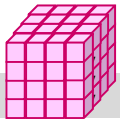
1. Data Warehouses - Einführung

- Definitionen und Merkmale
  - Grobdefinition
  - Einsatzbeispiele
  - DW-Merknmale nah Imnon
  - OLTP vs. OLAP
  - Grobarchitektur
- Virtuelle vs. physische Datenintegration

,

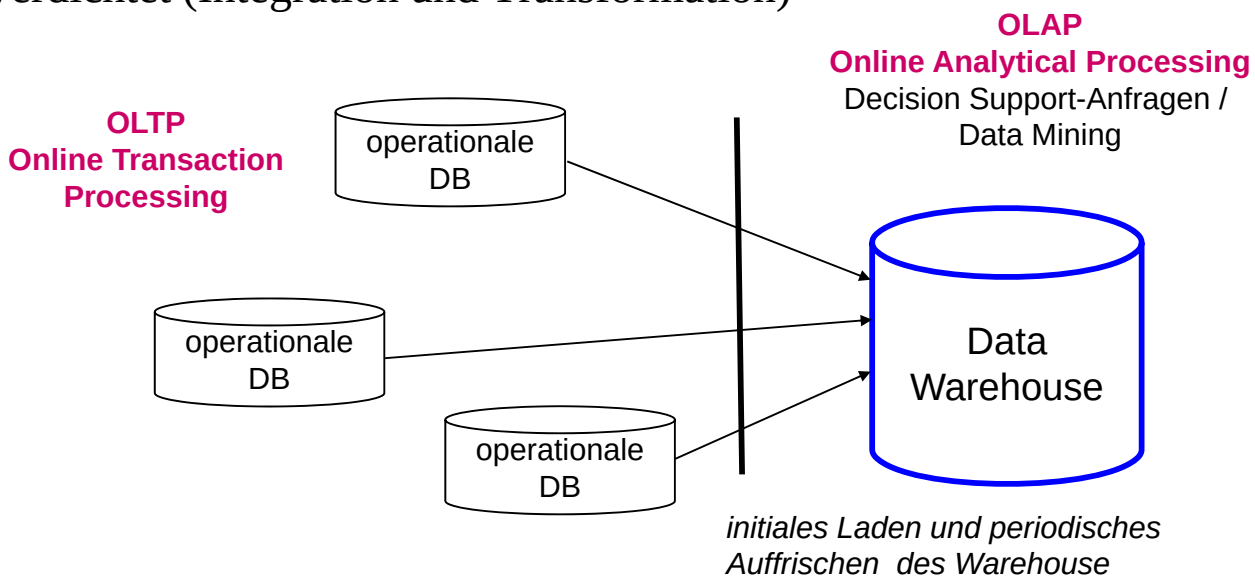
- Mehrdimensionale Datensicht
  - Stern-Schema und -Anfragen
- Data Mining
- Big Data

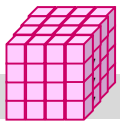


Data Warehouses

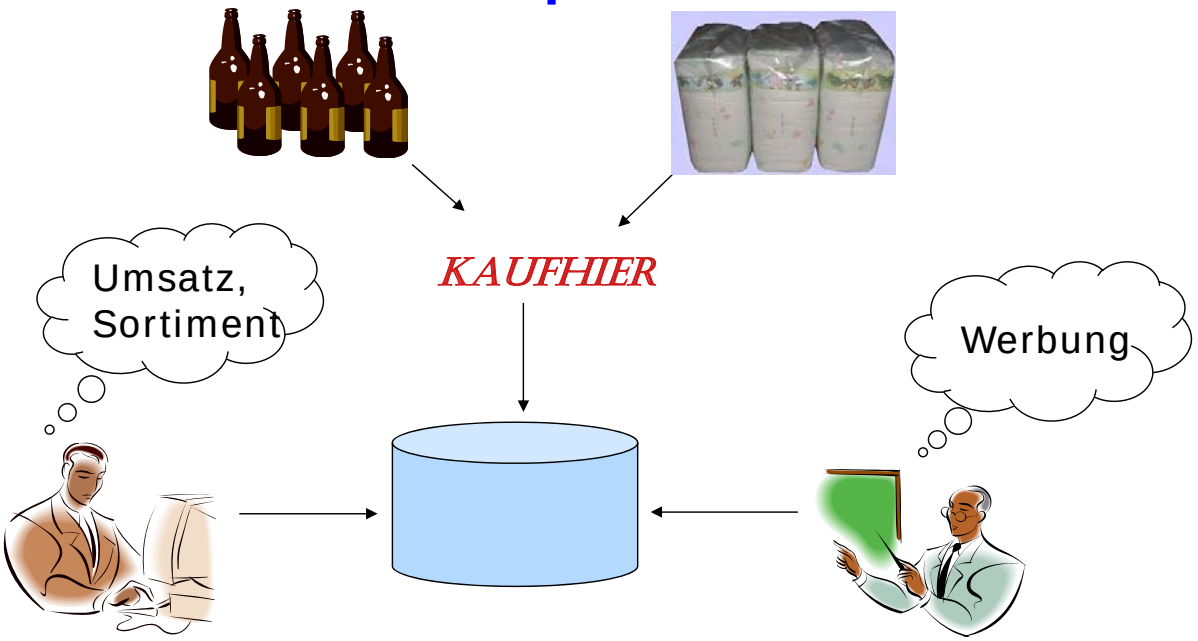
- Ausgangsproblem
  - viele Unternehmen haben Unmengen an Daten, ohne daraus ausreichend Informationen und Wissen für kritische Entscheidungsaufgaben ableiten zu können
- ***Data Warehouse (Def.):*** für Analysezwecke optimierte zentrale Datenbank, die Daten aus mehreren, i.a. heterogenen Quellen zusammenführt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

,





Szenario: Supermarktkette



- Anfragen:
  - Wie viele Pakete Windeln wurden letzten Monat verkauft?
  - Wie hat sich der Verkauf von Bier und Wasser im letzten Jahr entwickelt?
  - Wo sind unsere Top-Filialen?
  - Von welchem Lieferanten beziehen wir das meiste Bier?
  - Wie wirkten sich die Wernepreise für Produkt X aus? …

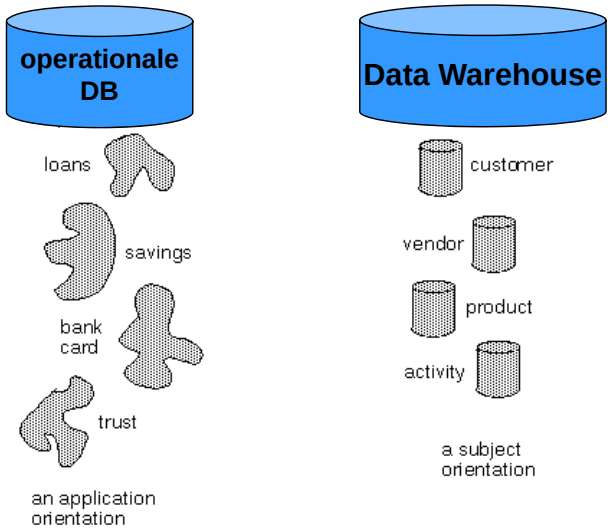
Einsatzbeispiele

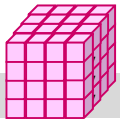
- Warenhauskette
  - Verkaufszahlen und Lagerbestände aller Warenhäuser
  - mehrdimensionale Analysen: Verkaufszahlen nach Produkten, Regionen, Warenhäusern
  - Ermittlung von Kassenschlagern und Ladenhütern
  - Analyse des Kaufverhaltens von Kunden (Warenkorbanalyse)
  - Erfolgskontrolle von Marketing-Aktivitäten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

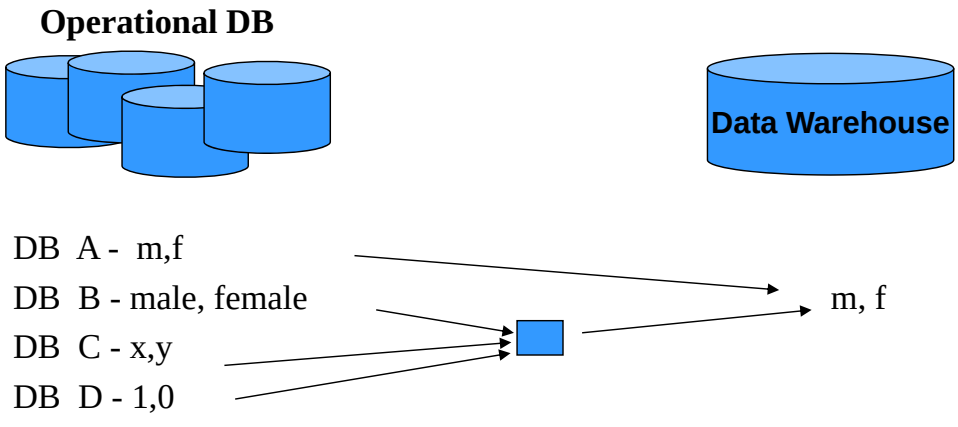
- Themenorientiert (subject-oriented):
  - Zweck des Systems ist nicht Erfüllung einer dedizierten Aufgabe (z.B. Personaldatenverwaltung), sondern Unterstützung übergreifender Auswertungsmöglichkeiten aus verschiedenen Perspektiven
  - alle Daten - unternehmensweit - über ein Subjekt (Kunden, Produkte,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

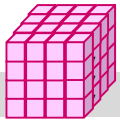




DW-Eigenschaften nach Inmon (2)

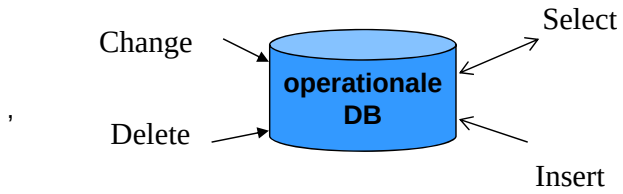
- Integrierte Datenbasis (integrated): Daten aus mehreren verschiedenen Datenquellen

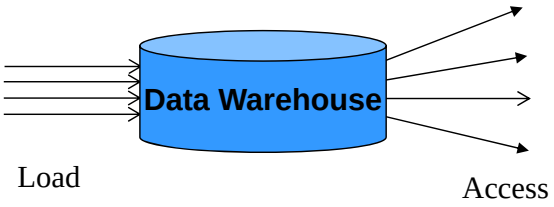




DW-Eigenschaften nach Inmon (3)

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

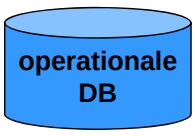




regelmäßige Änderungen von Sätzen

DW-Eigenschaften nach Inmon (4)

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



Time Variancy

aktuelle Datenwerte:

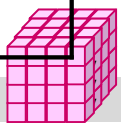
- Zeitbezug optional
- Zeithorizont: 60-90 Tage
- Daten änderbar

Schnappschuß-Daten

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

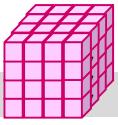
Operationale Datenbanken vs. Data Warehouses (OLTP vs. OLAP)

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

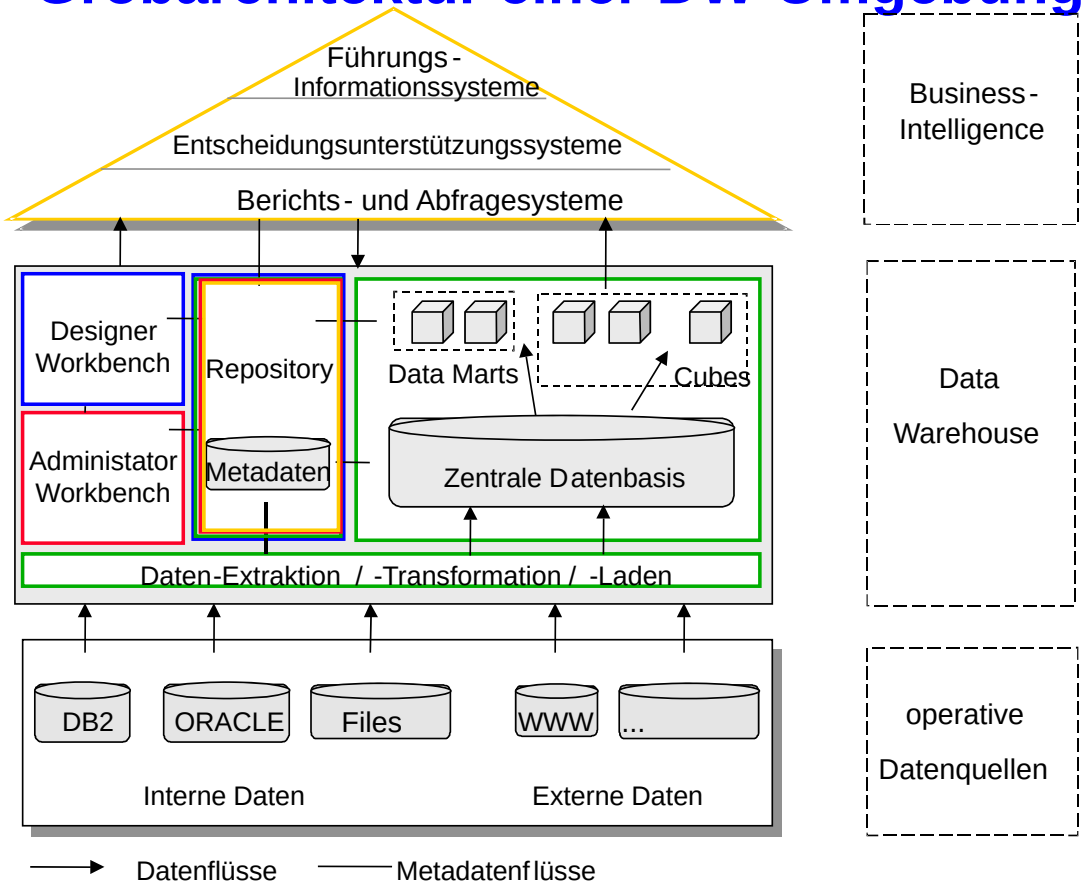


Warum separates Data Warehouse?

- 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

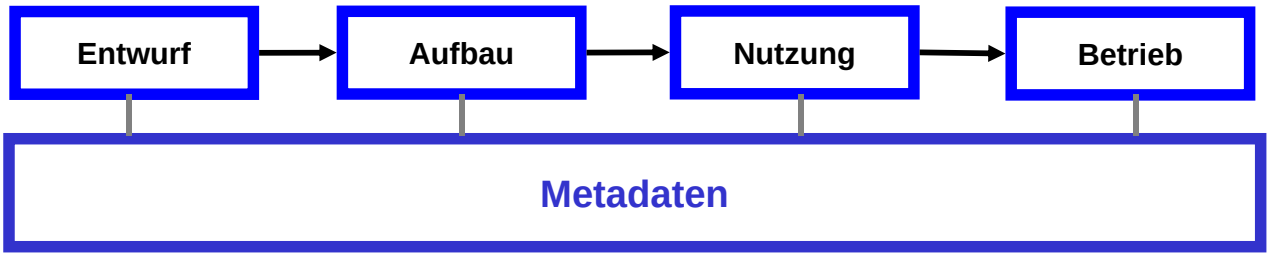


Grobarchitektur einer DW-Umgebung



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")

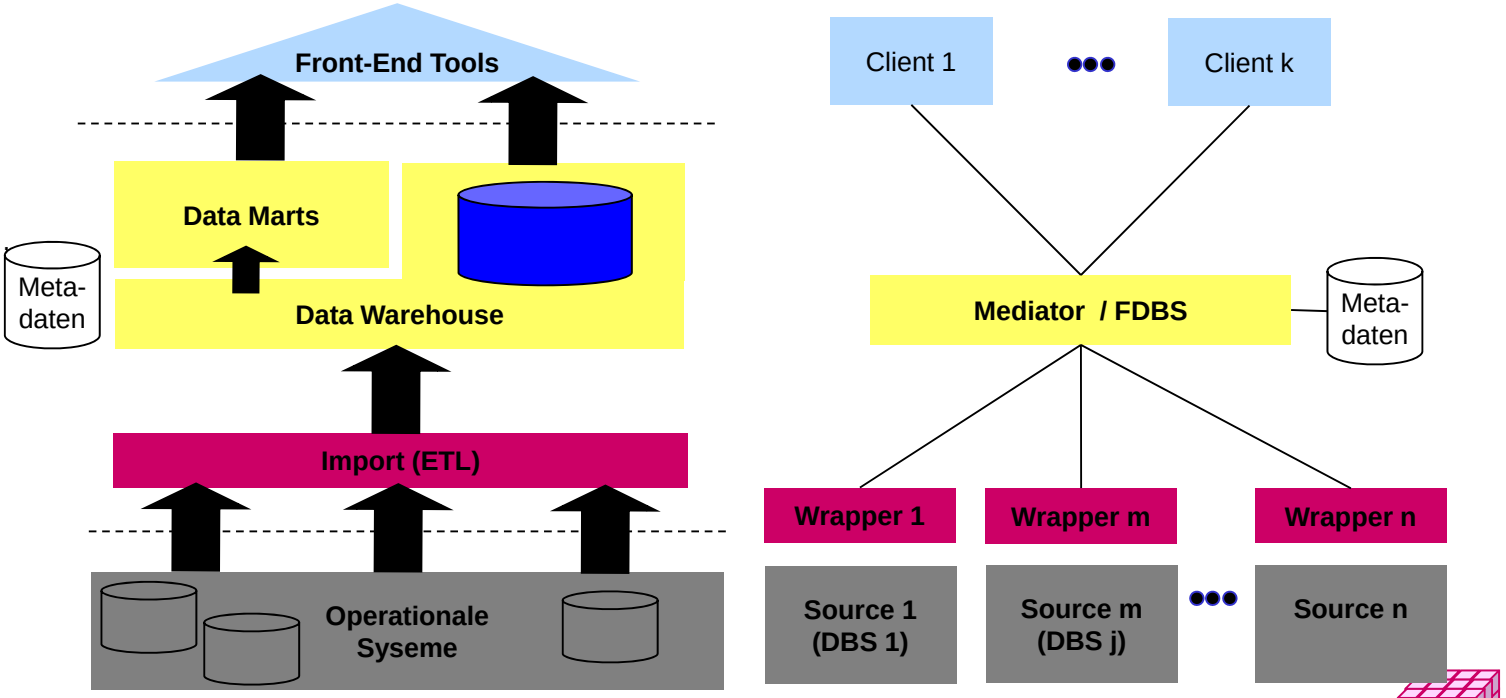


- DW ist meist kein monolithisches System
  - meist Nutzung von Tools / Komponenten unterschiedlicher Hersteller sowie eigenprogrammierten Anteilen
- zentrale Bedeutung der Metadaten, jedoch oft unzureichend unterstützt

Datenintegration: physisch vs. virtuell

Physische (Vor-) Integration
(Data Warehousing)

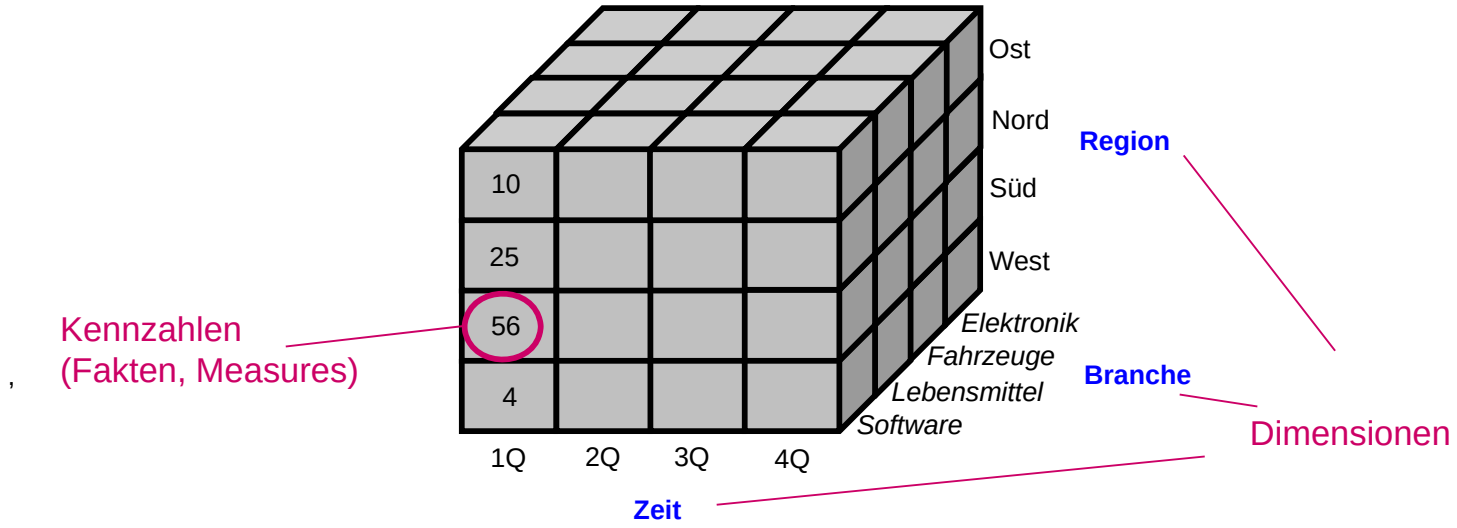
Virtuelle Integration
(Mediator/Wrapper-Architekturen, föderierte DBS)



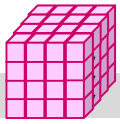
Datenintegration: physisch vs. virtuell (2)

	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		

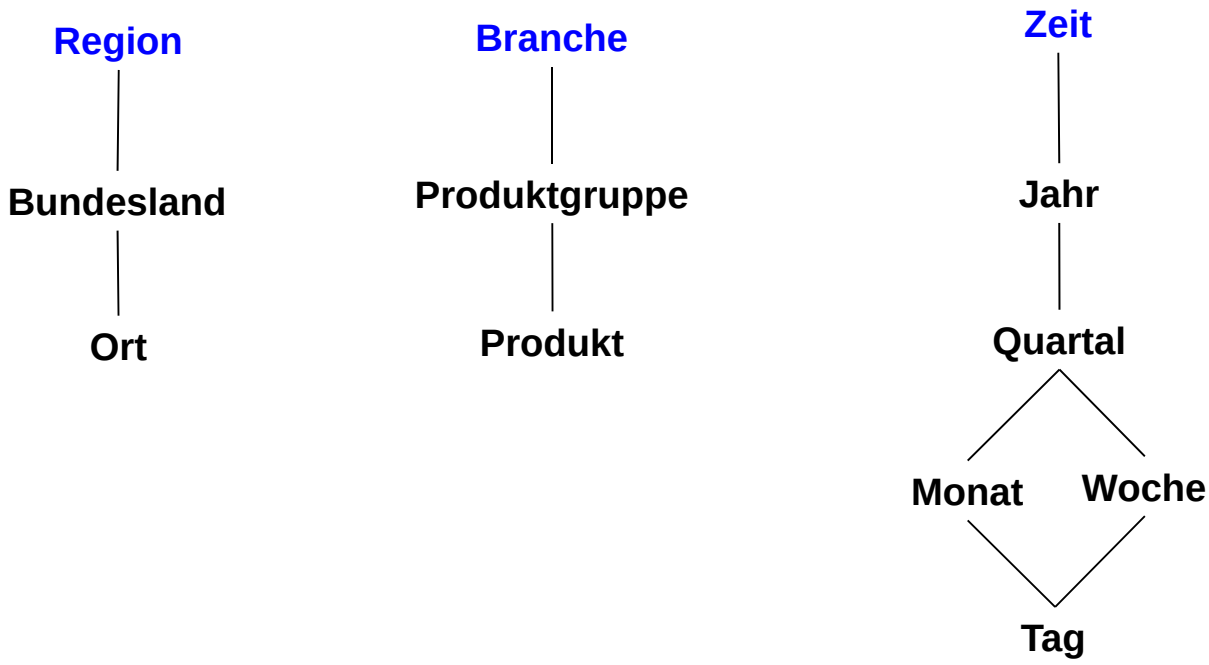
Mehrdimensionale Datensicht



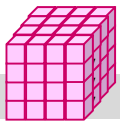
- Kennzahlen: numerische Werte als Grundlage für Aggregationen/Berechnungen (z.B. Absatzzahlen, Umsatz, etc.)
- Dimensionen: beschreibende Eigenschaften
- Operationen:
  - Aggregierung der Kennzahlen über eine oder mehrere Dimension(en)
  - Slicing and Dicing: Bereichseinschränkungen auf Dimensionen



Hierarchische Dimensionierung

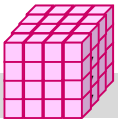


- Operationen zum Wechsel der Dimensionsebenen
  - Drill-Down
  - Roll-Up

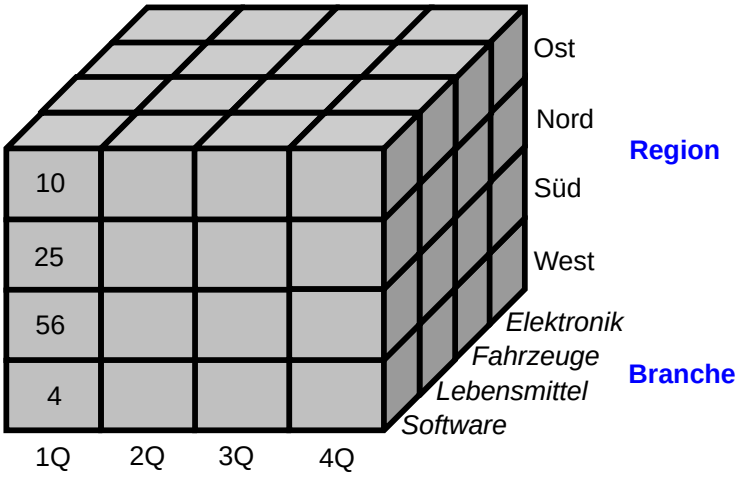


OLAP (Online Analytical Processing)

- 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

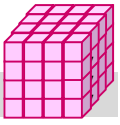


Multidimensional vs. relational



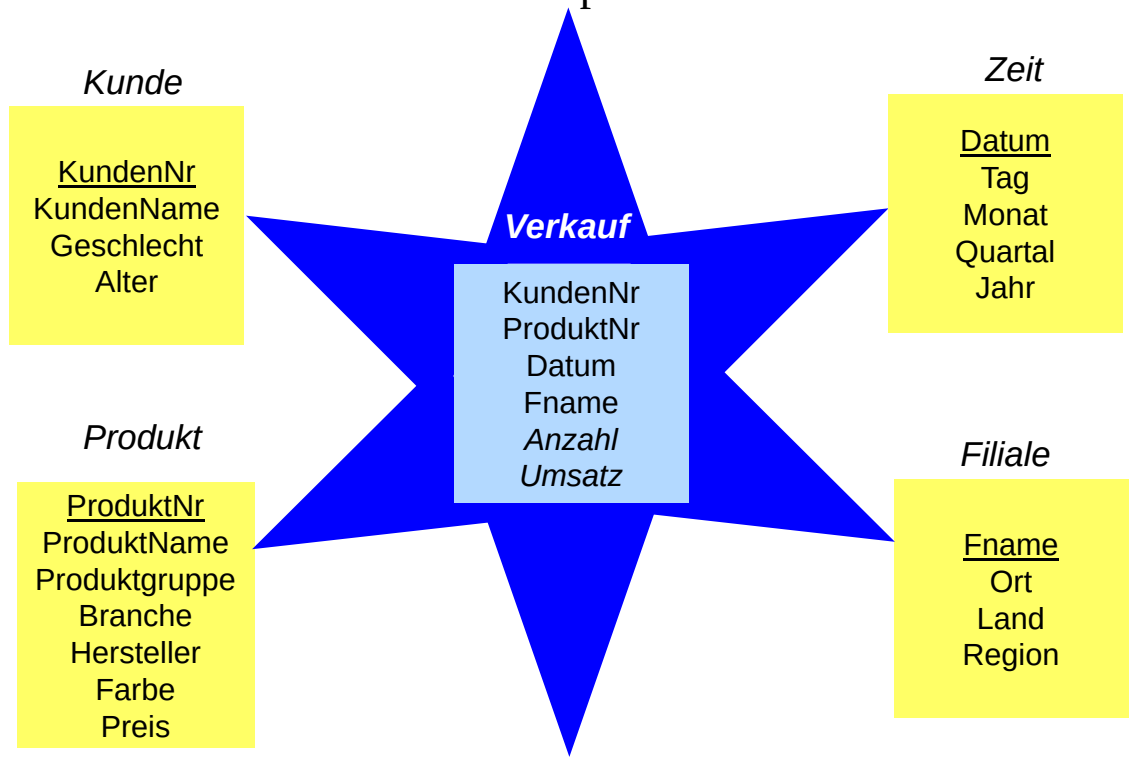
Bestellnr	Region	Branche	Zeit	Menge
1406	Ost	Fahrzeuge	2Q	5
4123	West	Elektronik	1Q	58
7829	Süd	Fahrzeuge	2Q	30
5327	Ost	Lebensmittel	4Q	3000
9306	Nord	Software	1Q	25
2574	Ost	Elektronik	4Q	2

- 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 2013 favorisiert?

```
select p.Hersteller, sum (v.Anzahl)
from Verkauf v, Filialen f, Produkt p, Zeit z, Kunden k
where z.Jahr = 2013 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;
```

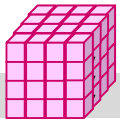
- Star-Join
  - sternförmiger Join der (relevanten) Dimensionstabellen mit der Faktentabelle
  - Einschränkung der Dimensionen
  - Verdichtung der Kennzahlen durch Gruppierung und Aggregation

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, Speadsheet-Integration

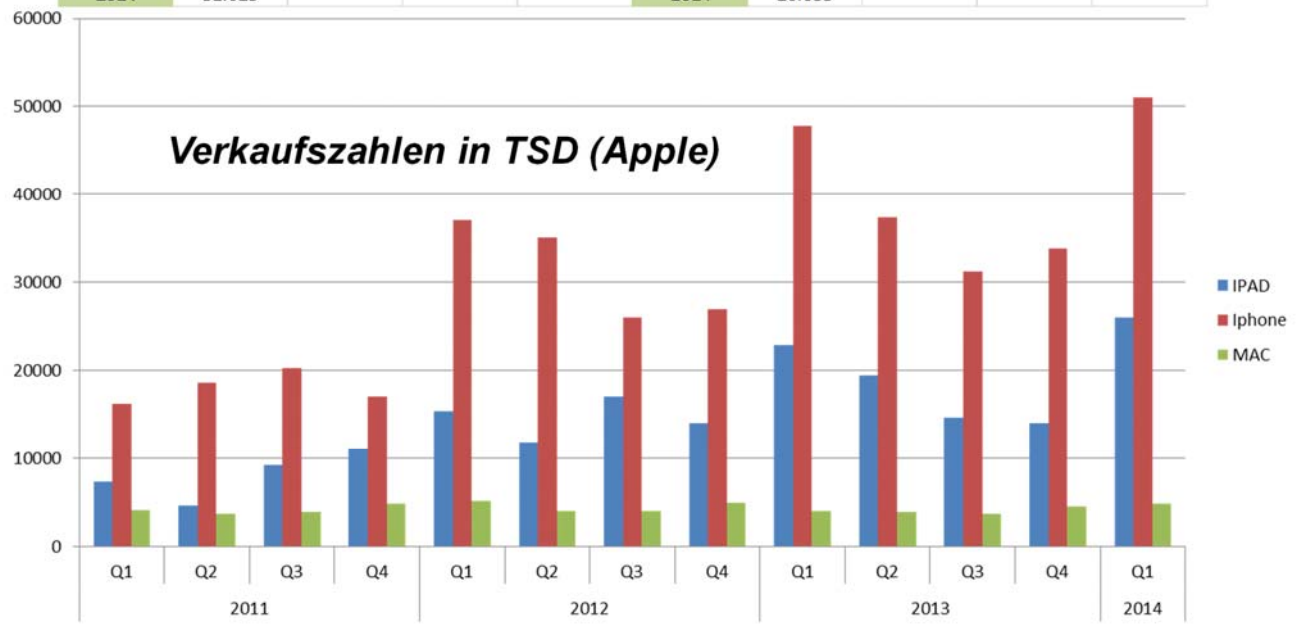


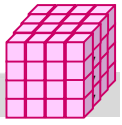
Beispiel: OLAP-Ausgabe (Excel)

,

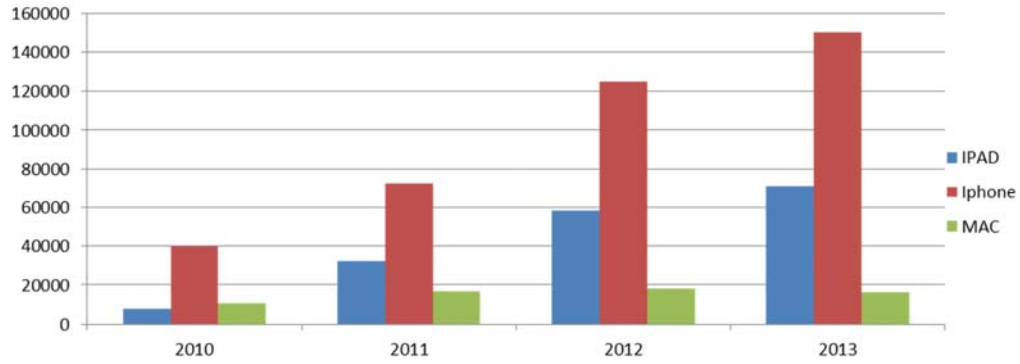
IPHONE	Oct-Dec Q1	Jan-Mar Q2	Apr-Jun Q3	Jul-Sep Q4
2010	8.737	8.752	8.398	14.102
2011	16.240	18.650	20.340	17.070
2012	37.044	35.064	26.028	26.910
2013	47.789	37.430	31.241	33.797
2014	51.025			

IPAD	Oct-Dec Q1	Jan-Mar Q2	Apr-Jun Q3	Jul-Sep Q4
2010			3.270	4.188
2011	7.331	4.694	9.246	11.123
2012	15.434	11.798	17.042	14.036
2013	22.860	19.477	14.617	14.079
2014	26.035			



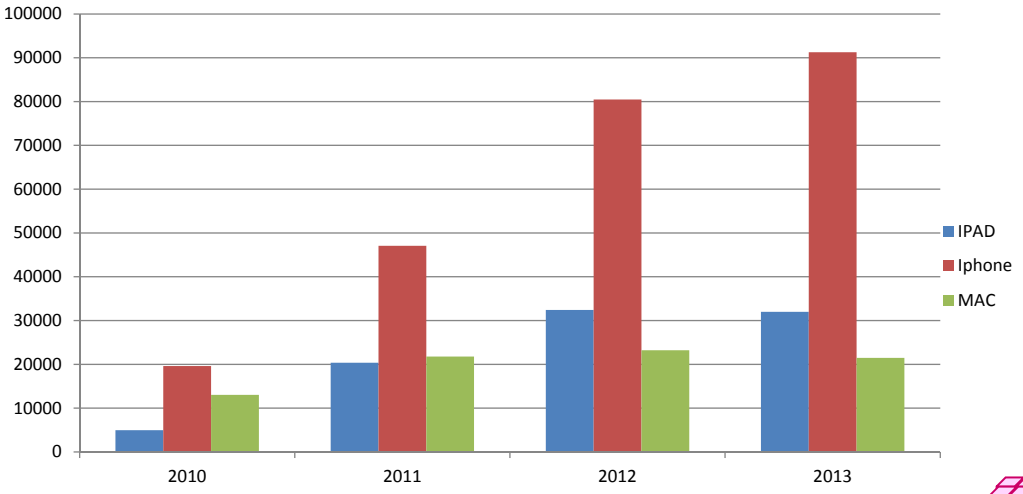


Jahresabsatz #Produkte (in TSD)

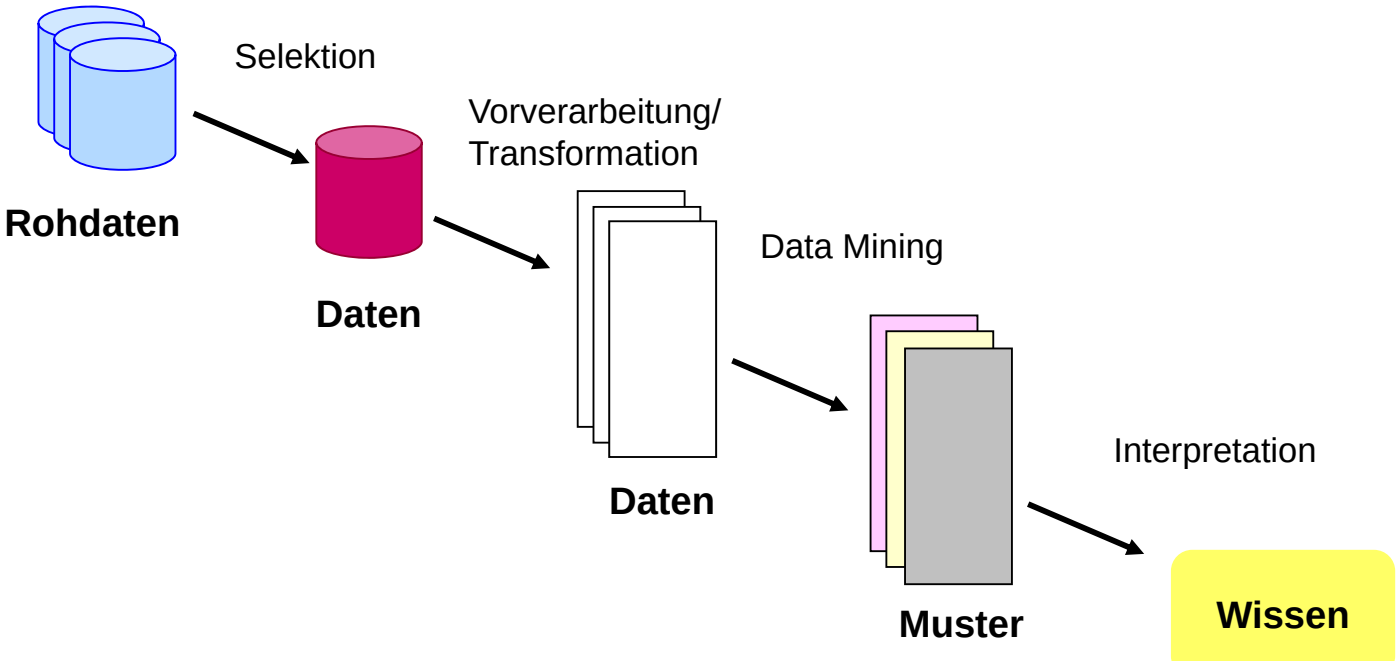


,

Jahresumsatz nach Produkttyp (in Mio $)



Knowledge Discovery

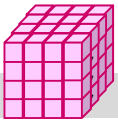


,

Data Mining: Techniken

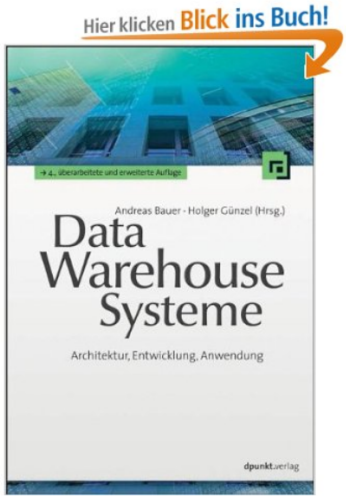
- Data Mining: Einsatz statistischer und wissensbasierter Methoden auf Basis von Data Warehouses
  - Auffinden von Korrelationen, Mustern und Trends in Daten
  - “Knowledge Discovery”: setzt im Gegensatz zu OLAP (“knowledge verification”) kein formales Modell voraus

- 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



Für eine größere Ansicht klicken Sie auf das Bild
Für Kunden: Stellen Sie Ihre eigenen Bilder ein.
Hier reinlesen und suchen

Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung
Ausgabe]
Andreas Bauer (Autor), Holger Günzel (Autor)
Geben Sie die erste Bewertung für diesen Artikel ab

Preis: EUR 49,90 kostenlose Lieferung. Siehe Details.
Alle Preisangaben inkl. MwSt.

Nur noch 11 auf Lager (mehr ist unterwegs).
Verkauf und Versand durch **Amazon**. Geschenkverpackung verfügbar.

Lieferung bis Freitag, 4. April: Bestellen Sie innerhalb **1 Stunde und 36 Minuten** per **Morning-E**

73 neu ab EUR 49,90 8 gebraucht ab EUR 29,

GRATIS-LIEFERUNG AM NÄCH
› Mehr erfahren

Weitere Ausgaben	Amazon-Prei
Kindle Edition	EUR 39,99
Gebundene Ausgabe	EUR 49,90

Nutzen Sie bewährte Cloud-R
Mit der Cloud-Lösung Amazon W
mehrere Dev- & Test-Umgebung
you-go-Modell betreiben. Jetzt
› Weitere Hinweise und Aktione

Kunden, die diesen Artikel gekauft haben, kauften auch


Business Intelligence - Grundlagen und ...
› Hans-Georg Kemper
(7)
Taschenbuch
EUR 32,99

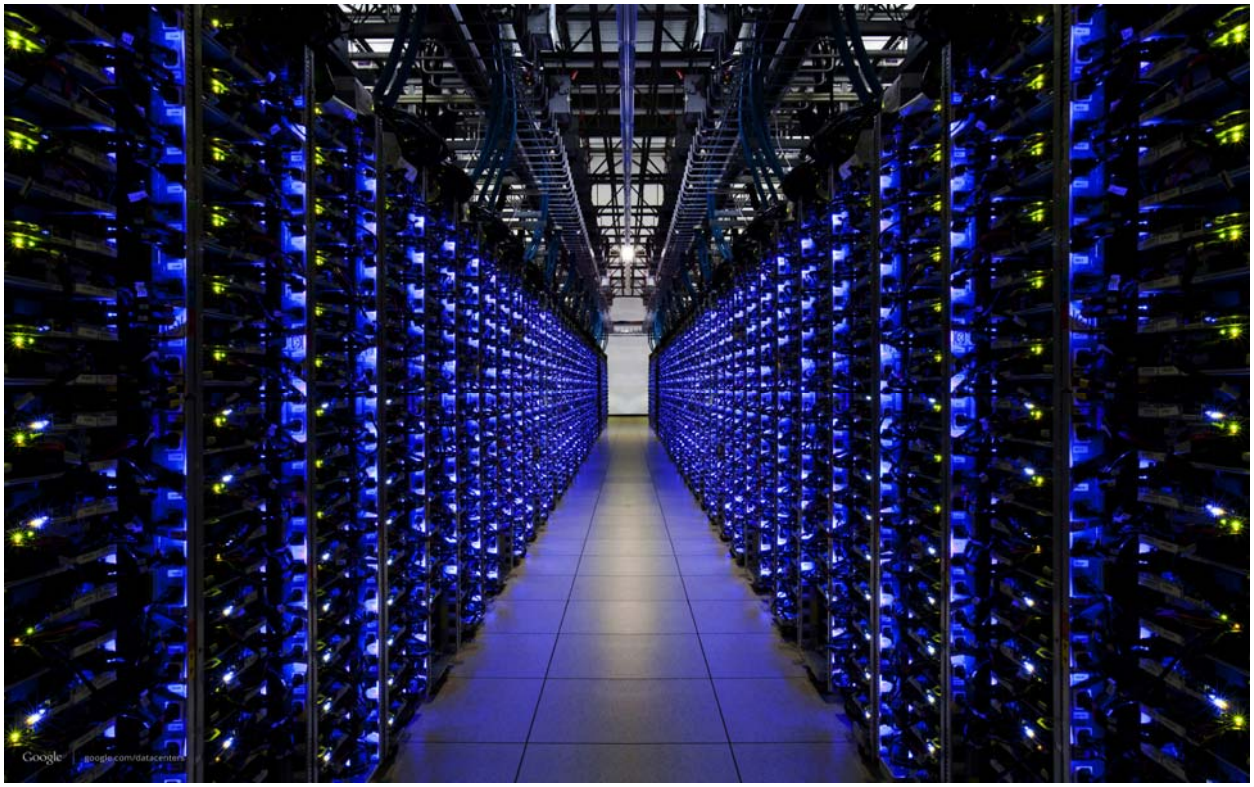

Data Warehouse Technologien: ...
› Veit Köppen
(2)
Broschiert
EUR 29,95

The Data Warehouse Toolkit: The Definitive ...
› Ralph Kimball
(2)
Taschenbuch
EUR 46,69

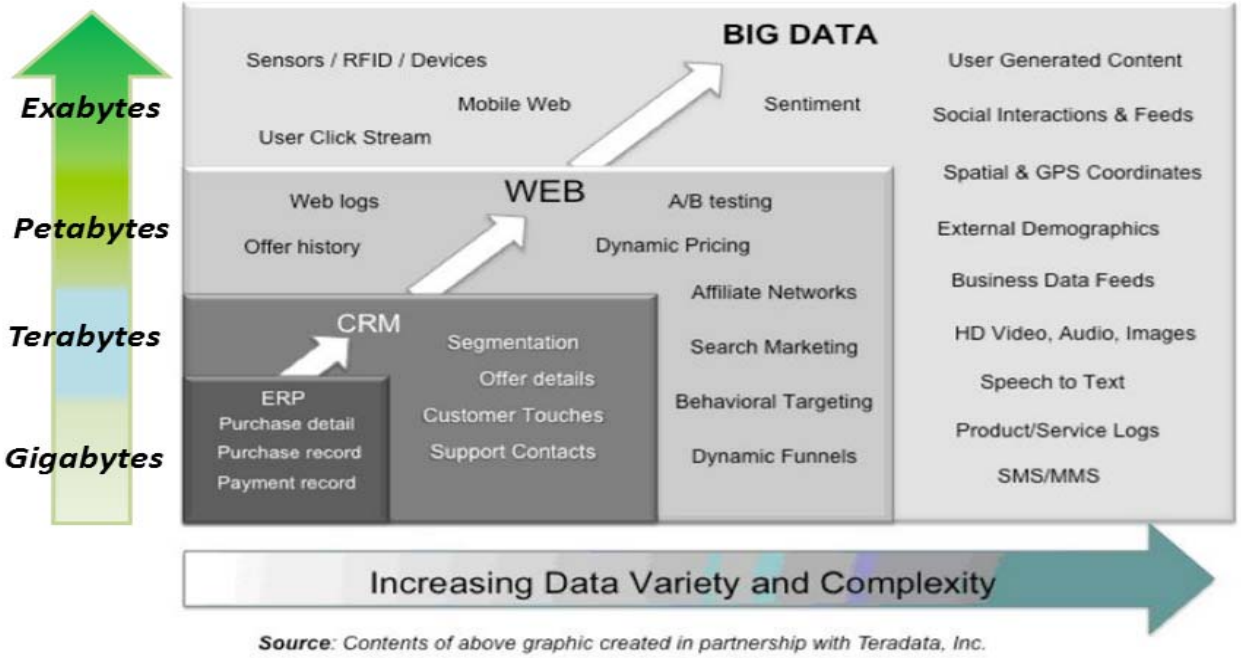
Wird oft zusammen gekauft

Preis für alle drei: EUR 107,84
Alle drei in den Einkaufswagen
Verfügbarkeit und Versanddetails anzeigen

Big Data



Massives Wachstum an Daten: Big Data



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Gartner-Schätzungen:

- pro Tag werden 2.5 Exabytes an Daten generiert
- 90% aller Daten weltweit wurden in den 2 letzten Jahren erzeugt.

Big Data Challenges

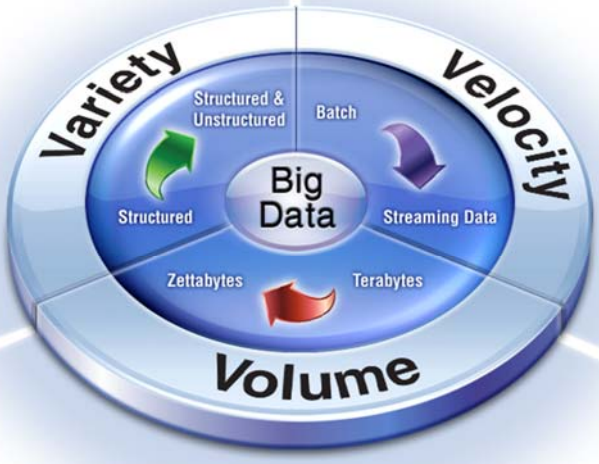
Volume Skalierbarkeit von Terabytes nach Petabytes (1K TBs) bis Zettabytes (1 Milliarde TBs)

Variety variierende Komplexität: strukturiert, teilstrukturiert, Text / Bild / Video

Velocity: Near-Realtime, Streaming

Veracity: Vertrauenswürdigkeit

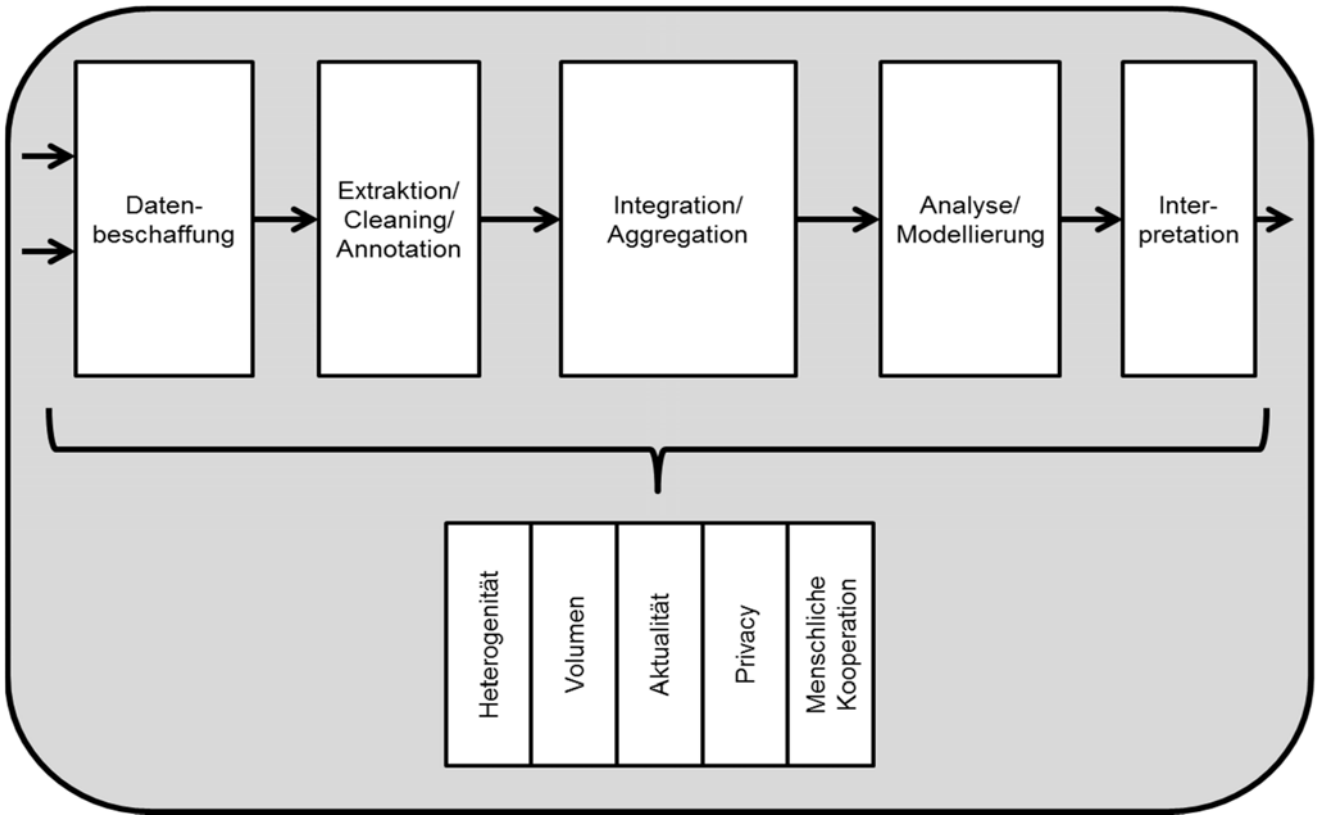
Value Erzielen des (wirtschaftl.) Nutzens durch Analysen



Potentiale für Big Data-Technologien

- Daten sind Produktionsfaktor ähnlich Betriebsmitteln und "Humankapital "
- Essentiell für viele Branchen und Wissenschaftsbereiche
- Valide Grundlage für zahlreiche Entscheidungsprozesse
  - Vorhersage/Bewertung/Kausalität von Ereignissen
- Kurzfristige Analysen von Realdaten im Geschäftsleben
- Beispiele
  - Nutzungsanalyse auf Web-Sites
  - Empfehlungsdienste (Live Recommendations)
  - Analyse/Optimierung von Werbe-Massnahmen
  - Smart Cities, Smart House
  - Industrie 4.0
  - Personalisierte Medizin
  - Kriminalistik …

Analyse-Pipeline für Big Data



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 integrierter Daten