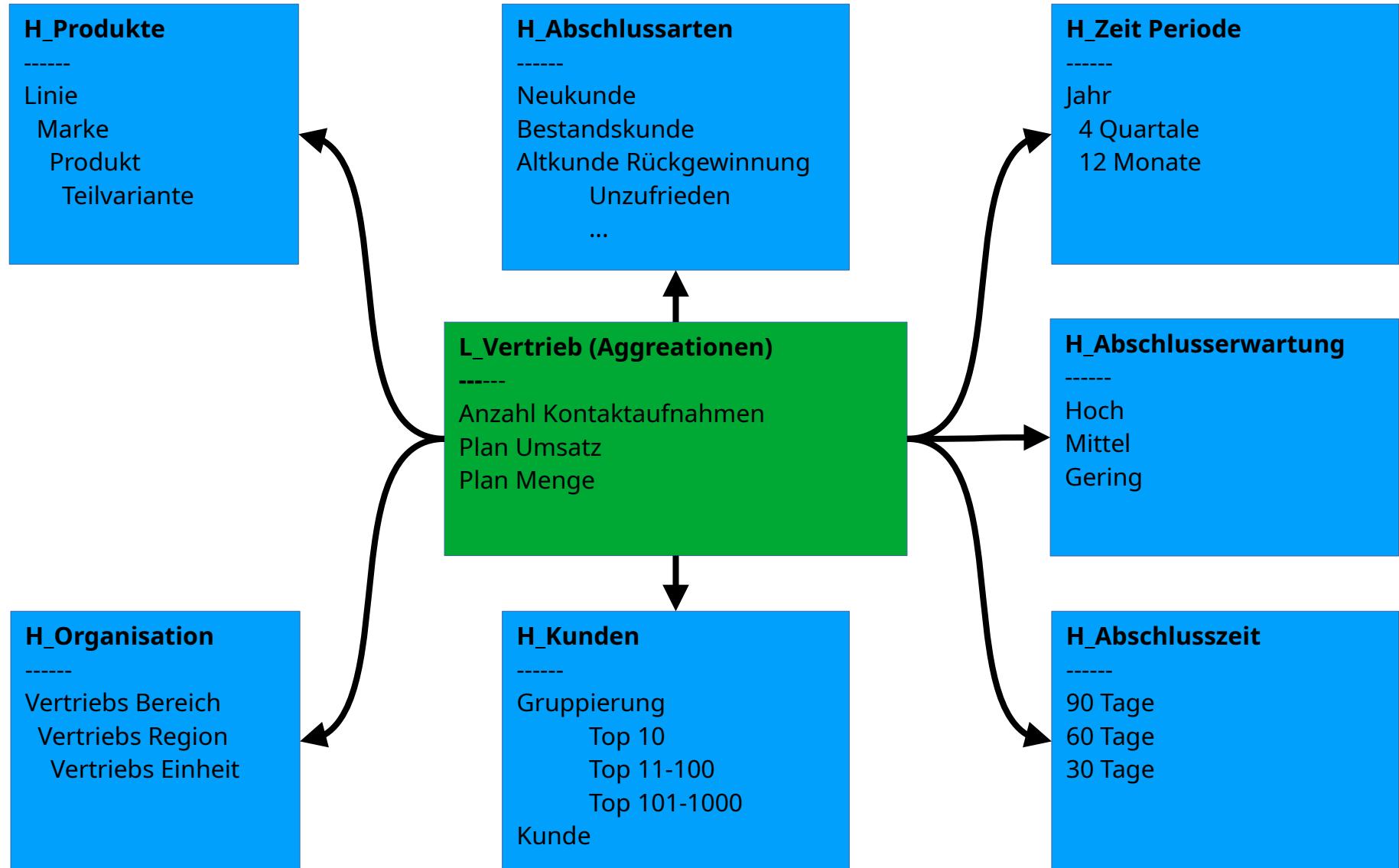
# Beispiele für OLAP Schemen

## Vertriebsanalyse Planung und Ergebnisziele



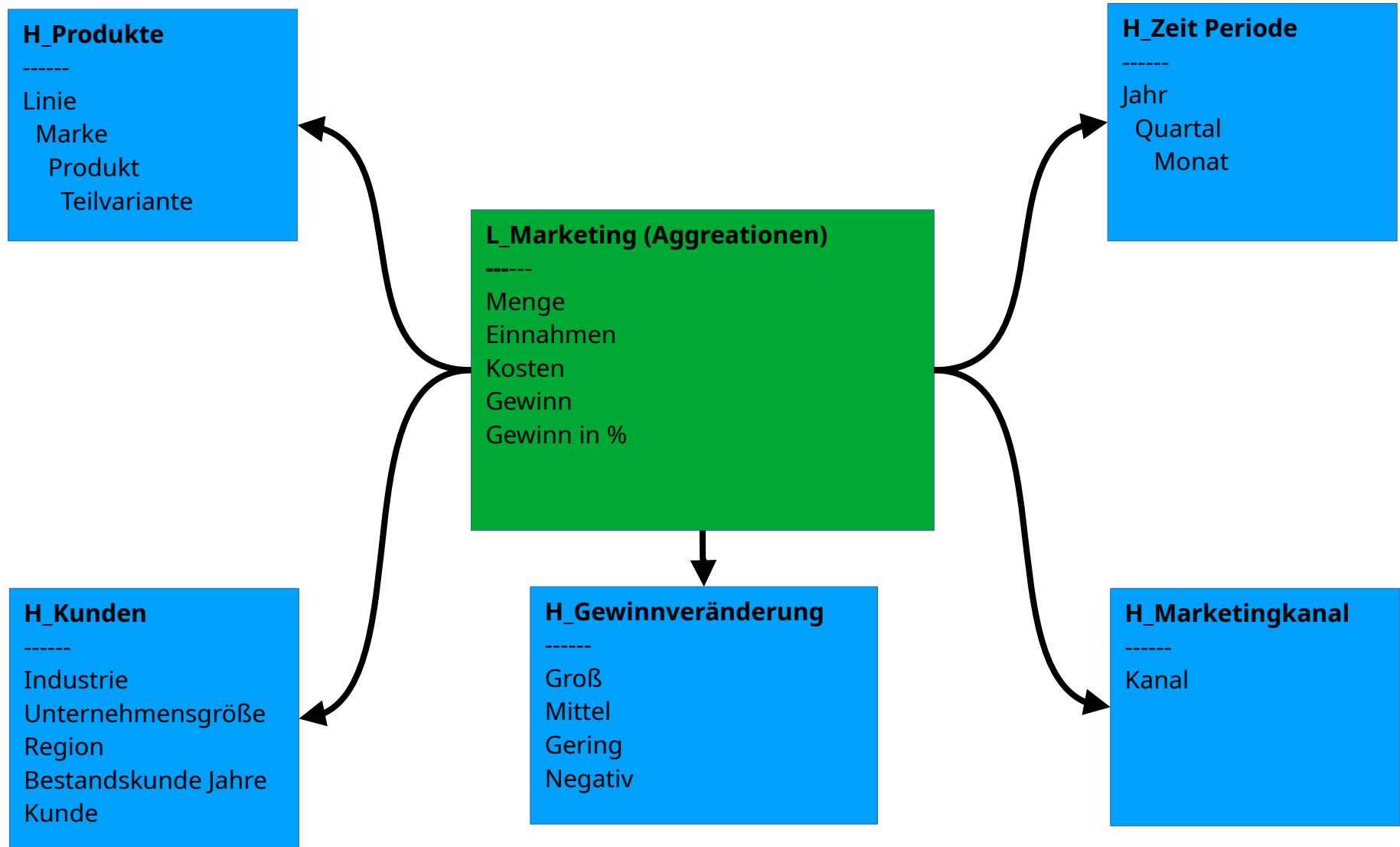
# Beispiele für OLAP Schemen

## Vertriebsanalyse Abschlüsse



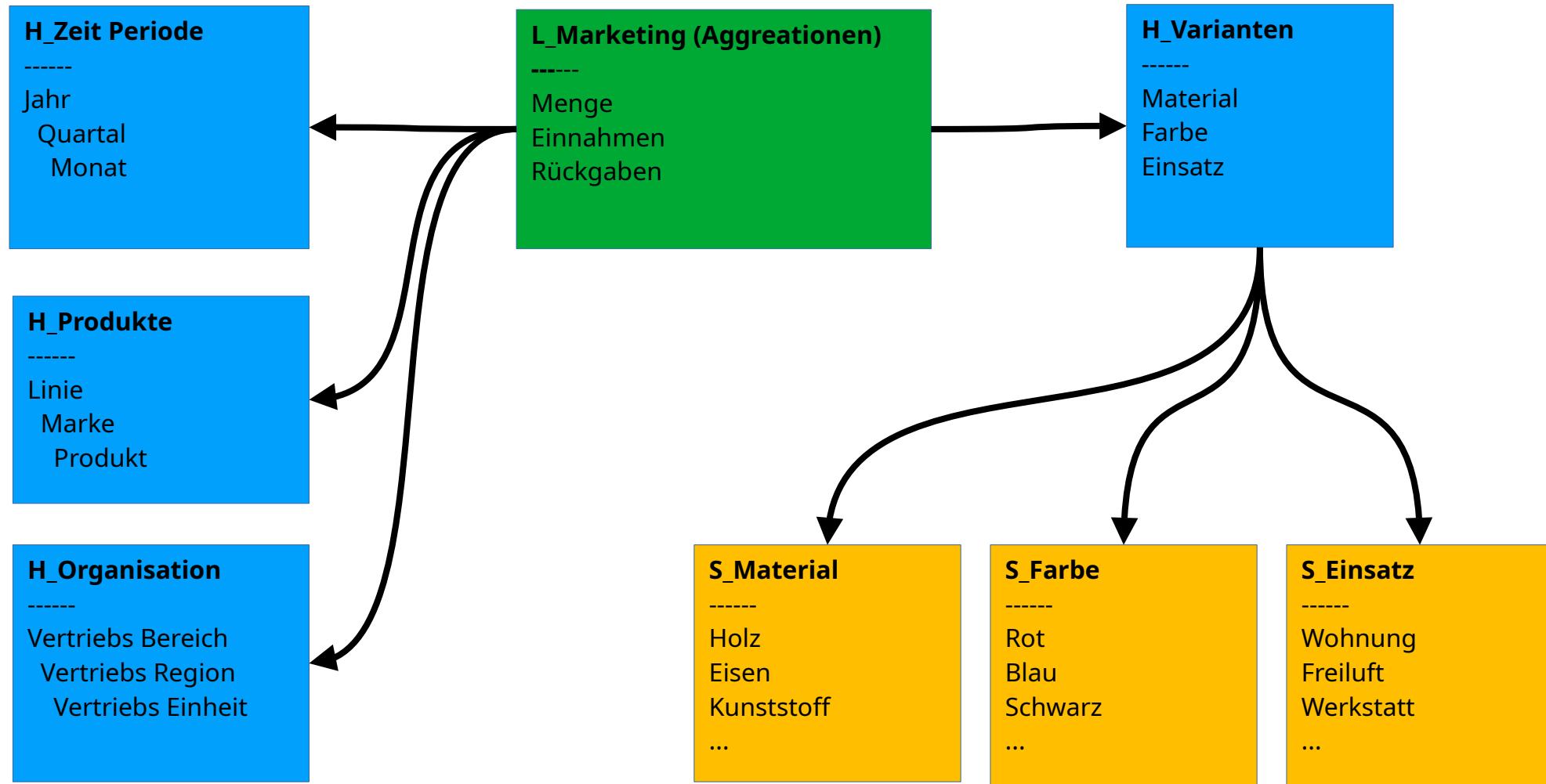
# Beispiele für OLAP Schemen

## Marketinganalyse Strategisch



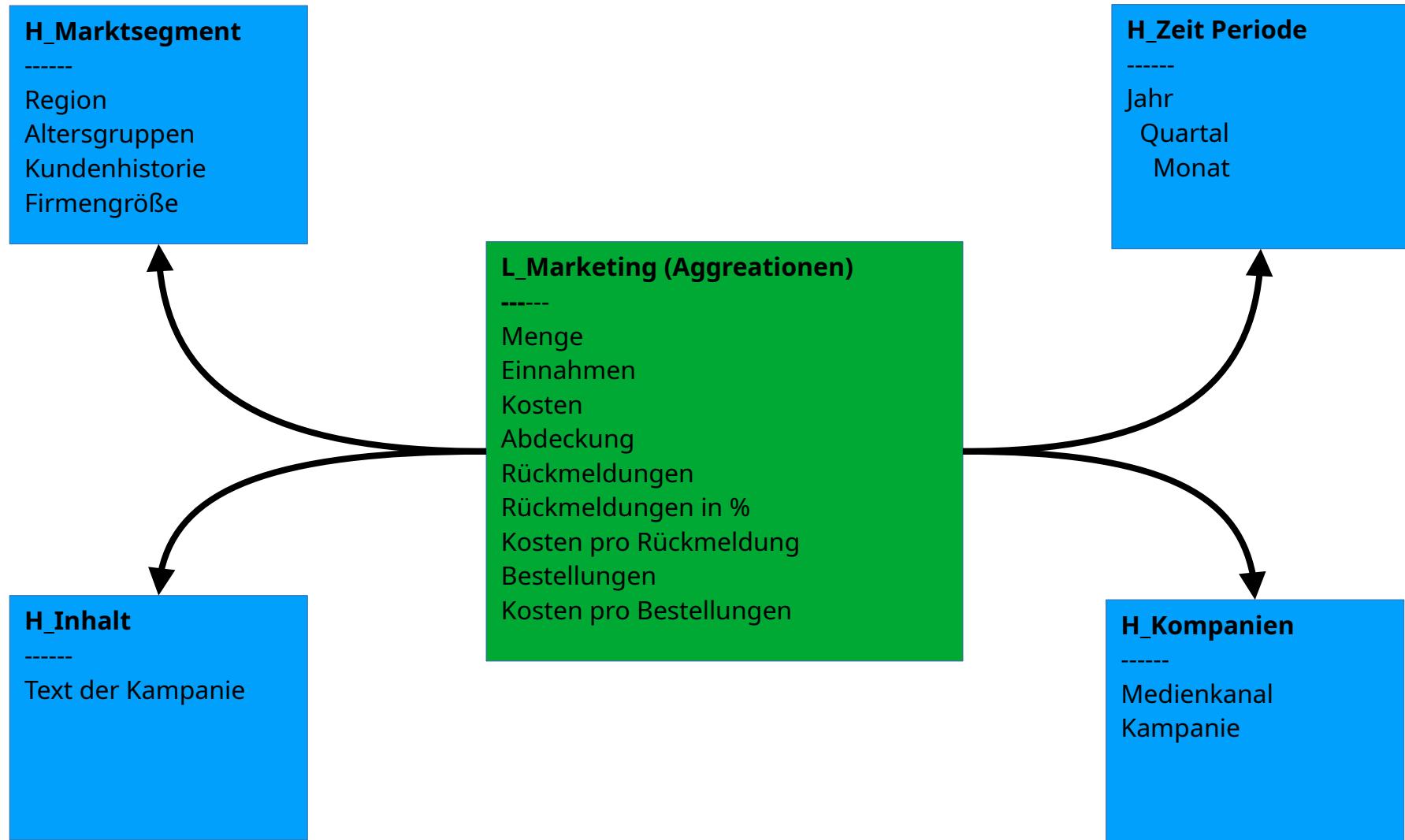
# Beispiele für OLAP Schemen

## Marketinganalyse Produktvariationen



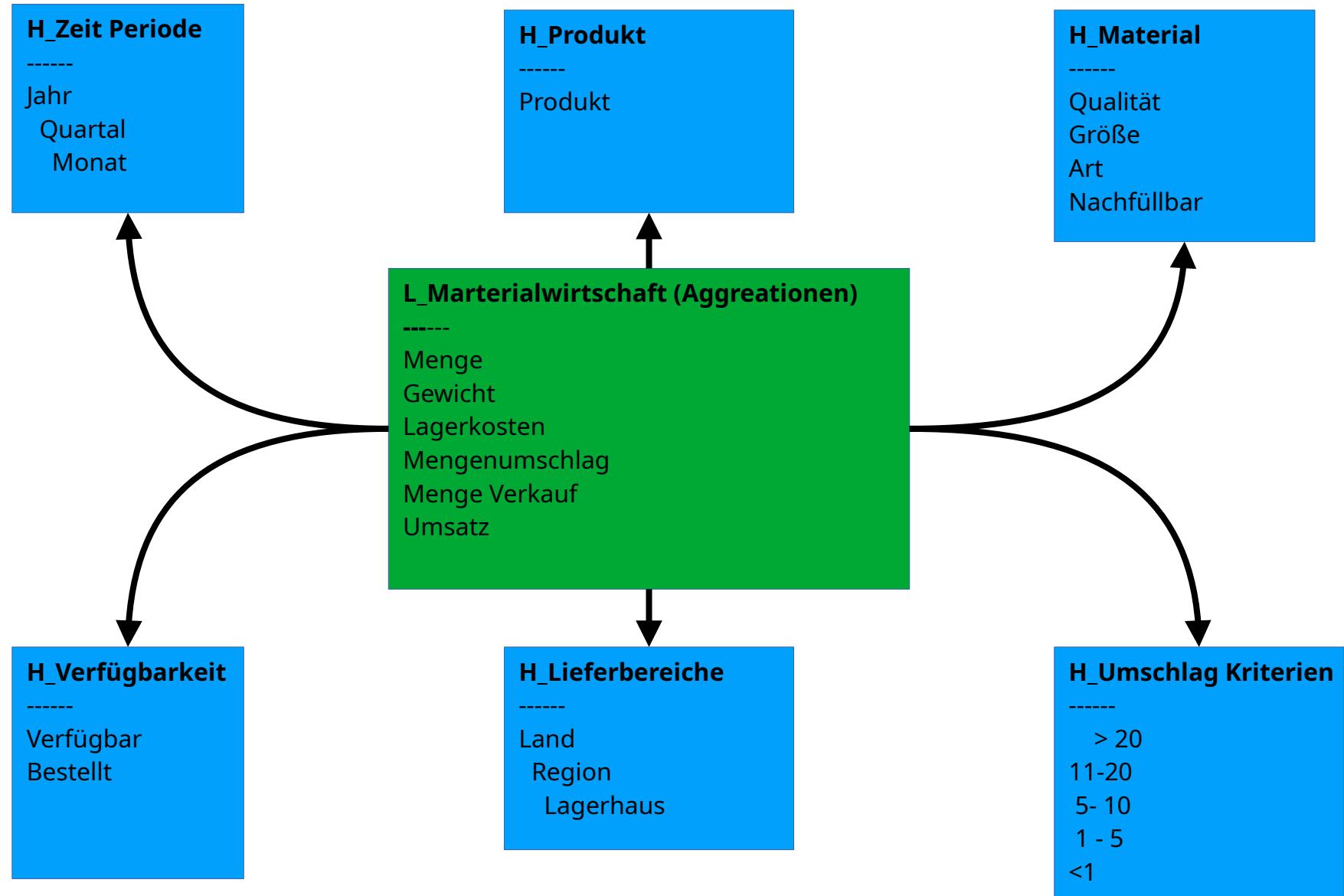
# Beispiele für OLAP Schemen

## Marketinganalyse Taktische Kampanienanalyse



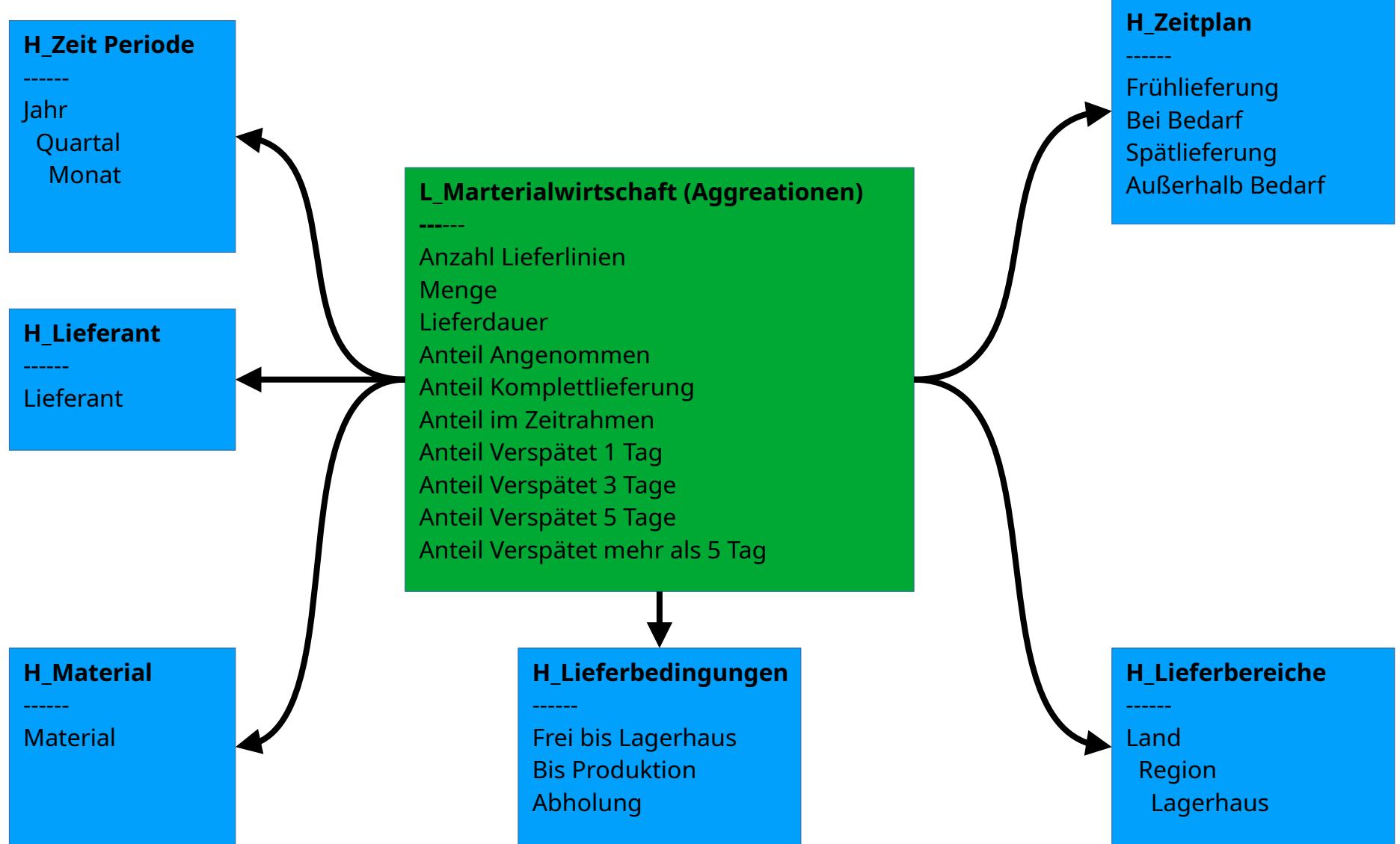
# Beispiele für OLAP Schemen

## Materialwirtschaftsanalyse Material Durchlauf



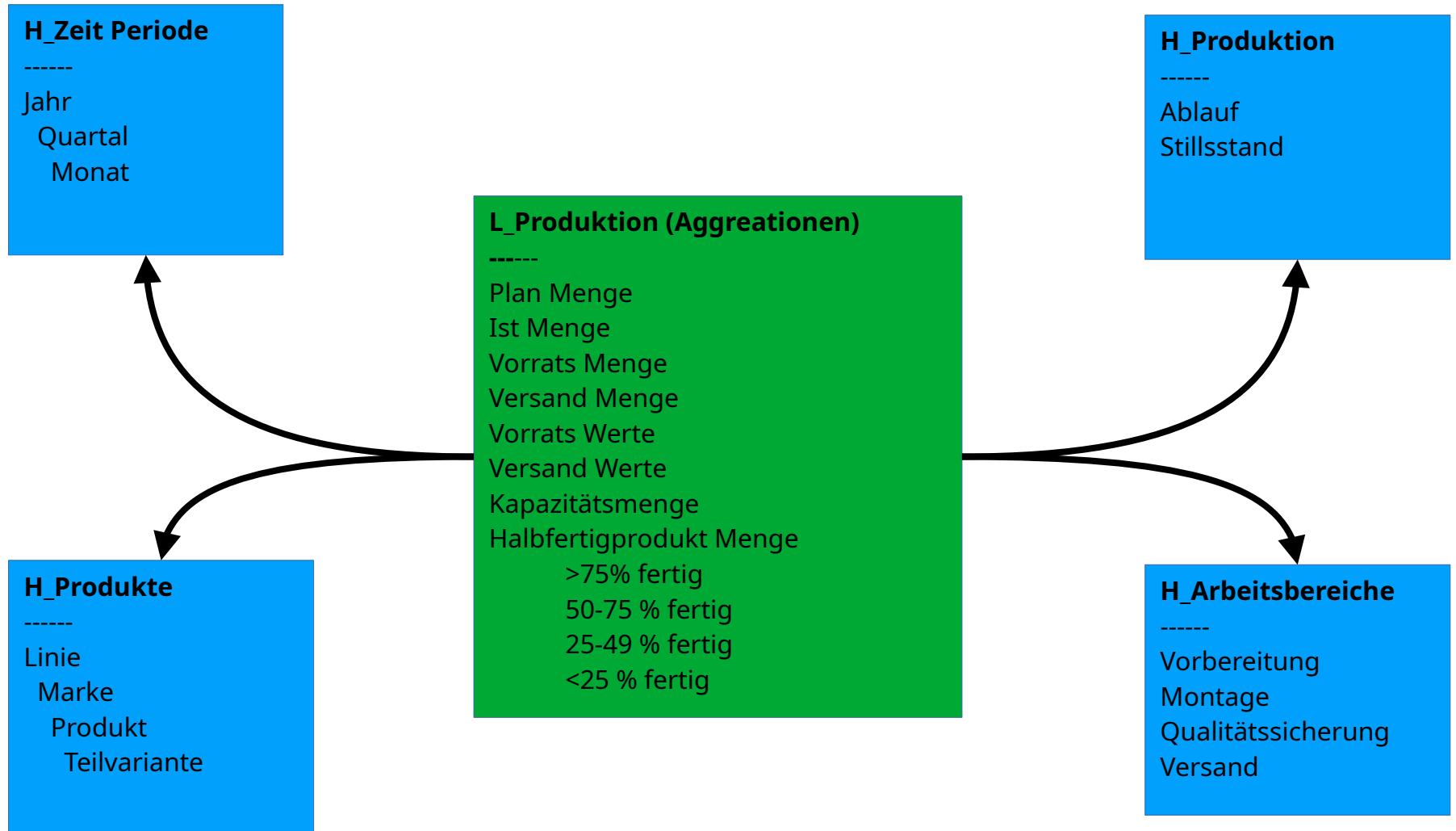
# Beispiele für OLAP Schemen

## Materialwirtschaftsanalyse Lieferanten



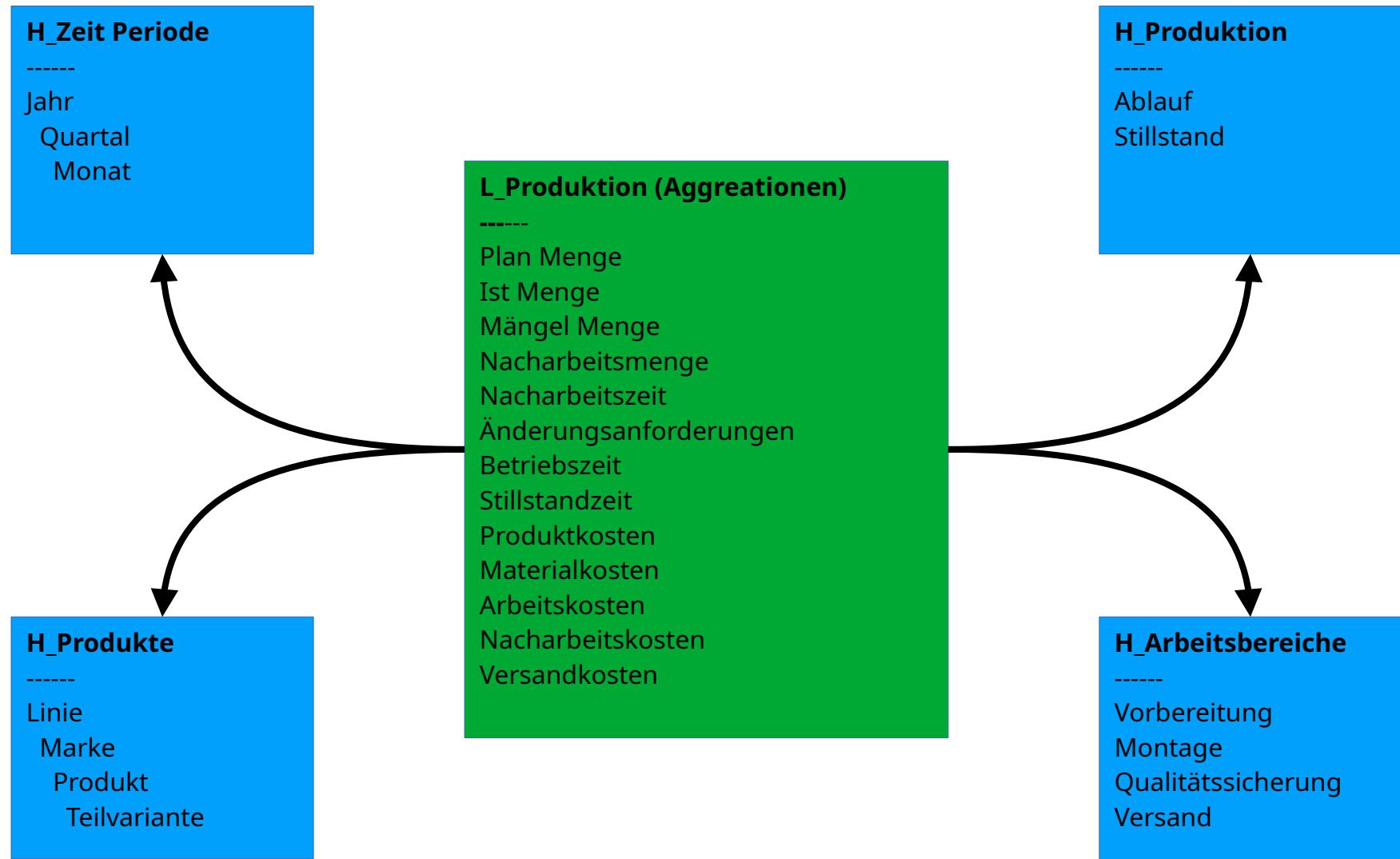
# Beispiele für OLAP Schemen

## Produktionsanalyse Kapazität



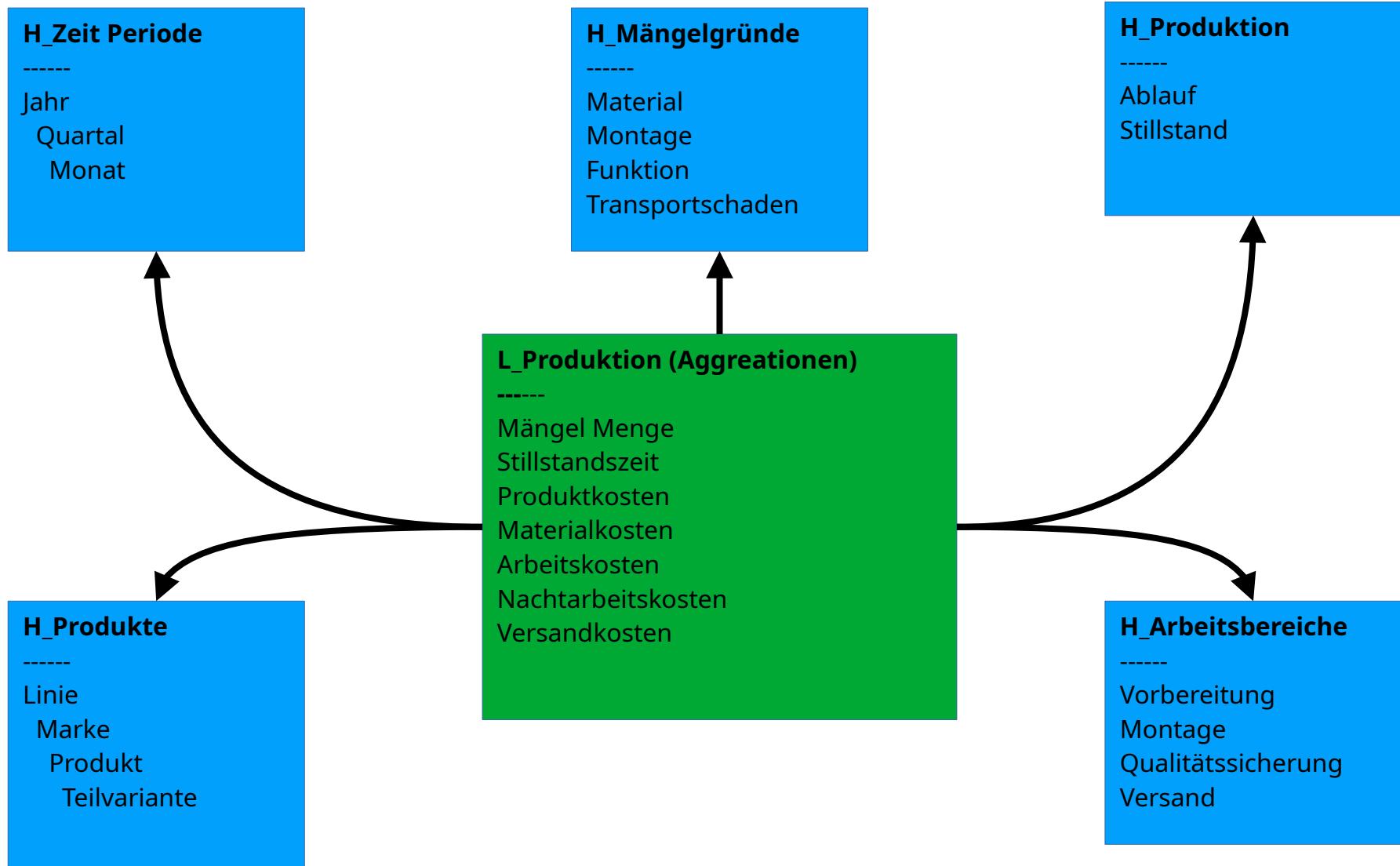
# Beispiele für OLAP Schemen

## Produktionsanalyse Kosten und Qualität



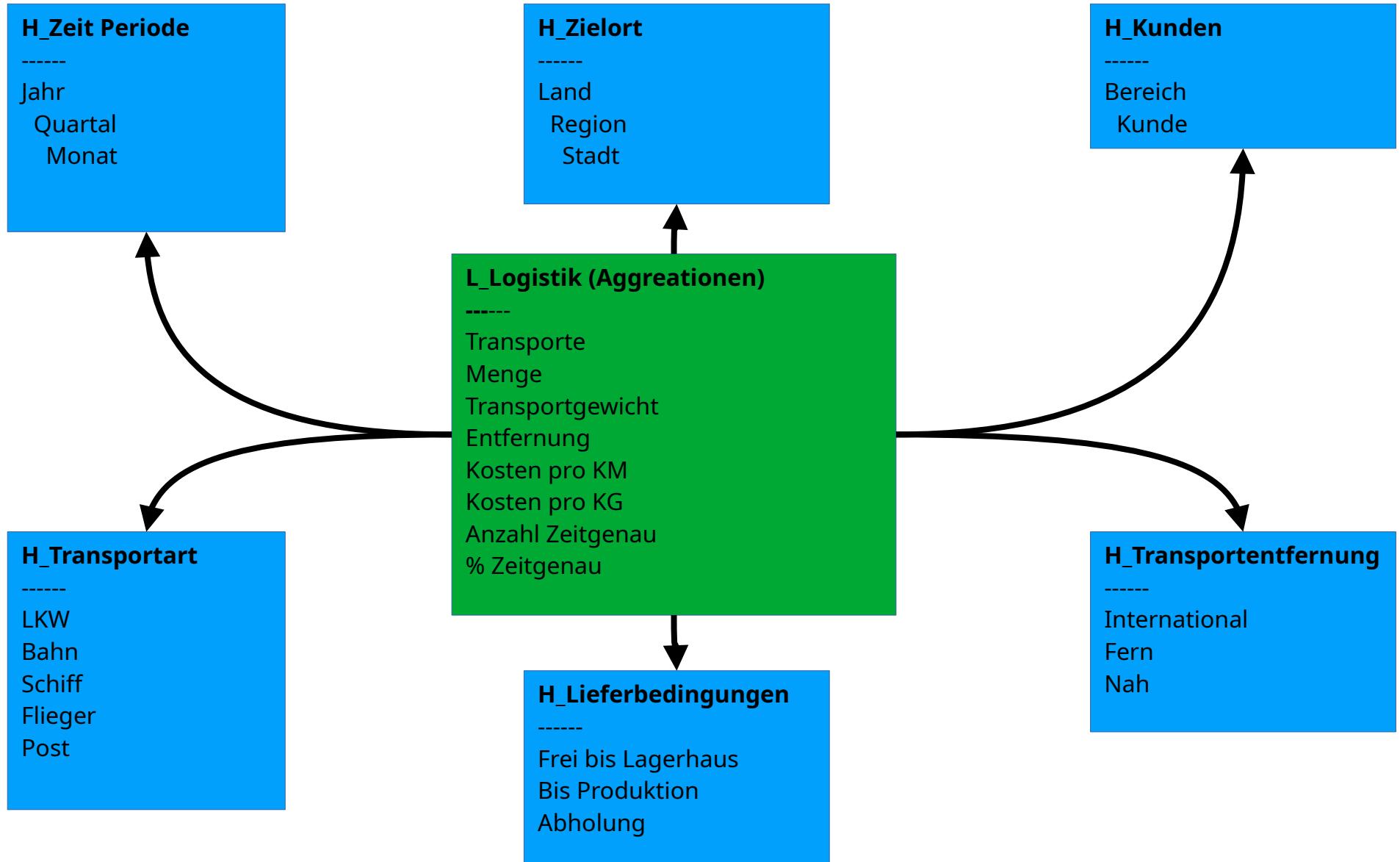
# Beispiele für OLAP Schemen

## Produktionsanalyse Mängelkosten



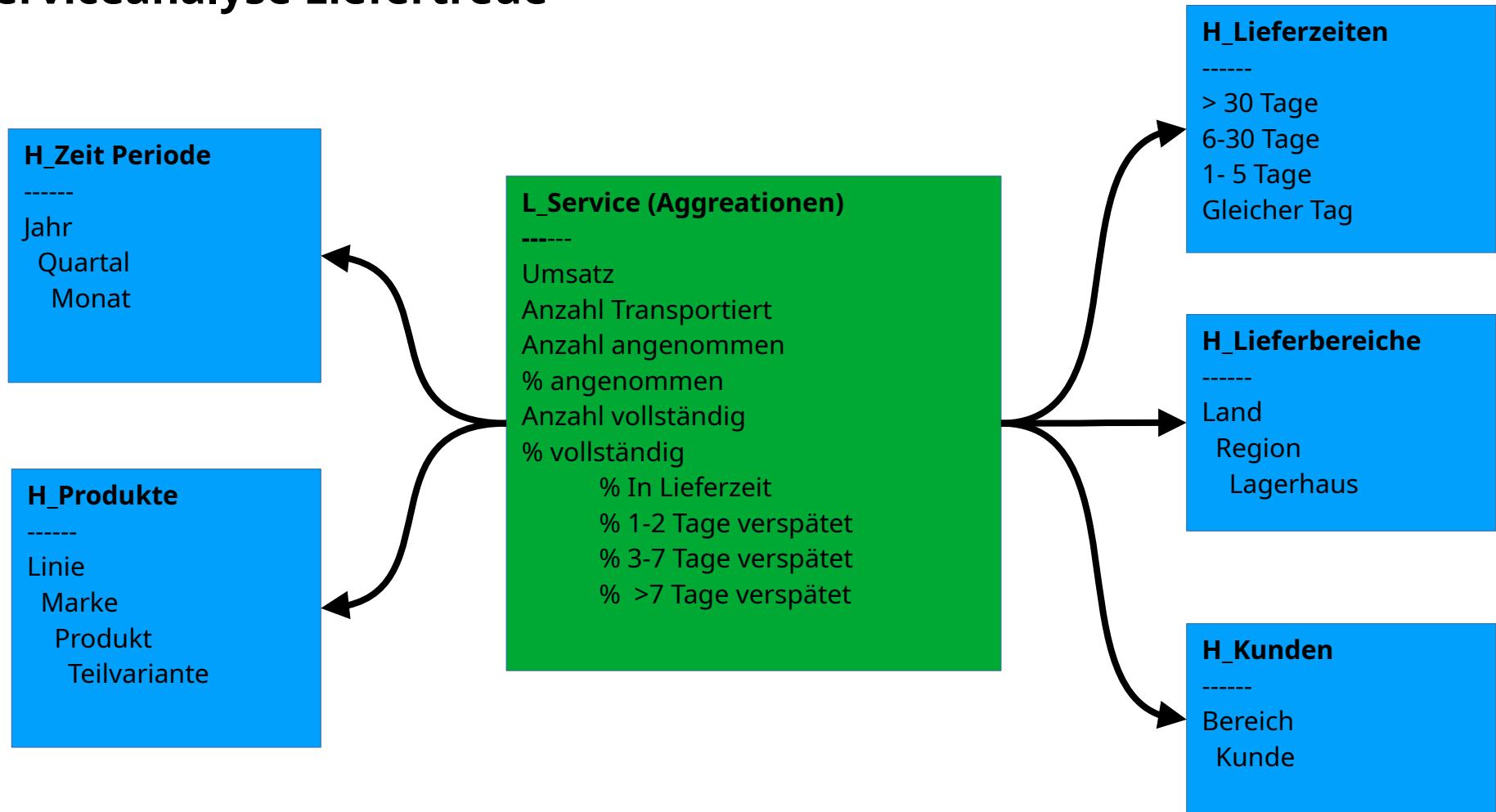
# Beispiele für OLAP Schemen

## Lieferkettenanalyse Transport



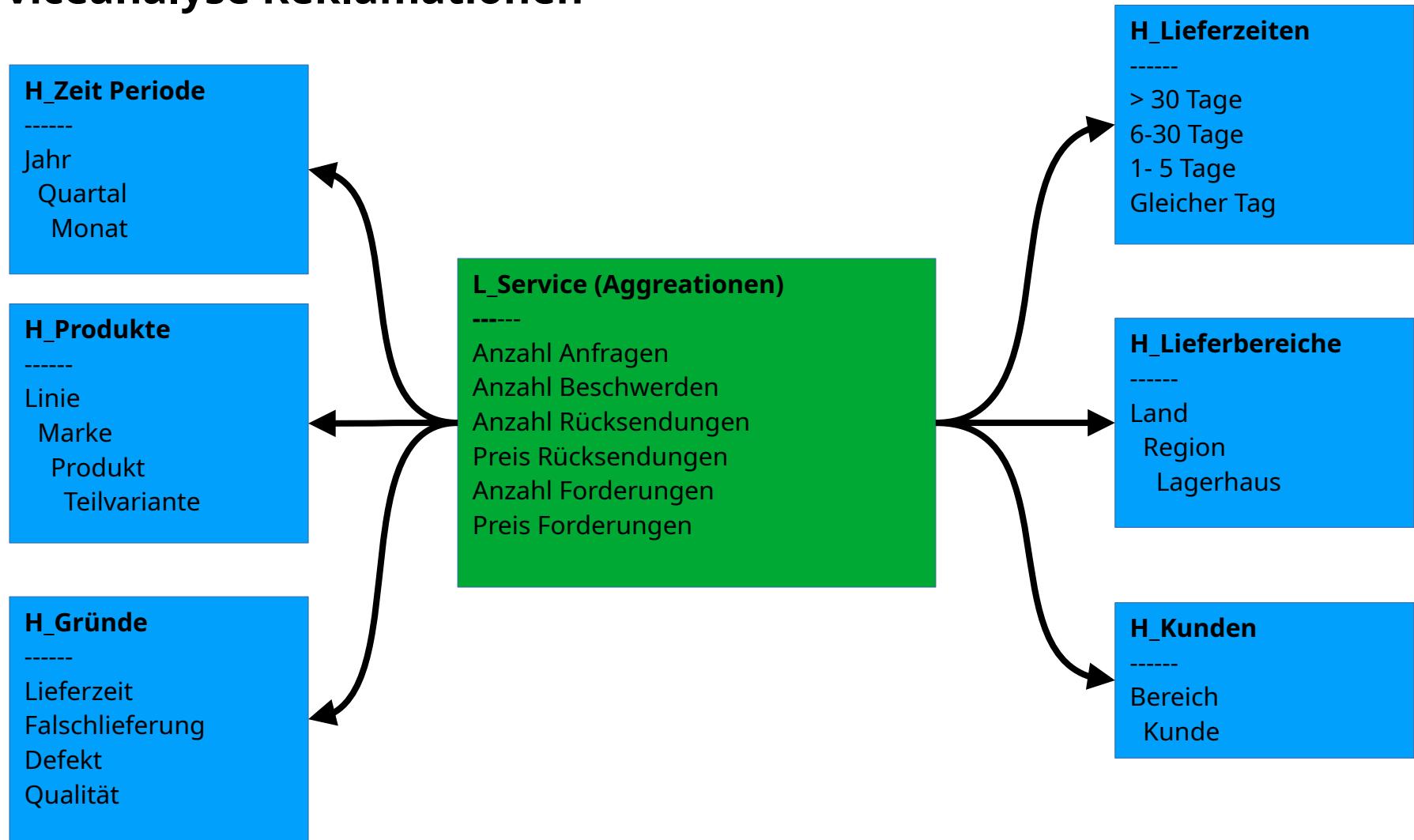
# Beispiele für OLAP Schemen

## Serviceanalyse Liefertreue



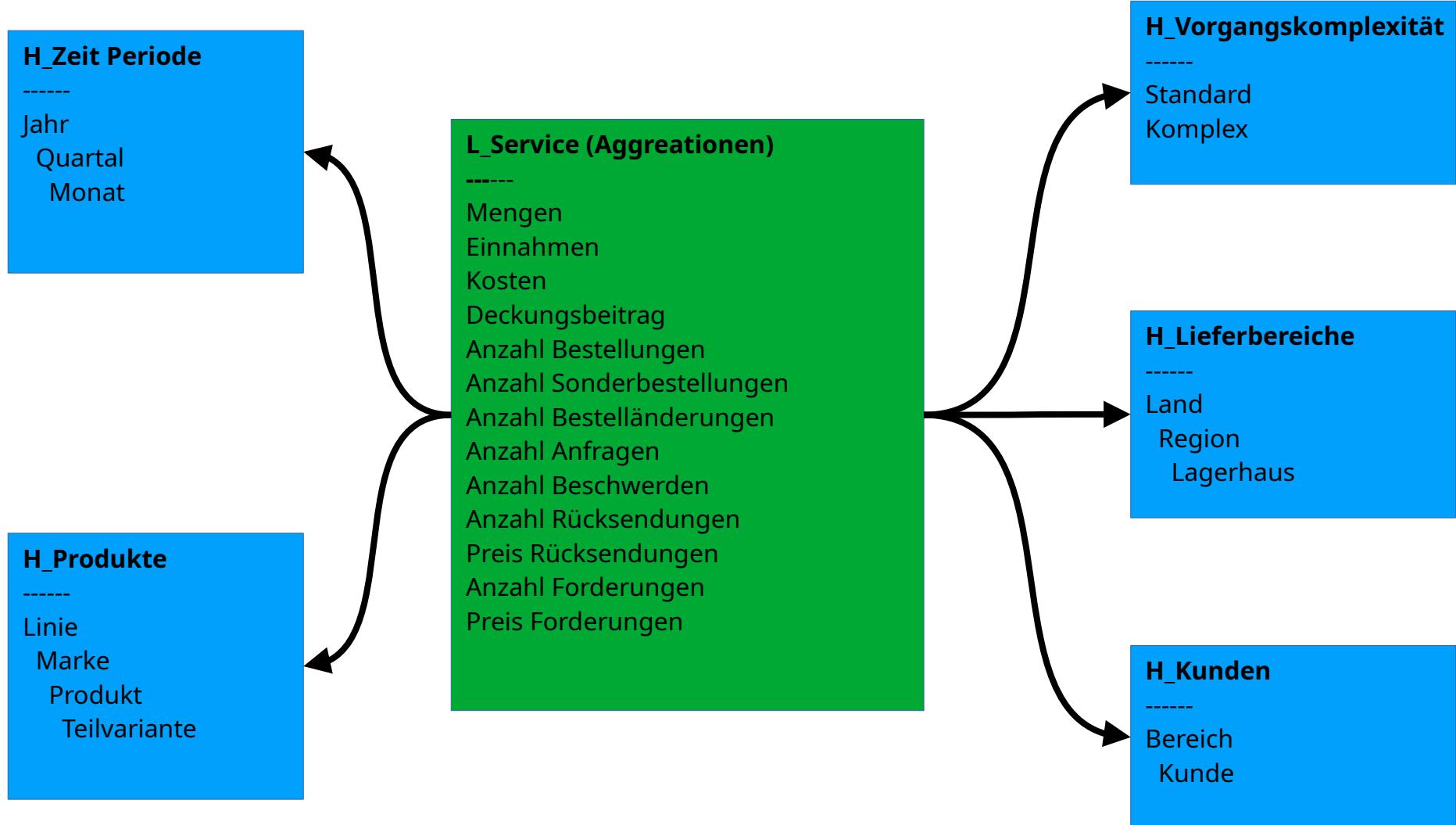
# Beispiele für OLAP Schemen

## Serviceanalyse Reklamationen



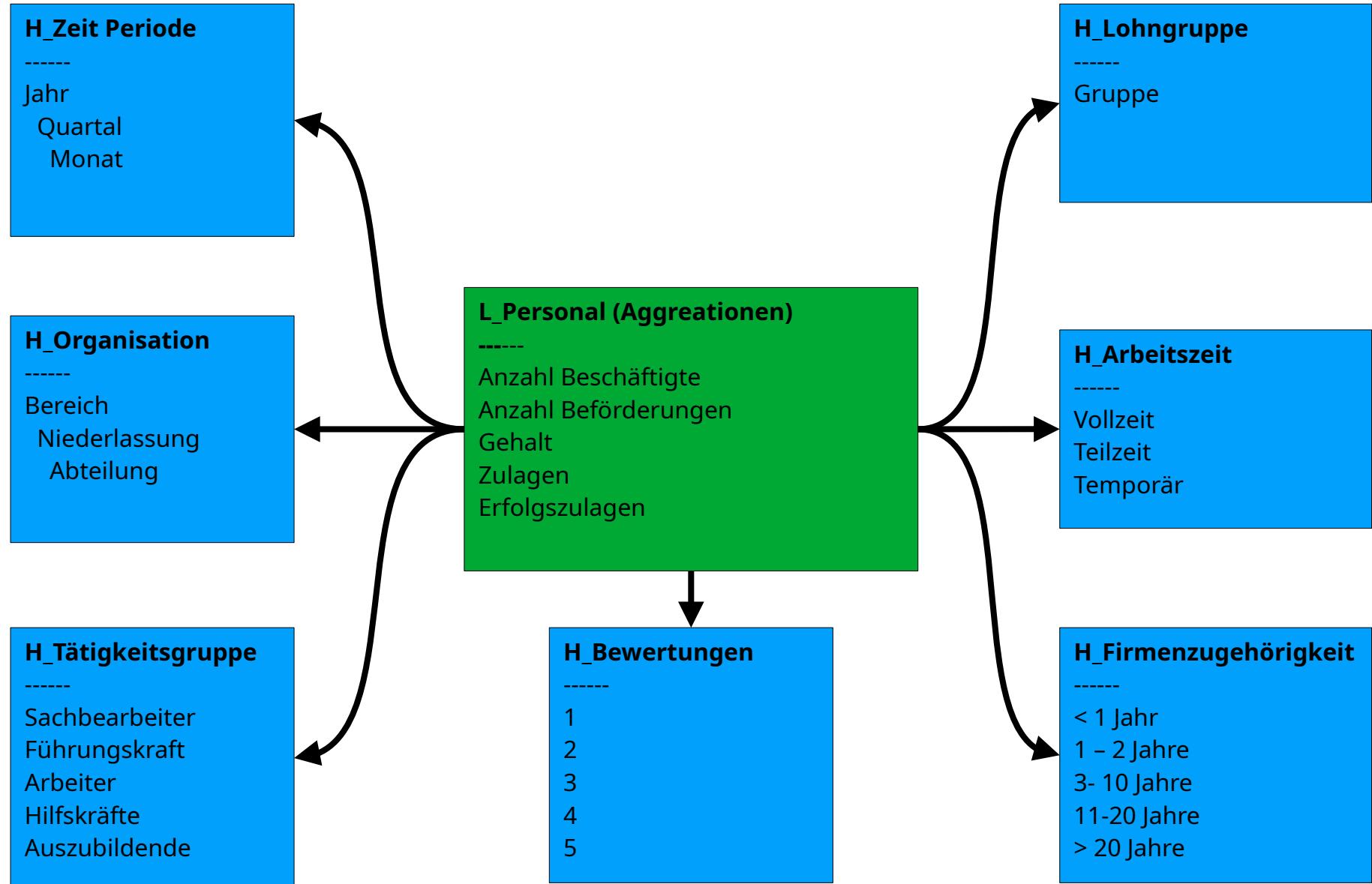
# Beispiele für OLAP Schemen

## Serviceanalyse Kundenbeziehung



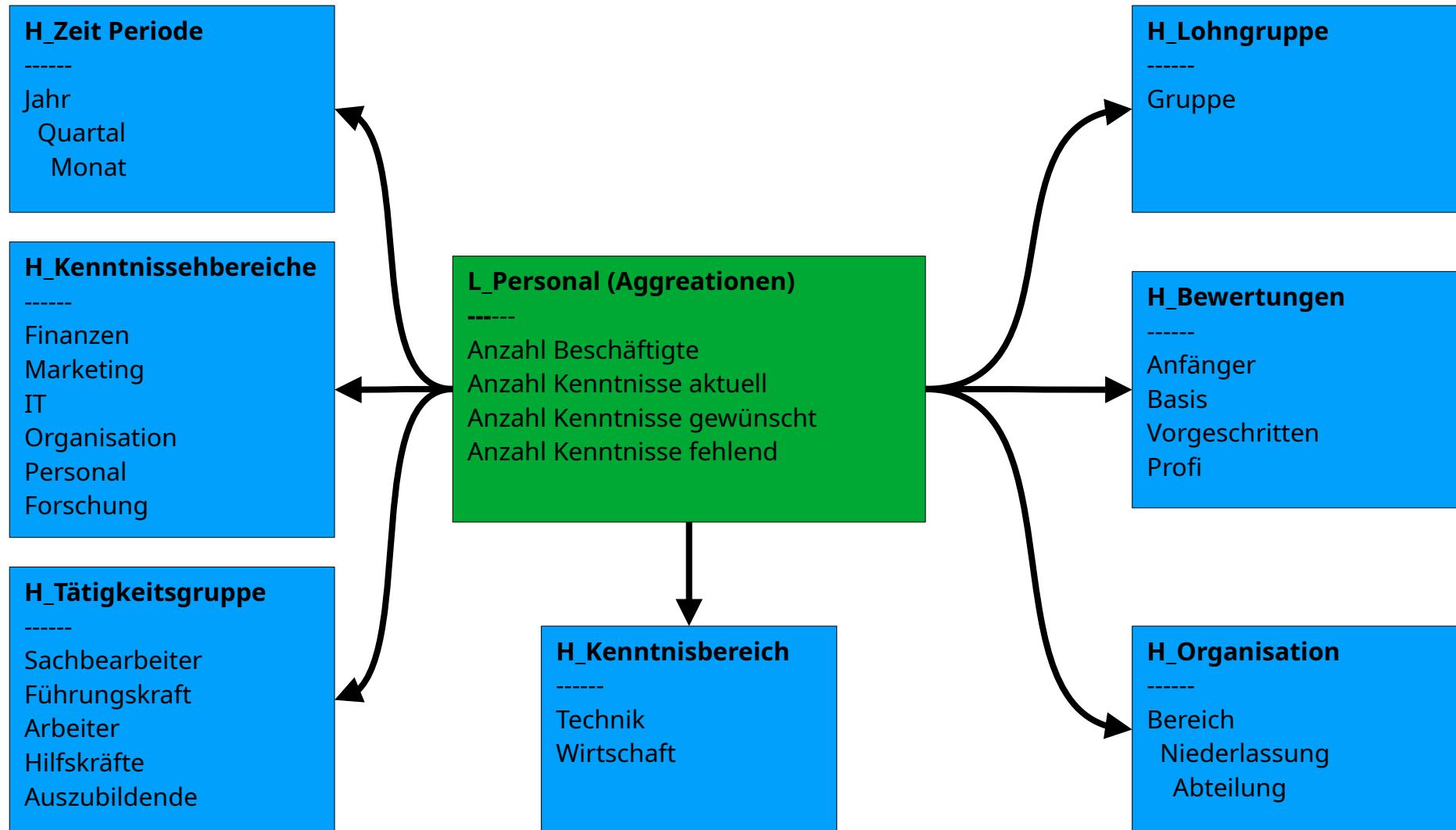
# Beispiele für OLAP Schemen

## Personalanalyse Verwaltung



# Beispiele für OLAP Schemen

## Personalanalyse Know How





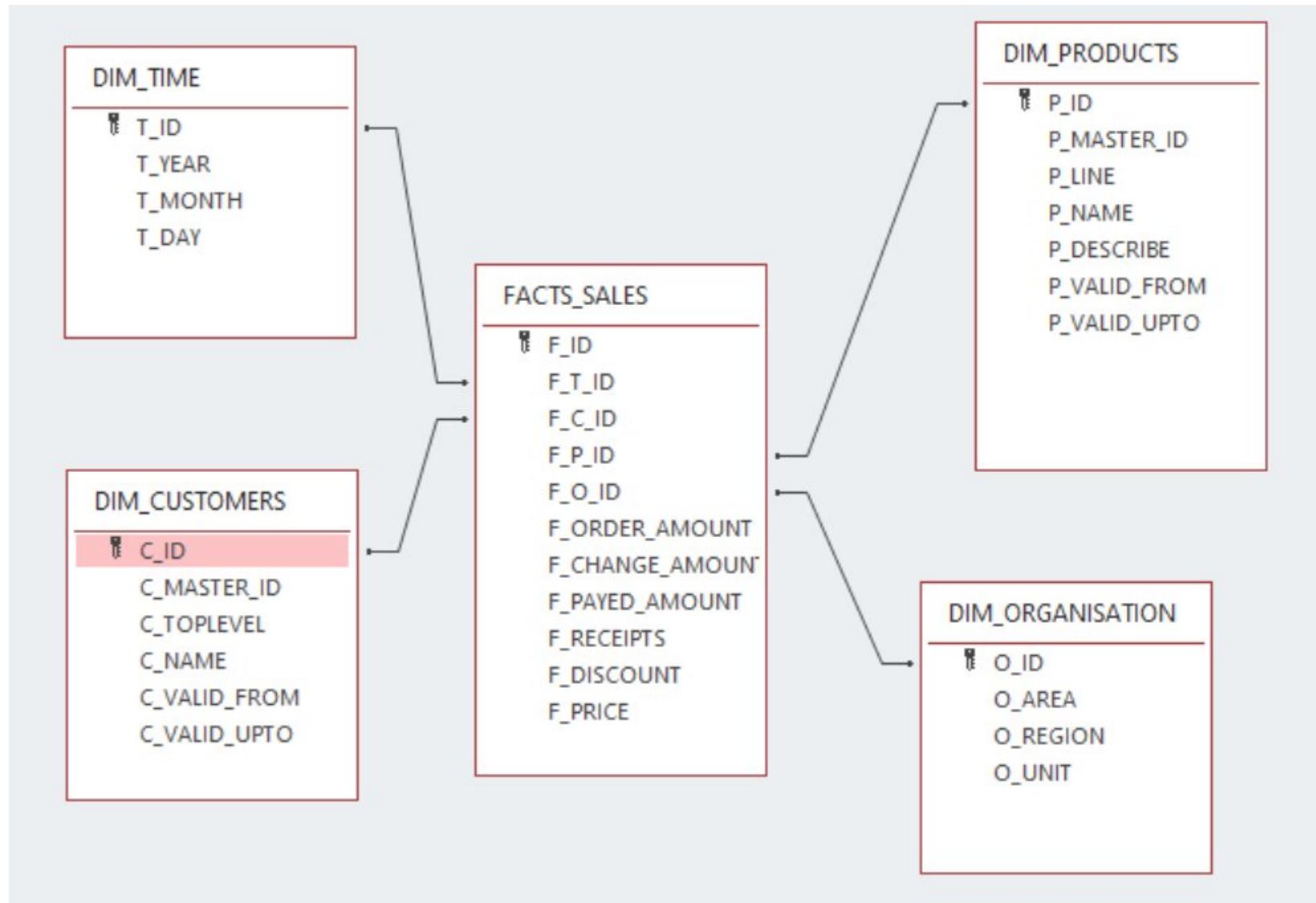
# Beispiele für OLAP Schemen

## Vergleich Beispielmodelle mit bekannten Modellen

- Vergleichen Sie die einzelnen Vorschläge der Analyse mit Ihren Unternehmen und finden Sie Unterschiede, die bei Ihnen anders aussehen würden.
- Wie würden die Kategorisierungen (Dimensionen) in Ihrem Unternehmen aussehen?
- Welche anderen Kennzahlen werden in Ihrem Unternehmen gelebt?
- Entwerfen Sie ein Star Datenschema für den Finanzbereich mit min. 5 Kennzahlen und min. 4 Dimensionen. Bitte achten Sie darauf, dass die Dimensionen min. Stufen haben. Erstellen Sie zusätzlich hierzu eine SQL Abfrage, mit Gruppenwechsel auf alle 3 Aggregationsstufen einer Dimension.
- Erweitern Sie Ihr Datenmodell um eine 2D-Temporalisierung für eine Dimension und passen Sie Ihre SQL Abfrage entsprechend an.

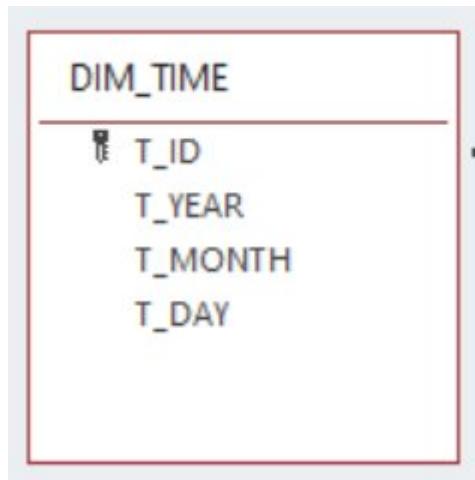
# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema



# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Zeit Dimension



Feldname	Felddatentyp
T_ID	Zahl
T_YEAR	Zahl
T_MONTH	Zahl
T_DAY	Zahl

[+]	44212	2021	1	16
[+]	44213	2021	1	17
[+]	44214	2021	1	18
[+]	44215	2021	1	19
[+]	44216	2021	1	20
[+]	44217	2021	1	21
[+]	44218	2021	1	22
[+]	44219	2021	1	23
[+]	44220	2021	1	24
[+]	44221	2021	1	25
[+]	44222	2021	1	26
[+]	44223	2021	1	27
[+]	44224	2021	1	28
[+]	44225	2021	1	29
[+]	44226	2021	1	30
[+]	44227	2021	1	31
[+]	44228	2021	2	1
[+]	44229	2021	2	2
[+]	44230	2021	2	3
[+]	44231	2021	2	4
[+]	44232	2021	2	5
[+]	44233	2021	2	6
[+]	44234	2021	2	7
[+]	44235	2021	2	8
[+]	44236	2021	2	9
[+]	44237	2021	2	10
[+]	44238	2021	2	11
[+]	44239	2021	2	12
[+]	44240	2021	2	13
[+]	44241	2021	2	14

Datensatz: 1 von 365 Kein Filter Suchen

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Kunden Dimension mit Temporalisierung

DIM_CUSTOMERS	
C_ID	
C_MASTER_ID	
C_TOPLVEL	
C_NAME	
C_VALID_FROM	
C_VALID_UPTO	

Feldname	Felddatentyp
C_ID	Zahl
C_MASTER_ID	Zahl
C_TOPLVEL	Kurzer Text
C_NAME	Kurzer Text
C_VALID_FROM	Datum/Uhrzeit
C_VALID_UPTO	Datum/Uhrzeit

Kunde A und C haben den Level gewechselt

C_ID	C_MASTER_ID	C_TOPLVEL	C_NAME	C_VALID_FROM	C_VALID_UPTO
1000		1000 1000	Kunde G	01.01.2021	31.12.2999
1001		1001 1000	Kunde F	01.01.2021	31.12.2999
1002		1002 1000	Kunde C	01.01.2021	01.06.2021
1003		1003 1000	Kunde A	01.01.2021	05.03.2021
1004		1004 100	Kunde E	01.01.2021	31.12.2999
1005		1005 10	Kunde B	01.01.2021	31.12.2999
1006		1003 100	Kunde A	05.03.2021	15.06.2021
1007		1002 100	Kunde C	01.06.2021	31.12.2999
1008		1003 10	Kunde A	15.06.2021	31.12.2999

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Produkt Dimension mit Temporalisierung

DIM_PRODUCTS	
+	P_ID
	P_MASTER_ID
	P_LINE
	P_NAME
	P_DESCRIBE
	P_VALID_FROM
	P_VALID_UPTO

Feldname	Felddatentyp
P_ID	Zahl
P_MASTER_ID	Zahl
P_LINE	Kurzer Text
P_NAME	Kurzer Text
P_DESCRIBE	Kurzer Text
P_VALID_FROM	Datum/Uhrzeit
P_VALID_UPTO	Datum/Uhrzeit

Das Buch wurde zweimal korrigiert

P_ID	P_MASTER_ID	P_LINE	P_NAME	P_DESCRIBE	P_VALID_FR	P_VALID_UPTO
1000	1000	Buch	Per Mitfahrer durch die Galaxis	cooles Buch, falsch geschrieben	01.01.2021	05.05.2021
1001	1001	Film	Per Anhalter durch die Galaxis	cooler Film	01.01.2021	31.12.2999
1002	1002	Serie	Per Anhalter durch die Galaxis	Coole Serie	01.01.2021	31.12.2999
1003	1000	Buch	Per Anhalter durch die Galaxis	cooles älteres Buch	05.05.2021	10.06.2021
1004	1000	Buch	Per Anhalter durch die Galaxis	cooles Buch	10.06.2021	31.12.2999

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Organisations Dimension

DIM_ORGANISATION	
O_ID	
O_AREA	
O_REGION	
O_UNIT	

O_ID	O_AREA	O_REGION	O_UNIT
1000	SÜD	Hochland	Einheit A
1001	SÜD	Tiefland	Einheit B
1002	Nord	Seeland	Einheit C
1003	Nord	Küste	Einheit D
1004	Nord	Hinterland	Einheit E

Feldname	Felddatentyp
O_ID	Zahl
O_AREA	Kurzer Text
O_REGION	Kurzer Text
O_UNIT	Kurzer Text

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten Würfel an Vertriebs Vorlage angelehnt

FACTS_SALES											
	F_ID	F_T_ID	F_C_ID	F_P_ID	F_O_ID	F_ORDER_A	F_CHANGE_A	F_PAYED_A	F_RECEIPTS	F_DISCOUNT	F_PRICE
.	1000	44197	1001	1001	1001	56	0	56	2.973,03 €	2,04 €	12,02 €
.	1001	44208	1001	1001	1001	202	-44	158	1.984,56 €	0,17 €	5,59 €
.	1002	44211	1001	1002	1001	800	-127	673	3.084,24 €	3,58 €	18,85 €
.	1003	44214	1001	1002	1001	734	-611	123	217,15 €	1,97 €	12,31 €
.	1004	44216	1001	1000	1001	505	-182	323	140,71 €	1,03 €	5,72 €
.	1005	44223	1001	1002	1001	733	-227	506	3.340,01 €	0,39 €	6,58 €
.	1006	44226	1001	1002	1001	466	-436	30	4.454,47 €	1,57 €	17,42 €
.	1007	44228	1001	1002	1001	852	-486	366	875,78 €	0,62 €	3,11 €
.	1008	44230	1001	1000	1001	165	-63	102	3.332,09 €	0,82 €	16,39 €
.	1009	44233	1001	1002	1001	672	-479	193	1.234,40 €	2,19 €	12,14 €
.	1010	44238	1001	1001	1001	15	-10	5	20,48 €	2,72 €	12,96 €
.	1011	44247	1001	1000	1001	721	-652	69	985,94 €	0,93 €	4,41 €
.	1012	44248	1001	1002	1001	409	-170	239	3.709,68 €	1,69 €	13,00 €
.	1013	44251	1001	1000	1001	832	-225	607	2.714,33 €	1,44 €	17,99 €
.	1014	44252	1001	1002	1001	630	-93	537	1.700,17 €	3,41 €	17,94 €
.	1015	44255	1001	1002	1001	71	-25	46	2.962,43 €	0,07 €	3,37 €
.	1016	44262	1001	1000	1001	141	-48	93	186,60 €	1,87 €	9,33 €
.	1017	44277	1001	1000	1001	838	-763	75	4.210,33 €	1,76 €	11,71 €
.	1018	44279	1001	1001	1001	362	-197	165	2.691,53 €	0,19 €	18,88 €
.	1019	44283	1001	1001	1001	166	-20	146	274,62 €	2,15 €	9,78 €
.	1020	44284	1001	1002	1001	779	-119	660	1.506,82 €	1,80 €	19,95 €
.	1021	44285	1001	1000	1001	124	-94	30	428,05 €	0,87 €	7,26 €
.	1022	44297	1001	1001	1001	171	-21	150	6,78 €	0,00 €	6,78 €
.	1023	44302	1001	1002	1001	349	-149	200	1.888,27 €	0,51 €	12,69 €
.	1024	44304	1001	1002	1001	445	-175	270	15.089,91 €	0,99 €	19,88 €
.	1025	44305	1001	1001	1001	302	-107	195	4.799,22 €	2,24 €	10,17 €
.	1026	44307	1001	1002	1001	688	-211	477	922,10 €	0,66 €	6,61 €
.	1027	44314	1001	1001	1001	139	-56	83	1.638,95 €	0,14 €	3,47 €
.	1028	44318	1001	1001	1001	202	-145	57	423,26 €	2,53 €	19,46 €
.	1029	44319	1001	1002	1001	105	-83	22	1.537,99 €	2,06 €	10,85 €
.	1030	44324	1001	1002	1001	408	-210	198	116,10 €	0,32 €	2,25 €
.	1031	44328	1001	1002	1001	495	-127	368	2.180,80 €	0,00 €	4,70 €
.	1032	44330	1001	1001	1001	248	-120	128	4.291,47 €	3,48 €	15,81 €

Stensatz: 14 53 von 673 Kein Filter Suchen

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Kunden gültig am 3.5.2021

```
SELECT C_ID, C_MASTER_ID, C_TOPLEVEL, C_NAME, C_VALID_FROM, C_VALID_UPTO  
FROM DIM_CUSTOMERS  
WHERE C_VALID_FROM <=#3/5/2021# AND C_VALID_UPTO > #4/5/2021#  
ORDER BY C_NAME;
```

C_ID	C_MASTER_ID	C_TOPLEVEL	C_NAME	C_VALID_FR	C_VALID_UF
1006	1003	100	Kunde A	05.03.2021	15.06.2021
1005	1005	10	Kunde B	01.01.2021	31.12.2999
1002	1002	1000	Kunde C	01.01.2021	01.06.2021
1004	1004	100	Kunde E	01.01.2021	31.12.2999
1001	1001	1000	Kunde F	01.01.2021	31.12.2999
1000	1000	1000	Kunde G	01.01.2021	31.12.2999

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Organisation einfach

```
SELECT O_AREA, O_REGION, O_UNIT, sum(F_ORDER_AMOUNT) AS ORDERS,
       sum(F_CHANGE_AMOUNT) AS CHANGES, sum(F_PAYED_AMOUNT) AS PAYEDS
FROM FACTS_SALES, DIM_ORGANISATION
WHERE F_O_ID = O_ID
GROUP BY O_AREA, O_REGION, O_UNIT;
```

Keine Temporalisierung verwendet

Jede Kombination ein Gruppenwechsel

O_AREA	O_REGION	O_UNIT	ORDERS	CHANGES	PAYEDS
Nord	Hinterland	Einheit E	80833	-40342	40491
SÜD	Hochland	Einheit A	127368	-68449	58919
SÜD	Tiefland	Einheit B	128195	-62960	65235

Modernerer Join

```
SELECT O_AREA, O_REGION, O_UNIT, sum(F_ORDER_AMOUNT) AS ORDERS,
       sum(F_CHANGE_AMOUNT) AS CHANGES, sum(F_PAYED_AMOUNT) AS PAYEDS
FROM FACTS_SALES
INNER JOIN DIM_ORGANISATION ON F_O_ID = O_ID
GROUP BY O_AREA, O_REGION, O_UNIT;
```

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Organisation dreistufig

```

SELECT * FROM
(
  SELECT O_AREA, '' AS O_REGION, '' AS O_UNIT, sum(F_ORDER_AMOUNT) AS ORDERS,
         sum(F_CHANGE_AMOUNT) AS CHANGES, sum(F_PAYED_AMOUNT) AS PAYEDS
    FROM FACTS_SALES, DIM_ORGANISATION
   WHERE F_O_ID = O_ID
  GROUP BY O_AREA
UNION
  SELECT O_AREA, O_REGION, '' AS O_UNIT, sum(F_ORDER_AMOUNT) AS ORDERS,
         sum(F_CHANGE_AMOUNT) AS CHANGES, sum(F_PAYED_AMOUNT) AS PAYEDS
    FROM FACTS_SALES, DIM_ORGANISATION
   WHERE F_O_ID = O_ID
  GROUP BY O_AREA, O_REGION
UNION
  SELECT O_AREA, O_REGION, O_UNIT, sum(F_ORDER_AMOUNT) AS ORDERS,
         sum(F_CHANGE_AMOUNT) AS CHANGES, sum(F_PAYED_AMOUNT) AS PAYEDS
    FROM FACTS_SALES, DIM_ORGANISATION
   WHERE F_O_ID = O_ID
  GROUP BY O_AREA, O_REGION, O_UNIT
)
ORDER BY O_AREA, O_REGION, O_UNIT;
  
```

Gruppenwechsel auf Bereich

Gruppenwechsel auf Bereich & Region

Gruppenwechsel auf Bereich & Region & Einheit

O_AREA	O_REGION	O_UNIT	ORDERS	CHANGES	PAYEDS
Nord			80833	-40342	40491
Nord	Hinterland		80833	-40342	40491
Nord	Hinterland	Einheit E	80833	-40342	40491
SÜD			255563	-131409	124154
SÜD	Hochland		127368	-68449	58919
SÜD	Hochland	Einheit A	127368	-68449	58919
SÜD	Tiefland		128195	-62960	65235
SÜD	Tiefland	Einheit B	128195	-62960	65235

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Organisation einfach mit Umsatz

```

SELECT T_YEAR, T_MONTH, O_AREA, O_REGION, O_UNIT,
       sum(F_ORDER_AMOUNT) AS ORDERS,
       sum(F_CHANGE_AMOUNT) AS CHANGES,
       sum(F_PAYED_AMOUNT) AS PAYEDS
  FROM FACTS_SALES, DIM_ORGANISATION, DIM_TIME
 WHERE F_O_ID = O_ID
   AND F_T_ID = T_ID
 GROUP BY T_YEAR, T_MONTH,
          O_AREA, O_REGION, O_UNIT
ORDER BY T_YEAR, T_MONTH,
          O_AREA, O_REGION, O_UNIT;
    
```

T_YEAR	T_MONTH	O_AREA	O_REGION	O_UNIT	ORDERS	CHANGES	PAYEDS
2021	1	Nord	Hinterland	Einheit E	7459	-3393	4066
2021	1	SÜD	Hochland	Einheit A	10578	-5996	4582
2021	1	SÜD	Tiefland	Einheit B	13460	-6479	6981
2021	2	Nord	Hinterland	Einheit E	4142	-1882	2260
2021	2	SÜD	Hochland	Einheit A	10425	-4730	5695
2021	2	SÜD	Tiefland	Einheit B	10772	-5529	5243
2021	3	Nord	Hinterland	Einheit E	9257	-4876	4381
2021	3	SÜD	Hochland	Einheit A	6263	-3735	2528
2021	3	SÜD	Tiefland	Einheit B	10892	-5723	5169
2021	4	Nord	Hinterland	Einheit E	6064	-3187	2877
2021	4	SÜD	Hochland	Einheit A	6761	-2962	3799
2021	4	SÜD	Tiefland	Einheit B	9314	-4932	4382
2021	5	Nord	Hinterland	Einheit E	6845	-3158	3687
2021	5	SÜD	Hochland	Einheit A	6980	-4227	2753
2021	5	SÜD	Tiefland	Einheit B	13334	-7399	5935
2021	6	Nord	Hinterland	Einheit E	4134	-2001	2133
2021	6	SÜD	Hochland	Einheit A	16855	-11023	5832
2021	6	SÜD	Tiefland	Einheit B	13863	-7022	6841
2021	7	Nord	Hinterland	Einheit E	5852	-3371	2481
2021	7	SÜD	Hochland	Einheit A	12553	-6080	6473
2021	7	SÜD	Tiefland	Einheit B	9028	-4608	4420
2021	8	Nord	Hinterland	Einheit E	5294	-2542	2752
2021	8	SÜD	Hochland	Einheit A	17510	-9965	7545
2021	8	SÜD	Tiefland	Einheit B	7392	-3436	3956
2021	9	Nord	Hinterland	Einheit E	7658	-3446	4212
2021	9	SÜD	Hochland	Einheit A	6883	-4119	2764
2021	9	SÜD	Tiefland	Einheit B	7799	-3059	4740
2021	10	Nord	Hinterland	Einheit E	9861	-4778	5083
2021	10	SÜD	Hochland	Einheit A	11859	-5909	5950
2021	10	SÜD	Tiefland	Einheit B	11628	-4849	6779
2021	11	Nord	Hinterland	Einheit E	7867	-4280	3587
2021	11	SÜD	Hochland	Einheit A	10080	-4585	5495
2021	11	SÜD	Tiefland	Einheit B	7238	-3672	3566

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Kunden mit relativen Rabatt

```

SELECT T_YEAR, T_MONTH, C_TOPLVEL, C_NAME,
       sum(F_RECEIPTS) AS RECEIPTS,
       sum(F_DISCOUNT) AS DISCOUNT,
       format(sum(F_DISCOUNT)/sum(F_RECEIPTS), "Percent") AS DISCOUNT_REL
FROM FACTS_SALES,
     DIM_CUSTOMERS,
     DIM_TIME
WHERE F_C_ID = C_ID
  AND F_T_ID = T_ID
GROUP BY T_YEAR, T_MONTH,
         C_TOPLVEL, C_NAME
ORDER BY T_YEAR, T_MONTH,
         C_TOPLVEL, C_NAME;
    
```

T_YEAR	T_MONTH	C_TOPLVEL	C_NAME	RECEIPTS	DISCOUNT	DISCOUNT_REL
2021	1	10	Kunde B	26.218,67 €	17,98 €	0,07%
2021	1	100	Kunde E	9.129,60 €	6,75 €	0,07%
2021	1	1000	Kunde A	9.291,71 €	10,72 €	0,12%
2021	1	1000	Kunde C	21.542,24 €	15,65 €	0,07%
2021	1	1000	Kunde F	16.194,17 €	10,75 €	0,07%
2021	1	1000	Kunde G	21.415,26 €	17,65 €	0,08%
2021	2	10	Kunde B	40.640,08 €	19,53 €	0,05%
2021	2	100	Kunde E	7.560,90 €	5,01 €	0,07%
2021	2	1000	Kunde A	45.512,41 €	17,99 €	0,04%
2021	2	1000	Kunde C	23.053,06 €	9,15 €	0,04%
2021	2	1000	Kunde F	17.535,30 €	13,89 €	0,08%
2021	2	1000	Kunde G	19.127,95 €	14,73 €	0,08%
2021	3	10	Kunde B	11.621,86 €	14,97 €	0,13%
2021	3	100	Kunde E	17.569,58 €	11,15 €	0,06%
2021	3	1000	Kunde A	9.553,82 €	6,80 €	0,07%
2021	3	1000	Kunde C	31.127,42 €	9,30 €	0,03%
2021	3	1000	Kunde F	9.297,95 €	8,64 €	0,09%
2021	3	1000	Kunde G	49.561,96 €	23,10 €	0,05%
2021	4	10	Kunde B	21.733,92 €	12,88 €	0,06%
2021	4	100	Kunde E	24.468,76 €	9,40 €	0,04%
2021	4	1000	Kunde A	12.393,42 €	8,11 €	0,07%
2021	4	1000	Kunde C	4.552,49 €	8,98 €	0,20%
2021	4	1000	Kunde F	24.345,23 €	4,54 €	0,02%
2021	4	1000	Kunde G	18.535,08 €	13,12 €	0,07%
2021	5	10	Kunde B	20.064,33 €	14,58 €	0,07%
2021	5	100	Kunde E	12.745,59 €	2,58 €	0,02%
2021	5	1000	Kunde A	6.143,49 €	8,02 €	0,13%
2021	5	1000	Kunde C	32.668,33 €	14,00 €	0,04%
2021	5	1000	Kunde F	22.656,11 €	14,25 €	0,06%
2021	5	1000	Kunde G	39.526,31 €	18,06 €	0,05%
2021	6	10	Kunde B	39.192,51 €	13,30 €	0,03%
2021	6	100	Kunde E	2.395,74 €	0,50 €	0,02%
2021	6	1000	Kunde A	37.790,95 €	13,06 €	0,03%

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Top-Level mit relativen Rabatt

```

SELECT T_YEAR, T_MONTH, C_TOPLVEL, sum(F_RECEIPTS) AS RECEIPTS,
       sum(F_DISCOUNT) AS DISCOUNT,
       format(sum(F_DISCOUNT)/sum(F_RECEIPTS), "Percent") AS DISCOUNT_REL
  FROM FACTS_SALES, DIM_CUSTOMERS, DIM_TIME
 WHERE F_C_ID = C_ID AND F_T_ID = T_ID
 GROUP BY T_YEAR, T_MONTH, C_TOPLVEL
 ORDER BY T_YEAR, T_MONTH, C_TOPLVEL;
    
```

T_YEAR	T_MONTH	C_TOPLVEL	RECEIPTS	DISCOUNT	DISCOUNT_REL
2021	1 10	26.218,67 €	17,98 €	0,07%	
2021	1 100	9.129,60 €	6,75 €	0,07%	
2021	1 1000	68.443,38 €	54,77 €	0,08%	
2021	2 10	40.640,08 €	19,53 €	0,05%	
2021	2 100	7.560,90 €	5,01 €	0,07%	
2021	2 1000	105.228,72 €	55,76 €	0,05%	
2021	3 10	11.621,86 €	14,97 €	0,13%	
2021	3 100	17.569,58 €	11,15 €	0,06%	
2021	3 1000	99.541,15 €	47,84 €	0,05%	
2021	4 10	21.733,92 €	12,88 €	0,06%	
2021	4 100	24.468,76 €	9,40 €	0,04%	
2021	4 1000	59.826,22 €	34,75 €	0,06%	
2021	5 10	20.064,33 €	14,58 €	0,07%	
2021	5 100	12.745,59 €	2,58 €	0,02%	
2021	5 1000	100.994,24 €	54,33 €	0,05%	
2021	6 10	39.192,51 €	13,30 €	0,03%	
2021	6 100	2.395,74 €	0,50 €	0,02%	
2021	6 1000	115.790,22 €	48,06 €	0,04%	
2021	7 10	30.049,49 €	13,01 €	0,04%	
2021	7 100	3.504,67 €	0,28 €	0,01%	
2021	7 1000	76.739,52 €	47,20 €	0,06%	
2021	8 10	54.725,75 €	20,65 €	0,04%	
2021	8 100	3.136,39 €	7,70 €	0,25%	
2021	8 1000	93.629,10 €	49,89 €	0,05%	
2021	9 10	15.268,61 €	14,18 €	0,09%	
2021	9 100	7.195,01 €	8,51 €	0,12%	
2021	9 1000	86.978,51 €	50,36 €	0,06%	
2021	10 10	20.611,58 €	16,46 €	0,08%	
2021	10 100	858,12 €	0,08 €	0,01%	
2021	10 1000	121.107,52 €	48,17 €	0,04%	
2021	11 10	36.416,26 €	15,41 €	0,04%	
2021	11 100	1.158,14 €	1,73 €	0,15%	
2021	11 1000	65.052,13 €	46,44 €	0,07%	

Datensatz: 14 | 36 von 36 | Kein Filter Suchen

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Produkte mit Umsatz

```
SELECT T_YEAR, T_MONTH, P_LINE, P_NAME, sum(F_RECEIPTS) AS RECEIPTS
FROM FACTS_SALES, DIM_PRODUCTS, DIM_TIME
WHERE F_P_ID = P_ID AND F_T_ID = T_ID
GROUP BY T_YEAR, T_MONTH, P_LINE, P_NAME
ORDER BY T_YEAR, T_MONTH, P_LINE, P_NAME;
```

Erste Version des Produkts für alle Ergebnisse

T_YEAR	T_MONTH	P_LINE	P_NAME	RECEIPTS
2021		1 Buch	Per Mitfahrer d	23.436,59 €
2021		1 Film	Per Anhalter d	39.329,83 €
2021		1 Serie	Per Anhalter d	41.025,23 €
2021		2 Buch	Per Mitfahrer d	77.673,13 €
2021		2 Film	Per Anhalter d	26.418,71 €
2021		2 Serie	Per Anhalter d	49.337,86 €
2021		3 Buch	Per Mitfahrer d	34.749,35 €
2021		3 Film	Per Anhalter d	59.709,07 €
2021		3 Serie	Per Anhalter d	34.274,17 €
2021		4 Buch	Per Mitfahrer d	25.951,05 €
2021		4 Film	Per Anhalter d	30.406,75 €
2021		4 Serie	Per Anhalter d	49.671,10 €
2021		5 Buch	Per Mitfahrer d	30.160,68 €
2021		5 Film	Per Anhalter d	50.217,61 €
2021		5 Serie	Per Anhalter d	53.425,87 €
2021		6 Buch	Per Mitfahrer d	67.168,63 €
2021		6 Film	Per Anhalter d	30.868,20 €
2021		6 Serie	Per Anhalter d	59.341,64 €
2021		7 Buch	Per Mitfahrer d	36.018,30 €
2021		7 Film	Per Anhalter d	48.751,74 €
2021		7 Serie	Per Anhalter d	25.523,64 €
2021		8 Buch	Per Mitfahrer d	44.125,10 €
2021		8 Film	Per Anhalter d	67.910,65 €
2021		8 Serie	Per Anhalter d	39.455,49 €
2021		9 Buch	Per Mitfahrer d	31.229,19 €
2021		9 Film	Per Anhalter d	49.756,73 €
2021		9 Serie	Per Anhalter d	28.456,21 €
2021		10 Buch	Per Mitfahrer d	45.353,81 €
2021		10 Film	Per Anhalter d	52.726,89 €
2021		10 Serie	Per Anhalter d	44.496,52 €
2021		11 Buch	Per Mitfahrer d	43.794,08 €
2021		11 Film	Per Anhalter d	30.469,70 €
2021		11 Serie	Per Anhalter d	28.362,75 €

Datensatz: 14 36 von 36 > >> Kein Filter Suchen

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Produkte, unschöne Temporalisierung

```
SELECT T_YEAR, P_LINE, P_NAME, sum(F_RECEIPTS) AS RECEIPTS  
FROM FACTS_SALES, DIM_PRODUCTS, DIM_TIME  
WHERE F_P_ID = P_ID AND F_T_ID = T_ID  
GROUP BY T_YEAR, P_LINE, P_NAME  
ORDER BY T_YEAR, P_LINE, P_NAME;
```

Erste Version des Produkt mit Fehler

T_YEAR	P_LINE	P_NAME	RECEIPTS
2021	Buch	Per Mitfahrer durch die Galaxis	481.374,52 €
2021	Film	Per Anhalter durch die Galaxis	585.152,38 €
2021	Serie	Per Anhalter durch die Galaxis	490.397,96 €

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Produkte mit Temporalisierung

```
SELECT T_YEAR, P_LINE, P_NAME, sum(F_RECEIPTS) AS RECEIPTS  
FROM FACTS_SALES, DIM_PRODUCTS, DIM_TIME  
WHERE F_P_ID = P_MASTER_ID  
    AND P_VALID_FROM <= #11/3/2021# AND P_VALID_UPTO > #11/3/2021#  
    AND F_T_ID = T_ID  
GROUP BY T_YEAR, P_LINE, P_NAME  
ORDER BY T_YEAR, P_LINE, P_NAME;
```

Produkt Version an einem bestimmten Tag

Korrigierte Version des Produkts

T_YEAR	P_LINE	P_NAME	RECEIPTS
2021	Buch	Per Anhalter durch die Galaxis	481.374,52 €
2021	Film	Per Anhalter durch die Galaxis	585.152,38 €
2021	Serie	Per Anhalter durch die Galaxis	490.397,96 €

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Produkte mit Temporalisierung

```

SELECT T_YEAR, T_MONTH, P_LINE, P_NAME, sum(F_RECEIPTS) AS RECEIPTS
FROM FACTS_SALES, DIM_PRODUCTS, DIM_TIME
WHERE F_P_ID = P_MASTER_ID
    AND P_VALID_FROM <= DateSerial(T_YEAR, T_MONTH, T_DAY)
    AND P_VALID_UPTO > DateSerial(T_YEAR, T_MONTH, T_DAY)
    AND F_T_ID = T_ID
GROUP BY T_YEAR, T_MONTH, P_LINE, P_NAME
ORDER BY T_YEAR, T_MONTH, P_LINE, P_NAME;

```

Produkt Version am Umsatztag

Im Mai gab es zwei Versionen

T_YEAR	T_MONTH	P_LINE	P_NAME	RECEIPTS
2021	1	Buch	Per Mitfahrer durch die Galaxis	23.436,59 €
2021	1	Film	Per Anhalter durch die Galaxis	39.329,83 €
2021	1	Serie	Per Anhalter durch die Galaxis	41.025,23 €
2021	2	Buch	Per Mitfahrer durch die Galaxis	77.673,13 €
2021	2	Film	Per Anhalter durch die Galaxis	26.418,71 €
2021	2	Serie	Per Anhalter durch die Galaxis	49.337,86 €
2021	3	Buch	Per Mitfahrer durch die Galaxis	34.749,35 €
2021	3	Film	Per Anhalter durch die Galaxis	59.709,07 €
2021	3	Serie	Per Anhalter durch die Galaxis	34.274,17 €
2021	4	Buch	Per Mitfahrer durch die Galaxis	25.951,05 €
2021	4	Film	Per Anhalter durch die Galaxis	30.406,75 €
2021	4	Serie	Per Anhalter durch die Galaxis	49.671,10 €
2021	5	Buch	Per Anhalter durch die Galaxis	29.311,75 €
2021	5	Film	Per Mitfahrer durch die Galaxis	848,93 €
2021	5	Serie	Per Anhalter durch die Galaxis	50.217,61 €
2021	6	Buch	Per Anhalter durch die Galaxis	67.168,63 €
2021	6	Film	Per Anhalter durch die Galaxis	30.868,20 €
2021	6	Serie	Per Anhalter durch die Galaxis	59.341,64 €
2021	7	Buch	Per Anhalter durch die Galaxis	36.018,30 €
2021	7	Film	Per Anhalter durch die Galaxis	48.751,74 €
2021	7	Serie	Per Anhalter durch die Galaxis	25.523,64 €
2021	8	Buch	Per Anhalter durch die Galaxis	44.125,10 €
2021	8	Film	Per Anhalter durch die Galaxis	67.910,65 €
2021	8	Serie	Per Anhalter durch die Galaxis	39.455,49 €
2021	9	Buch	Per Anhalter durch die Galaxis	31.229,19 €
2021	9	Film	Per Anhalter durch die Galaxis	49.756,73 €
2021	9	Serie	Per Anhalter durch die Galaxis	28.456,21 €
2021	10	Buch	Per Anhalter durch die Galaxis	45.353,81 €
2021	10	Film	Per Anhalter durch die Galaxis	52.726,89 €
2021	10	Serie	Per Anhalter durch die Galaxis	44.496,52 €
2021	11	Buch	Per Anhalter durch die Galaxis	43.794,08 €
2021	11	Film	Per Anhalter durch die Galaxis	30.469,70 €

Datensatz: 1 13 von 37 Kein Filter Suchen

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Produkte mit Temporalisierung

```

SELECT T_YEAR, T_MONTH, P_LINE, P_NAME, sum(F_RECEIPTS) AS RECEIPTS
FROM FACTS_SALES, DIM_PRODUCTS, DIM_TIME
WHERE F_P_ID = P_MASTER_ID
    AND P_VALID_FROM <= DateSerial(T_YEAR, T_MONTH, 1)
    AND P_VALID_UPTO > DateSerial(T_YEAR, T_MONTH, 1)
    AND F_T_ID = T_ID
GROUP BY T_YEAR, T_MONTH, P_LINE, P_NAME
ORDER BY T_YEAR, T_MONTH, P_LINE, P_NAME;

```

Produkt Version am 1. eines Monats

Korrigierte Version erst am 3. Mai

T_YEAR	T_MONTH	P_LINE	P_NAME	RECEIPTS
2021	1 Buch	Per Mitfahrer durch die Galaxis	23.436,59 €	
2021	1 Film	Per Anhalter durch die Galaxis	39.329,83 €	
2021	1 Serie	Per Anhalter durch die Galaxis	41.025,23 €	
2021	2 Buch	Per Mitfahrer durch die Galaxis	77.673,13 €	
2021	2 Film	Per Anhalter durch die Galaxis	26.418,71 €	
2021	2 Serie	Per Anhalter durch die Galaxis	49.337,86 €	
2021	3 Buch	Per Mitfahrer durch die Galaxis	34.749,35 €	
2021	3 Film	Per Anhalter durch die Galaxis	59.709,07 €	
2021	3 Serie	Per Anhalter durch die Galaxis	34.274,17 €	
2021	4 Buch	Per Mitfahrer durch die Galaxis	25.951,05 €	
2021	4 Film	Per Anhalter durch die Galaxis	30.406,75 €	
2021	4 Serie	Per Anhalter durch die Galaxis	49.671,10 €	
2021	5 Buch	Per Mitfahrer durch die Galaxis	30.160,68 €	
2021	5 Film	Per Anhalter durch die Galaxis	50.217,61 €	
2021	5 Serie	Per Anhalter durch die Galaxis	53.425,87 €	
2021	6 Buch	Per Anhalter durch die Galaxis	67.168,63 €	
2021	6 Film	Per Anhalter durch die Galaxis	30.868,20 €	
2021	6 Serie	Per Anhalter durch die Galaxis	59.341,64 €	
2021	7 Buch	Per Anhalter durch die Galaxis	36.018,30 €	
2021	7 Film	Per Anhalter durch die Galaxis	48.751,74 €	
2021	7 Serie	Per Anhalter durch die Galaxis	25.523,64 €	
2021	8 Buch	Per Anhalter durch die Galaxis	44.125,10 €	
2021	8 Film	Per Anhalter durch die Galaxis	67.910,65 €	
2021	8 Serie	Per Anhalter durch die Galaxis	39.455,49 €	
2021	9 Buch	Per Anhalter durch die Galaxis	31.229,19 €	
2021	9 Film	Per Anhalter durch die Galaxis	49.756,73 €	
2021	9 Serie	Per Anhalter durch die Galaxis	28.456,21 €	
2021	10 Buch	Per Anhalter durch die Galaxis	45.353,81 €	
2021	10 Film	Per Anhalter durch die Galaxis	52.726,89 €	
2021	10 Serie	Per Anhalter durch die Galaxis	44.496,52 €	
2021	11 Buch	Per Anhalter durch die Galaxis	43.794,08 €	
2021	11 Film	Per Anhalter durch die Galaxis	30.469,70 €	
2021	11 Serie	Per Anhalter durch die Galaxis	28.362,75 €	

Datensatz: 1 4 36 von 36 > >> Kein Filter Suchen

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten Würfel an Vertriebs temporalisiert

FACTS_SALES_TEMPO	
F_ID	
F_T_ID	
F_C_ID	
F_P_ID	
F_O_ID	
F_ORDER_AMOUNT	
F_CHANGE_AMOUNT	
F_PAYED_AMOUNT	
F_RECEIPTS	
F_DISCOUNT	
F_PRICE	
F_VALID_FROM	
F_VALID_UPTO	

Feldname	Felddatentyp
F_ID	Zahl
F_T_ID	Zahl
F_C_ID	Zahl
F_P_ID	Zahl
F_O_ID	Zahl
F_ORDER_AMOUNT	Zahl
F_CHANGE_AMOUNT	Zahl
F_PAYED_AMOUNT	Zahl
F_RECEIPTS	Währung
F_DISCOUNT	Währung
F_PRICE	Währung
F_VALID_FROM	Datum/Uhrzeit
F_VALID_UPTO	Datum/Uhrzeit

F_ID	F_T_ID	F_C_ID	F_P_ID	F_O_ID	F_ORDER_A	F_CHANGE_A	F_PAYED_AI	F_RECEIPTS	F_DISCOUNT	F_PRICE	F_VALID_FR	F_VALID_UF
1147	44355	1002	1000	1001	148	-103	45	6.484,19 €	2,58 €	17,22 €	01.01.2021	01.01.2999
1240	44356	1003	1000	1000	895	-705	190	2.983,50 €	0,82 €	13,74 €	01.01.2021	01.01.2999
1529	44356	1005	1001	1000	370	-223	147	140,68 €	0,50 €	3,31 €	01.01.2021	01.01.2999
1382	44357	1000	1001	1004	293	-250	43	869,16 €	0,98 €	6,97 €	01.01.2021	01.01.2999
1148	44357	1002	1002	1001	636	-112	524	169,20 €	1,49 €	11,44 €	01.01.2021	01.01.2999
1383	44358	1000	1000	1004	461	-332	129	847,34 €	0,14 €	2,77 €	01.01.2021	01.01.2999
1041	44359	1001	1002	1001	966	-375	591	775,75 €	1,40 €	6,68 €	01.01.2021	01.01.2999
1241	44359	1003	1000	1000	946	-936	10	11,19 €	0,24 €	3,04 €	01.01.2021	01.01.2999
1530	44359	1005	1002	1000	809	-661	148	1.280,97 €	0,37 €	18,41 €	01.01.2021	01.01.2999
1384	44360	1000	1000	1004	527	-368	159	339,02 €	0,14 €	7,06 €	01.01.2021	01.01.2999
1042	44360	1001	1001	1001	616	-36	580	2.646,40 €	1,23 €	7,23 €	01.01.2021	01.01.2999
1149	44360	1002	1000	1001	981	-850	131	8.035,82 €	2,34 €	17,97 €	01.01.2021	20.06.2021
1700	44360	1002	1000	1001	2000	0	2000	35.940,00 €	0,00 €	17,97 €	20.06.2021	01.01.2999
1385	44361	1000	1000	1004	122	-51	71	2.383,83 €	0,00 €	7,29 €	01.01.2021	01.01.2999
1043	44361	1001	1001	1001	825	-503	322	506,09 €	0,14 €	1,42 €	01.01.2021	01.01.2999
1150	44361	1002	1000	1001	898	-479	419	1.243,44 €	0,22 €	11,13 €	01.01.2021	01.01.2999
1386	44362	1000	1001	1004	463	-93	370	11,75 €	1,16 €	12,91 €	01.01.2021	01.01.2999
1044	44363	1001	1001	1001	871	-681	190	8,70 €	0,61 €	2,79 €	01.01.2021	01.01.2999
1242	44363	1003	1001	1000	735	-590	145	2.106,91 €	0,05 €	4,87 €	01.01.2021	01.01.2999
1531	44363	1005	1002	1000	629	-478	151	2.320,24 €	0,39 €	13,00 €	01.01.2021	01.01.2999
1532	44364	1005	1000	1000	857	-835	22	3.956,58 €	2,85 €	15,82 €	01.01.2021	01.01.2999
1243	44365	1003	1002	1000	883	-738	145	1.620,82 €	0,99 €	12,32 €	01.01.2021	01.01.2999
1533	44365	1005	1002	1000	367	-61	306	3.396,56 €	1,66 €	7,56 €	01.01.2021	01.01.2999
1045	44366	1001	1000	1001	930	-291	639	384,52 €	0,09 €	4,51 €	01.01.2021	01.01.2999
1244	44366	1003	1000	1000	493	-304	189	3.015,34 €	0,20 €	6,60 €	01.01.2021	01.01.2999
1534	44366	1005	1000	1000	362	-113	249	375,56 €	0,46 €	2,29 €	01.01.2021	01.01.2999
1046	44367	1001	1001	1001	860	-851	9	11,93 €	0,22 €	2,21 €	01.01.2021	01.01.2999
1245	44367	1003	1002	1000	895	-748	147	2.563,16 €	0,52 €	5,15 €	01.01.2021	01.01.2999
1535	44367	1005	1000	1000	560	-20	540	3.312,51 €	1,03 €	5,73 €	01.01.2021	01.01.2999
1387	44368	1000	1001	1004	123	-43	80	5.173,28 €	0,59 €	19,68 €	01.01.2021	01.01.2999
1246	44369	1003	1001	1000	475	-143	332	2.551,19 €	0,50 €	4,55 €	01.01.2021	01.01.2999
1536	44369	1005	1001	1000	303	-81	222	504,10 €	0,71 €	7,09 €	01.01.2021	01.01.2999
1388	44370	1000	1000	1004	132	-8	124	1.041,66 €	1,17 €	10,64 €	01.01.2021	01.01.2999
1047	44370	1001	1000	1001	444	-82	362	628,08 €	0,57 €	3,77 €	01.01.2021	01.01.2999
1151	44370	1002	1002	1001	156	-82	74	301,22 €	0,13 €	2,21 €	01.01.2021	01.01.2999
1300	44371	1000	1000	1004	446	0	420	3.476,34 €	0,63 €	6,20 €	01.01.2021	01.01.2999

# OLAP Schema Demo-Abfrage

## Vertrieb Star-Schema: Fakten und Organisation dreistufig

```

SELECT * FROM
(
  SELECT O_AREA, '' AS O_REGION, '' AS O_UNIT, sum(F_ORDER_AMOUNT) AS ORDERS,
         sum(F_CHANGE_AMOUNT) AS CHANGES, sum(F_PAYED_AMOUNT) AS PAYEDS
    FROM FACTS_SALES_TEMPO, DIM_ORGANISATION
   WHERE F_O_ID = O_ID AND F_VALID_UPTO = #01/01/2999#
  GROUP BY O_AREA
UNION
  SELECT O_AREA, O_REGION, '' AS O_UNIT, sum(F_ORDER_AMOUNT) AS ORDERS,
         sum(F_CHANGE_AMOUNT) AS CHANGES, sum(F_PAYED_AMOUNT) AS PAYEDS
    FROM FACTS_SALES_TEMPO, DIM_ORGANISATION
   WHERE F_O_ID = O_ID AND F_VALID_UPTO = #01/01/2999#
  GROUP BY O_AREA, O_REGION
UNION
  SELECT O_AREA, O_REGION, O_UNIT, sum(F_ORDER_AMOUNT) AS ORDERS,
         sum(F_CHANGE_AMOUNT) AS CHANGES, sum(F_PAYED_AMOUNT) AS PAYEDS
    FROM FACTS_SALES_TEMPO, DIM_ORGANISATION
   WHERE F_O_ID = O_ID AND F_VALID_UPTO = #01/01/2999#
  GROUP BY O_AREA, O_REGION, O_UNIT
)
ORDER BY O_AREA, O_REGION, O_UNIT;
  
```

Nur letzte Version

Nur letzte Version

Nur letzte Version

O_AREA	O_REGION	O_UNIT	ORDERS	CHANGES	PAYEDS
Nord			80833	-40342	40491
Nord	Hinterland		80833	-40342	40491
Nord	Hinterland	Einheit E	80833	-40342	40491
SÜD			256582	-130559	126023
SÜD	Hochland		127368	-68449	58919
SÜD	Hochland	Einheit A	127368	-68449	58919
SÜD	Tiefland		129214	-62110	67104
SÜD	Tiefland	Einheit B	129214	-62110	67104

# Kurs Einführung In das Thema Data Ware House & Business Intelligence

5. Kapitel 5: Analyse und Marktübersicht

# OLAP Analysen

## Meistgenutzte Formen der Datenanalyse

- Formen der Datenanalyse
  - Dashboard
  - SQL
  - MDX
  - R, SPSS, Jupyter, ...
  - Proprietär Abfragesprache

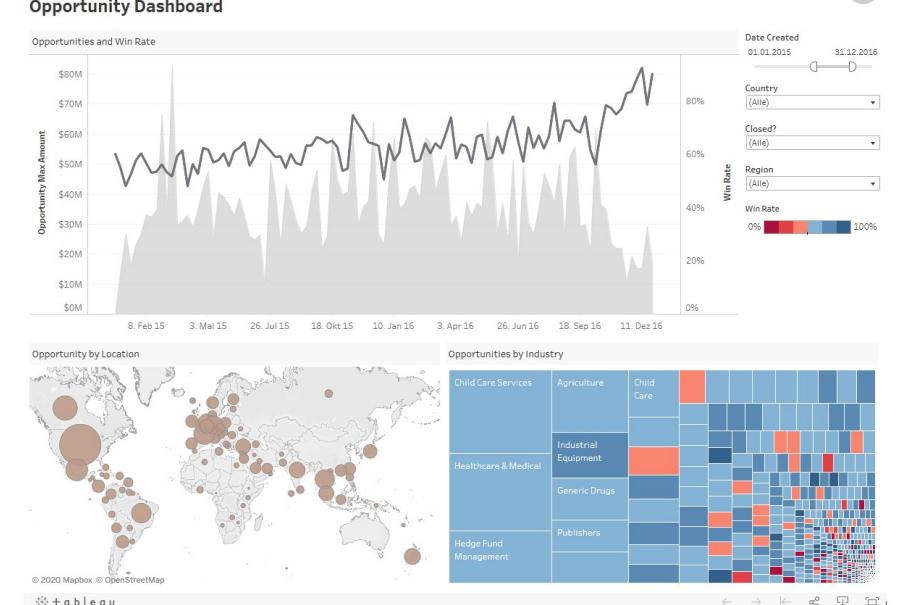
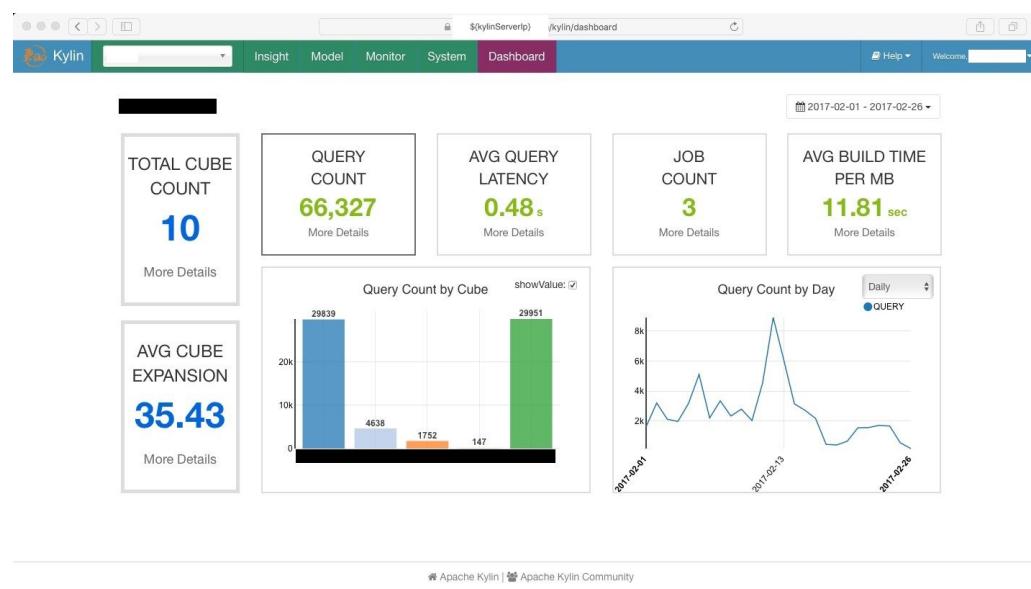
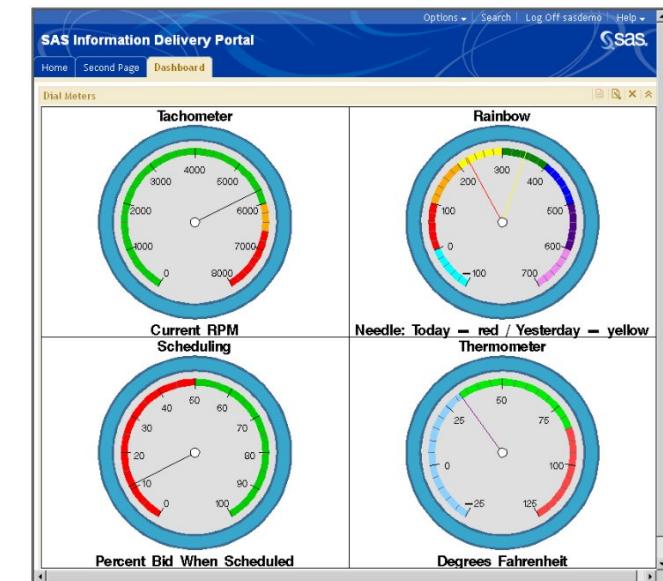
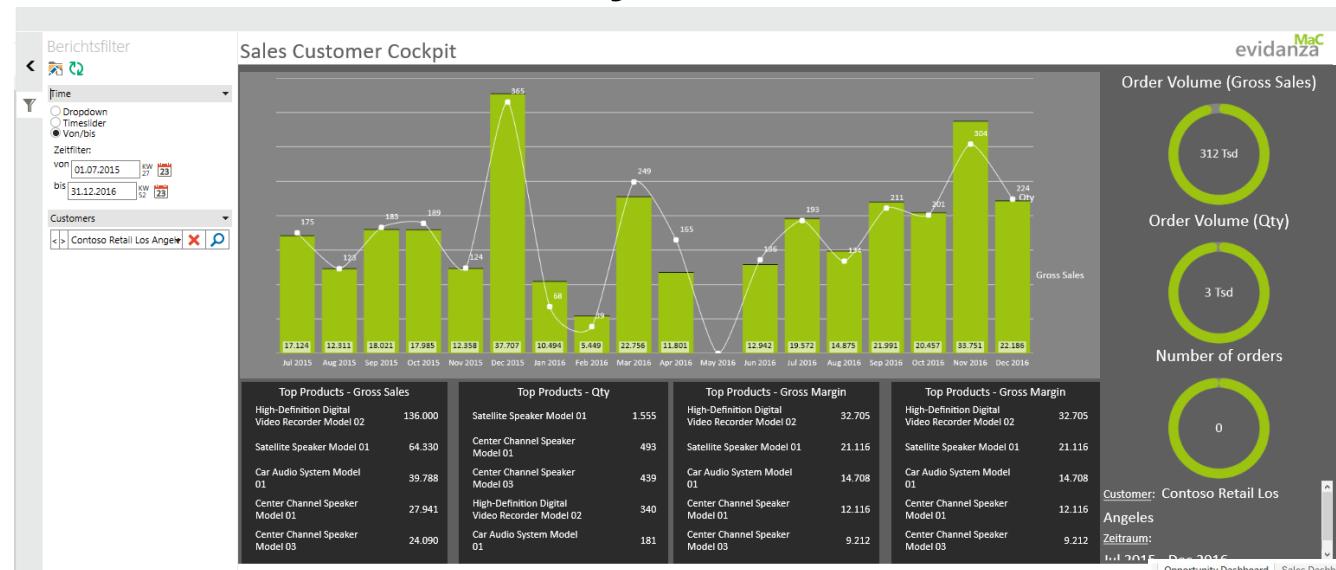
# OLAP Analysen

## Form der Datenanalyse: Dashboard

- Grafische Darstellung von Analyseergebnissen
- Einstellbare unterschiedlich Darstellungsformen
- Meist Basisabfrage auf hohem Aggregationsgrad eingestellt
- Navigation mittels Drill-Down
- Aufgabe komplexe Inhalte vereinfacht Darzustellen
- Entscheider taugliches Informationsmanagement
- Dynamische Darstellung mit Focus auf Warnungen möglich
- Exportmöglichkeiten
- Integration in andere Tools, z.B. Excel, PowerPoint, ...
- Pufferung von Daten bis zur Offline Lösung

# OLAP Analysen

# Form der Datenanalyse: Dashboard



# OLAP Analysen

## Produkt der Datenanalyse: AWS Services

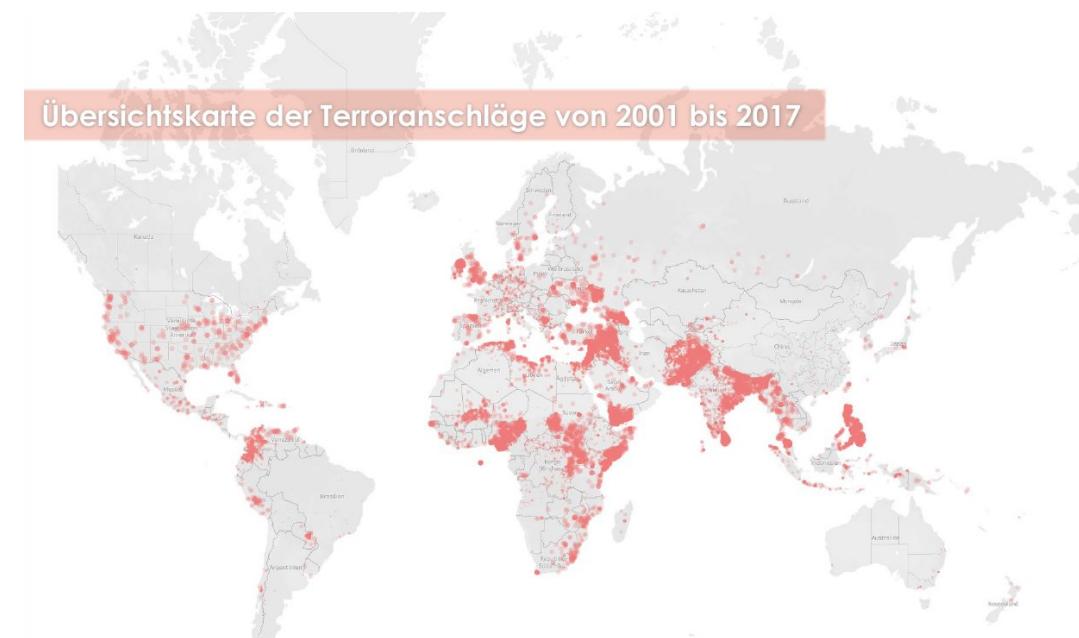
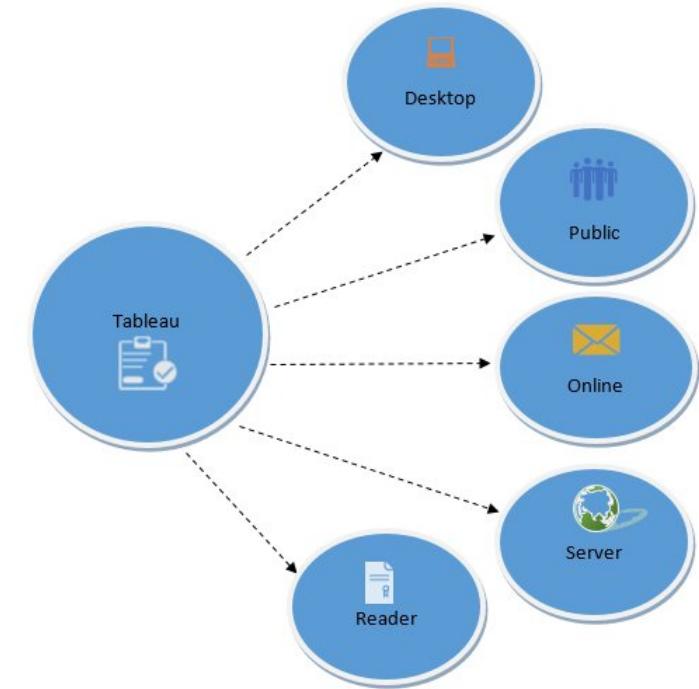
- Dienste sind in der Cloud
- Dienste können u.a. per WebServices angebunden werden
- Auch KI-Analyse Tools verfügbar



# OLAP Analysen

## Produkt der Datenanalyse: Tableau

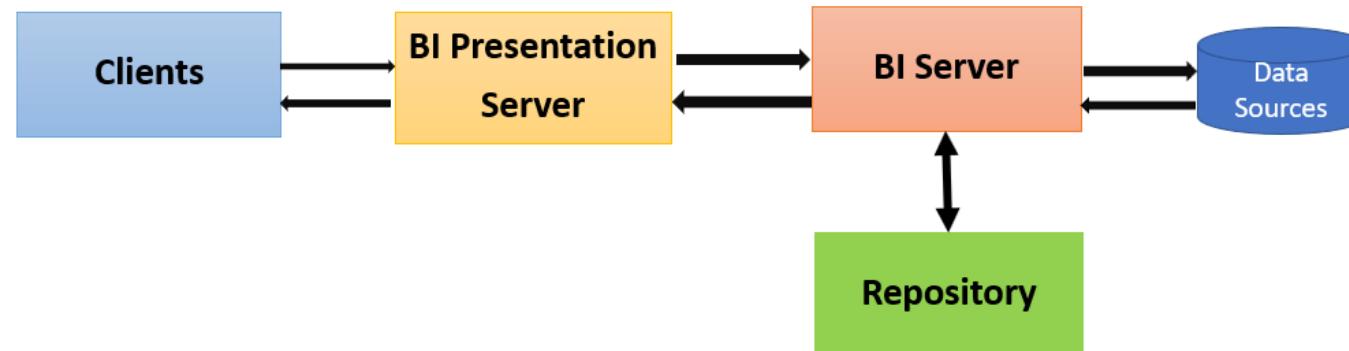
- Spezialisiert zur schnellen Analyse
- Leicht grafische Darstellung möglich
- Umfangreiche Datenverbindungen



# OLAP Analysen

**Produkt der Datenanalyse: OBIEE** (Oracle Business Intelligence Enterprise Edition)

- Abdeckung des vollen OLAP-Systems samt Datenbank
- Alte Komponenten wurden zu einem Produkt umgebaut

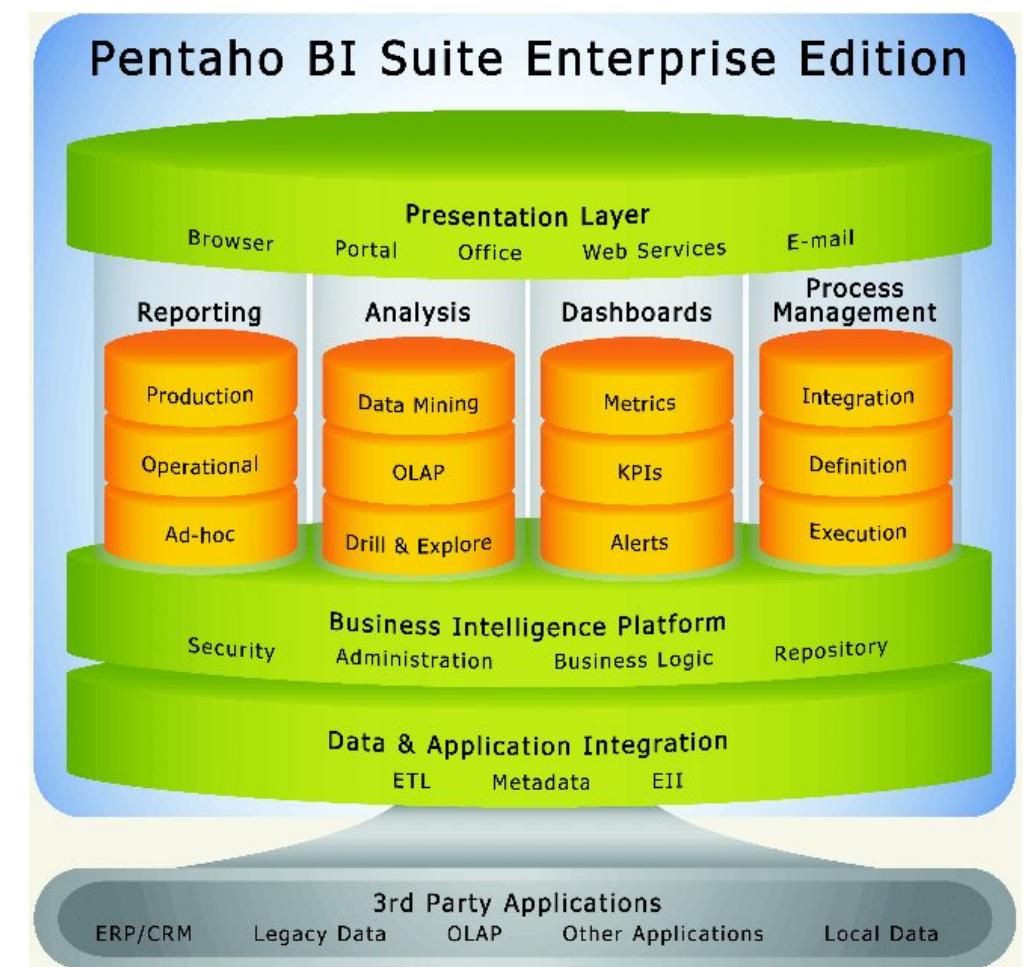


© guru99.com

# OLAP Analysen

## Produkt der Datenanalyse: Hitachi Pentaho BI

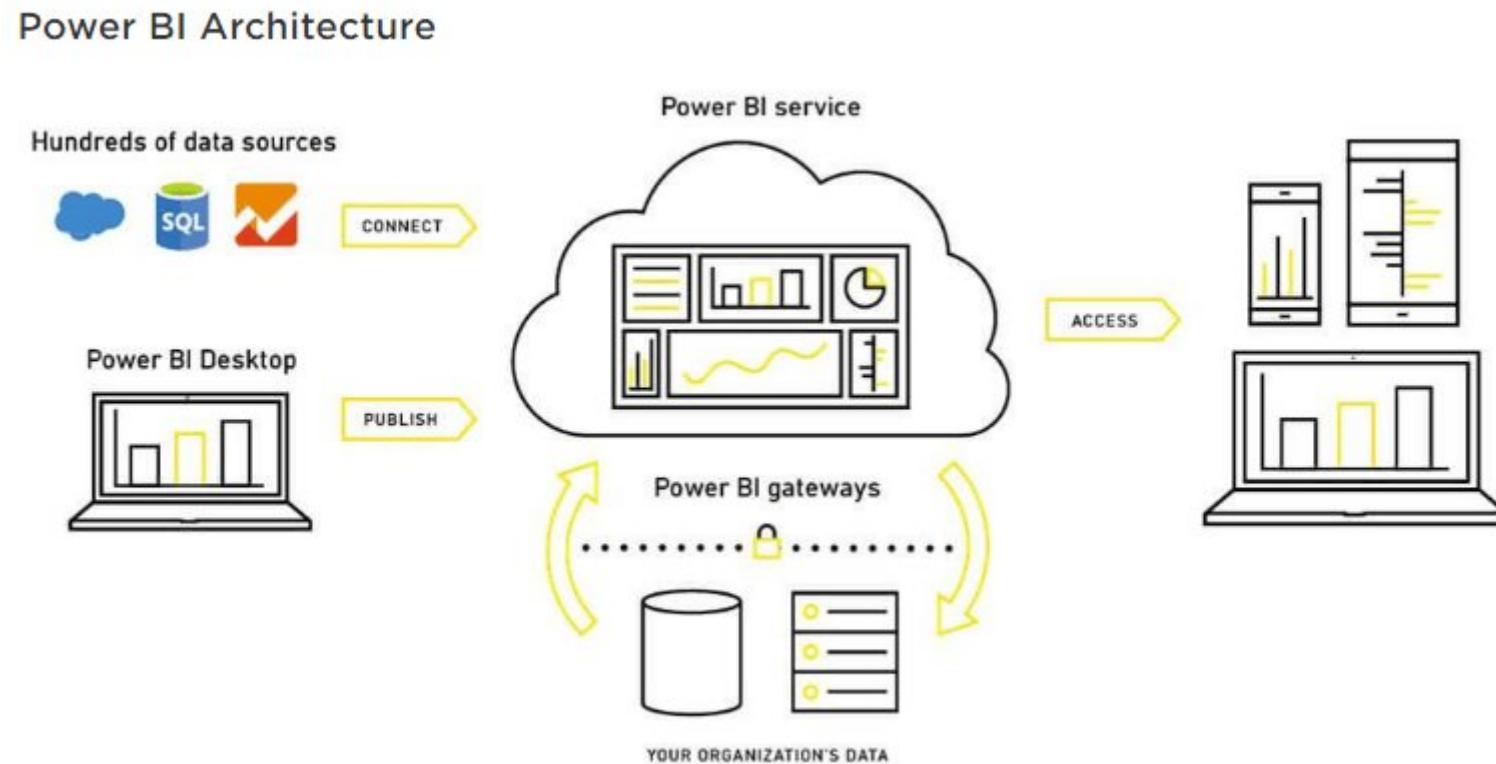
- Abdeckung des vollen OLAP-Systems
- Community Version verfügbar



# OLAP Analysen

## Produkt der Datenanalyse: Microsoft Power BI

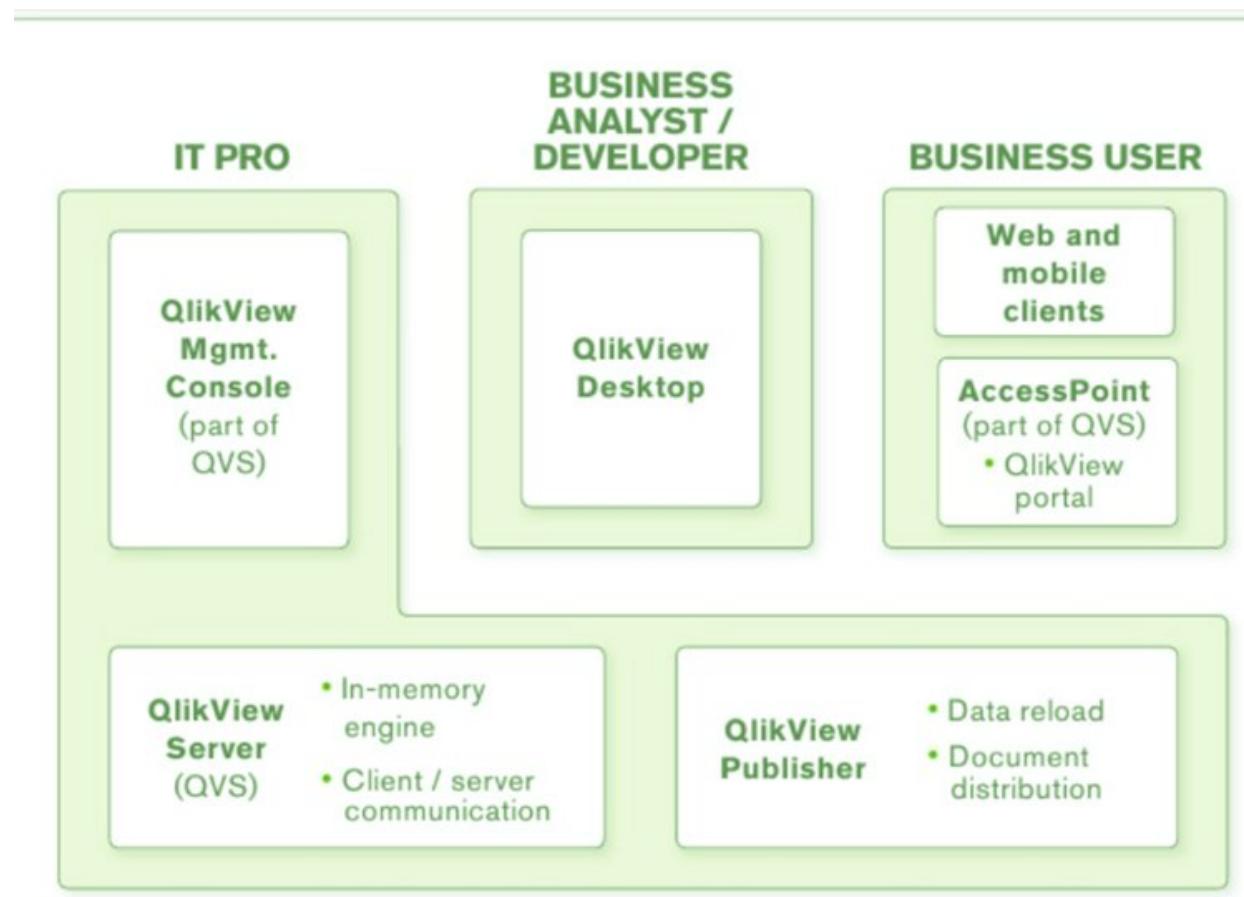
- Abdeckung des vollen OLAP-Systems samt Datenbank
- Auch als AZURE Cloud verfügbar



# OLAP Analysen

## Produkt der Datenanalyse: Qlik

- Ursprünglich reines DOLAP System
- Mit Qlik View Server auch Client-Server tauglich



# OLAP Analysen

## Form der Datenanalyse: SQL

- CLI
- Abfrage Tool mit Listendarstellung
- Hilfe meist über grafische Darstellung des Schemas
- Data Mart übergreifende Analyse möglich
- Direkte Übernahme in Dashboard Elemente
- Verleitung zu proprietären Tool Funktionen im SQL
- Nachteile:
  - Analysen sind nicht Management tauglich
  - Darstellung von mehr als zwei Dimensionen problematisch

# OLAP Analysen

## Form der Datenanalyse: MDX

- Multidimensionale Erweiterung von SQL
  - Entstanden aus Microsofts OLE DB for OLAP API
- Nicht von allen Olap-Server unterstützt
- Analysen sind nicht Management tauglich
- Beispiel:
  - ```
SELECT {[Measures].[Unit Sales], [Measures].[Store Sales]} ON COLUMNS,
{[Product].members} ON ROWS
FROM [Sales]
WHERE [Time].[1997].[Q2]
```
- Integriert in MS Power BI und in Mondrian
- <https://mondrian.pentaho.com/documentation/mdx.php>

# OLAP Analysen

## Form der Datenanalyse: R, SPSS,

- Was ist R?
  - Ein Statistik Projekt mit eigener Entwicklungsumgebung und Laufzeitumgebung
  - Besitzt eine eigene Programmiersprache, alles ist eine Liste
  - Umfangreiche Anbindungen möglich
  - Import und Export in vielen Formaten
  - Berechnungen, Berichterstellung, Grafikerstellung sind integriert
- Was ist SPSS?
  - Statistik Tool aus Uni Stanford, nun IBM
  - Viele Anbindungen und Austauschformate
  - Modularer Aufbau für Erweiterungen
  - Berechnungen, Entscheidungsbäume, Grafikerstellung sind integriert
- Weitere Tools: <https://www.capterra.com/business-intelligence-software/>

# OLAP Analysen

## **Form der Datenanalyse: Proprietär Abfragesprache**

- REST Abfragen
- Webservices mit WSDL und REST
- Json Querys
- Proprietäre APIs
- Excel Integration (z.B. Excel PowerPivot)
- Apache Spark
- DOLAP Tools, z.B. QlikView

# OLAP Marktübersicht

## Marktbetrachtung, Bereiche

- Infrastruktur
  - Datenintegration, Datenqualität, Datenmanagement, Data Streaming, Verteilung
- Datenbankplattformen
  - Speicherkonzepte, Analytische Funktionen, Datenmodellierung
- ETL und Data Integration
  - Prozessteuerung, Transformationsmöglichkeiten
- Olap Analyse Werkzeuge
  - Navigationsmöglichkeiten, Berichtswesen, Analyseverfahren, Planungsfunktionen, Finanzanbindungen, Governance, Riskomanagement, Regelkontrolle, Strategieunterstützung
- Data Mining
  - Lern-Methoden, Cluster und Muster Erkennung, Statistikfunktionen, Betrugserkennung

# OLAP Marktübersicht

## Marktbetrachtung, Übergreifende Kriterien

- Funktionsvielfalt und Erweiterbarkeit
- Sicherheit
- Entwicklungsfähigkeit und Skalierbarkeit
- Betriebsfähigkeit und Stabilität
- Lieferbarkeit
- Verbreitung
- Schnittstellen
- Benutzerfreundlichkeit und Einarbeitungsaufwand
- Ressourcenbedarf
- Arbeitsgeschwindigkeit

# OLAP Marktübersicht

## Marktbetrachtung, Links

→ Ein paar interessante Sammlungen

- <https://www.digital-ratio.de/bi-software/kennen-sie-die-leader-im-aktuellen-gartners-magic-quadrant-v-1-0/>
- <https://www.olap.it/Products.htm>
- <https://www.bi-scout.com/home>
- <https://db-engines.com/de/ranking>
- <http://www.business-intelligence24.com/home>

# OLAP Marktübersicht

## Gartner Quadrat



- [https://de.looker.com/learn/gartner-magic-quadrant?  
utm\\_campaign=7012R000001JGcp&gclid=EAIIaIQobChMIZ4ap\\_oKG6gIVyIeyCh3YBQxHEAAYASAAEgIzavD\\_BwE](https://de.looker.com/learn/gartner-magic-quadrant?utm_campaign=7012R000001JGcp&gclid=EAIIaIQobChMIZ4ap_oKG6gIVyIeyCh3YBQxHEAAYASAAEgIzavD_BwE)



# Beispiele für OLAP Schemen

## Konzeption mit Marktbetrachtung

- Ein Unternehmen des Mittelstands (bis 100\*106 € Umsatz) mit 8 Niederlassungen beabsichtigt ein OLAP System einzuführen. Dabei entstehen aus 5 unterschiedlichen operativen Systemen im jeweiligen Rechenzentrum der Niederlassungen ca. 500 GByte Daten pro Tag. Das Management ist dabei dezentralisiert und entscheidet selbstständig vor Ort. Welche Software und Struktur empfehlen Sie dem Unternehmen. Bitte begründen Sie Ihre Entscheidungen.
- Wie Ihre Empfehlung aussehen wenn das Unternehmen mit nur einem zentralen Management gesteuert wird. Die Zentralisierung geht hierbei mit einer implementierungsstarken Entwicklungsabteilung einher. Begründen Sie auch hier Ihre Entscheidung.
- Sie sollen für ihr Unternehmen ein OLAP System auswählen. Nennen Sie min. 10 Kriterien nach denen das System ausgewählt werden soll. Vergeben Sie eine Gewichtung der Kriterien und Begründen Sie Ihre Gewichtungen.

# Kurs Einführung In das Thema Data Ware House & Business Intelligence

Kapitel 6: Organisation von BI&DWH Projekten

# Organisation von BI&DWH Projekten

## Überblick

- Motive für Business Intelligence: Controlling, Sales, Marketing, Business Development, Top-Management & Aufsichtsrat
- IT in der Defensive oder die Probleme der IT bei Business Intelligence
- Probleme aus den Fachabteilungen
- Unternehmensstrategie und Business Intelligence
- Projektumsetzung
- Erfolgsfaktoren

# Organisation von BI&DWH Projekten

## Motive für Business Intelligence

- Permanente Forecast Beobachtung
- Zeitnahe, qualitativ gesicherte aktuelle Information
- Information auf allen Hierarchieebenen nutzen
- Herausforderung der wachsenden Datenmengen
- Konzentration auf entscheidungsrelevanten Informationen
- Gewinnen neuer Erkenntnisse und Chancen
- Zeitnahe Feedback von Maßnahmen
- Steigerung von Produktivität
- Steigerung Wettbewerbsfähigkeit

# Organisation von BI&DWH Projekten

## Motive für Business Intelligence

- **Controlling:** Die IT schafft es nicht konsolidiertes Material aus den Rohdaten rechtzeitig bereitzustellen
- **Sales:** Aus den Rohdaten lassen unterschiedlichste Ergebnisse erstellen. Die Diskussion über die korrekte Auswertung kostet eine ummenge Zeit
- **Sales:** Mit dem Wissen welche Produkte, wann wie viele verkauft werden könnten wir den Umsatz erheblich steigern
- **Sales:** Alle KPI sind über viele Excel-Dateien verteilt. Geht das nicht einfacher?
- **Marketing:** Für Zielgerichtete Kampagnen brauchen wir min. Tagesaktuelle harmonisierte Datenauswertungen
- **Business Development:** Da die Kollegen unterschiedliche Zahlen aus unterschiedlichen Tools bringen ist ein Produktplanung sehr wage.
- **Aufsichtsrat:** Wieso reagieren unsere Wettbewerber immer vor uns?

# Organisation von BI&DWH Projekten

## Die IT in der Defensive: Erfolgsschwierigkeiten

- Keine Strategie für Business Intelligence
- Fehlende Prio im Management
- Geringe Datenqualität
- Fehlende Stammdaten Data-Management oder Meta-Data-Management
- Politische Konflikte
- Business Intelligence als reines Technik Projekt sehen

# Organisation von BI&DWH Projekten

## Die IT in der Defensive: Widerstandsgründe

- Seit den 70er kündigt die IT aussagekräftige Reports an
- Gewachsene heterogene Datenstrukturen, Datensysteme und Informationsquellen
- Vorstellung Datenauswertung ist intrinsische Aufgabe der IT
- Organisatorische Veränderung eilt der IT immer schneller vorraus  
(vgl. Conways-Law)
- Teufelskreis:  
Heterogen-> Kein Mehrwert-> Geringe Akzeptanz-> Geringe Ressourcen

# Organisation von BI&DWH Projekten

## Die IT in der Defensive: Strukturprobleme

- Tagesgeschäft frißt Innovation (Axt-Dielema)
- Fehlende Ressourcen für Wachstum
- Monopol auf BI-Know-How, Inselwissen des BI in der IT
- Fehlendes Informationsmanagement für Fachabteilungen
- Technische Fokussierung der Lösungen
- IT als Owner der BI Einführung, nicht nur für Systeme und Daten
- Fehlendes Anforderungsmanagement
- Outsourcen der BI Einführung

# Organisation von BI&DWH Projekten

## Die IT in der Defensive: Architektur

- Heterogene Datenschnittstellen
- Fehlende und Instabile Datenschnittstellen
- Veraltete Architekturkonzepte
- Zu hohe Forderung des echtezeit DWH

# Organisation von BI&DWH Projekten

## Probleme aus den Fachbereiche

- By-Pass-Reporting, Reporting direkt auf Rohdaten
  - Fehlende Transparenz
  - Fehlende Vergleichbarkeit
  - Uneinheitliche Kennzahlen
  - Höhere Kosten
  - Maßnahmen schlecht übergreifend ableitbar
- Umgehen von Fehler verhindert die Fehlerbeseitigung
- Politische Machtspiele zwischen den Fachabteilungen

# Organisation von BI&DWH Projekten

## **Unternehmensstrategie und BI**

- Primärziele als Ableitung der Vision als Mission festlegen
- Ziele als BI festlegen in der Unternehmensstrategie festlegen um Mission umzusetzen
- Bedürfnisse der Geschäftstätigkeit ermitteln und BI-Unterstützung ermitteln
- Operative Prozesse mit Analysen unterstützen
- Business Intelligence muß "Single Point of Truth" sein
- Erkenntniss Rückfluß in Operative Systeme

# Organisation von BI&DWH Projekten

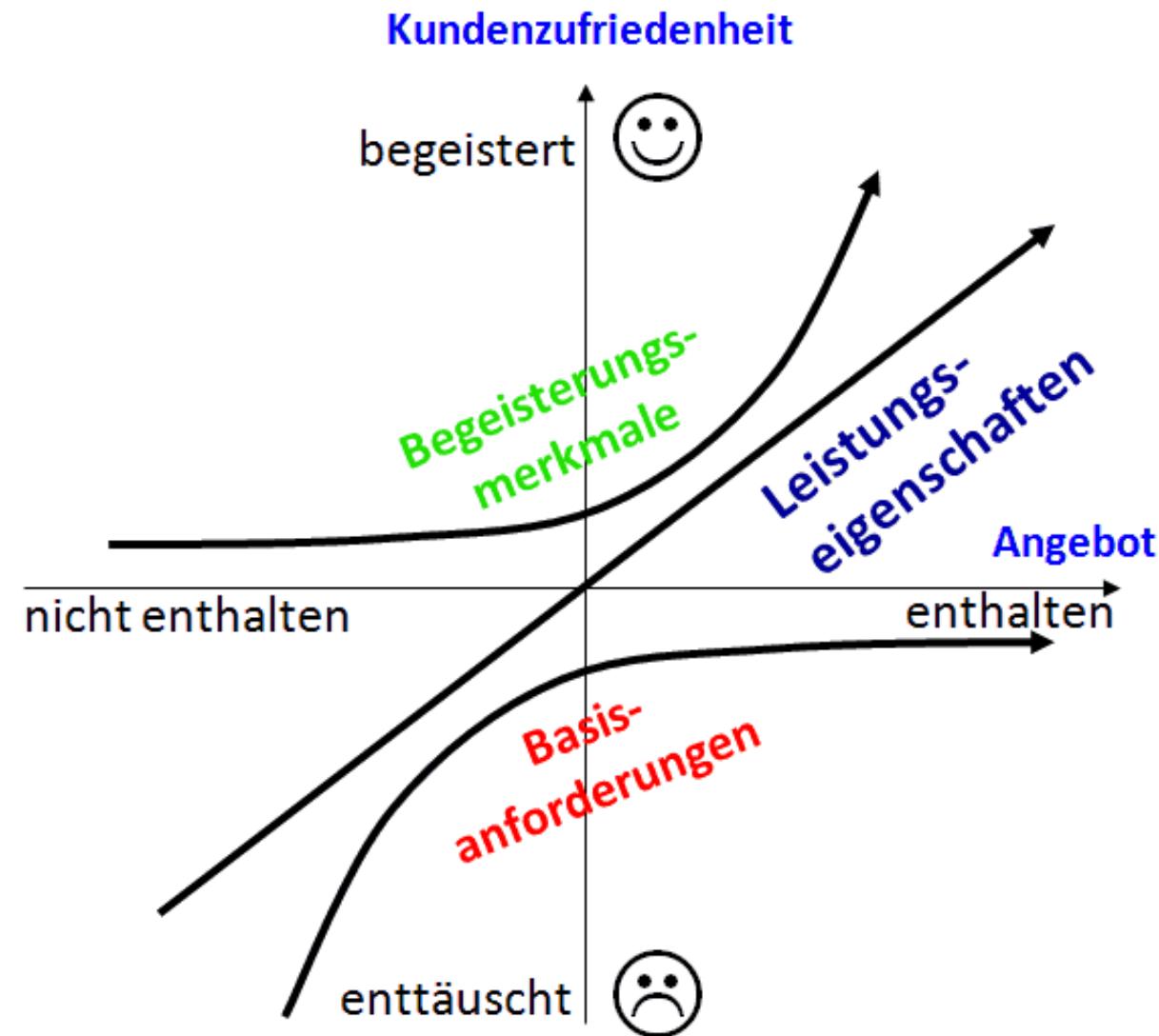
## Projektumsetzung: Vorprojekt

- Sammeln von Informationen über
  - Existierende Prozesse
  - Operative Systeme und deren Schnittstellen
  - Stammdaten
  - Metadaten
  - Existierende Strategien
  - Existirendes Berichtswesen und Kennzahlen
- Bedürfnisse der Geschäftstätigkeit/Wertschöpfungskette ermitteln
- Unterstützung im Top-Management erreichen
- Anforderungen einnsammeln

# Organisation von BI&DWH Projekten

## Projektumsetzung: Vorprojekt, Anforderungen

→ Kano-Modell



# Organisation von BI&DWH Projekten

## **Projektumsetzung: Planung des Projekts**

- Festlegen der Kennzahlen und der Taktung
- Festlegen der Systemanalyse und Architektur
- Planung des Backends
- Planung des Frontends ggf. auch Erprobung von UX Prototypen
- Evaluierung der Meta-, Stamm- und Bewegdaten
- Evaluierung der Bewegdaten bzw. Faktendaten
- Evaluierung der Architektur
- Planung von Dimensionen und Fakten
- Evaluierung der Prozessanpassungen
- Planung der Systemstabilisierung und dem Betriebskonzept
- Planung des Rückluss von Erkenntnisse in den Wertschöpfungsprozess
- Evaluierung der Anforderungserfüllung
- Unterstützung im Top-Management erreichen

# Organisation von BI&DWH Projekten

## **Projektumsetzung: Umsetzung**

- Realisierung der evaluierten Aufgaben
- Umsetzen eines Betriebskonzepts samt Überwachung
- Angehen von Kultur-Veränderungen
- Ablösen der By-Pass Aktionen
- Umsetzen von IT-Governance
- Einführung BI als "Single-Point of Truth"
- Unterstützung im Top-Management erreichen
- Integration in die Unternehmensstrategie erreichen

# Organisation von BI&DWH Projekten

## **Projektumsetzung: Nacharbeiten**

- Kontrolle der Zielerreichung der Einführung
- Kontrolle der dauerhaften Datenqualität
- Nutzen auf allen Unternehmesebenen erreichen
- Nutzen des Rücklusses in die Wertschöpfung sicherstellen
- Big Picture, d.h. Bereichsübergreifender Nutzen erreichen
- Weiterpflege des Systems optimieren
- Training aller Anwender
- Unterstützung im Top-Management erreichen

# Organisation von BI&DWH Projekten

## Erfolgsfaktoren für Business Intelligence

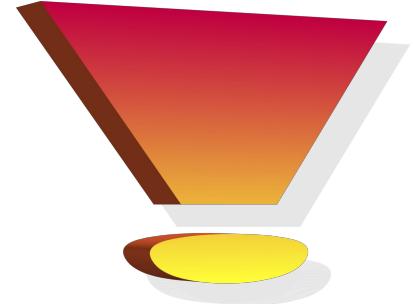
- Integration in die Unternehmensstrategie
- Integration in die Geschäftsmodelle
- Aufbau von Steuerungslogiken
- Organisation und Prosssteuerung
- Aufsetzen von Arbeitsabläufen
- Schafen von Firmenkulterellen Rahmenbedingungen
- Aufbau eines erfolgreichen Qualitätsmanagement für Stamm-, Meta-, und Bewegdaten



# Organisation von BI&DWH Projekten

## Erfolgsfaktoren für Business Intelligence

- Stellen Sie als Projektleiter einen Projektplan auf, um bei einem Kunden ein OLAP-System einzuführen. Der Kunde hat keinerlei Erfahrung im Bereich OLAP und besitzt nur traditionelle Berichte in operativen Systemen erstellt. Ihnen stehen ein Anforderungsmanager, ein ETL-Spezialist und DWH-Spezialist zur Verfügung. Bringen Sie die Schritte in eine Reihenfolge, die Sie begründen können. Beschreiben Sie an welchen Stellen Sie besondere Aufmerksamkeit benötigen.



# Weitere Links

## Weiterführende Informationen

- Zehn SQL-Tricks: <https://youtu.be/mgipNdAgQ3o?feature=shared>