

UNIVERSITÄT
LEIPZIG

DATA WAREHOUSING

Kapitel 1: Einführung

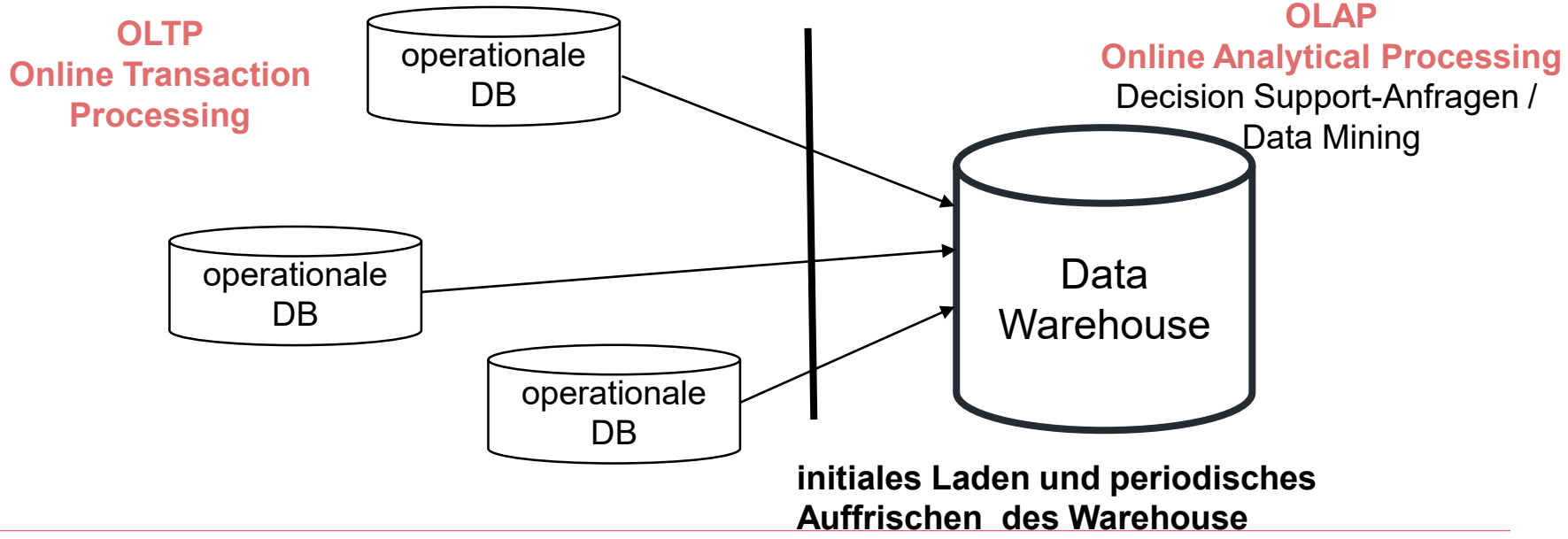
Victor Christen
Abteilung Datenbanken
Institut für Informatik
Universität Leipzig

AGENDA

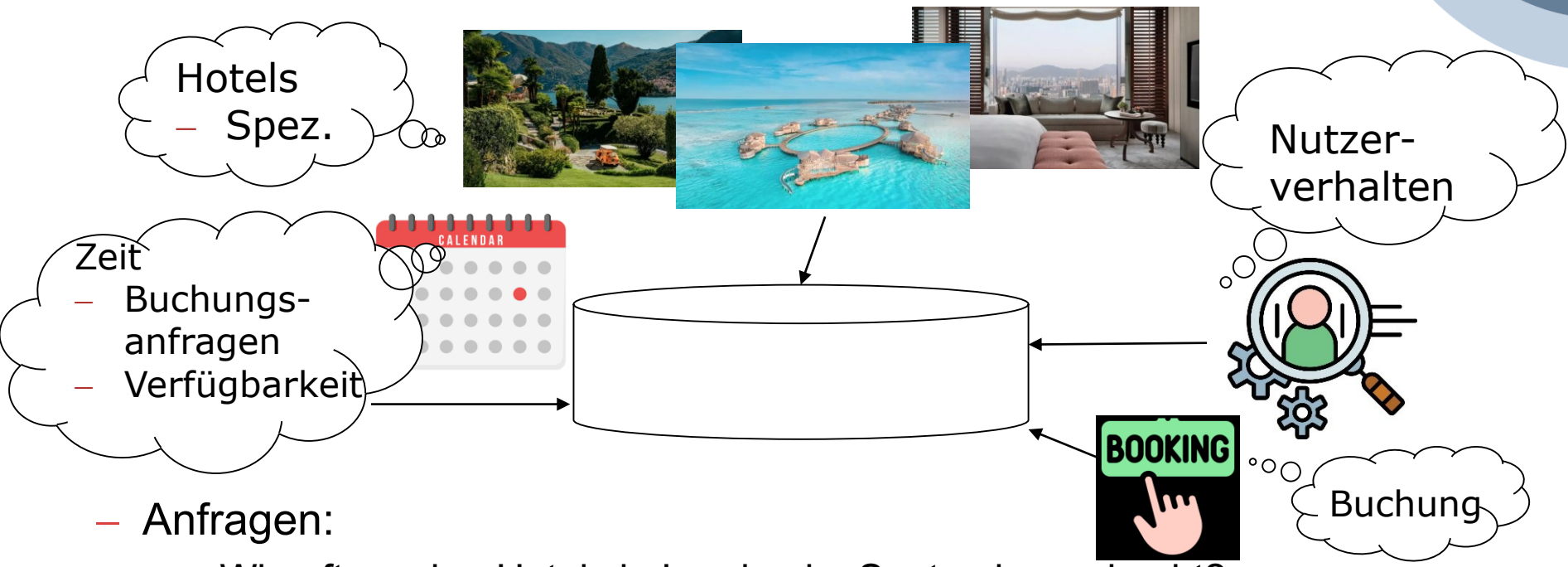
- Definitionen und Merkmale
  - Grobdefinition
  - Einsatzbeispiele
  - DW-Merkmale nach Imnon
  - OLTP vs. OLAP
  - Grobarchitektur
  - virtuelle vs. physische Datenintegration
- Mehrdimensionale Datensicht
  - Stern-Schema und -Anfragen
- Analysearten (OLAP, Data Mining)
- Big Data

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: REISEPORTAL



- Anfragen:
  - Wie oft wurden Hotels in London im September gebucht?
  - Wie hat sich die Bewertung des Adlon in Berlin über die 10 Jahre entwickelt
  - Welche Hotels sind ähnlich bzgl. Hotelausstattung in Barcelona?
  - Welche Region hat von Juni bis August nur 10 Hotels verfügbar?
  - Wie wirkte sich die neue Suche auf die durchschnittliche Zeit von der Suche bis zur Buchung aus?

EINSATZBEISPIELE

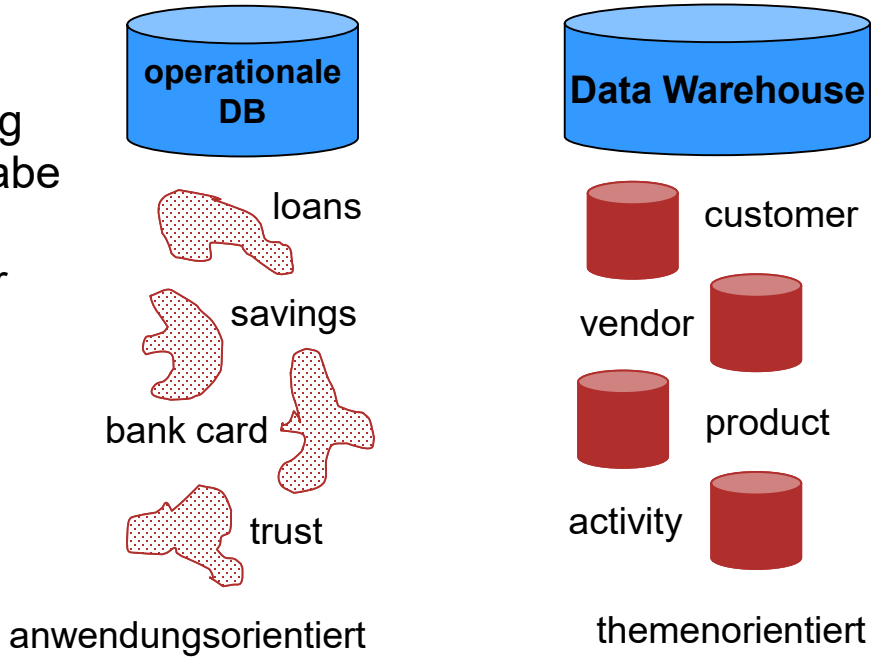
- Warenhauskette
  - Verkaufszahlen und Lagerbestände aller Warenhäuser
  - mehrdimensionale Analysen: Verkaufszahlen nach Produkten, Regionen, Warenhäusern
  - Ermittlung von Kassenschlagern und Ladenhütern
  - Analyse des Kaufverhaltens von Kunden (Warenkorbanalyse)
  - Erfolgskontrolle von Marketing-Aktivitäten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

EINSATZBEISPIELE (2)

- Versicherungsunternehmen
  - Bewertung von Filialen, Vertriebsbereichen, Schadensverlauf, ...
  - automatische Risikoanalyse
  - schnellere Entscheidung über Kreditkarten, Lebensversicherung, Krankenversicherung ...
- Banken, Versandhäuser, Restaurant-Ketten
- wissenschaftliche Einsatzfälle • • •

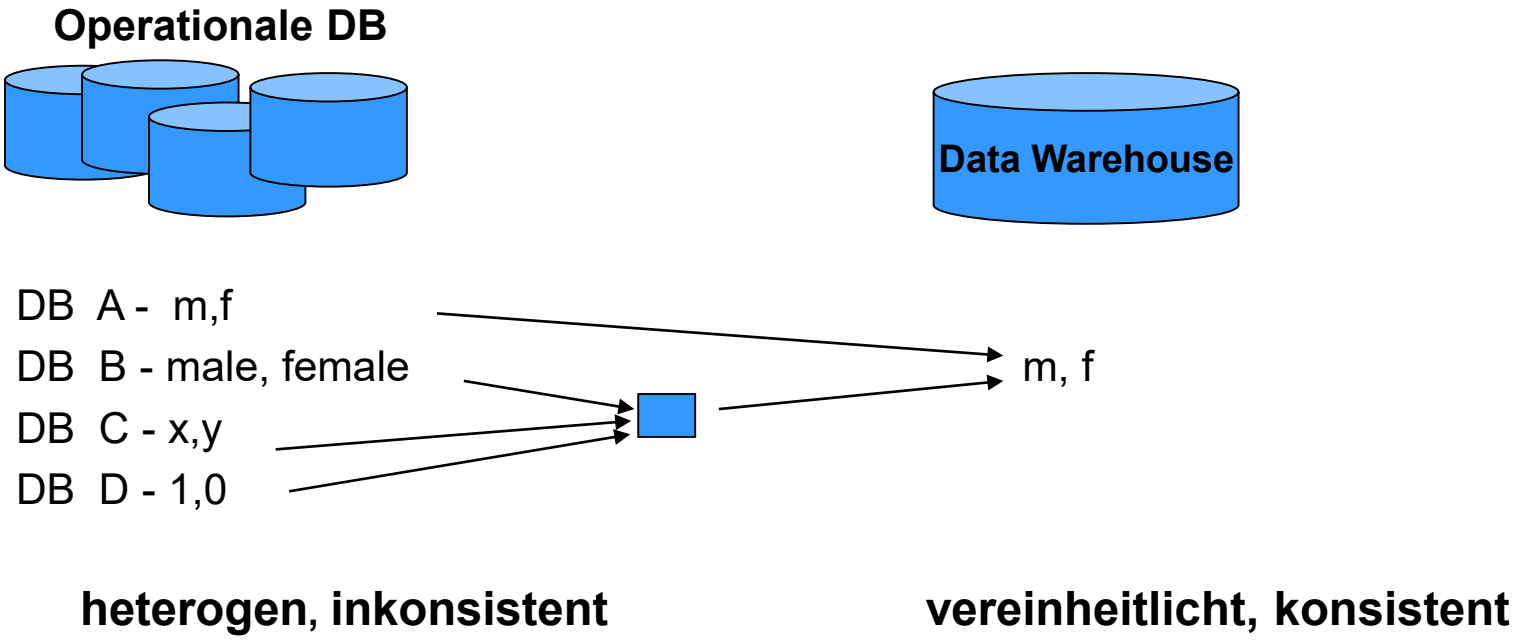
DW-EIGENSCHAFTEN NACH INMON

- A Data Warehouse is a subject-oriented, integrated, non-volatile, and time variant collection of data in support of management decisions (W. H. Inmon, Building the Data Warehouse, 1996)
- Subjekt-orientiert
  - Zweck des Systems ist nicht Erfüllung einer dedizierten operationalen Aufgabe (z.B. Personaldatenverwaltung),
  - alle Daten - unternehmensweit - über ein Subjekt (Kunden, Produkte, Regionen ...) und nicht „versteckt“ in verschiedenen Anwendungen
  - Unterstützung übergreifender Auswertungsmöglichkeiten aus verschiedenen Perspektiven



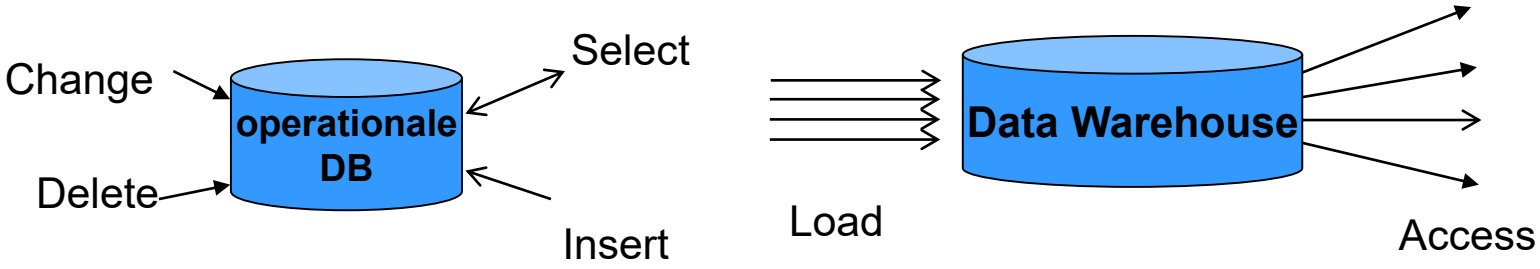
DW-EIGENSCHAFTEN NACH INMON (2)

– integrierte Datenbasis (integrated): konsolidierte Daten aus mehreren verschiedenen Datenquellen



DW-EIGENSCHAFTEN NACH INMON (3)

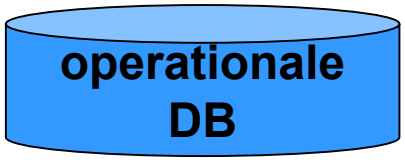
- dauerhafte Datenbasis (non-volatile):
  - Daten im DW werden i.a. nicht mehr geändert
  - stabile, persistente Datenbasis



regelmäßige Änderungen von Sätzen

DW-EIGENSCHAFTEN NACH INMON (4)

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





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	für je eine Applikation / eine Perspektive	mehrere Perspektiven / anwendungs-übergreifend
Bedeutung	Tagesgeschäft	Entscheidungs-/Planungsaufgaben
Nutzer	Sachbearbeiter, Online-Nutzer	Analysten / Manager
Datenzugriff	sehr häufiger Zugriff, kleine Datenmengen pro Operation, Lesen, Schreiben, Modifizieren, Löschen	moderate Zugriffsfrequenz, große Datenmengen, vorwiegend lesender Zugriff
Änderungen	sehr häufig	periodisches Auffrischen
#Datenquellen	meist eine pro Anwendung	mehrere
Datenmerkmale	nicht abgeleitet, autonom, zeitaktuell, dynamisch	abgeleitet, integriert, i.a. leicht veraltet, stabil
Optimierungsziele	hoher Durchsatz, sehr kurze Antwortzeiten (ms .. s), hohe Verfügbarkeit	gute Antwortzeiten für komplexe Analysen

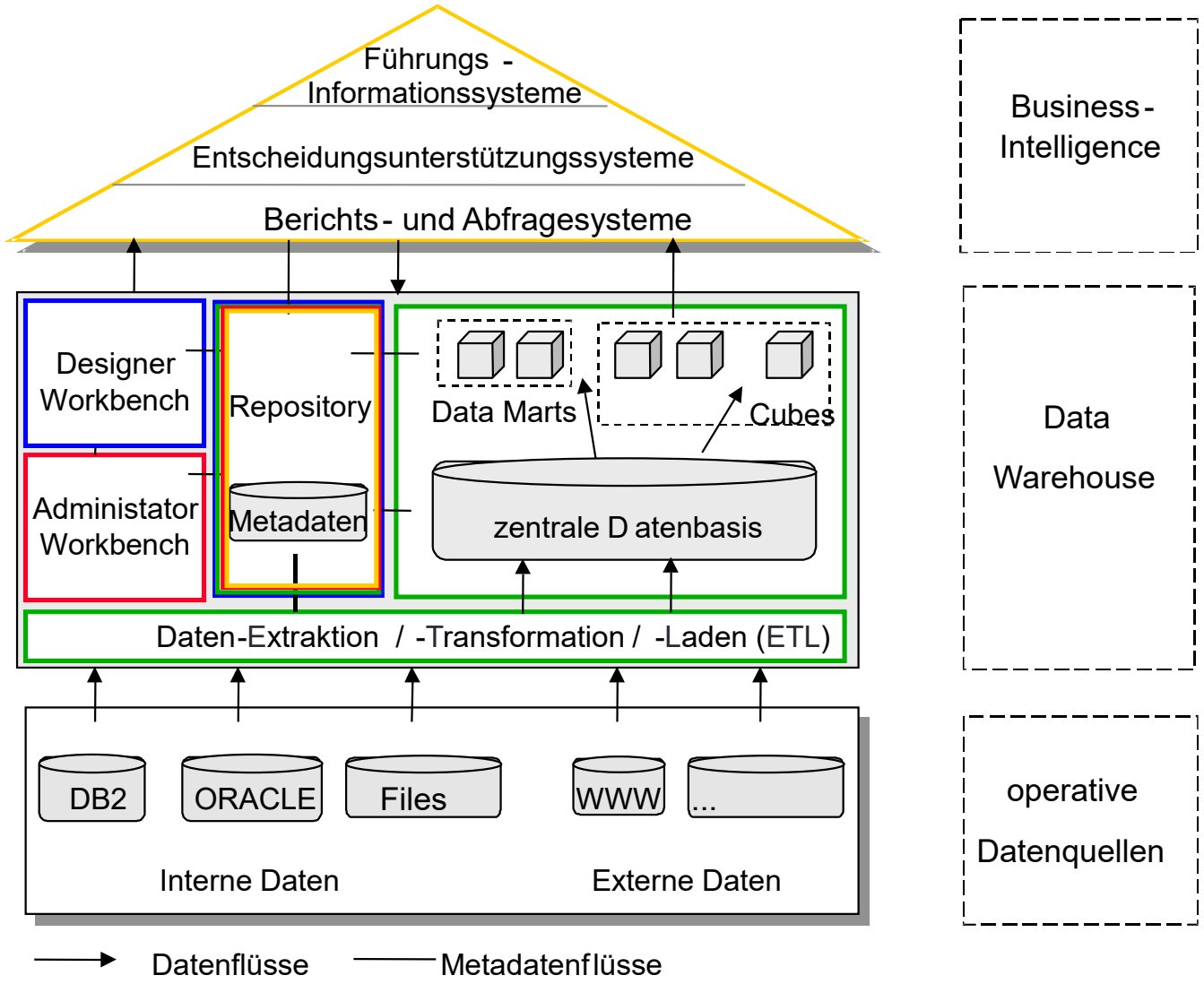
WARUM SEPARATES DATA WAREHOUSE?

- unterschiedliche Nutzung und Datenstrukturierung
- unterschiedliche Funktionalität
  - historische Daten
  - Konsolidierung (Integration, Bereinigung und Aggregation) von Daten aus heterogenen Datenquellen
- Performance
  - OLTP optimiert für kurze Transaktionen und bekannte Lastprofile
  - komplexe OLAP-Anfragen würden gleichzeitige OLTP-Transaktionen ausbremsen
  - OLAP erfordert speziellen logischen / physischen DB-Entwurf für mehrdimensionale Anfragen
  - Transaktionseigenschaften (ACID) für OLAP weniger wichtig

WARUM SEPARATES DATA WAREHOUSE?

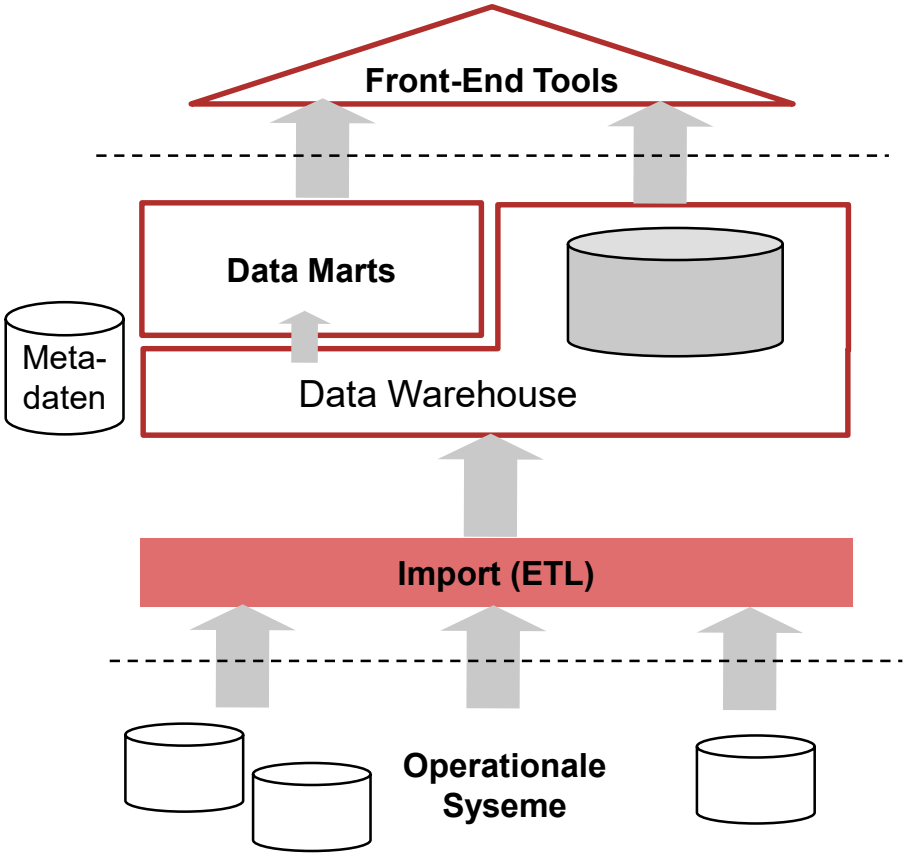
- Sicherheit
- Nachteile der separaten Lösung
  - Datenredundanz
  - Daten nicht vollständig aktuell
  - hoher Administrationsaufwand
  - hohe Kosten

GROBARCHITEKTUR EINER DW-UMGEBUNG

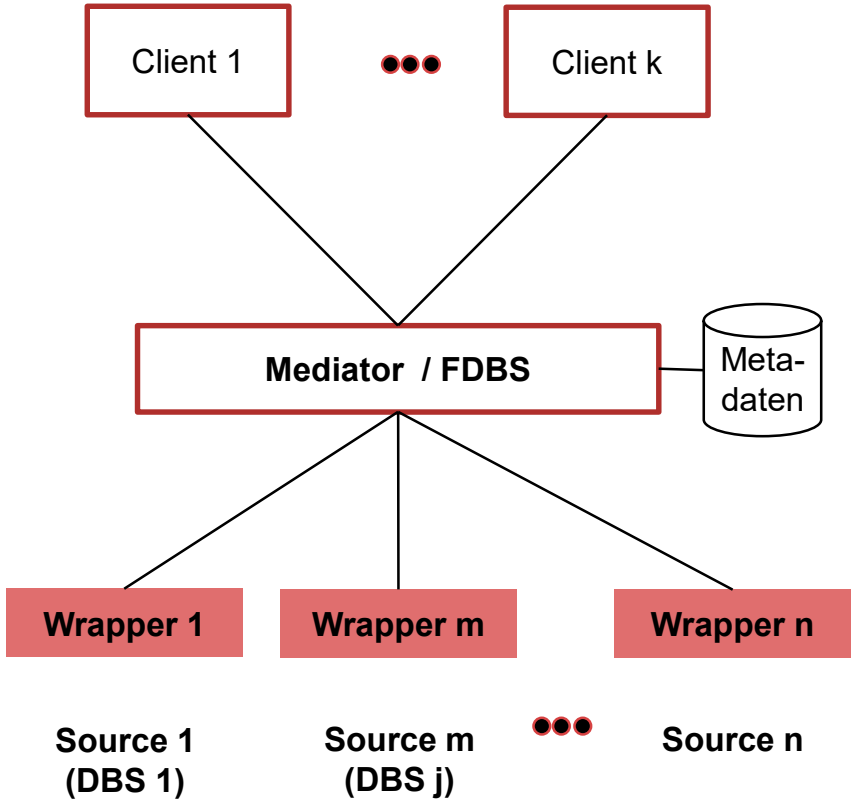


DATENINTEGRATION: PHYSISCH VS. VIRTUELL

Physische (Vor-) Integration
(Data Warehousing)



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



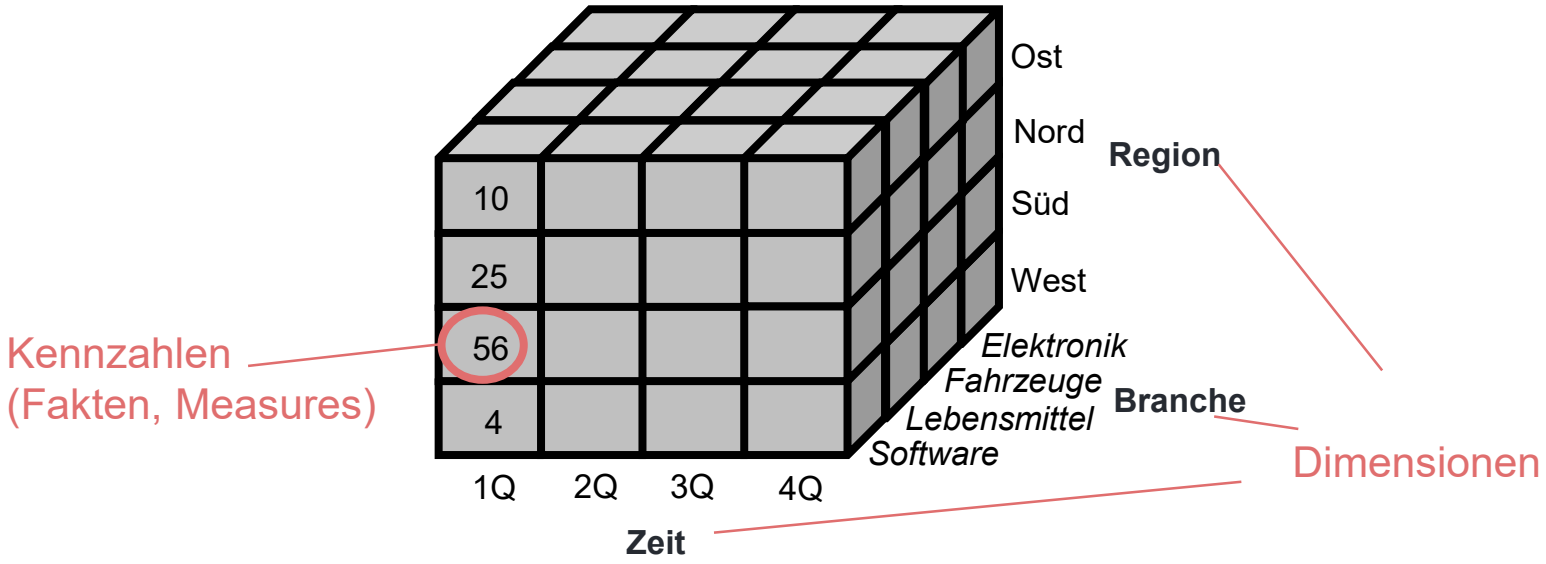
DATENINTEGRATION: PHYSISCH VS. VIRTUELL (2)

	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		

OLAP (ONLINE ANALYTICAL PROCESSING)

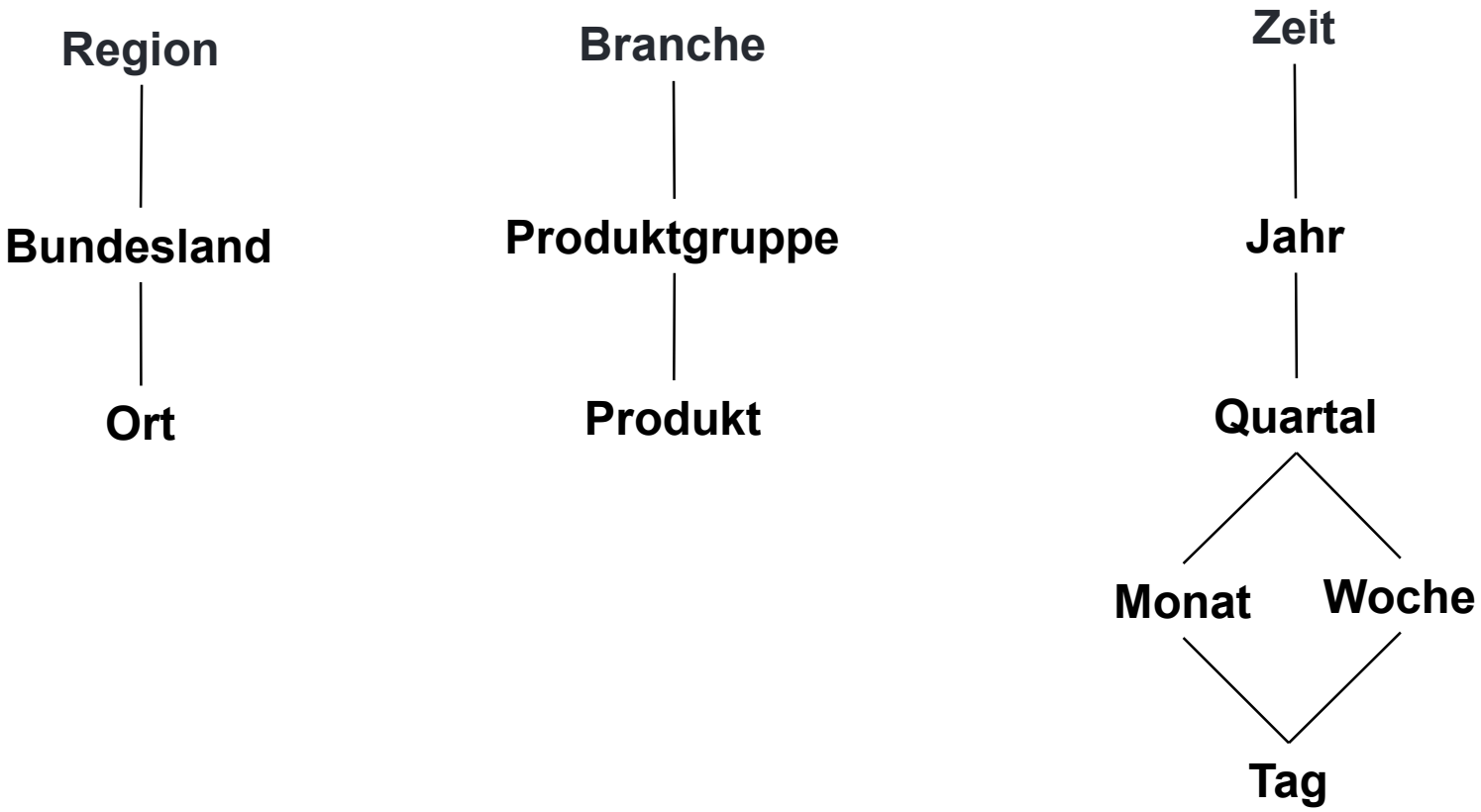
- interaktive multidimensionale Analyse auf konsolidierten Unternehmensdaten
- FASMI-Merkmale (Fast Analysis of Shared Multidimensional Information)
  - Skalierbarkeit auf große Datenmengen
    - stabile, volumenunabhängige Antwortzeiten
  - intuitive, interaktive Analyse und Visualisierung
  - Mehrbenutzerunterstützung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

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

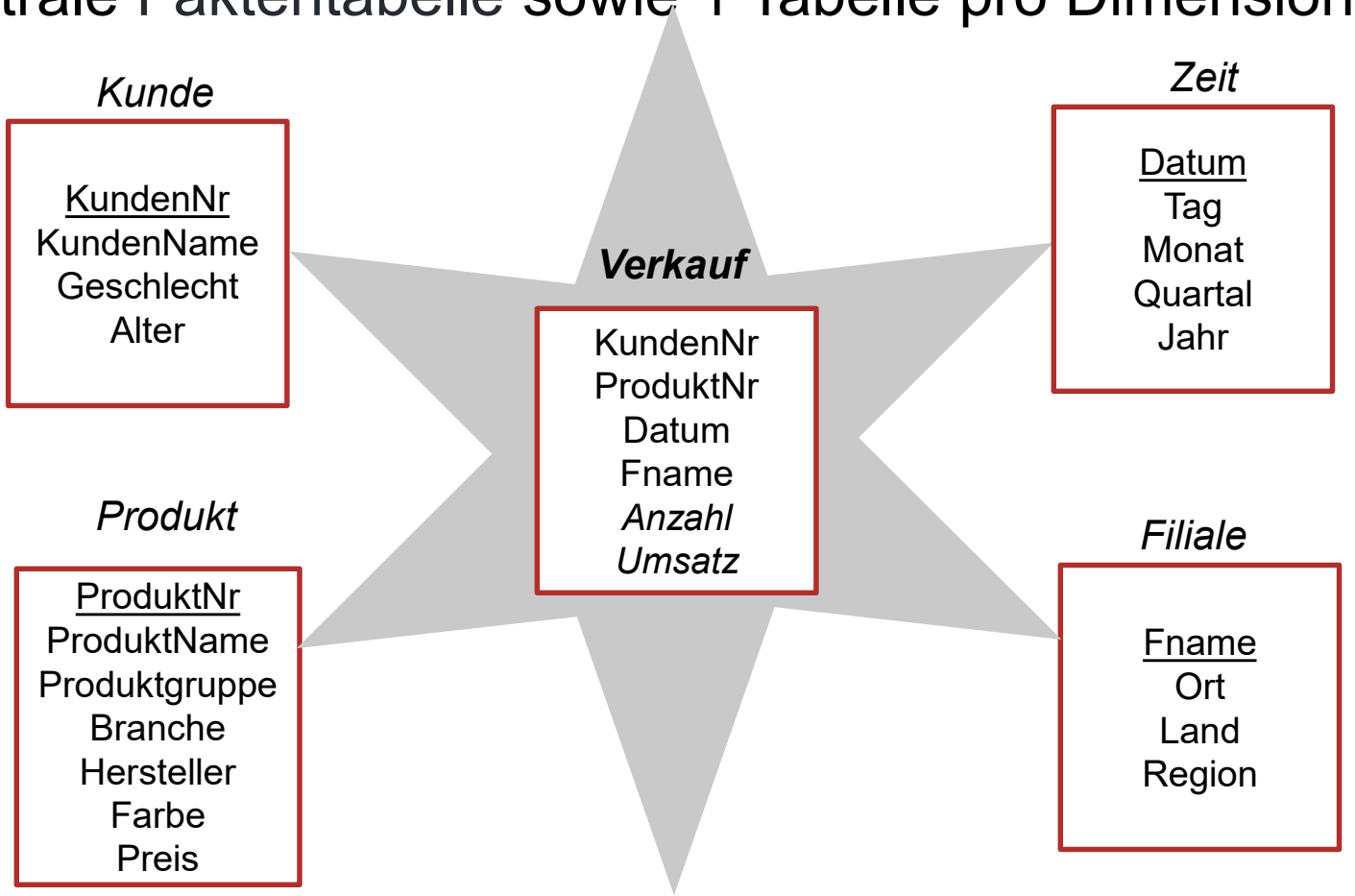
HIERARCHISCHE DIMENSIONIERUNG



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

STAR-SCHEMA

– zentrale Faktentabelle sowie 1 Tabelle pro Dimension



ANFRAGEN

Star-Join

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

Beispielanfrage:

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

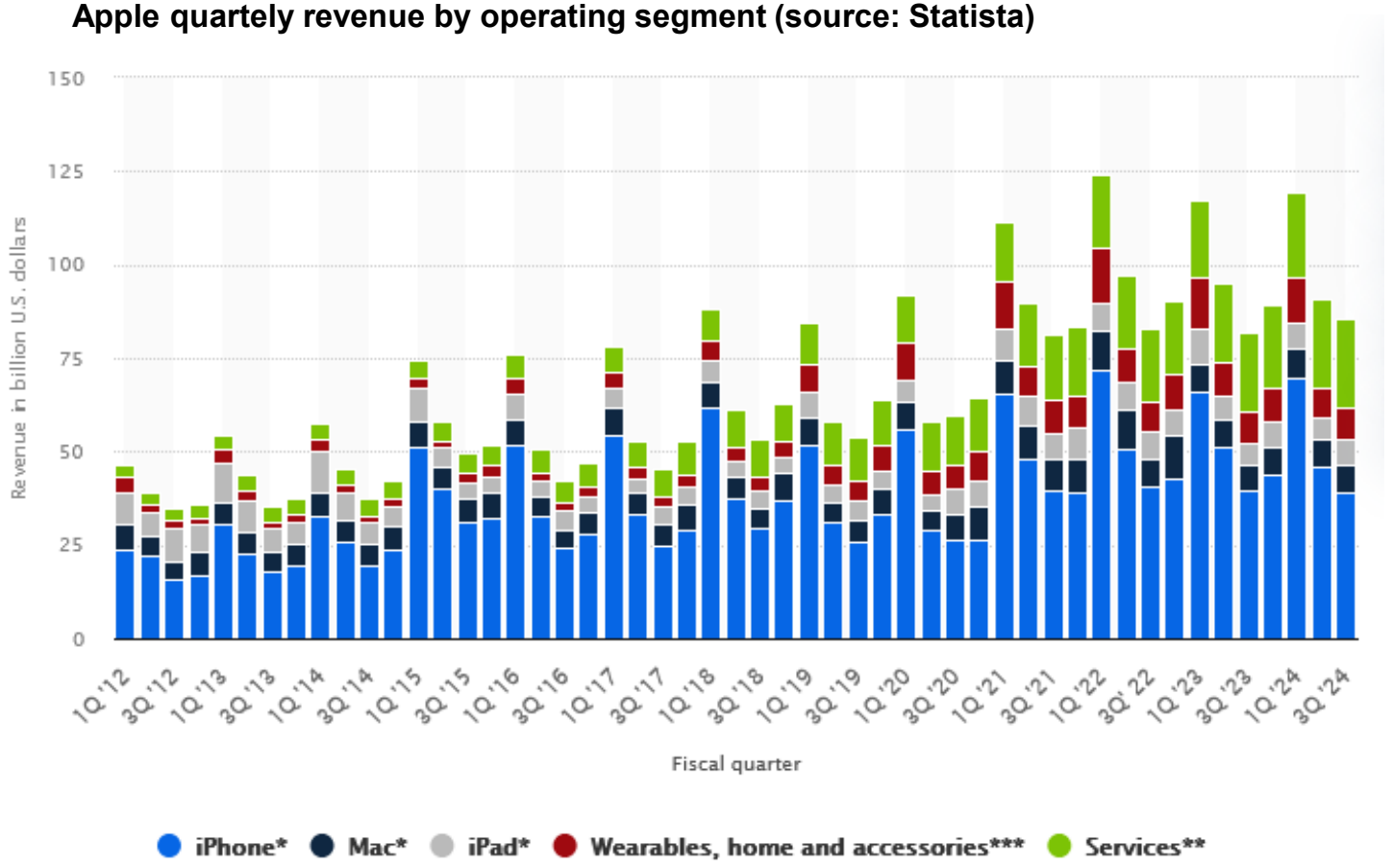
```
SELECT p.Hersteller, SUM (v.Anzahl)
FROM Verkauf v, Filialen f,
Produkt p, Zeit z, Kunden k
WHERE z.Quartal = 1 and
k.Geschlecht = 'W' and
p.Produkttyp = 'Auto' and
f.Land = 'Sachsen' and
v.Datum = z.Datum and
v.ProduktNr = p.ProduktNr and
v.Filiale = f. FName and
v.KundenNr = k.KundenNr
GROUP BY p.Hersteller
ORDER BY 2 desc;
```

ANALYSEWERKZEUGE

- (Ad Hoc-) Query-Tools
- Reporting-Werkzeuge, Berichte mit flexiblen Formatierungsmöglichkeiten
- OLAP-Tools
  - OLAP-Unterstützung in Speadsheet-Tools bzw. im Web-Browser
  - oft Datendarstellung als Pivot-Tabellen (Kreuztabellen)
  - interaktive mehrdimensionale Analyse und Navigation (Drill Down, Roll Up, ...)
  - Gruppierungen, statistische Berechnungen,
  - unterschiedlichste Visualisierungen
- Tools/Verfahren für Data Mining und maschinelles Lernen

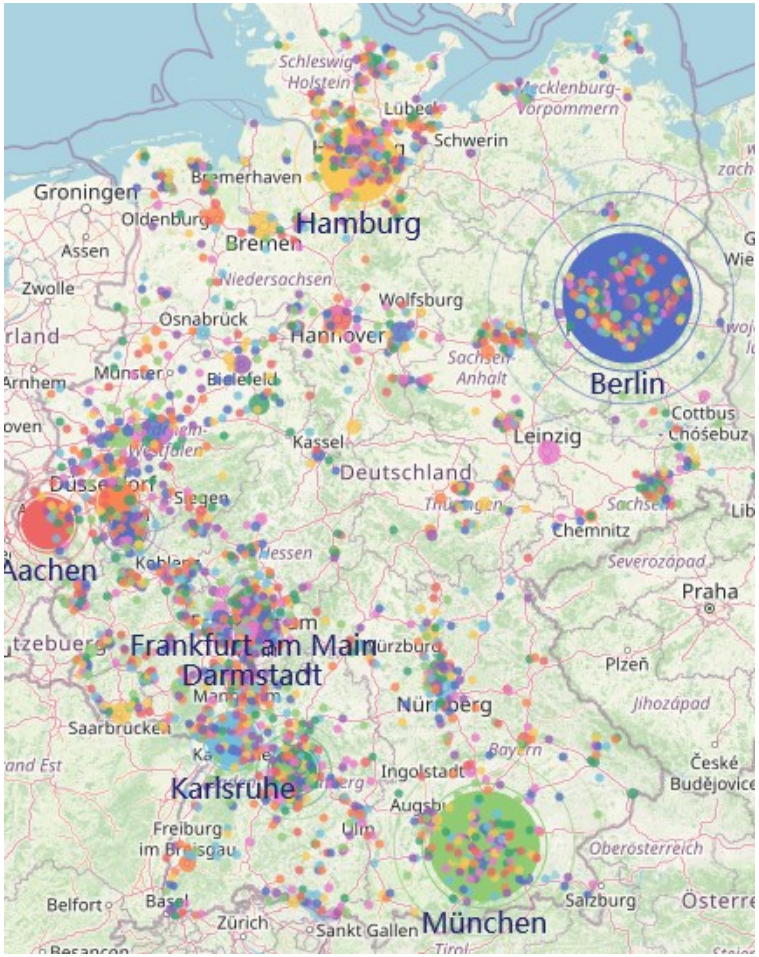
BEISPIEL: OLAP-AUSGABE

Apple quartely revenue by operating segment (source: Statista)



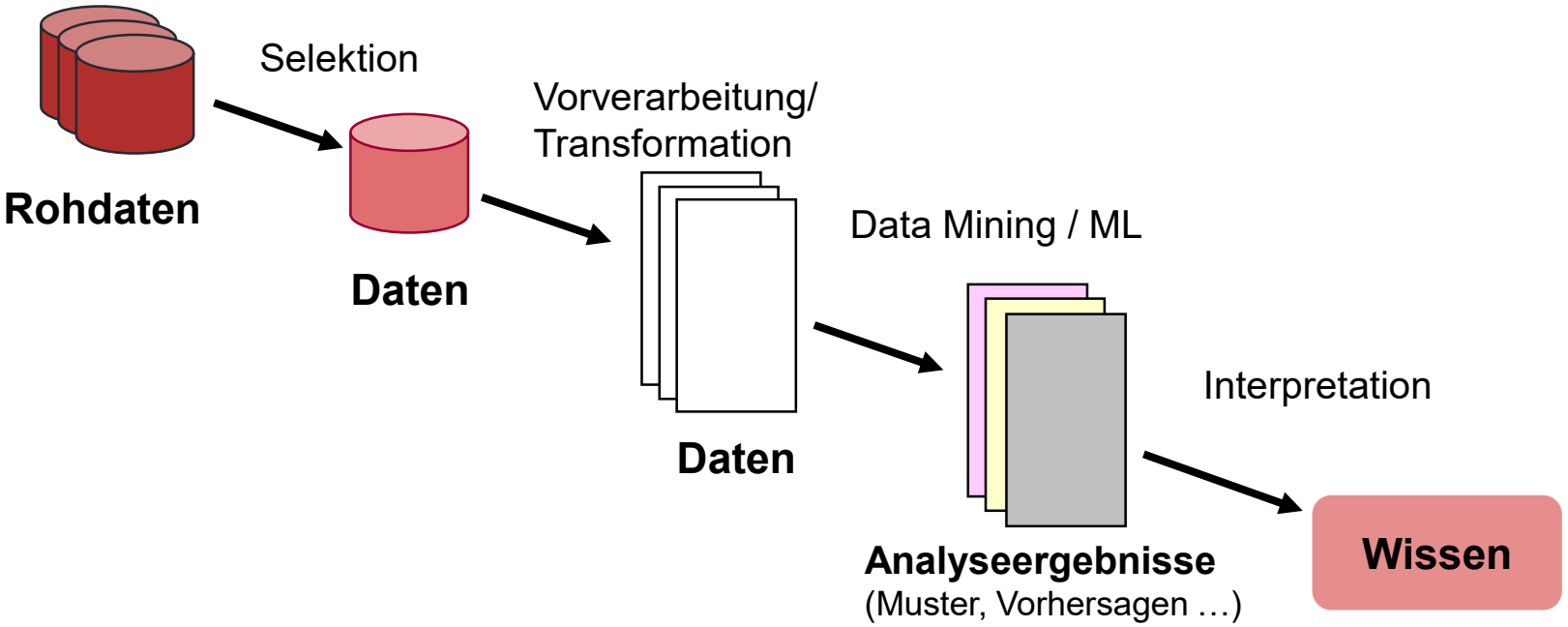
ERGEBNISVISUALISIERUNG

(MITGLIEDER IN GESELLSCHAFT FÜR INFORMATIK PRO PLZ-BEZIRK)



KNOWLEDGE DISCOVERY

- (semi-) automatische Extraktion von Wissen aus Daten
- Kombination von Verfahren zu Datenbanken, Statistik (Data Mining) und KI (maschinelles Lernen)

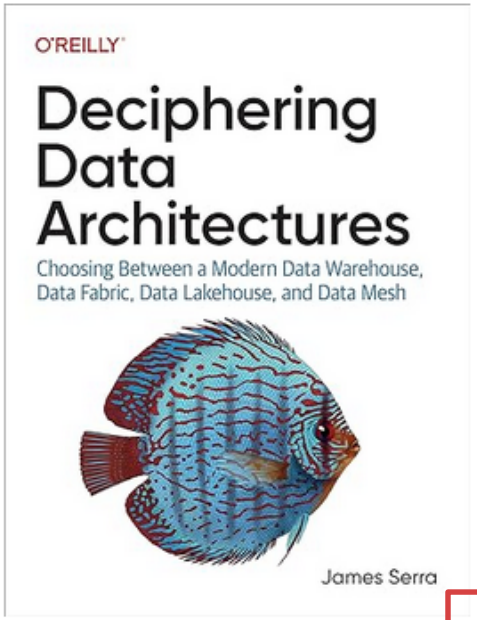


TECHNIKEN DES DATA MINING/ML

- Einsatz statistischer, wissens- und lernbasierter Methoden zur Datenanalyse
  - Auffinden von Korrelationen, Mustern und Trends in Daten, Vorhersagen
- 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)
  - Nutzung für Kaufvorhersagen / Recommendations, …
- Klassifikation
  - Zuordnung von Objekten zu Gruppen/Klassen mit gemeinsamen Eigenschaften bzw. Vorhersage von Attributwerten
  - Verwendung von Stichproben (Trainingsdaten)
  - Ansätze: Entscheidungsbaum-Verfahren, neuronale Netze, statistische Auswertungen

BEISPIEL WARENKORBANALYSE

amazon.de



Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh Taschenbuch – 12. März 2024

Englisch Ausgabe von James Serra (Autor)

4,4 ★★★★½ 61 Sternebewertungen

Data fabric, data lakehouse, and data mesh hav
These new architectures have solid benefits, bu
book provides a guided tour of these architectu

James Serra, big data and data warehousing so
including how data warehouses have had to ev
help you achieve, as well as how to distinguish
appropriate data architecture for your needs. W

- Gain a working understanding of several dat
- Learn the strengths and weaknesses of each approach
- Distinguish data architecture theory from reality

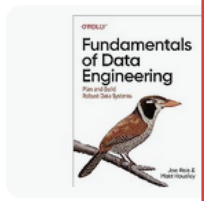
Wird oft zusammen gekauft



+



+

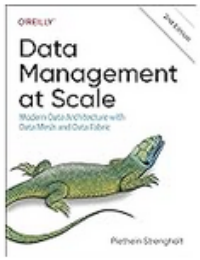


Dieser Artikel: Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Dat… 54,99 €

Data Management at Scale: Modern Data Architecture with Data Mesh and Data Fabric 51,99 €

Fundamentals of Data Engineering: Plan and Bu Robust Data Systems 46,64 € prime

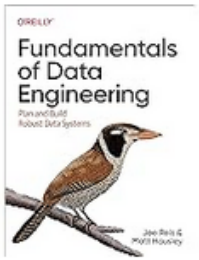
Kunden, die diesen Artikel angesehen haben, haben auch angesehen



Data Management at Scale: Modern Data Architecture with Data… › Piethein Strengholt



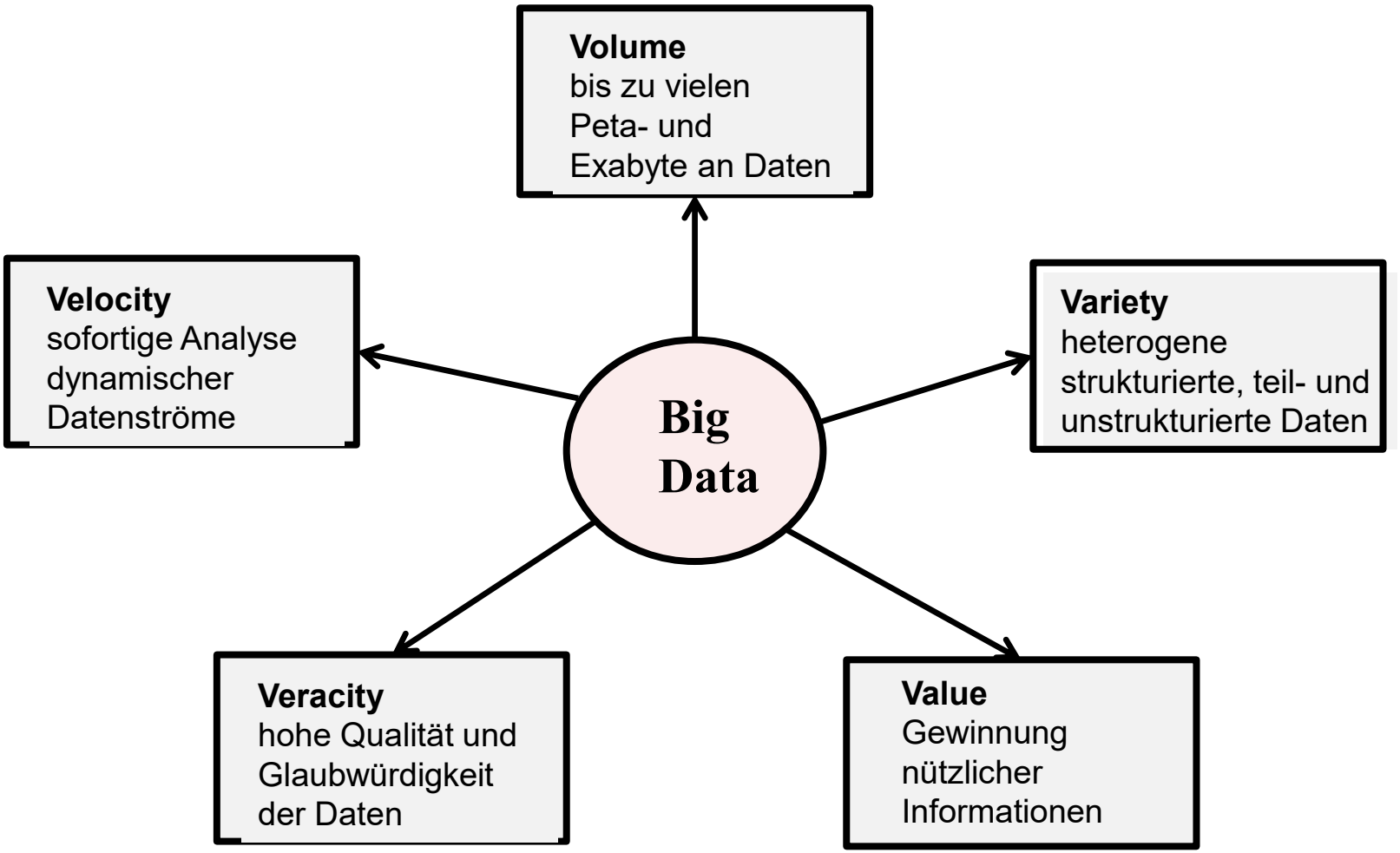
Architecting Data and Machine Learning Platforms: Enable… › Marco Tranquillin



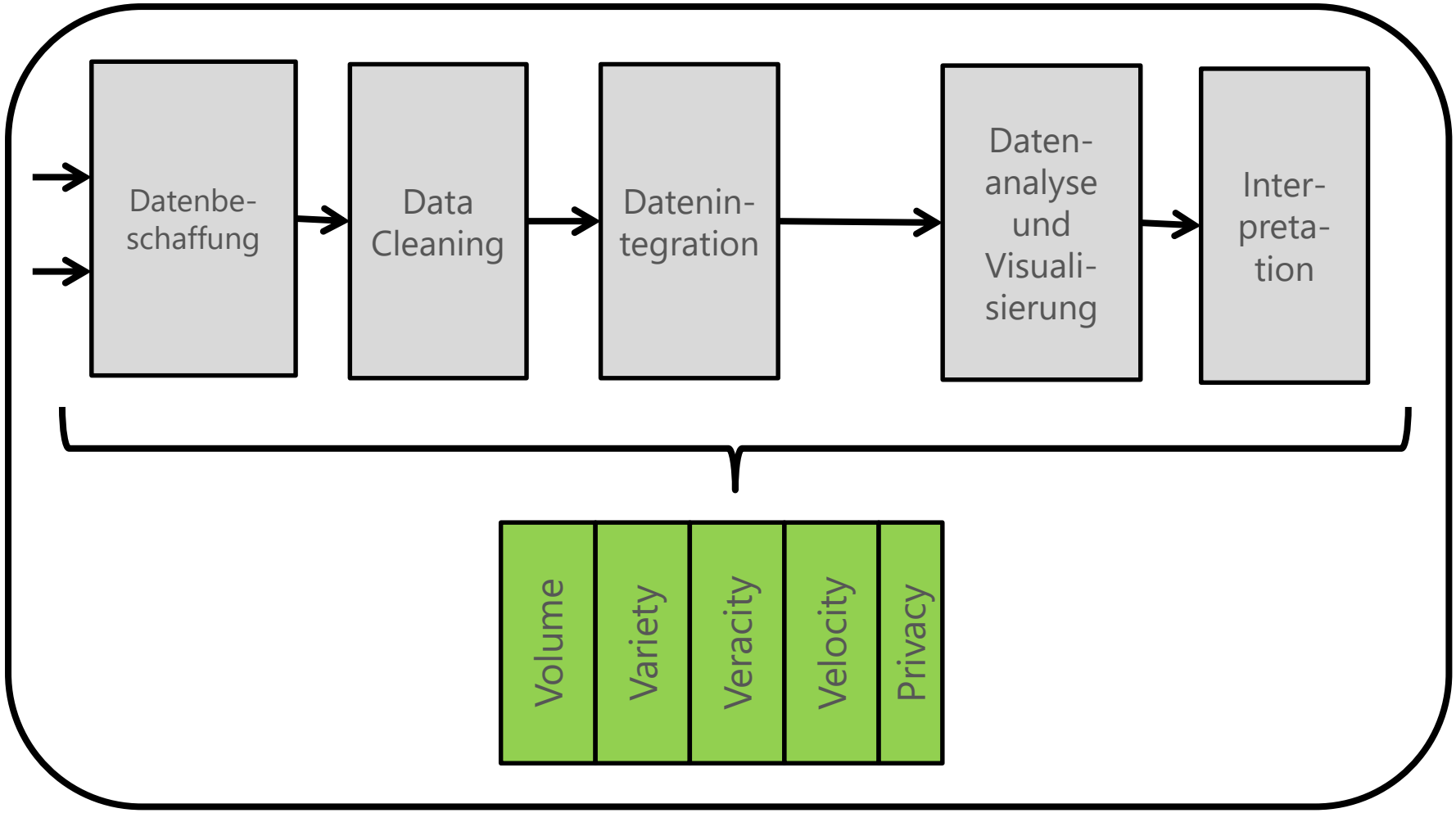
Fundamentals of Data Engineering: Plan and Build Robust Data… › Joe Reis


Kunden, die
dieses Produkt
gekauft haben
kauften
auch
DÖN MARIO
DÖN MARIO

ANFORDERUNGEN FÜR „BIG DATA“



BIG DATA ANALYSE-PIPELINE



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
  - effiziente Analyse auf großen Datenmengen
- mehrdimensionale Datenmodellierung und -organisation
- breites Spektrum an Auswertungs- und Analysemöglichkeiten
- Data Mining: selbständiges Aufspüren relevanter Muster in Daten
- Big Data: Datenanalysen auf großen Mengen auch unstrukturierter Daten