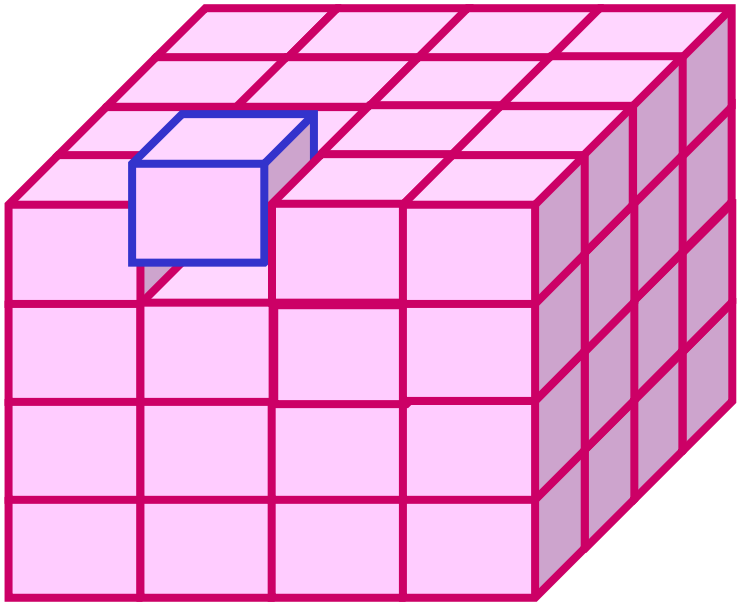
Data Warehousing

Kapitel 1: Einführung

Michael Hartung
Sommersemester 2011

Universität Leipzig
Institut für Informatik

http://dbs.uni-leipzig.de

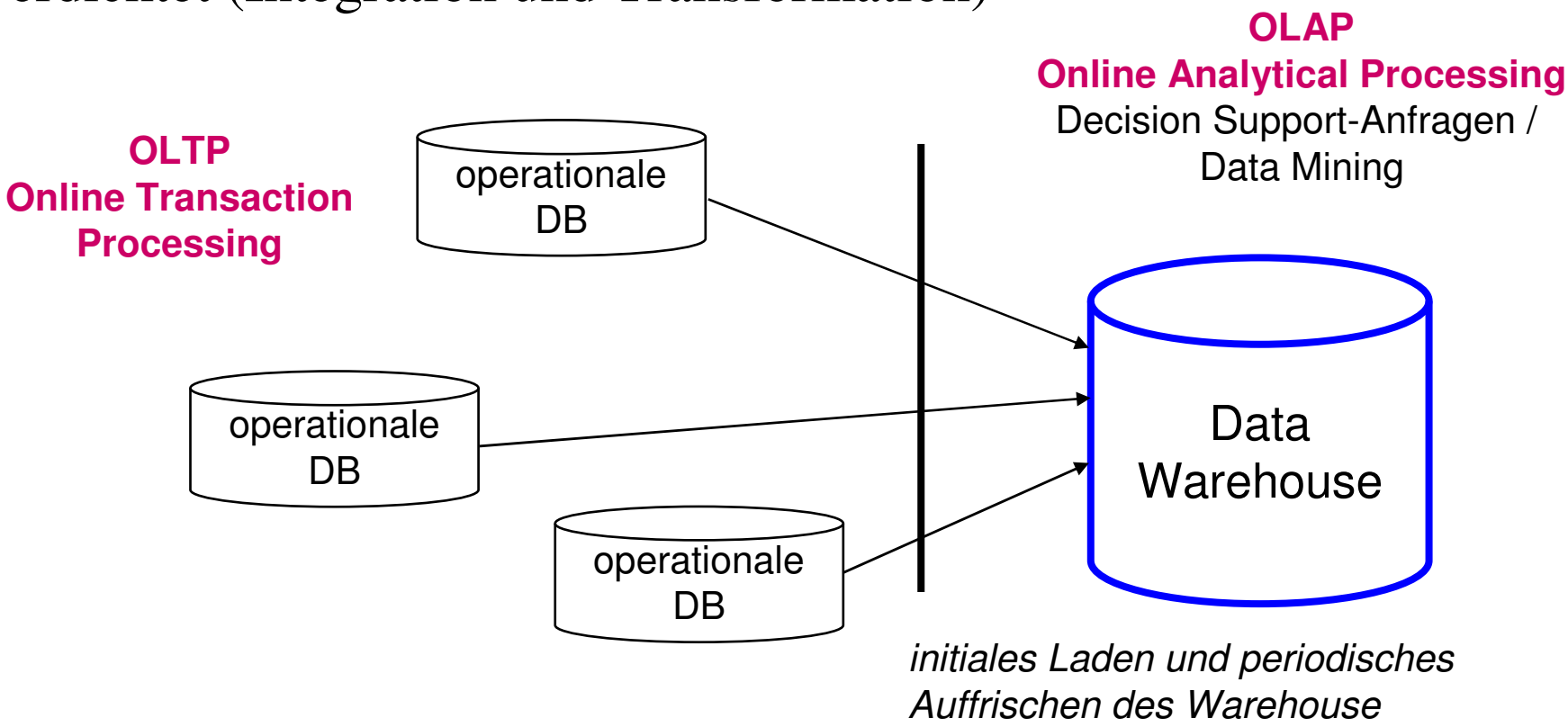


1. Data Warehouses - Einführung

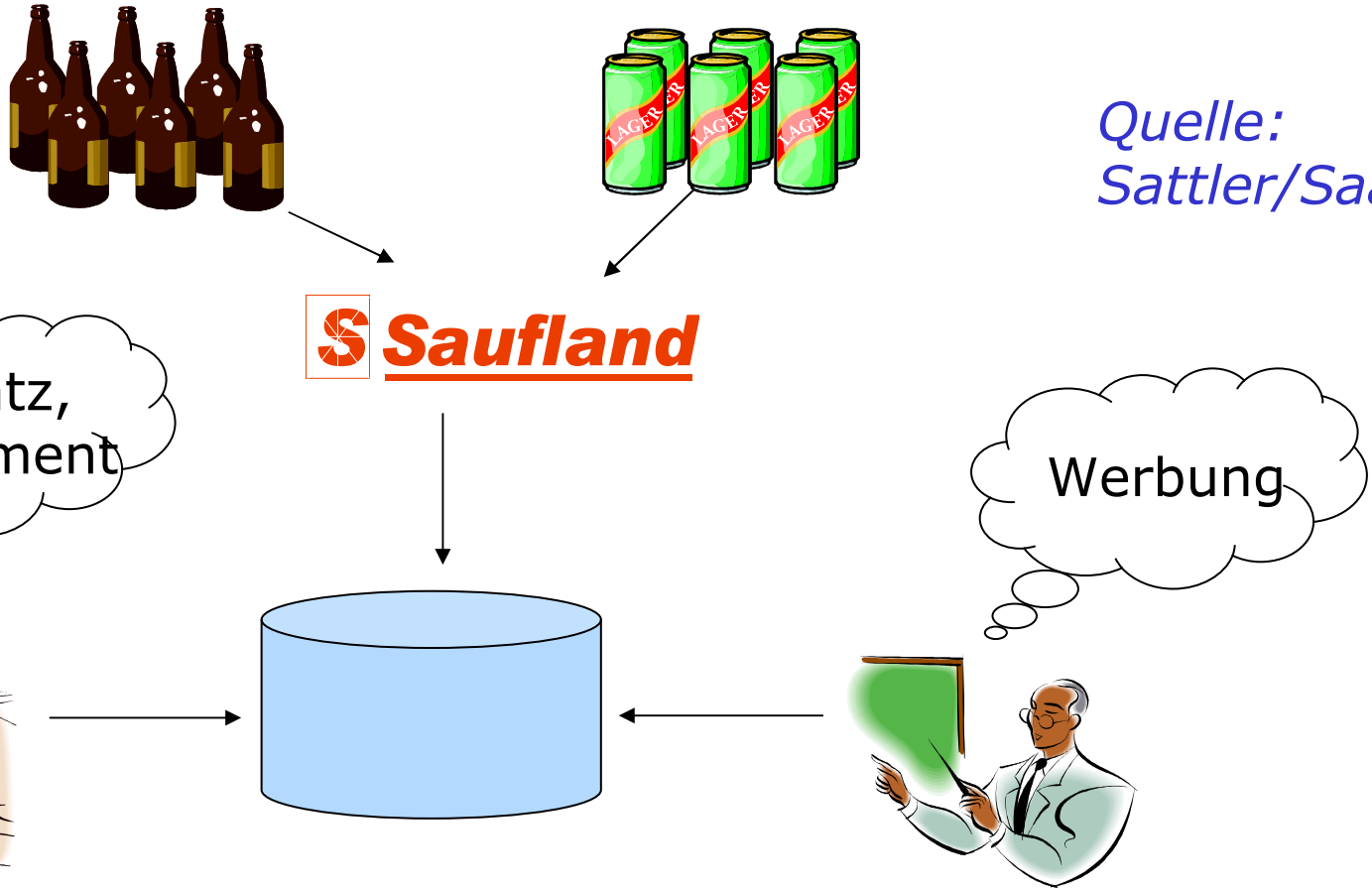
- Definition *Data Warehouse*
- Einsatzbeispiele
- OLTP vs. OLAP
- Grobarchitektur
- Virtuelle vs. physische Datenintegration
- Mehrdimensionale Datensicht
- Star-Schema, -Anfragen
- Data Mining

Data Warehouses

- Ausgangsproblem
  - viele Unternehmen haben Unmengen an Daten, ohne daraus ausreichend Informationen und Wissen für kritische Entscheidungsaufgaben ableiten zu können
- ***Data Warehouse (Def.):*** für Analysezwecke optimierte zentrale Datenbank, die Daten aus mehreren, i.a. heterogenen Quellen zusammenführt und verdichtet (Integration und Transformation)



Szenario: Getränkemarkt

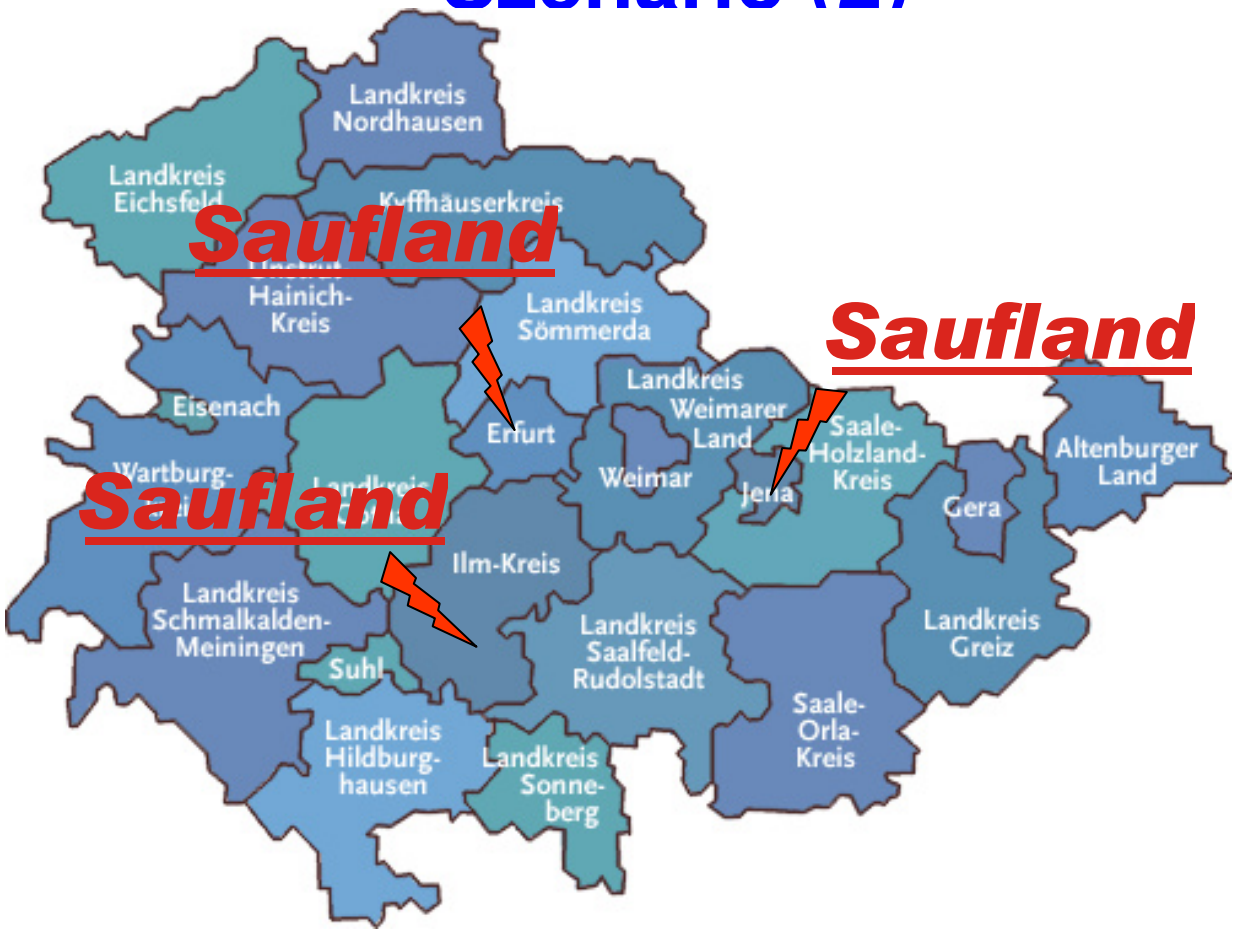


Quelle: Sattler/Saake

,

- Anfragen:
  - Wie viele Flaschen Cola wurden letzten Monat verkauft?
  - Wie hat sich der Verkauf von Rotwein im letzten Jahr entwickelt?
  - Wer sind unsere Top-Kunden?
  - Von welchem Lieferanten beziehen wir die meisten Kisten?

Szenario (2)



Quelle: Sattler/Saake

,

- Anfragen
  - Verkaufen wir in Ilmenau mehr Bier als in Erfurt?
  - Wie viel Cola wurde im Sommer in ganz Thüringen verkauft?
  - Mehr als Wasser?

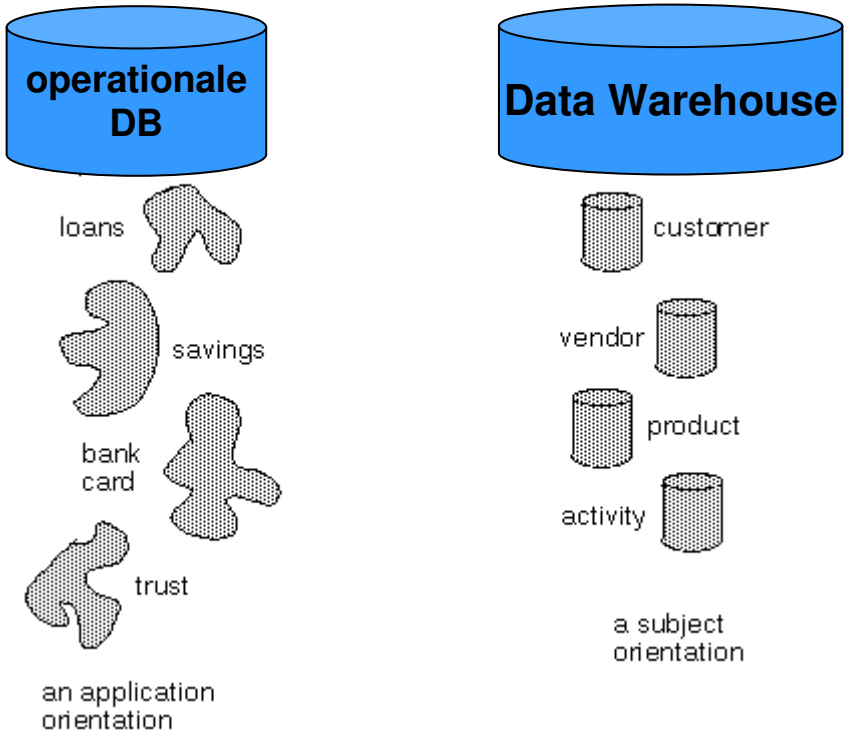
Einsatzbeispiele

- Warenhauskette
  - Verkaufszahlen und Lagerbestände aller Warenhäuser
  - mehrdimensionale Analysen: Verkaufszahlen nach Produkten, Regionen, Warenhäusern
  - Ermittlung von Kassenschlagern und Ladenhütern
  - Analyse des Kaufverhaltens von Kunden (Warenkorbanalyse)
  - Erfolgskontrolle von Marketing-Aktivitäten
  - Minimierung von Beständen
  - Optimierung der Produktpalette
  - Optimierung der Preisgestaltung • • •
- Versicherungsunternehmen
  - Bewertung von Filialen, Vertriebsbereichen, Schadensverlauf, ...
  - automatische Risikoanalyse
  - schnellere Entscheidung über Kreditkarten, Lebensversicherung; Krankenversicherung ...
- Banken, Versandhäuser, Restaurant-Ketten
- wissenschaftliche Einsatzfälle (z.B. Bioinformatik) • • •

DW-Eigenschaften nach Inmon

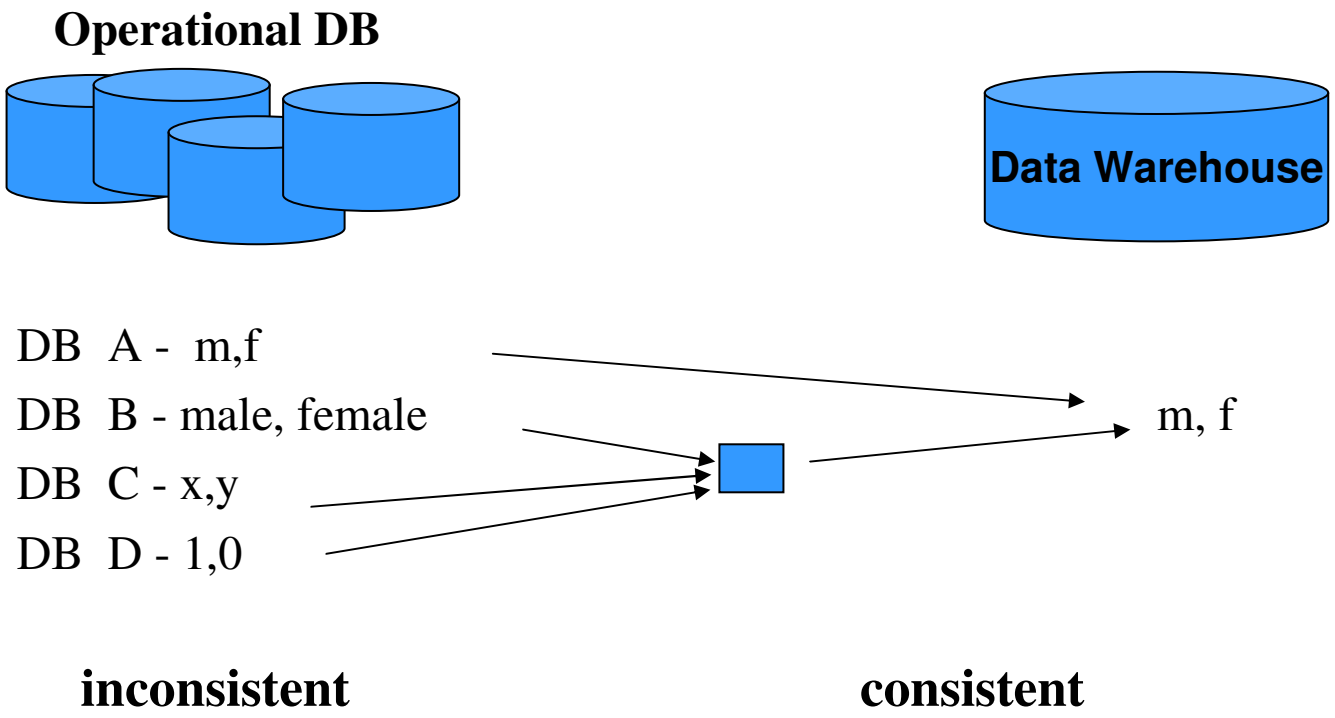
A Data Warehouse is a subject-oriented, integrated, non-volatile, and time variant collection of data in support of managements decisions (W. H. Inmon, Building the Data Warehouse, 1996)

- Themenorientiert (subject-oriented):
  - Zweck des Systems ist nicht Erfüllung einer dedizierten Aufgabe (z.B. Personaldatenverwaltung), sondern Unterstützung übergreifender Auswertungsmöglichkeiten aus verschiedenen Perspektiven
  - alle Daten - unternehmensweit - über ein Subjekt (Kunden, Produkte, Regionen ...) und nicht „versteckt“ in verschiedenen Anwendungen



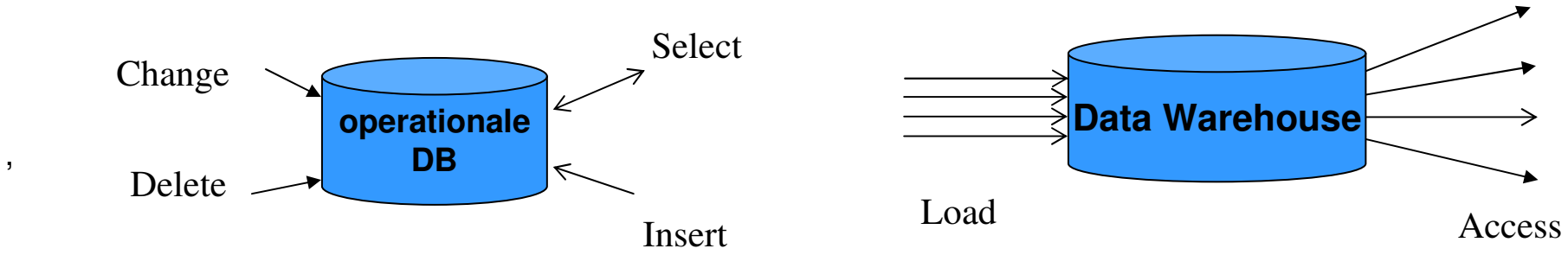
DW-Eigenschaften nach Inmon (2)

- Integrierte Datenbasis (integrated): Daten aus mehreren verschiedenen Datenquellen



DW-Eigenschaften nach Inmon (3)

- Nicht-flüchtige Datenbasis (non-volatile):
  - Daten im DW werden i.a. nicht mehr geändert
  - stabile, persistente Datenbasis

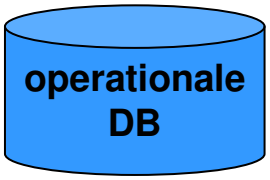


,

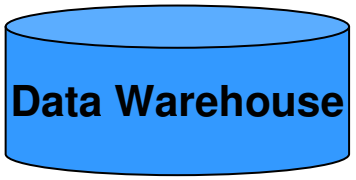
regelmäßige Änderungen von Sätzen

DW-Eigenschaften nach Inmon (4)

- Historische Daten (time-variant):
  - Vergleich der Daten über Zeit möglich (Zeitreihenanalyse)
  - Speicherung über längeren Zeitraum



Time Variancy



,

aktuelle Datenwerte:

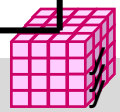
- Zeitbezug optional
- Zeithorizont: 60-90 Tage
- Daten änderbar

Schnappschuss-Daten

- Zeitbezug aller Objekte
- Zeithorizont: 2-10 Jahre
- keine Änderung nach Schnappschuss-Erstellung

Operationale Datenbanken vs. Data Warehouses (OLTP vs. OLAP)

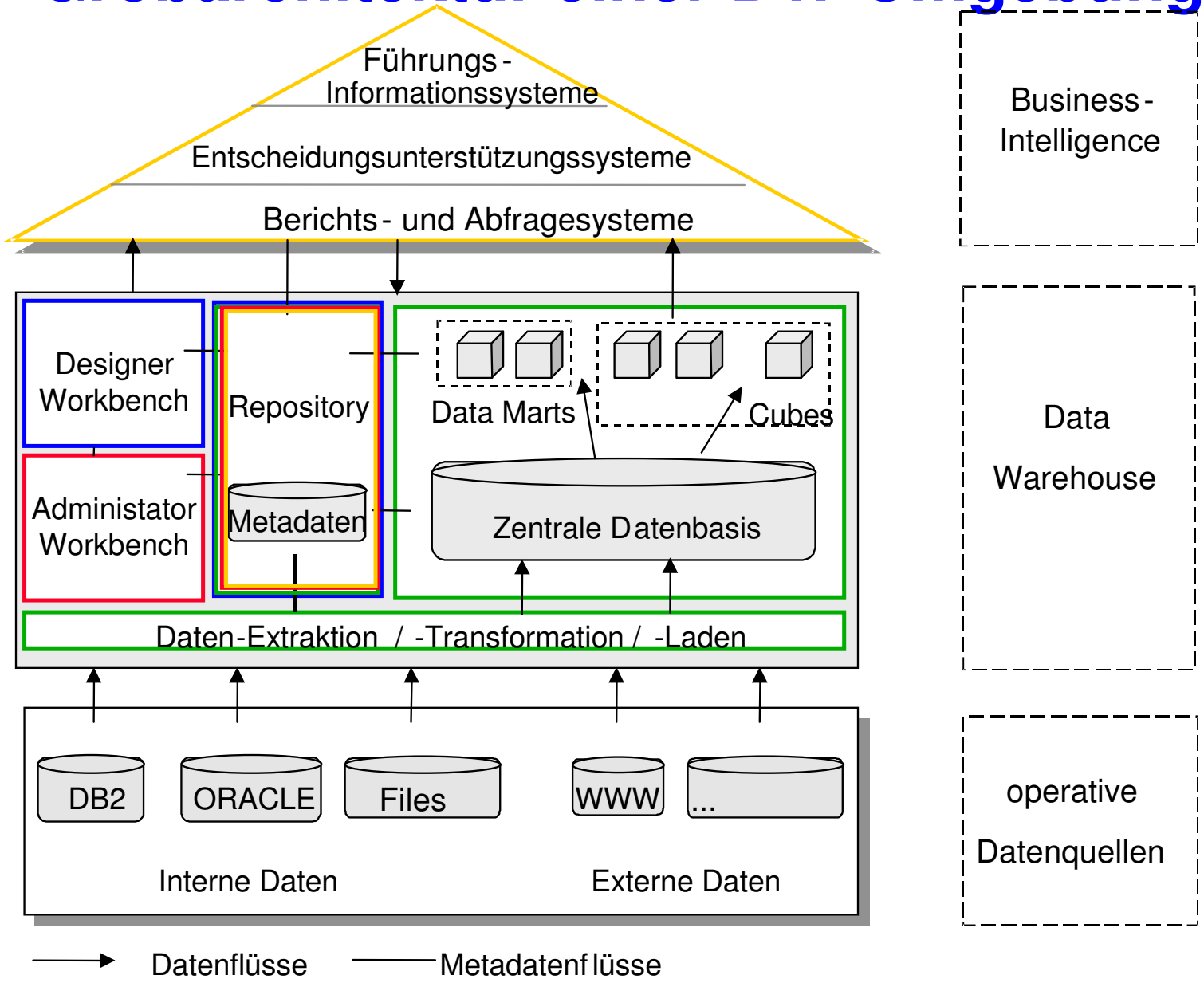
	Operationale Datenbanken /OLTP	**Data Warehouses/OLAP**
Entstehung	jeweils für eine Applikation oder aus einer bestimmten Perspektive heraus	
Bedeutung	Tagesgeschäft	
Nutzer	Sachbearbeiter, Online-Nutzer	
Datenzugriff	sehr häufiger Zugriff; kleine Datenmengen pro Operation; Lesen, Schreiben, Modifizieren, Löschen	
Änderungen/Aktualität	sehr häufig / stets aktuell	
#Datenquellen	meist eine	
Datenmerkmale	nicht abgeleitet, zeitaktuell, autonom, dynamisch	
Optimierungsziele	hoher Durchsatz, sehr kurze Antwortzeiten (ms .. s), hohe Verfügbarkeit	



Warum separates Data Warehouse?

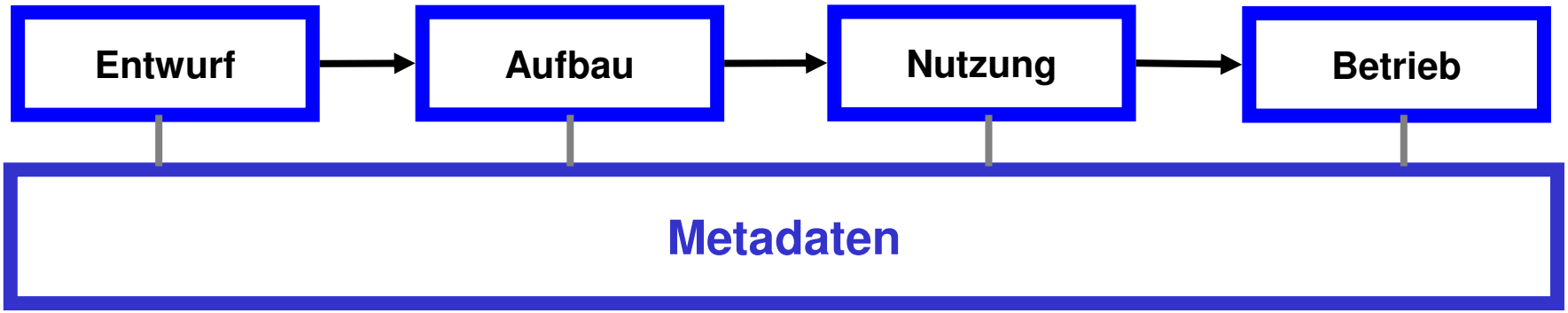
- Unterschiedliche Nutzung und Datenstrukturierung
- Performance
  - OLTP optimiert für kurze Transaktionen und bekannte Lastprofile
  - komplexe OLAP-Anfragen würden gleichzeitige OLTP-Transaktionen des operationalen Betriebs drastisch verschlechtern
  - spezieller physischer und logischer Datenbankentwurf für multidimensionale Sichten/Anfragen notwendig
  - Transaktionseigenschaften (ACID) nicht wichtig
- Funktionalität
  - historische Daten
  - Konsolidierung (Integration, Bereinigung und Aggregation) von Daten aus heterogenen Datenquellen
- Sicherheit
- Nachteile der separaten Lösung
  - Datenredundanz
  - Daten nicht vollständig aktuell
  - hoher Administrationsaufwand
  - hohe Kosten

Grobarchitektur einer DW-Umgebung



DW-Prozesse

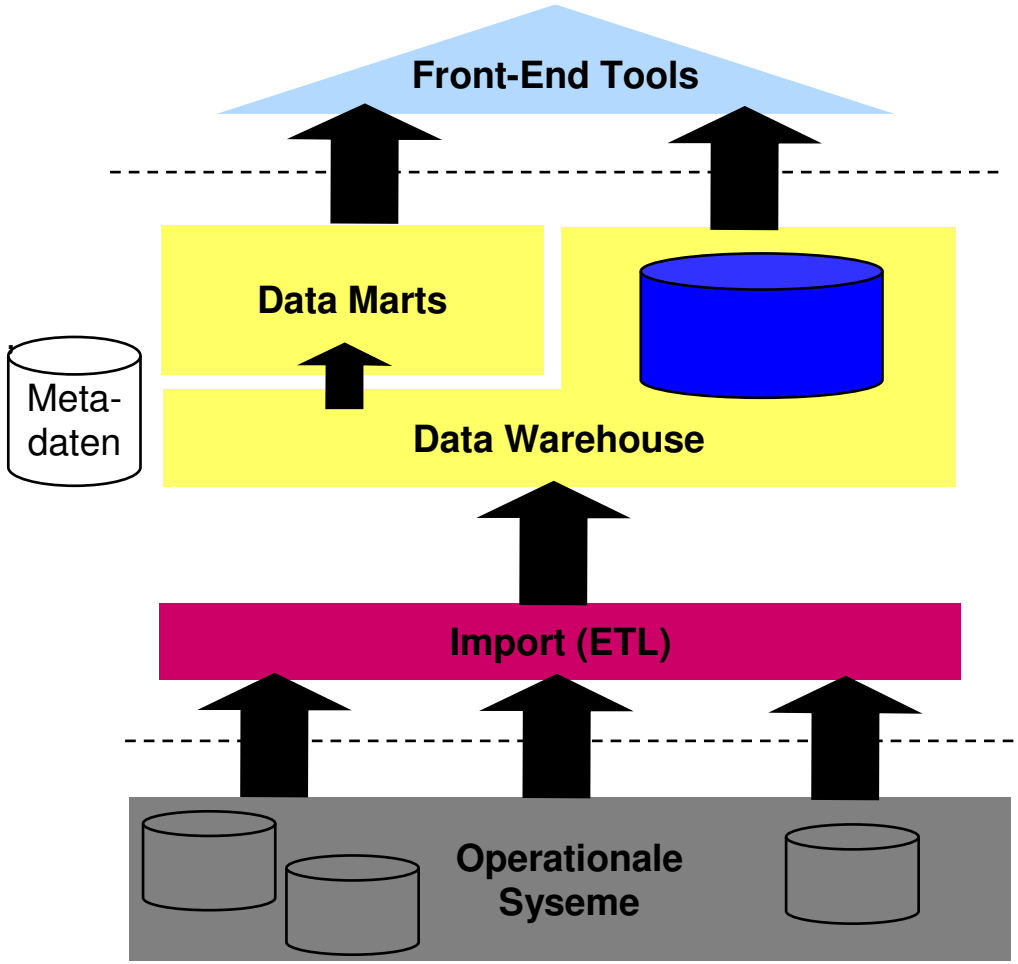
- Data Warehousing umfaßt mehrere Teilprozesse
  - Entwurf ("design it"),
  - Aufbau ("build it", „populate"),
  - Nutzung ("use it", „analyze") sowie
  - Betrieb und Administration („maintain it" / „administer")



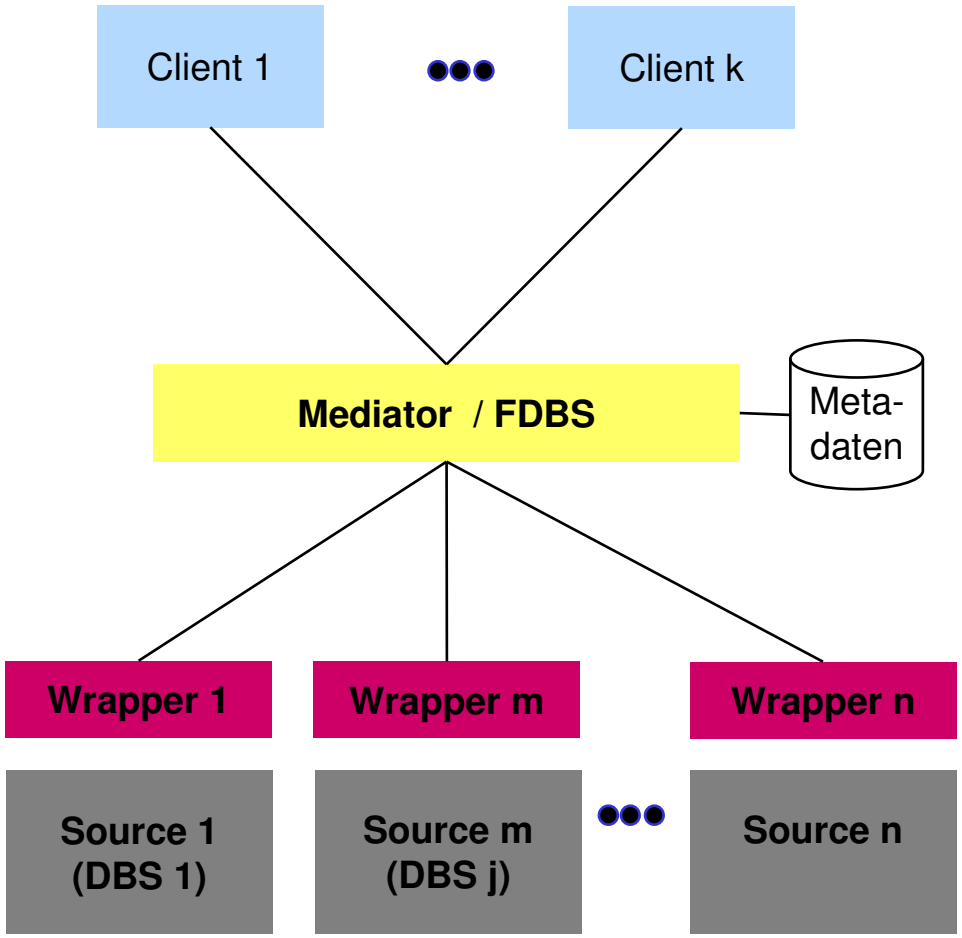
- DW ist meist kein monolithisches System
  - meist Nutzung von Tools / Komponenten unterschiedlicher Hersteller sowie eigenprogrammierten Anteilen
- zentrale Bedeutung der Metadaten, jedoch oft unzureichend unterstützt

Datenintegration: physisch vs. virtuell

Physische (Vor-) Integration
(Data Warehousing)

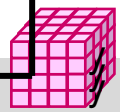


Virtuelle Integration
(Mediator/Wrapper-Architekturen, förderierte DBS)

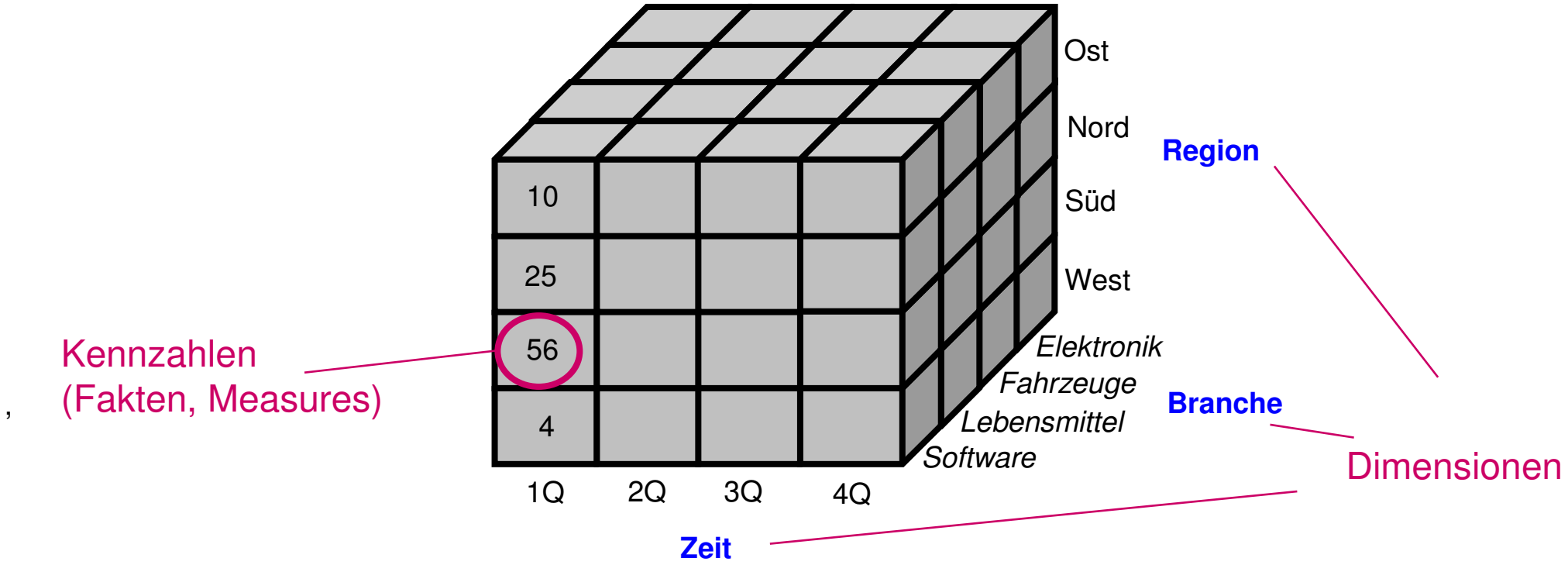


Datenintegration: physisch vs. virtuell (2)

	Physisch (Data Warehouse)	Virtuell
Integrationszeitpunkt: Metadaten	Vorab (DW-Schema)	Vorab (globale Sicht)
Integrationszeitpunkt: Daten	vorab	Dynamisch (zur Anfragezeit)
Aktualität der Daten		
Autonomie der Datenquellen		
Erreichbare Datenqualität		
Analysezeitbedarf für große Datenmengen		
Hardwareaufwand		
Skalierbarkeit auf viele Datenquellen		

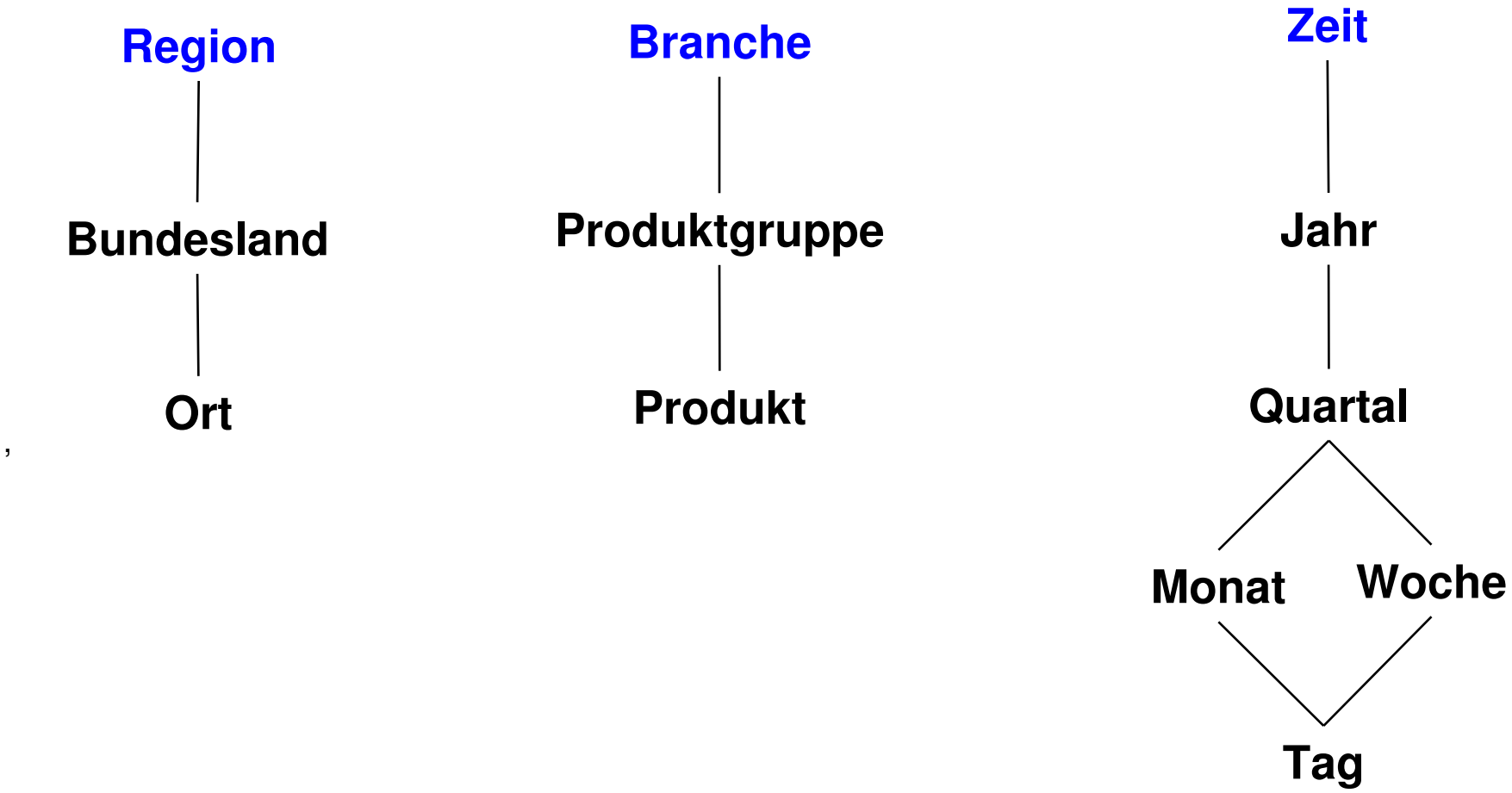


Mehrdimensionale Datensicht



- Kennzahlen: numerische Werte als Grundlage für Aggregationen/Berechnungen (z.B. Absatzzahlen, Umsatz, etc.)
- Dimensionen: beschreibende Eigenschaften
- Operationen:
  - Aggregierung der Kennzahlen über eine oder mehrere Dimension(en)
  - Slicing and Dicing: Bereichseinschränkungen auf Dimensionen

Hierarchische Dimensionierung



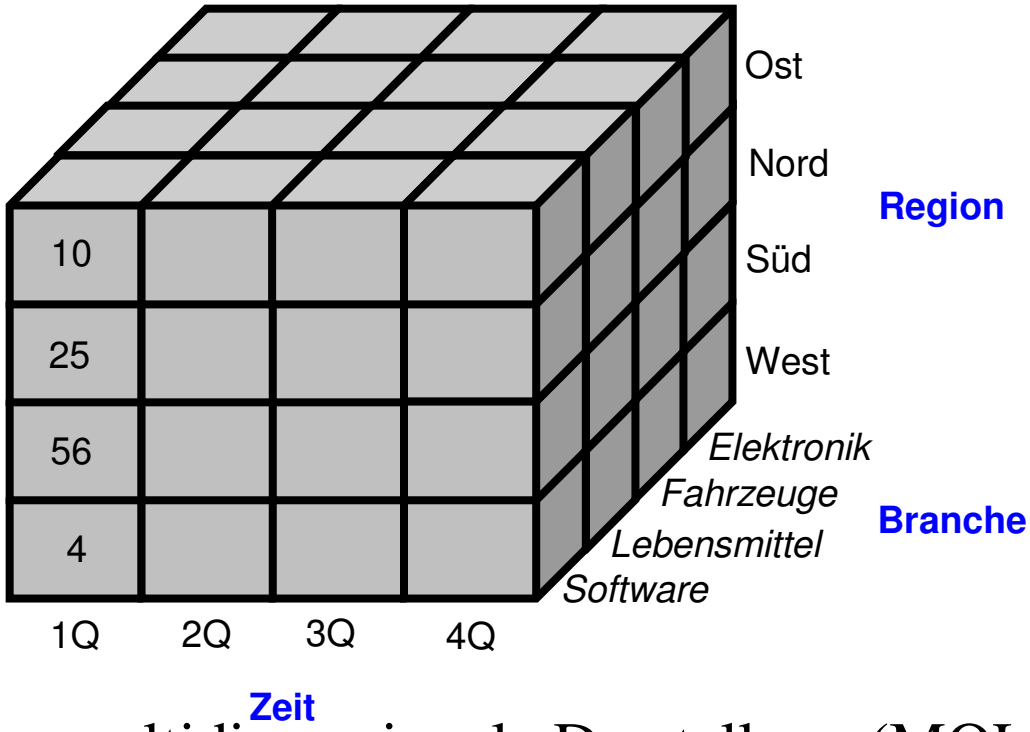
,

- Operationen zum Wechsel der Dimensionsebenen
  - Drill-Down
  - Roll-Up

OLAP (Online Analytical Processing)

- interaktive multidimensionale Analyse auf konsolidierten Unternehmensdaten
- Merkmale / Anforderungen
  - multidimensionale, konzeptionelle Sicht auf die Daten
  - unbegrenzte Anzahl an Dimensionen und Aggregationsebenen
  - unbeschränkte dimensionsübergreifende Operationen
  - intuitive, interaktive Datenmanipulation und Visualisierung
  - transparenter (integrierter) Zugang zu heterogenen Datenbeständen mit logischer Gesamtsicht
  - Skalierbarkeit auf große Datenmengen
  - stabile, volumenunabhängige Antwortzeiten
  - Mehrbenutzerunterstützung
  - Client/Server-Architektur

Multidimensional vs. relational

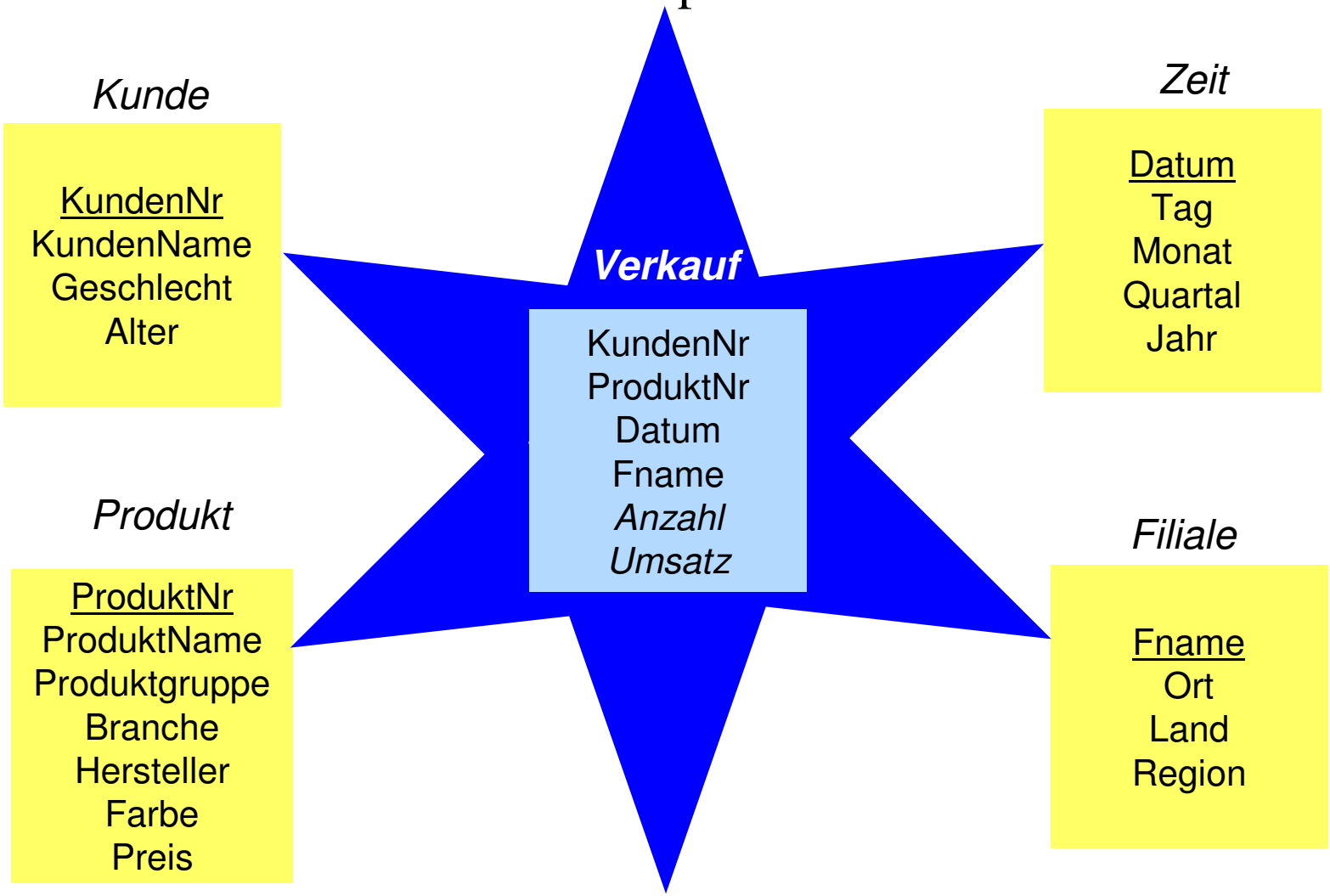


Bestellnr	Region	Branche	Zeit	Menge
1406	Ost	Fahrzeuge	2Q	5
4123	West	Elektronik	1Q	58
7829	Süd	Fahrzeuge	2Q	30
5327	Ost	Lebensmittel	4Q	3000
9306	Nord	Software	1Q	25
2574	Ost	Elektronik	4Q	2

- multidimensionale Darstellung (MOLAP): Kreuzprodukt aller Wertebereiche mit aggregiertem Wert pro Kombination
  - Annahme: fast alle Kombinationen kommen vor
- relationale Darstellung (ROLAP):
  - Relation: Untermenge des Kreuzproduktes aller Wertebereiche
  - nur vorkommende Wertekombinationen werden gespeichert (Tupel)
- Hybrides OLAP (HOLAP): ROLAP + MOLAP

Star-Schema

- zentrale Faktentabelle sowie 1 Tabelle pro Dimension



Anfragen

Beispielanfrage: Welche Auto-Hersteller wurden von weiblichen Kunden in Sachsen im 1. Quartal 2008 favorisiert?

```
select p.Hersteller, sum (v.Anzahl)
from Verkauf v, Filialen f, Produkt p, Zeit z, Kunden k
where z.Jahr = 2008 and z.Quartal = 1 and k.Geschlecht = 'W' and
      p.Produkttyp = 'Auto' and f.Land = 'Sachsen' and
      v.Datum = z.Datum and v.ProduktNr = p.ProduktNr and
      v.FName = f. FName and v.KundenNr = k.KundenNr
group by p.Hersteller;
```

- Star-Join
  - sternförmiger Join der (relevanten) Dimensionstabellen mit der Faktentabelle
  - Einschränkung der Dimensionen
  - Verdichtung der Kennzahlen durch Gruppierung und Aggregation

Analysewerkzeuge

- (Ad Hoc-) Query-Tools
- Reporting-Werkzeuge, Berichte mit flexiblen Formatierungsmöglichkeiten
- OLAP-Tools
  - interaktive mehrdimensionale Analyse und Navigation (Drill Down, Roll Up, ...)
  - Gruppierungen, statistische Berechnungen, ...
- Data Mining-Tools

- Darstellung
  - Tabellen, insbesondere Pivot-Tabellen (Kreuztabellen)
  - Analyse durch Vertauschen von Zeilen und Spalten, Veränderung von Tabellendimensionen
  - Graphiken sowie Text und Multimedia-Elemente
- Nutzung über Web-Browser, Spreadsheet-Integration

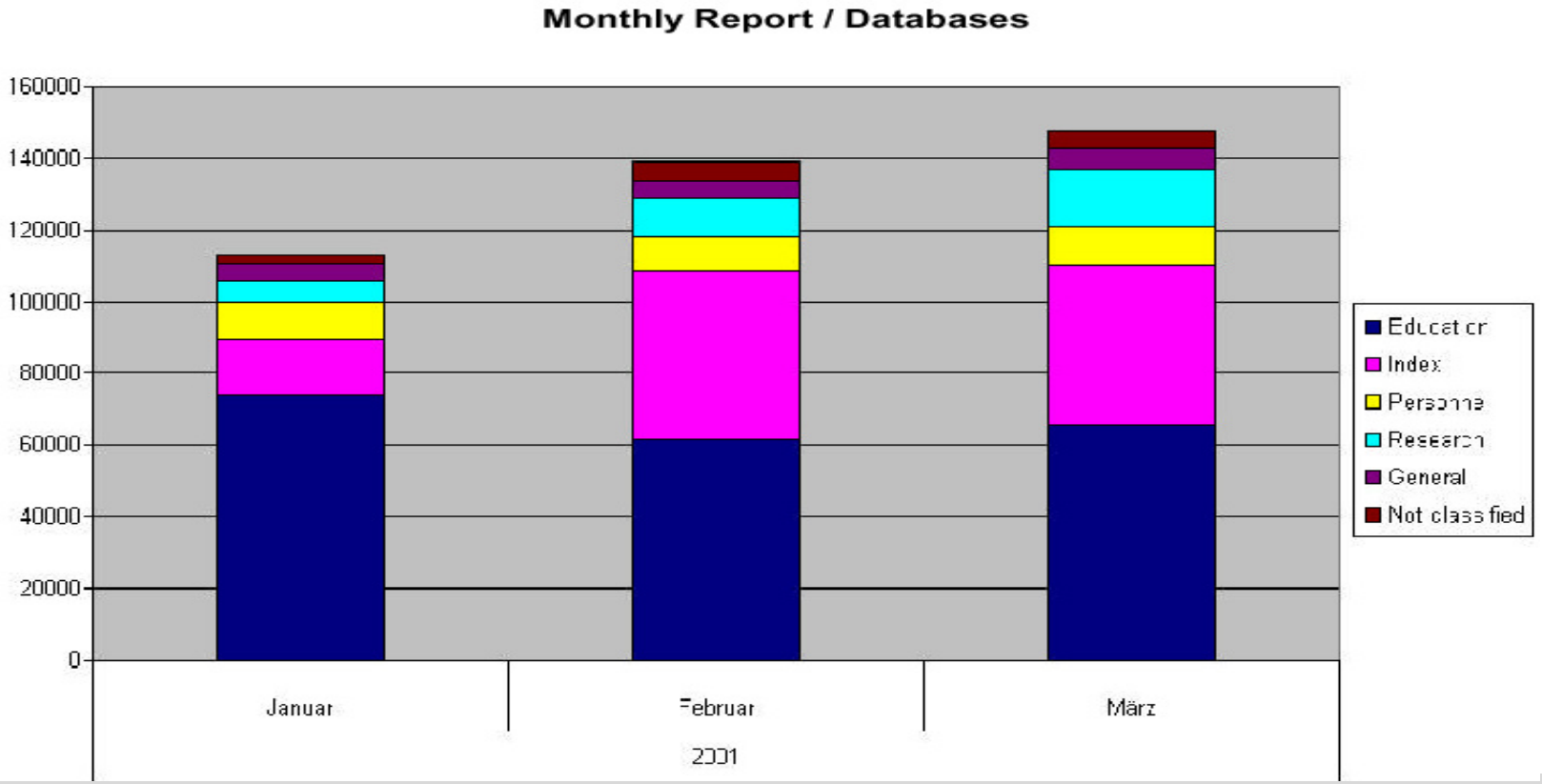
Beispiel: OLAP-Ausgabe (Excel)

Zugriffe

Source

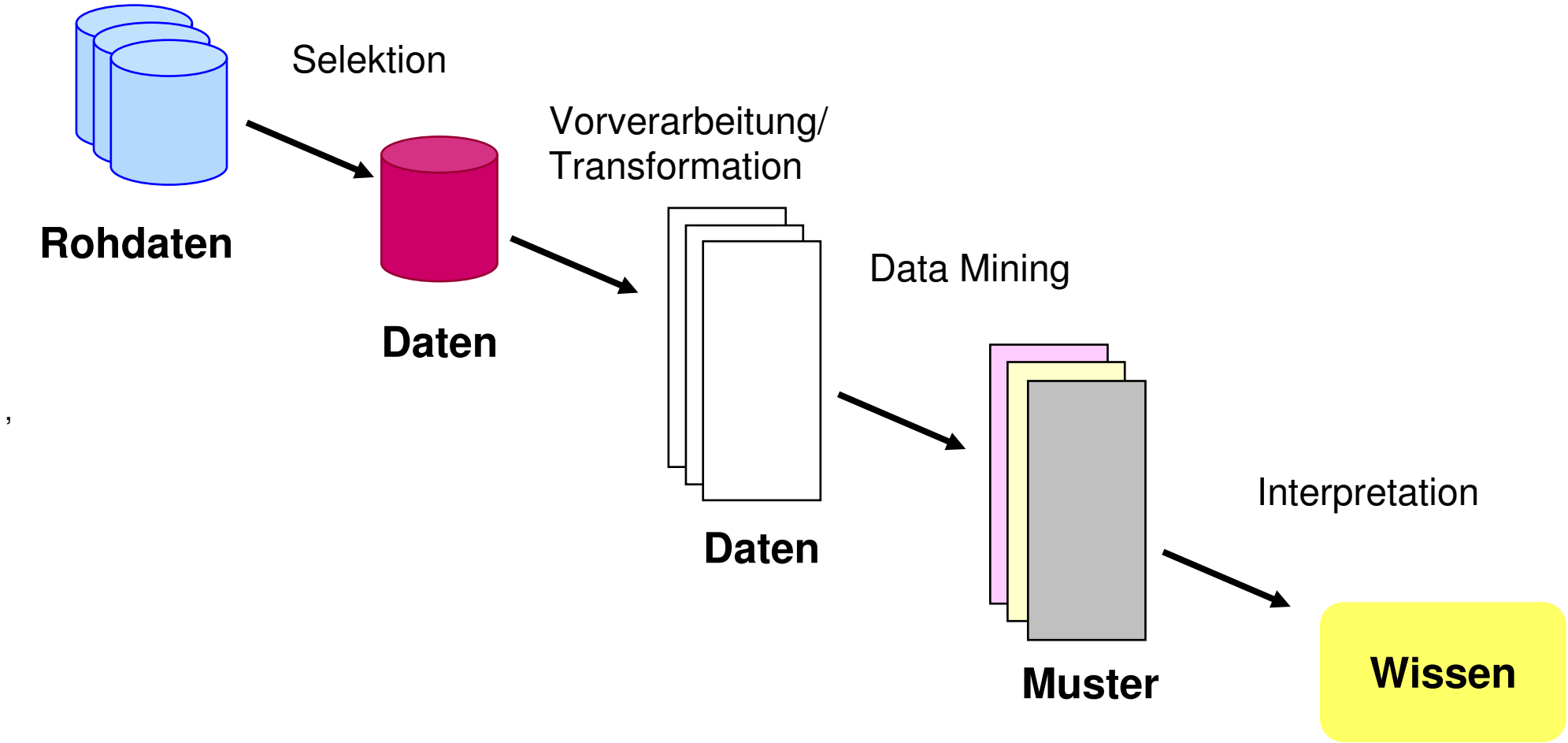
Alle Source

		Category					
		Education	Index	Personnel	Research	General	Not classified
Jahr	Monat Tag	Hits	Hits	Hits	Hits	Hits	Hits
⊟ 2001	⊞ Januar	73961	15494	10559	5079	4915	2060
	⊞ Februar	61697	46666	9880	10686	4558	5504
	⊞ März	65642	44708	10439	15837	5871	5334
	Total *	213494	115430	32867	33501	16054	14275
Gesamtergebnis *		1106493	189912	111213	84560	46708	39735



,

Knowledge Discovery



,

Data Mining: Techniken

- Data Mining: Einsatz statistischer und wissensbasierter Methoden auf Basis von Data Warehouses
  - Auffinden von Korrelationen, Mustern und Trends in Daten
  - “Knowledge Discovery”: setzt im Gegensatz zu OLAP (“knowledge verification”) kein formales Modell voraus
- Clusteranalyse
  - Objekte werden aufgrund von Ähnlichkeiten in Klassen eingeteilt (Segmentierung)
  - Bsp.: ähnliche Kunden, ähnliche Website-Nutzer ...
- Assoziationsregeln
  - Warenkorbanalyse (z.B. Kunde kauft A und B => Kunde kauft C)
- Klassifikation
  - Klassifikation von Objekten
  - Erstellung von Klassifikationsregeln / Vorhersage von Attributwerten (z.B. “guter Kunde” wenn Alter > 25 und ...)
  - mögliche Realisierung: Entscheidungsbaum

Beispiel Warenkorbanalyse

amazon.de

Data-Warehouse-Systeme Architektur, Entwicklung, Anwendung
von Andreas Bauer, Holger Günzel

Amazon-Preis: EUR 49,00 Kostenlose Lieferung. Siehe Details.

Gewöhnlich versandfertig bei Amazon in 24 Stunden.

Nur noch 5 Stück verfügbar -- jetzt bestellen. (Warenneulieferung i
Sie möchten dieses Produkt morgen bis 12 Uhr geliefert bekom
nächsten 2 Stunden und 9 Minuten und wählen Sie DHL **Overnig**

,

Kunden, die dieses Buch gekauft haben, haben auch diese Bücher gekauft:

- *The Data Warehouse Stag*
- *Der Data-Warehouse-Rah*

Kunden, die Bücher von A ekauft:

- Norbert Egger
- Ralph Kimball
- Hans-Georg Kemper
- Wolfgang Lehner
- Lothar Schirmer

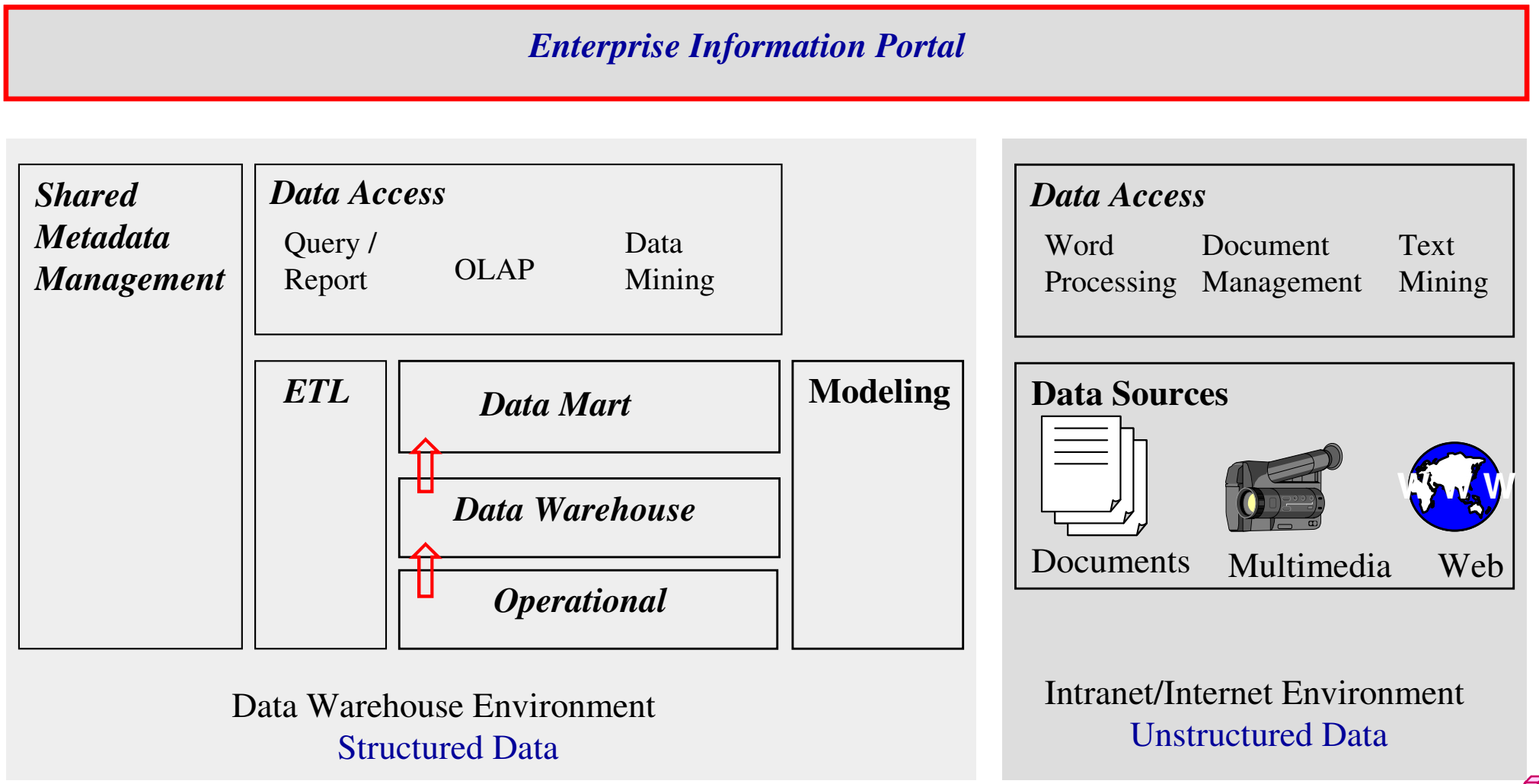
Kunden, die Artikel gekauft haben, welche Sie sich kürzlich angesehen haben, kauften auch:

Datenbanksysteme. Konzepte und Techniken der Implementierung.
von Theo Härder, Erhard Rahm

Grundlagen von Datenbanksystemen. Ausgabe Grundstudium
von Ramiz Elmasri, Shamkant B. Navathe

Enterprise Information Portale

- einheitlicher unternehmensweiter Zugang zu strukturierten und unstrukturierten Daten



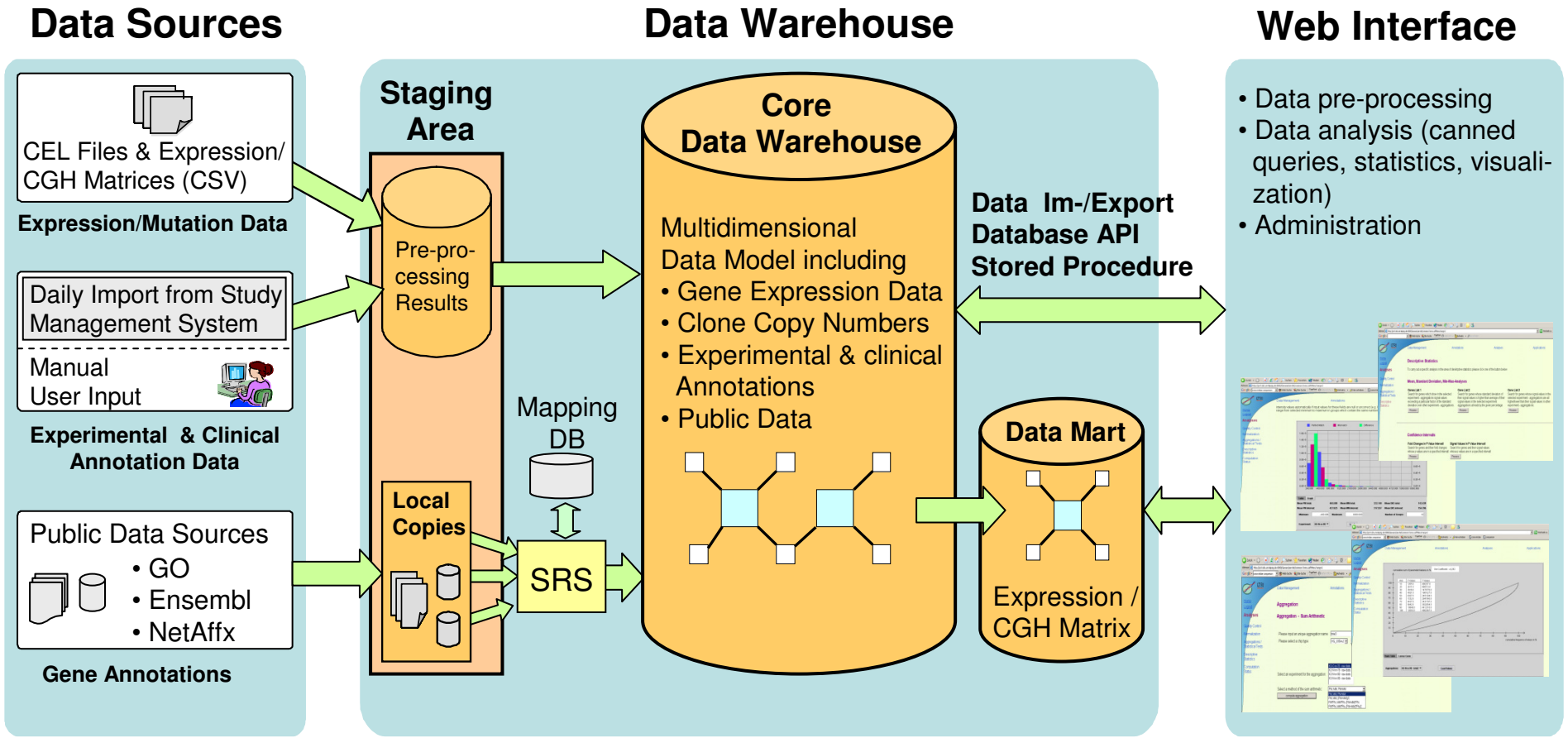
Data Warehouse Hype & Realität

- “turning data into knowledge”
- “360° view of customer”
- “a single version of the truth”
- “getting you closer to the customer”
- “Better decision making”

- Fragen
  - Wie werden welche Kundendaten genutzt?
  - Wie erfolgt die Sicherung der Datenqualität?

GeWare: Expression Data Warehouse*

- Verwaltung und Analyse großer Mengen von Genexpressionsdaten
- Integration weiterer Informationen zu Genen, Patienten, etc.



*E. Rahm, T. Kirsten, J. Lange: *The GeWare data warehouse platform for the analysis of molecular-biological and clinical data*. Journal of Integrative Bioinformatics, 4(1):47, 2007.

Zusammenfassung

- Data Warehousing: DB-Anfrageverarbeitung und Analysen auf integriertem Datenbestand für Decision Support (OLAP)
- riesige Datenvolumina
- Hauptschwierigkeit: Integration heterogener Datenbestände sowie Bereinigung von Primärdaten
- Physische Datenintegration ermöglicht aufwändige Datenbereinigung und effiziente Analyse auf großen Datenmengen
- Mehrdimensionale Datenmodellierung und -organisation
- Breites Spektrum an Auswertungs- und Analysemöglichkeiten
- Data Mining: selbständiges Aufspüren relevanter Muster in Daten
- *Data Warehouse ist weit mehr als Datenbank*