

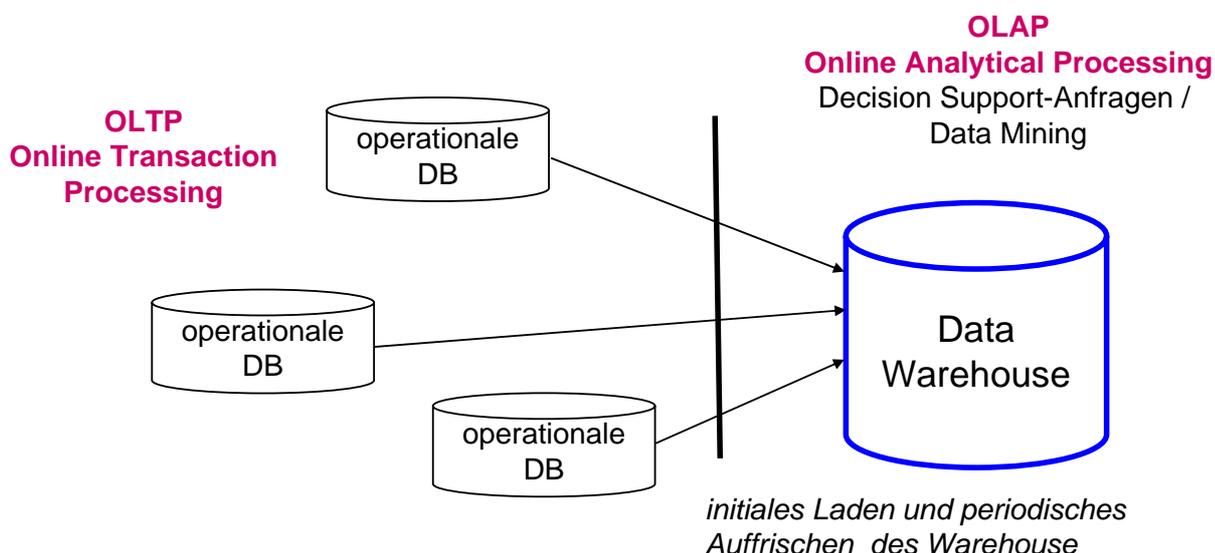
1. Data Warehouses - Einführung

- 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)



Einsatzbeispiele

■ Warenhauskette

- Verkaufszahlen und Lagerbestände aller Warenhäuser
- mehrdimensionale Analysen: Verkaufszahlen nach Produkten, Regionen, Warenhäusern
- Ermittlung von Kassenschlagern und Ladenhütern
- Analyse des Kaufverhaltens von Kunden
- Erfolgskontrolle von Marketing-Aktivitäten
- Minimierung von Beständen
- Optimierung der Produktpalette
- Optimierung der Preisgestaltung •••

■ Versicherungsunternehmen

- Bewertung von Filialen, Vertriebsbereichen, Schadensverlauf, ...
- automatische Risikoanalyse
- schnellere Entscheidung über Annahme eines Kunden (Lebensversicherung; Krankenversicherung ...)

■ Banken

■ Versandhäuser

■ Restaurant-Ketten

■ wissenschaftliche Einsatzfälle (z.B. Bioinformatik) •••

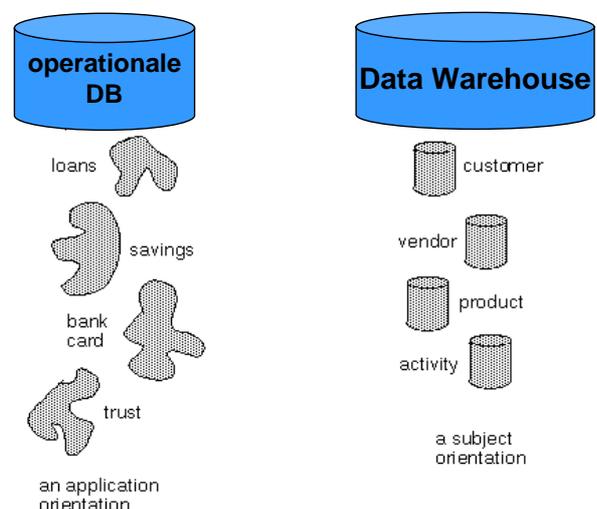


DW-Eigenschaften nach Inmon

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, 2nd ed., John Wiley 1996)

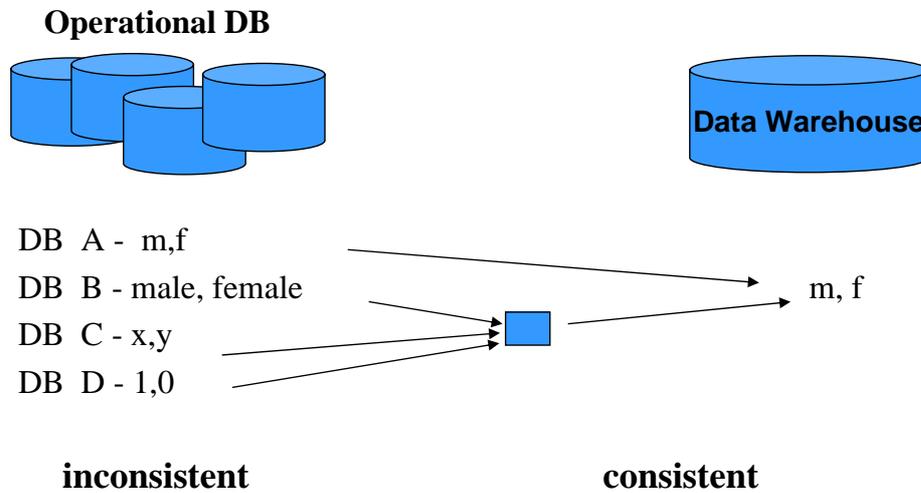
■ Themenorientiert (subject-oriented):

- 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)

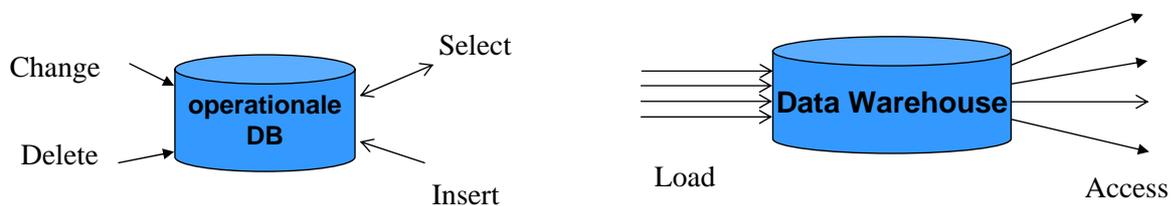
- Integrierte Datenbasis (integrated): Daten aus mehreren verschiedenen Datenquellen



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
<i>Entstehung</i>	jeweils für eine Applikation oder aus einer bestimmten Perspektive heraus	
<i>Bedeutung</i>	Tagesgeschäft	
<i>Nutzer</i>	Sachbearbeiter, Online-Nutzer	
<i>Datenzugriff</i>	sehr häufiger Zugriff; kleine Datenmengen pro Operation; Lesen, Schreiben, Modifizieren, Löschen	
<i>Änderungen/Aktualität</i>	sehr häufig / stets aktuell	
<i>Datenquellen</i>	meist eine	
<i>Datenmerkmale</i>	nicht abgeleitet, zeitaktuell, autonom, dynamisch	
<i>Optimierungsziele</i>	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

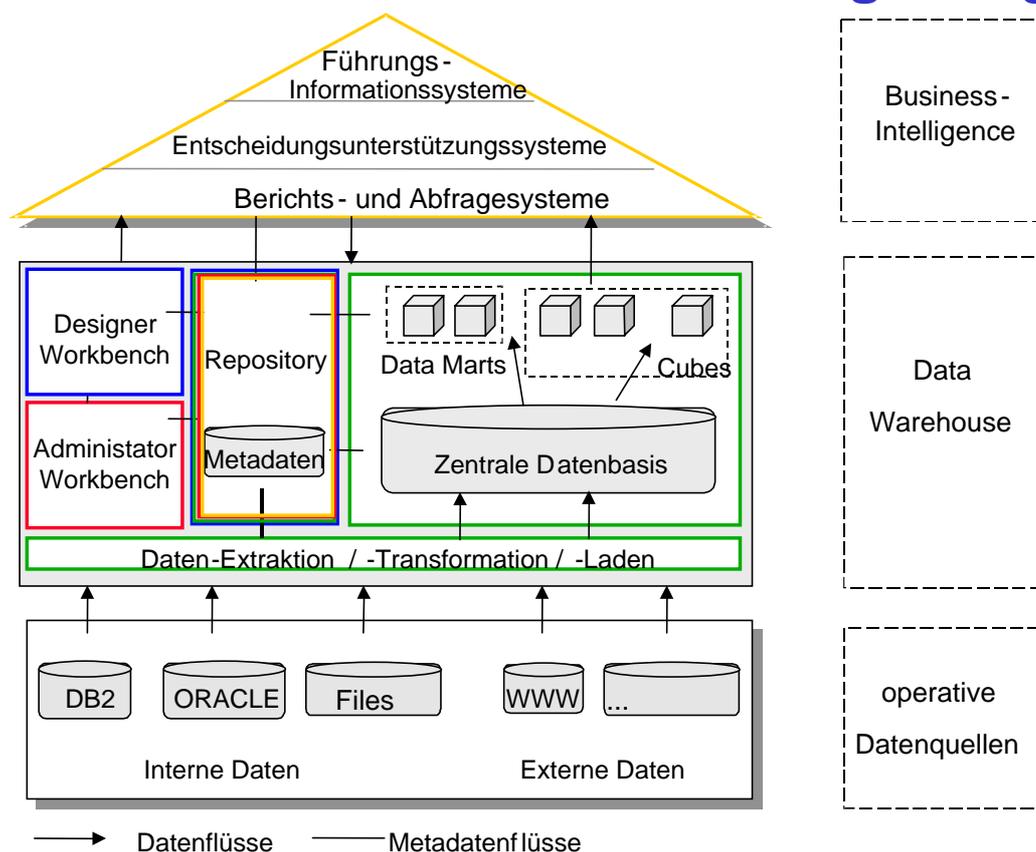
■ Sicherheit

■ Nachteile der separaten Lösung

- Datenredundanz
- Daten nicht vollständig aktuell
- hoher Administrationsaufwand
- hohe Kosten



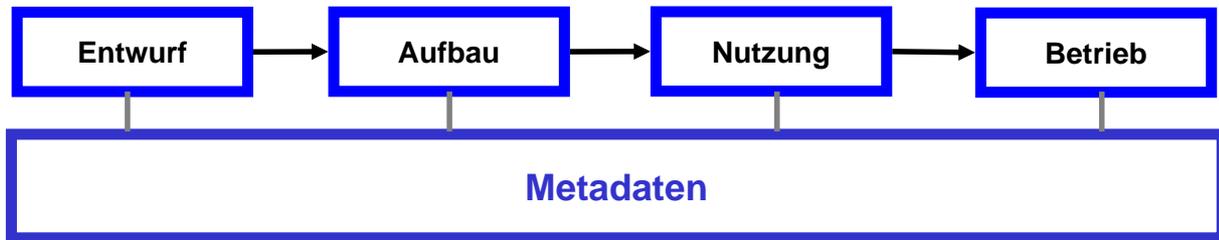
Grobarchitektur einer DW-Umgebung



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

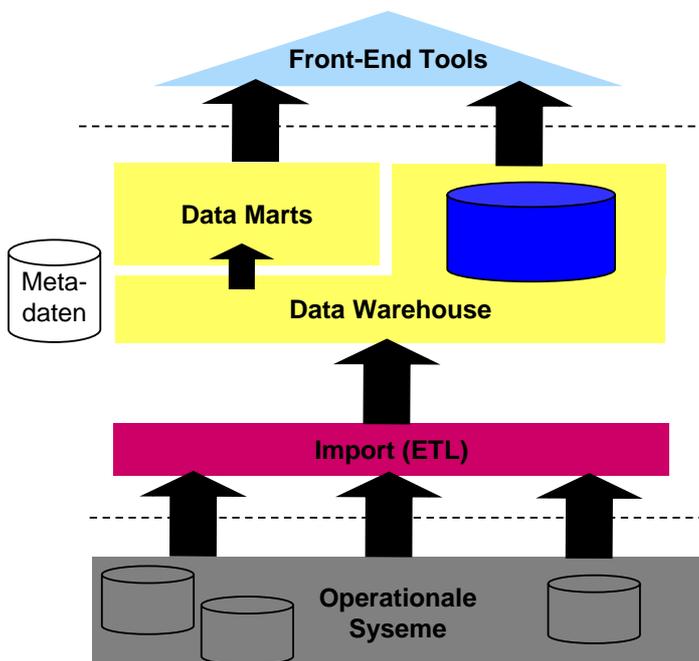
- meist Nutzung von Tools / Komponenten unterschiedlicher Hersteller sowie eigenprogrammierten Anteilen

■ zentrale Bedeutung der Metadaten, jedoch oft unzureichend unterstützt

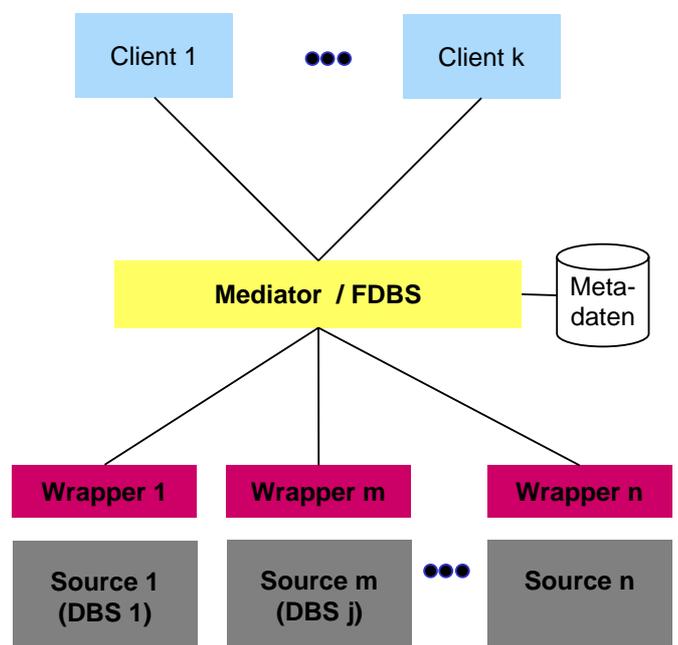


Datenintegration: physisch vs. virtuell

Physische (Vor-) Integration (Data Warehousing)



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



Datenintegration: physisch vs. virtuell (2)

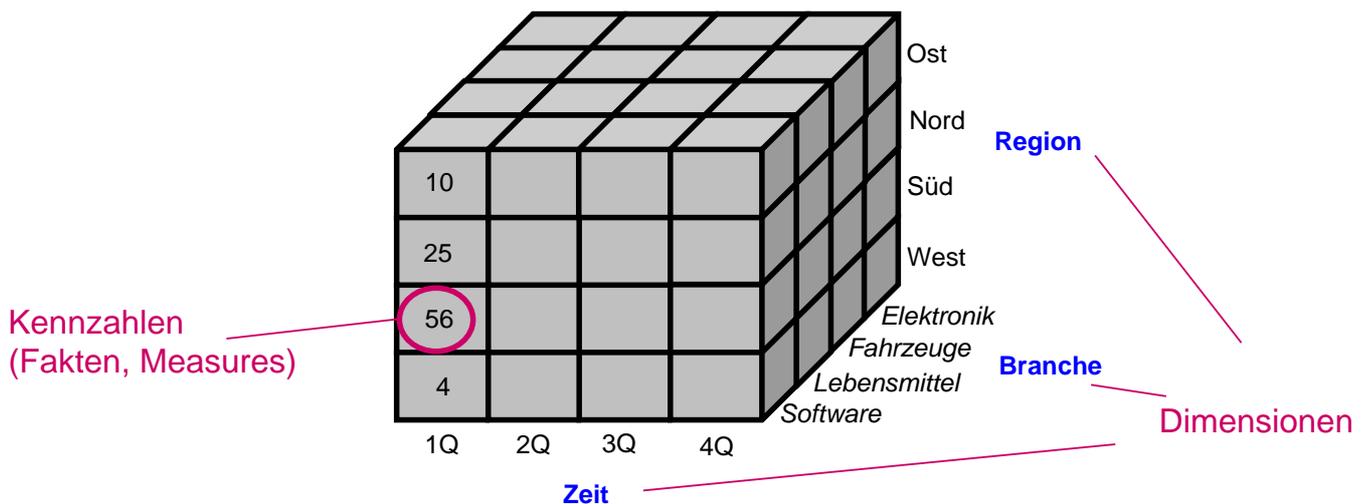
	Physisch (Data Warehouse)	Virtuell
Integrationszeitpunkt: Metadaten		
Integrationszeitpunkt: Daten		
Aktualität der Daten		
Autonomie der Datenquellen		
Erreichbare Datenqualität		
Analysezeitbedarf für große Datenmengen		
Hardwareaufwand		
Skalierbarkeit auf viele Datenquellen		

© Prof. E. Rahm

1 - 13



Mehrdimensionale Datensicht



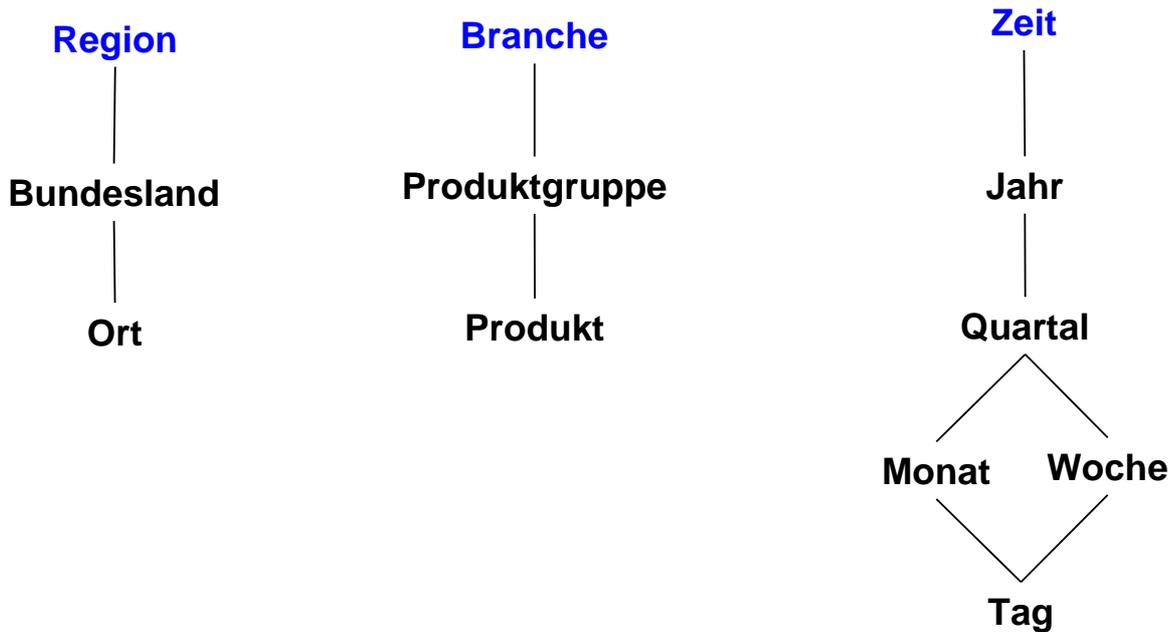
- **Kennzahlen:** numerische Werte als Grundlage für Aggregationen/Berechnungen (z.B. Absatzzahlen, Umsatz, etc.)
- **Dimensionen:** beschreibende Eigenschaften
- **Operationen:**
 - Aggregation der Kennzahlen über eine oder mehrere Dimension(en)
 - Slicing and Dicing: Bereichseinschränkungen auf Dimensionen

© Prof. E. Rahm

1 - 14



Hierarchische Dimensionierung



■ Operationen zum Wechsel der Dimensionsebenen

- Drill-Down
- Roll-Up

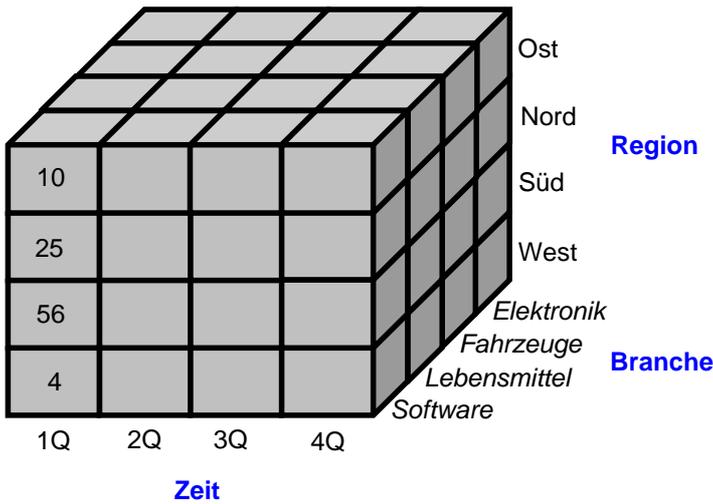


OLAP (Online Analytical Processing)

- 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



<u>Bestellnr</u>	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

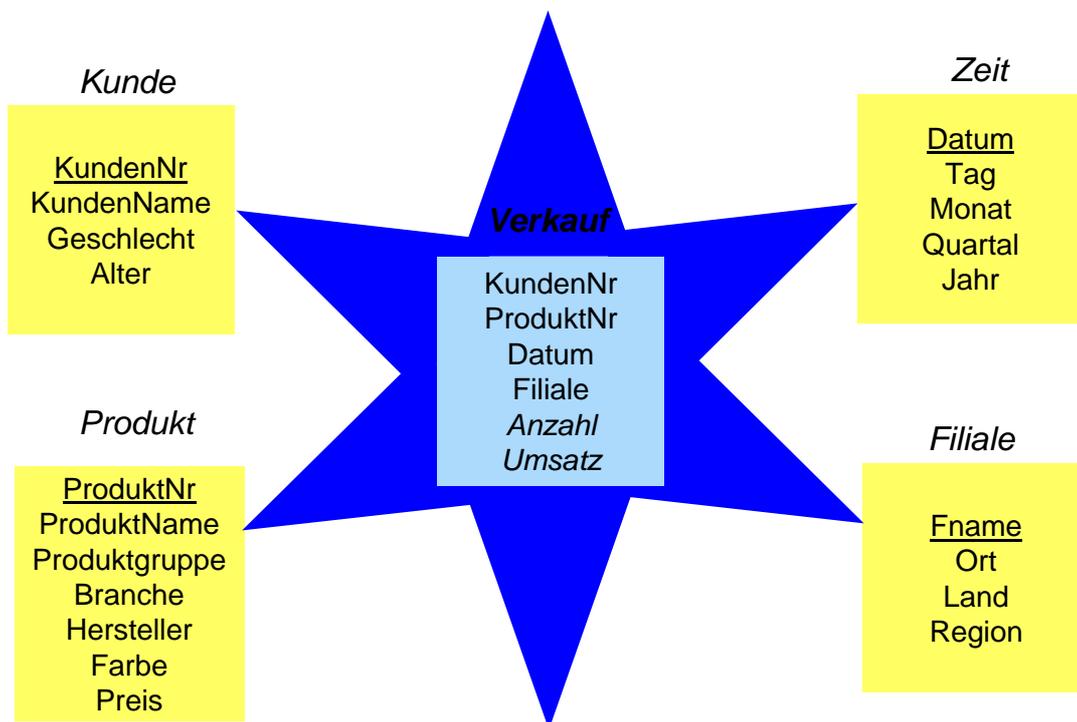
© Prof. E. Rahm

1 - 17



Star-Schema

- zentrale Faktentabelle sowie 1 Tabelle pro Dimension



© Prof. E. Rahm

1 - 18



Anfragen

Beispielanfrage: Welche Auto-Hersteller wurden von weiblichen Kunden in Sachsen im 1. Quartal 2006 favorisiert?

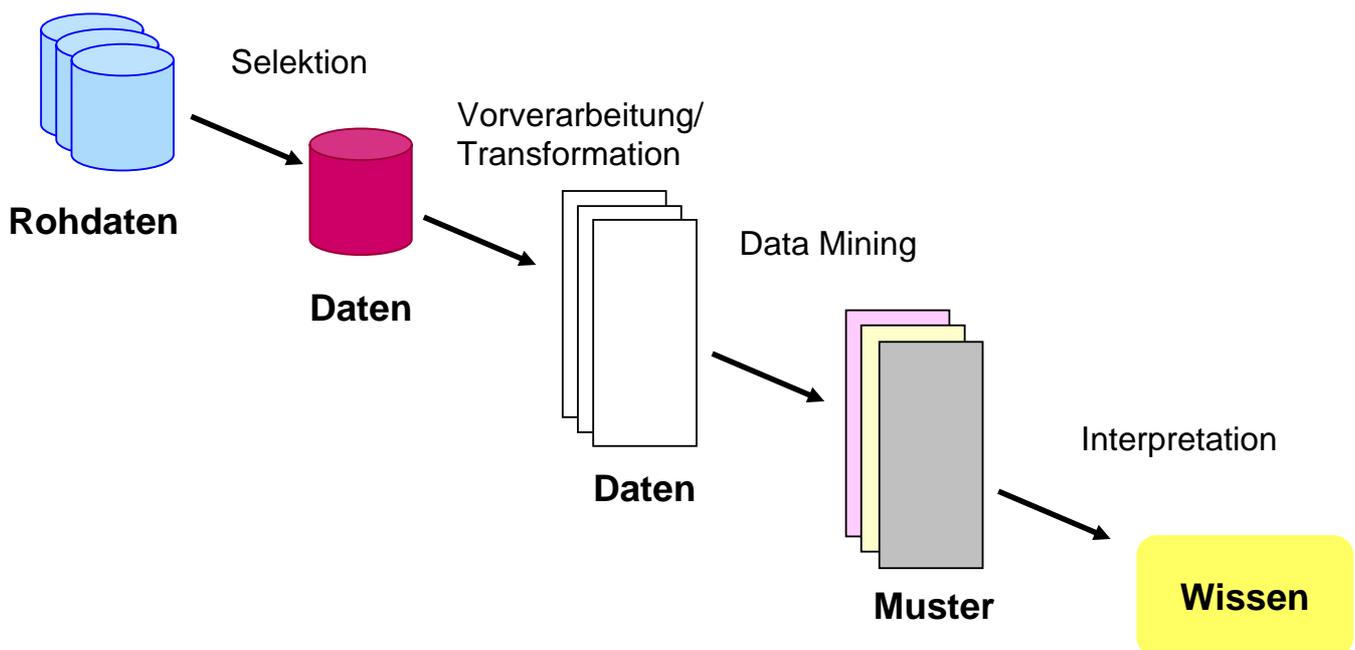
```
select p.Hersteller, sum (v.Anzahl)
from Verkauf v, Filialen f, Produkt p, Zeit z, Kunden k
where z.Jahr = 2006 and z.Quartal = 1 and k.Geschlecht = 'W' and
p.Produkttyp = 'Auto' and f.Land = 'Sachsen' and
v.Datum = z.Datum and v.ProduktNr = p.ProduktNr and
v.Filiale = f.FName and v.KundenNr = k.KundenNr
group by p.Hersteller;
```

■ Star-Join

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



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



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 in **Sie möchten dieses Produkt morgen bis 12 Uhr geliefert bekommen** 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 /

- ◆ [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

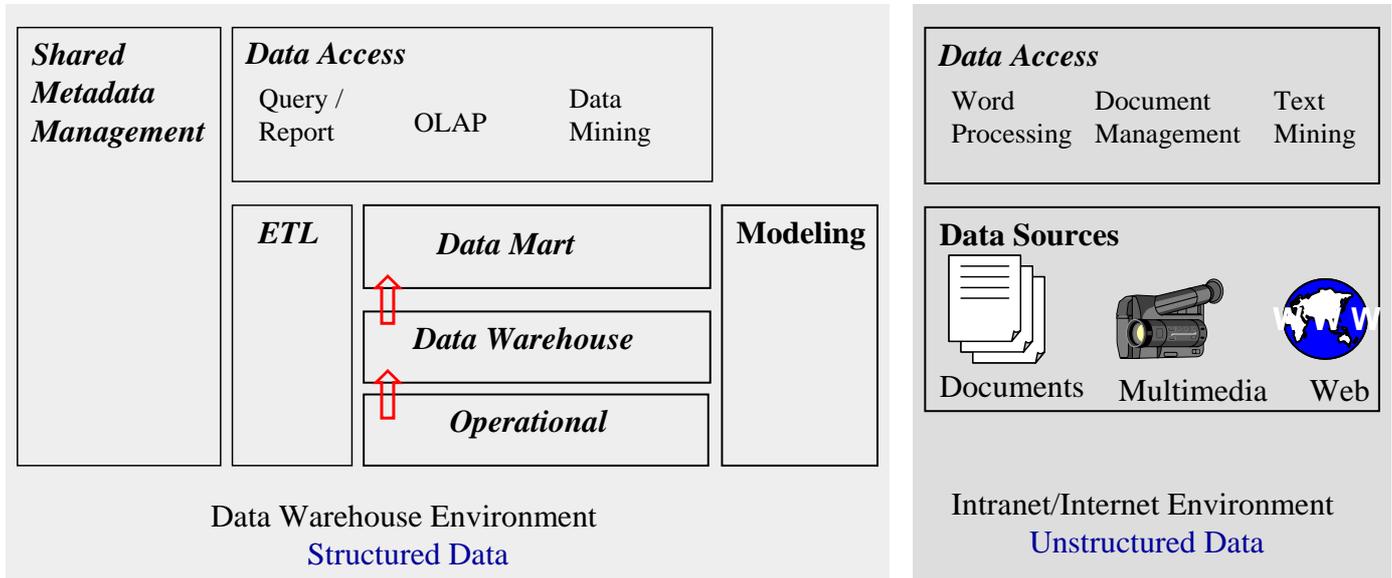
gekauft:



Enterprise Information Portale

- einheitlicher unternehmensweiter Zugang zu strukturierten und unstrukturierten Daten

Enterprise Information Portal



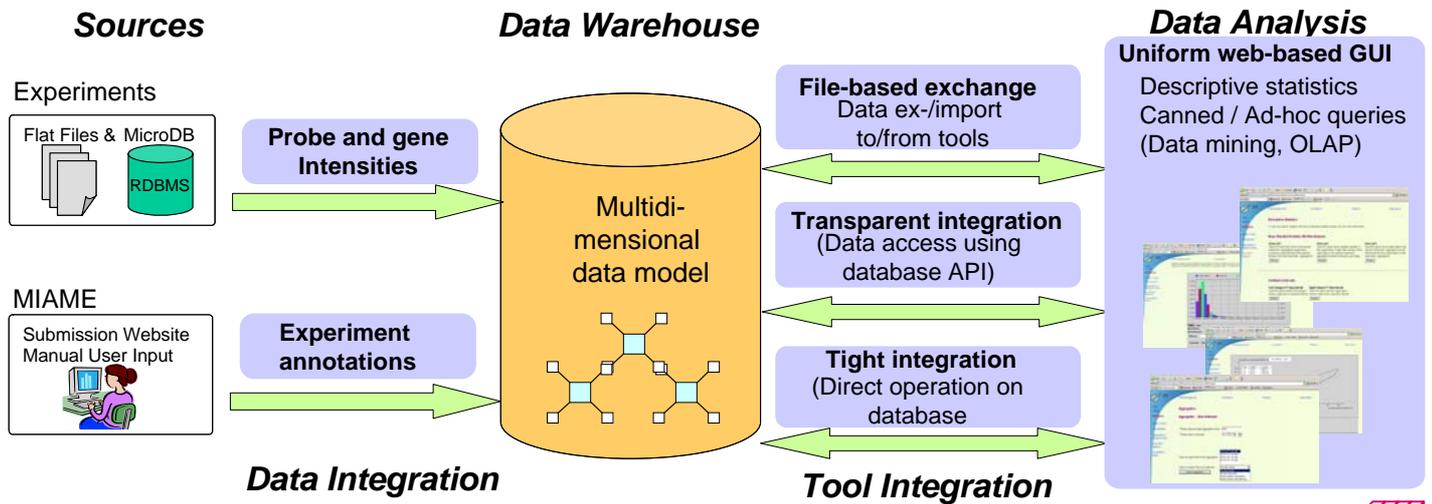
Data Warehouse Hype & Reality

- “turning data into knowledge”
 - “360° view of customer”
 - “a single version of the truth”
- How’s your data quality?
-
- “getting you closer to the customer”
- Is the information being leveraged? How?
-
- “Better decision making”
- Better access to important info to make decisions



GeWare: Expression Data Warehouse*

- Verwaltung und Analyse großer Mengen von Genexpressionsdaten (Microarray-Messungen)
- Einheitliche Spezifikation der Experimentannotationen
- Integration weiterer Informationen zu Genen etc.



*Kirsten, T., H.-H. Do, E. Rahm: *A Multidimensional Data Warehouse for Gene Expression Analysis*. GCB 2003
© Prof. E. Rahm 1 - 25



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*

