2. Grundlagen

Inhalt dieses Kapitels

2.1 Datenbanksysteme [Kemper & Eickler 1999]
Grundbegriffe, relationale Datenbanksysteme, Anfragesprache SQL, Methode der Anfragebearbeitung, physische Speicherung der Daten, Indexstrukturen zur effizienten Anfragebearbeitung

2.2 Statistik [Fahrmeier, Künstler, Pigeot & Tutz 1999]
univariate und multivariate Deskription, Wahrscheinlichkeitsrechnung, diskrete und stetige Zufallsvariablen, Approximation von Verteilungen, Parameterschätzung, Testen von Hypothesen

2.3 OLAP [S. Chaudhuri & U. Dayal, 1997]
OLTP, Kennzahlen, mutlidimensionales Datenmodell, Stern- und Schneeflockenschema, Cubes

2.4 Preprocessing [Pyle 1999]
Ziele der Vorverarbeitung, typische Vorverarbeitungsschritte, Beispiele

2. 1 Datenbanksysteme

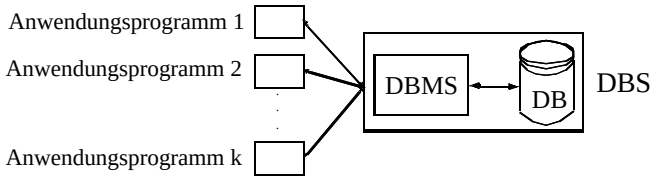
Definition

Ein *Datenbanksystem* (*DBS*) ist ein Softwaresystem zur dauerhaften Speicherung und zum effizienten Suchen in großen Datenmengen.

Komponenten

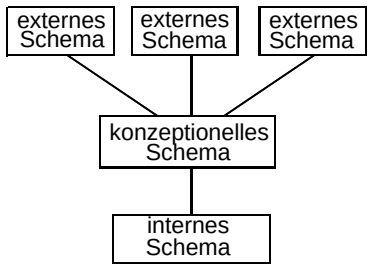
Datenbank (DB): Sammlung von Daten einer gegebenen Anwendung

Datenbank-Management-System (DBMS): Computerprogramm zum Management von Datenbanken beliebiger Anwendungen in einem spezifizierten Format



2.1 Datenbanksysteme

Drei-Ebenen-Architektur



Spezielle Sichten verschiedener Benutzer / Anwendungsprogramme auf dieselbe Datenbank

Logische Sicht der ganzen Datenbank

Physische Speicherung der Datenbank

2.1 Relationale Anfragesprache SQL

Beispiele

```
Kunde (KName, KAdr, Kto)
Auftrag (KName, Ware, Menge)
Lieferant (LName, LAdr, Ware, Preis)
```

```
select distinct Lname
from Lieferant, Auftrag
where Lieferant.Ware = Auftrag.Ware and KName =
'Huber'
```

```
select Ware, min (Preis), max (Preis), avg (Preis)
from Lieferant
group by Ware
order by Ware
```

2.1 Anfragebearbeitung

Prinzip

- eine SQL-Anfrage spezifiziert nur das „Was“
- der Anfrageoptimierer des DBMS bestimmt einen möglichst effizienten Anfrageplan, um die gegebene SQL-Anfrage zu beantworten
- Anfrageplan als *Operatorbaum*:
  - Die Blätter eines Operatorbaumes enthalten die auftretenden *Relationen*.
  - Die inneren Knoten repräsentieren die verwendeten *Operationen*.

Ablauf

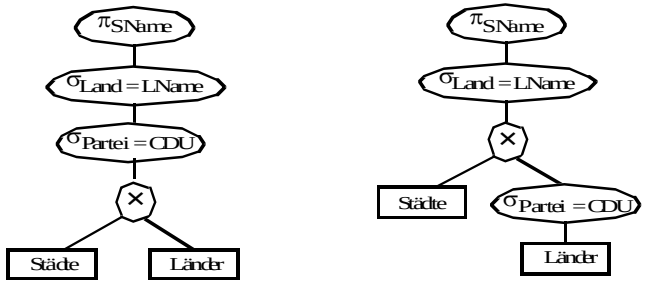
- Generierung von Anfrageplänen mit Hilfe von *heuristischen Regeln* (z.B. Selektionen vor Joins)
- Bewertung der Anfragepläne basierend auf einem *Kostenmodell* (Kostenmaß: Anzahl zu bearbeitender Tupel) und statistischen Angaben über die Ausprägung der Datenbank

2.1 Anfragebearbeitung

Beispiel

```
Städte (SName, SEinw, Land)
Länder (LName, LEinw, Partei)

select Sname from Städte,Länder
where Land=Lname and Partei=CDU
```



2.1 Physische Speicherung der Daten

Prinzip der Magnetplatten

- *Seiten* (Blöcke) als *kleinste Transfereinheit* zwischen Haupt- und Sekundärspeicher
- *Feste Größe* zwischen 128 Byte und 16 KByte
- *Direkter Zugriff* auf eine Seite mit gegebener Seitennummer

Wahlfreier Zugriff

- Positionierung des Schreib-/Lesekopfes
  Zeit für die Kammbewegung [6 ms]
- Warten auf den Sektor / die Seite
  im Durchschnitt die halbe Rotationszeit der Platte [4 ms]
- Übertragung der Seite
  Zeit für Schreiben bzw. Lesen [0,1 ms / KByte]

sehr teuer im Vergleich zu Hauptspeicher-Operationen

2.1 Physische Speicherung der Daten

Sequentieller Zugriff

- Zugriff auf eine Menge von Seiten mit aufeinanderfolgenden Adressen
- ab der zweiten Seite entfällt der große Aufwand zur Positionierung des Schreib-/Lesekopfes und für das Warten auf die Seite
- sequentieller Zugriff ist wesentlich effizienter als wahlfreier Zugriff

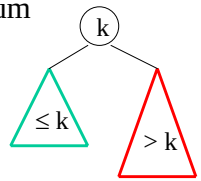
Kostenmaß für die Anfragebearbeitung

- Annahme: Zugriff auf Seiten erfolgt unabhängig voneinander
- sequentieller Zugriff ist dann nicht möglich
- Zeitaufwand für den wahlfreien Seitenzugriff ist um Größenordnungen höher als die Zeit für eine Operation im Hauptspeicher

Anzahl der Seitenzugriffe als Kostenmaß

Prinzipien

Suchbaum



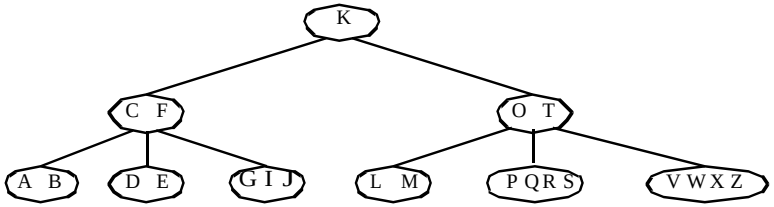
Balancierter Suchbaum

- alle Blätter des Baumes besitzen denselben Level
- die Höhe des Baumes ist O(log n) für n Datensätze
- die Operationen Einfügen, Entfernen und Suchen sind auf einen (oder wenige) Pfade beschränkt

➡ Knoten des Baums = Seite der Magnetplatte

B-Baum

- Jeder Knoten enthält höchstens 2*m* Schlüssel.
- Jeder Knoten außer der Wurzel enthält mindestens *m* Schlüssel, die Wurzel mindestens einen Schlüssel.
- Ein Knoten mit *k* Schlüsseln hat genau *k*+1 Söhne.
- Alle Blätter befinden sich auf demselben Level.



Punktanfrage im B⁺-Baum

```
PunktAnfrage (Seite s, Integer k);
  i:=1;
  while i < Anzahl der Einträge in s do
    if k ≤ i-ter Schlüssel in s then
      if s ist Datenseite then
        return i-ter Datensatz in s;
      else PunktAnfrage (i-ter Sohn von s, k);
    else i:= i + 1;
  if i = Anzahl der Einträge in s then
    PunktAnfrage (i-ter Sohn von s, k);
```


R-Baum

Vergleich mit B-Baum

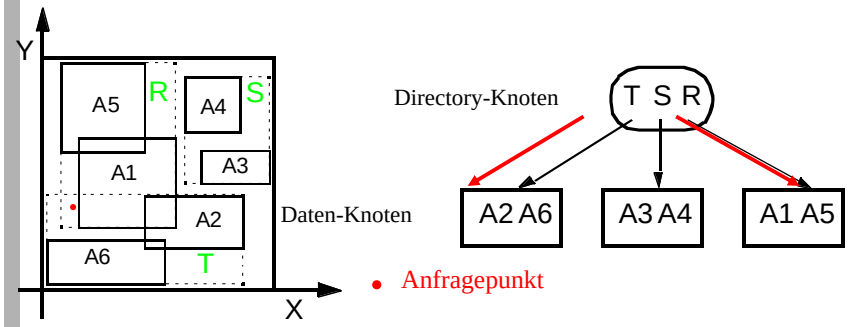
- B-Baum: eindimensionale Schlüssel (alphanumerische Werte)
- R-Baum: mehrdimensionale Schlüssel (Hyper-Rechtecke)

Definition

- Jeder Knoten außer der Wurzel besitzt zwischen *m* und *M* Einträge.
- Die Wurzel hat mindestens zwei Einträge, außer sie ist ein Blatt.
- Ein innerer Knoten mit *k* Einträgen hat genau *k* Söhne.
- Alle Blätter befinden sich auf demselben Level.

Punktanfrage im R-Baum

$M = 3, m = 1$



Anfragebearbeitung ist *nicht* mehr auf einen Pfad beschränkt

Grundaufgaben

deskriptive Statistik

- beschreibende und graphische Aufbereitung von Daten
- auch zur Validierung der Daten

explorative Statistik

- wenn die Wahl eines geeigneten statistischen Modells unklar ist
- sucht nach Strukturen und Besonderheiten in den Daten

induktive Statistik

- basiert auf stochastischen Modellen
- zieht aus den beobachteten Daten Schlüsse auf umfassendere Grundgesamtheiten
- vorbereitende deskriptive und explorative Analysen nötig

Grundbegriffe

Stichprobenerhebung

- n Untersuchungseinheiten
- Werte $x_1,..., x_n$ eines Merkmals X beobachtet
- $h(a)$ die *absolute Häufigkeit* und $f(a) = h(a)/n$ die *relative Häufigkeit* des Attributwerts a in der Stichprobe

Typen von Merkmalen

- numerisch (mit totaler Ordnung < und arithmetischen Operationen)
- ordinal (mit (totaler) Ordnung <)
- kategorisch/nominal (keine Ordnung und keine arithmetischen Operationen)

Lagemaße

- *arithmetisches Mittel* $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$
- *Median*
  (seien dazu die x_i aufsteigend sortiert)

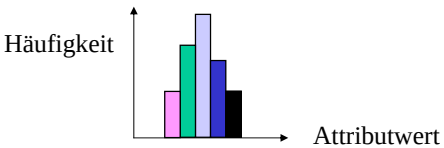
$$x_{med} = \begin{cases} x_{(n+1)/2} & \text{falls } n \text{ ungerade} \\ (x_{n/2} + x_{(n/2+1)})/2 & \text{falls } n \text{ gerade} \end{cases}$$

Streuungsmaße

- *Varianz* $\bar{s}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$
- *Standardabweichung* $\bar{s} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$

nur für numerische Merkmale

Histogramme



Kontingenztabelle

- für kategorische Merkmale X und Y
- repräsentiert für zwei Merkmale *X* und *Y* die absolute Häufigkeit h_{ik} jeder Kombination (x_i, y_k) und alle Randhäufigkeiten $h_{.k}$ und $h_{i.}$ von *X* und *Y*

	Mittelfristige Arbeitslosigkeit	Langfristige Arbeitslosigkeit	
Keine Ausbildung	19	18	37
Lehre	43	20	63
	62	38	100

Wie sollten die relativen Häufigkeiten verteilt sein, wenn die beiden Merkmale keinerlei Abhängigkeit besitzen? $\frac{h_{ik}}{n} = \frac{h_{i.}}{n} \cdot \frac{h_{.k}}{n}$

χ^2-Koeffizient

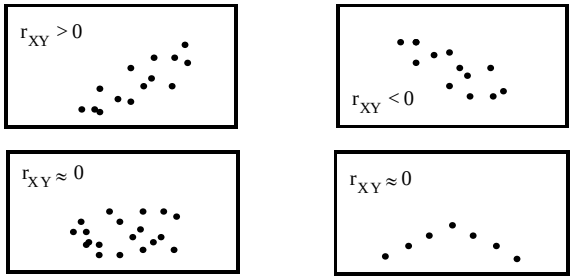
Differenz zwischen dem bei Unabhängigkeit erwarteten und dem tatsächlich beobachteten Wert von h_{ij} (Maß für die Stärke der Abhängigkeit)

Korrelationskoeffizient

- für numerische Merkmale X und Y
- wie stark sind die Abweichungen vom jeweiligen Mittelwert korreliert?

$$r_{XY} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

- Beispiele



Ereignisse und Wahrscheinlichkeitsmaße

- Ein *Zufallsvorgang* führt zu einem von mehreren sich gegenseitig ausschließenden Ergebnissen.
- Sei $\Omega = \{\omega_1, \ldots, \omega_n\}$ der *Ergebnisraum*, d.h. die Menge aller möglichen Ergebnisse eines Zufallsvorgangs.
- Teilmengen von Ω heißen *Ereignisse.*
- Ein *Wahrscheinlichkeitsmaß* ist eine Abbildung $P: 2^{\Omega} \rightarrow [0,1]$, die die folgenden Axiome erfüllt:

(A1) $P(A) \geq 0$ für alle $A \subseteq \Omega$,

(A2) $P(\Omega) = 1$,

(A3) $P(A \cup B) = P(A) + P(B)$ für alle $A, B \subseteq \Omega$ mit $A \cap B = \varnothing$.

Bedingte Wahrscheinlichkeiten

- Seien $A, B \subseteq \Omega$. Die *bedingte Wahrscheinlichkeit* von *A* unter *B*, $P(A|B)$, ist definiert für $P(B) \neq 0$ als $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
- *A* und *B* heißen *unabhängig*, wenn gilt $P(A|B) = P(A)$ und $P(B|A) = P(B)$.

Satz von Bayes

Sei $A_1, \ldots, A_k$ eine disjunkte Zerlegung von Ω , so daß für mindestens ein *i*, $1 \leq i \leq k$, gilt: $P(A_i) > 0$ und $P(B|A_i) > 0$. Dann gilt für alle $1 \leq j \leq k$:

$$P(A_j \mid B) = \frac{P(B \mid A_j) \cdot P(A_j)}{P(B)}$$

a-priori-Wahrscheinlichkeit: $P(A_i)$

a-posteriori-Wahrscheinlichkeit: $P(A_i|B)$

Grundbegriffe

- *Zufallsvariable*
  Merkmal, dessen Werte die Ergebnisse eines Zufallsvorgangs sind
- *diskrete Zufallsvariable*
  endlich oder abzählbar unendlich viele verschiedene Werte $x_1, x_2, \ldots, x_k, \ldots$
- *Wahrscheinlichkeitsfunktion* $f(x) = \begin{cases} P(x_i) \ \textit{falls} \ x = x_i \\ 0 \ \textit{sonst} \end{cases}$
- *Verteilungsfunktion (für ordinale oder numerische Werte)* $F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i)$
- *Erwartungswert* $E(X) = \sum_{i \ge 1} x_i \cdot f(x_i)$
- *Varianz* $Var(X) = \sum_{i \ge 1} (x_i - E(X))^2 \cdot f(x_i)$

Binomialverteilung

- *Bernoulli-Experiment*: nur zwei Ergebnisse (Treffer oder Nichttreffer), p die Wahrscheinlichkeit des Treffers
- n unabhängige Wiederholungen desselben Bernoulli-Experiments, die Gesamtanzahl k der Treffer wird beobachtet
- *binomialverteilte* Zufallsvariable mit den Parametern n und p besitzt folgende Wahrscheinlichkeitsfunktion:

$$f(k) = P(X = k) = \begin{cases} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \ \textit{falls} k \in \{0,1,\ldots,n\} \\ 0 \ \textit{sonst} \end{cases}$$

- Erwartungswert und Varianz einer binomialverteilten Zufallsvariablen:

$$E(X) = n \cdot p \qquad Var(X) = n \cdot p \cdot (1-p)$$

Beispiel einer Binomialverteilung

- Anwendung: Abschätzung des (auf einer Stichprobe bestimmten) Klassifikationsfehlers auf der Grundgesamtheit
- Bernoulli-Experiment: zufälliges Ziehen eines Objekts der Grundgesamtheit und Test, ob dieses Objekt von dem Klassifikator falsch klassifiziert wird
- Treffer: Objekt wird falsch klassifiziert
- Nichttreffer: Objekt wird korrekt klassifiziert
- p: Wahrscheinlichkeit einer Fehlklassifikation in der Grundgesamtheit
- n: Größe der Trainingsmenge

⇒ gesucht ist ein Intervall $[u,o]$, so daß mit einer Wahrscheinlichkeit von z.B. mindestens 95 % gilt

$$u \le p \le o$$

Grundbegriffe

- überabzählbar unendlich viele verschiedene Werte $x_1, x_2, \ldots, x_k, \ldots$
- Eine Zufallsvariable X heißt *stetig*, wenn es eine Funktion (*Wahrscheinlichkeits-Dichte*) $f(x) \ge 0$ gibt, so daß für jedes Intervall [a,b] gilt:

$$P(a \le X \le b) = \int_a^b f(x)dx$$

- *Verteilungsfunktion* $F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$
- *p-Quantil* x_p mit $F(x_p) = p$
- *Erwartungswert* $E(x) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$

Normalverteilung

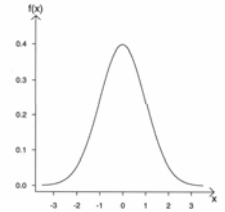
• Eine Zufallsvariable X heißt *normalverteilt* (bzw. *gaußverteilt*) mit den Parametern $\mu \in IR$ und $\sigma > 0$, wenn sie folgende Dichte besitzt:

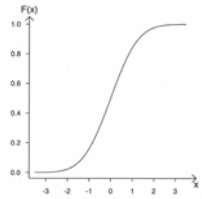
$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

• Erwartungswert $E(x) = \mu$

• *standardisierte Zufallsvariable* $Z = \frac{X - \mu}{\sigma}$

standardnormalverteilt (normalverteilt mit Parametern $\mu = 0$ und $\sigma = 1$)





Schwankungsintervall

• *Schwankungsintervall* $\mu - c \le X \le \mu + c$

• Es gilt $x_p = z_p \cdot \sigma + \mu$

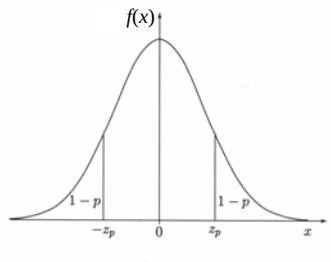
• Wahrscheinlichkeit dafür, daß der Wert von X im Schwankungsintervall liegt:

$$P(\mu - \sigma \cdot z_{1-\alpha/2} \le X \le \mu + \sigma \cdot z_{1-\alpha/2}) = 1 - \alpha$$

α *Irrtumswahrscheinlichkeit*

• es gilt z.B.

$$P(\mu - \sigma \le X \le \mu + \sigma) = 0{,}6827$$
$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0{,}9545$$
$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0{,}9973$$



Grundbegriffe

• *Nullhypothese* H_0 und *Alternative* H_1, die sich gegenseitig ausschließen

• Annahmen über die Verteilung oder bestimmte Parameter des interessierenden Merkmals in der Grundgesamtheit

• *Fehler 1. Art*

H_0 wird verworfen, obwohl H_0 wahr ist

• *Fehler 2. Art*

H_0 wird akzeptiert wird, obwohl H_1 wahr ist

• *Test zum Signifikanzniveau* α $(0 < \alpha < 1)$
ein Hypothesen-Test, bei dem die Wahrscheinlichkeit eines Fehlers 1. Art höchstens α beträgt

OLAP

2.3 OLAP

2.3 OLAP

2.3.1 Einführung in OLAP

Es gibt große Unterschiede zwischen operativen Systemen und einem Data Warehouse (DWh).

Entsprechend gibt es fundamentale Unterschiede auch zwischen den jeweiligen Zugriffsarten auf diese Datenquellen:

- **OLAP = On-Line Analytical Processing** benutzt DWh
- **OLTP = On-Line Transaction Processing** benutzt operative Systeme

2.3.1 Einführung in OLAP

OLTP

- hohe Zahl **kurzer**, atomarer, isolierter, wiederkehrender Transaktionen
  - z.B. Konto-Update, Flugbuchung, Telefon-Gespräch
- Transaktionen benötigen detaillierte, aktuelle Daten
- Daten werden (oft tupelweise) gelesen und relativ **häufig aktualisiert**
- Transaktionen dienen dem **Tagesgeschäft** und haben relativ hohe Ansprüche an die Bearbeitungsgeschwindigkeit

2.3.1 Einführung in OLAP

Definition von OLAP:

- **OLAP Systeme**
  - dienen der **Entscheidungs-Unterstützung** oder
  - können in den Phasen **„Data Understanding“** bzw. **„Data Preparation“** im Rahmen des Data-Mining-Prozesses eingesetzt werden.

- **OLAP-Funktionen** erlauben
  - den schnellen, **interaktiven** Zugriff auf Unternehmensdaten
  - unter „beliebigen“ unternehmensrelevanten Blickwinkeln (**Dimensionen**)
  - auf verschiedenen **Aggregationsstufen**
  - mit verschiedenen Techniken der Visualisierung

- Hauptmerkmal ist die **multi-dimensionale** Sichtweise auf Daten mit flexiblen interaktiven Aggregations- bzw. Verfeinerungsfunktionen entlang einer oder mehrerer Dimensionen.

2.3.1 Einführung in OLAP

Multi-Dimensionalität:

- Mehrdimensionale Sichtweise auf Daten ist sehr **natürlich**.
- Sichtweise der Analysten auf Unternehmen **ist** mehrdimensional.
  ⇒ Konzeptuelles Datenmodell sollte mehrdimensional sein, damit Analysten leicht und intuitiv Zugang finden.

- **Beispiel:** *Verkaufszahlen* können nach unterschiedlichen Kriterien / Dimensionen aggregiert und analysiert werden.
  - nach **Produkt:** *Produkt, Produktkategorie, Industriezweig*
  - nach **Region:** *Filiale, Stadt, Bundesland*
  - nach **Zeit:** *Tag, Woche, Monat, Jahr*
  - nach verschiedenen Dimensionen des Käufers: **Alter, Geschlecht, Einkommen** des Käufers
  - und nach **beliebigen Kombinationen von Dimensionen**, z.B.
    - nach *Produktkategorie, Stadt* und *Monat*

2.3.1 Einführung in OLAP

Kennzahlen:

- Die **Analyse-Gegenstände** von OLAP sind **numerische Werte,** typischerweise **Kennzahlen** genannt (oder auch Maße, Metriken oder Fakten).
  - **Beispiel:** *Verkaufszahlen, Umsatz, Gewinn, Lagerbestand,...*
- Diese numerischen Werte lassen sich auf verschiedene Weise verdichten, z.B.
  - Summenbildung
  - Mittelwertbildung
  - Minimum- oder Maximumbestimmung
  - Die zulässige Art der Verdichtung hängt vom **Skalenniveau** der Kennzahl ab.

2.3.1 Einführung in OLAP

Skalenniveaus

In der Statistik unterscheidet man die Attributausprägungen einer vorgegebenen Menge von Daten mittels Skalen mit unterschiedlichem Skalenniveau. Die wichtigsten Typen sind:

Nominalskalierte/kategorische Merkmale:
Ausprägungen sind ‘‘Namen‘‘, keine Ordnung möglich
→ keine Aggregation möglich

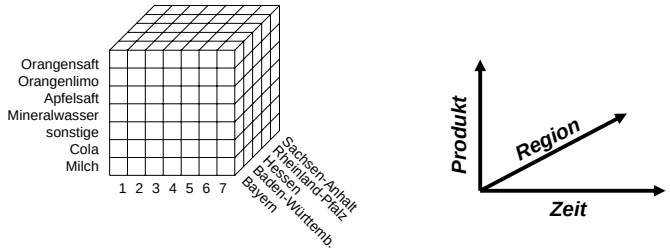
Ordinalskalierte Merkmale:
Ausprägungen können geordnet, aber Abstände nicht interpretiert werden.
→ Median macht Sinn, Mittelwert z.B. nicht

Kardinalskalierte/numerische Merkmale:
Ausprägungen sind Zahlen, Interpretation der Abstände möglich (metrisch)
→ Mittelwertbildung, Standardabweichung etc. sinnvoll

2.3.1 Einführung in OLAP

Dimensionen:

- Jede Kennzahl hängt von einer Menge von **Dimensionen** ab. Diese bilden den **Kontext der Kennzahlen.**
  - **Beispiel:** Die *Verkaufszahlen* (Kennzahl) hängen von den Dimensionen *Produkt, Region* und *Zeit* ab.
  - Die Dimensionen sind **orthogonal (unabhängig).**
  - Sie definieren einen sog. **Hyper-Würfel (hyper cube).**



- Es kann eine beliebige Zahl an Dimensionen geben (abhängig vom Zweck des OLAP-Systems und der enthaltenen Daten). In manchen Anwendungen treten bis zu 50 Dimensionen auf.

2.3.1 Einführung in OLAP

Dimension Zeit:

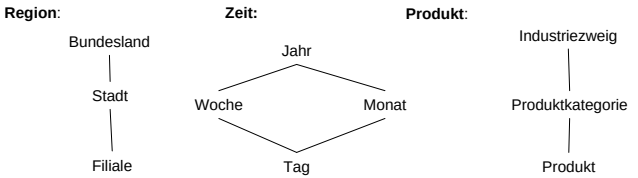
- **Spezielle Dimension,** die in jedem OLAP-System existiert, ist die ***Zeit.***
- Leistung eines Unternehmens wird immer anhand der Zeit bewertet:
  - aktueller Monat im Vergleich zu letztem Monat
  - aktueller Monat im Vergleich zum gleichen Monat des Vorjahres
- Dimension *Zeit* unterscheidet sich von allen anderen Dimensionen:
  - Zeit hat einen linearen Charakter:
    - Januar kommt vor Februar
  - Zeit hat Wiederholungscharakter: jeden Montag, werktags, ...
- OLAP-System muss Umgang mit der Dimension Zeit und den damit verbundenen Besonderheiten unterstützen.

Attribute und Attributelemente:
Jede Dimension ist durch eine **Menge von Attributen** charakterisiert.
- **Beispiel:** Die Dimension *Region* ist charakterisiert durch die Attribute *Filiale, Stadt* und *Bundesland.*

2.3.1 Einführung in OLAP

Attribute und Attributelemente:

•Diese Attribute können **hierarchisch** angeordnet sein (Aggregationsstufen)
- **Beispiel:**
  - Gesamtwert ergibt sich aus den Werten mehrerer *Bundesländer.*
  - Wert für ein *Bundesland* ergibt sich aus Werten mehrerer *Städte*.
  - Wert für eine Stadt ergibt sich aus Werten mehrerer *Filialen.*



2.3.1 Einführung in OLAP

- Ein Pfad in einer solchen **Attribut-Hierarchie** (z.B. *Tag, Monat, Jahr*) wird auch **consolidation path** genannt.

- Jedes Attribut einer Dimension wird durch **Attributelemente** instantiert.
  - **Beispiel:**
    - Das Attribut ***Produkt*** der Dimension *Produkt* hat die Attributelemente: *Coca-Cola, Pepsi-Cola, Afri-Cola, ...*
    - Das Attribut ***Produktkategorie*** hat die Attributelemente: *Orangensaft, Apfelsaft, Orangenlimo, Cola,...*
    - Das Attribut ***Industriezweig*** hat die Attributelemente: *Lebensmittelindustrie, Textilindustrie, Schwerindustrie,...*

2.3.2 OLAP Funktionalität

2.3.2 OLAP Funktionalität

- Bei der Analyse können beliebige Aggregationsstufen visualisiert werden:

  Drill-Down bzw. **Roll-Up-**Operationen

- Bedingungen an Dimensionen, Attribute und Attributelemente reduzieren Dimensionalität der visualisierten Daten:

  Slice & Dice - Operationen

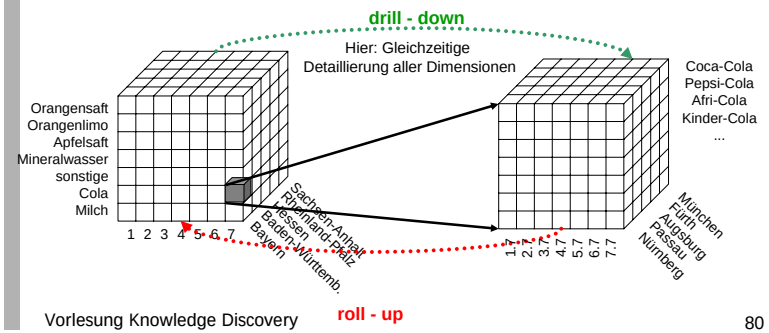
- Analyse wird durch Vielzahl von **Visualisierungstechniken** unterstützt. Bedingungen werden **interaktiv** gewählt (Buttons, Menüs, *drag & drop*), so dass Analysten und Manager keine komplizierte Anfragesprache lernen müssen.

2.3.2 OLAP-Funktionalität

Drill-Down und Roll-Up

- Entlang der Attribut-Hierarchien werden die Daten **verdichtet** bzw. wieder **detailliert** und sind so auf verschiedenen **Aggregationsstufen** für Analysen zugreifbar.

- Verdichtung/Detaillierung kann entlang einer, mehrerer oder aller Dimensionen geschehen - gleichzeitig oder in beliebiger Reihenfolge.



2.3.2 OLAP-Funktionalität

Slice & Dice:

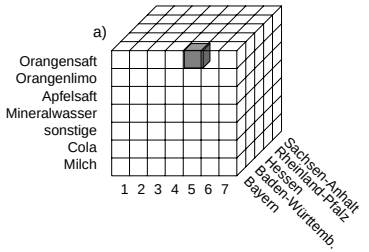
- Bei dieser Operation wird die **Dimensionalität** der visualisierten Daten **reduziert.**
- Zu einer Teilmenge der Dimensionen (sog. **page dimensions**) werden Bedingungen formuliert.
- Alle Daten in der resultierenden Tabelle genügen diesen Bedingungen.
- Die **page dimensions** tauchen in der neuen Tabelle nicht mehr explizit auf, sondern definieren implizit die Menge dargestellter Daten.

Slice & Dice entspricht dem Herausschneiden einer Scheibe (*slice*) aus dem Hyper-Würfel. Nur diese Scheibe wird weiterhin visualisiert.

2.3.2 OLAP-Funktionalität

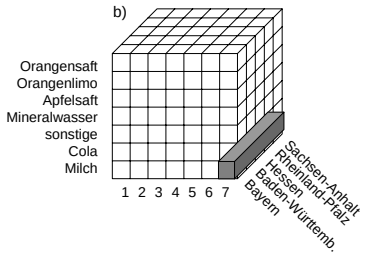
Beispiele:

Lokation bestimmter atomarer und aggregierter Werte im Hyper-Würfel.



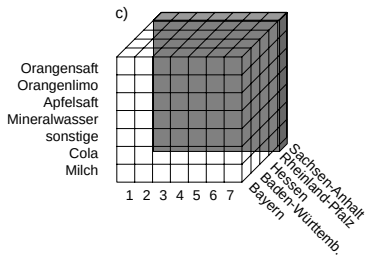
a) Verkaufszahlen für Orangensaft in Bayern im Mai

2.3.2 OLAP-Funktionalität



b) Verkaufszahlen für Milch in ganz Süddeutschland im Juli

2.3.2 OLAP-Funktionalität



c) Verkaufszahlen insgesamt für Sachsen-Anhalt

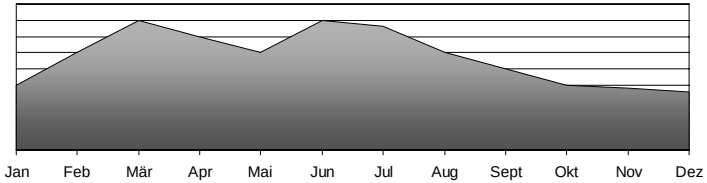
⇒ Aggregation der Verkaufszahlen über alle Monate **und** alle Produkte

2.3.2 OLAP-Funktionalität

- Analyse bezieht sich nur selten auf einen Wert:
  - sondern auf eine Folge von Werten
    ⇒ Entwicklungen und **Trends** erkennbar (d)
  - oder auf eine Menge von Werten
    ⇒ Vergleiche verschiedener Werte ermöglicht (e)

2.3.2 OLAP-Funktionalität

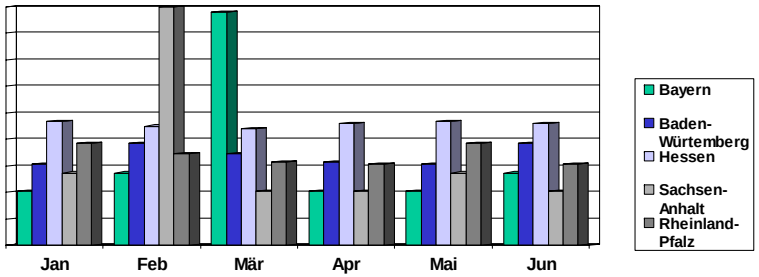
d) Entwicklung der Verkaufszahlen für Apfelsaft in Baden-Württemberg im letzten Jahr.



page dimensions: Produkt = Apfelsaft, Region = Baden-Württemberg

2.3.2 OLAP-Funktionalität

e) Vergleich der Verkaufszahlen für Apfelsaft in den Regionen Deutschlands für das erste Halbjahr



page dimensions: Produkt = Apfelsaft

2.3.3 Mehrdimensionales Datenmodell

2.3.3 Mehrdimensionales Datenmodell

Der beste Weg um zu einem OLAP-fähigen DWh zu kommen:

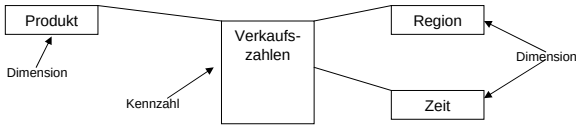
1. Erstellen eines **mehrdimensionalen** konzeptuellen Datenmodells.
2. Ableiten eines **relationalen** logischen Datenmodells.
   - Relationale DBS bilden die Implementierungsebene des DWh

Stern-Schema: (star schema)

- mehrdimensionales Datenmodell durch **Stern-Schema** realisierbar.
- Konstrukte eines Stern-Schemas:
  - **Kennzahlen:** Gegenstände der Analyse: Verkaufszahlen
  - **Dimensionen** definieren den Kontext der Kennzahlen: Produkt, Region, Zeit

2.3.3 Mehrdimensionales Datenmodell

Beispiel:



2.3.3 Mehrdimensionales Datenmodell

Vorteile des Stern-Schemas gegenüber herkömmlichen relationalen Schemata:

- Schema-Entwurf entspricht der **natürlichen Sichtweise** der Benutzer
  - Daten können in einer für Analysen adäquaten Weise zugegriffen werden.
- **Erweiterungen** und **Änderungen** am Schema sind leicht zu realisieren.
- **Beziehungen** zwischen den Tabellen sind **vordefiniert**
  - Join-Operationen können durch entsprechende Zugriffspfade unterstützt werden
  - Schnelle Antwortzeiten sind möglich
- Stern-Schema kann leicht in relationales DB-Schema umgesetzt werden.

2.3.3 Mehrdimensionales Datenmodell

- Umsetzung des Stern-Schemas in relationale Tabellen:
  - **Kennzahlentabelle (major table):** Die Gegenstände der Analyse (Kennzahlen) werden in dieser Tabelle gesichert
  - **Nebentabelle (minor tables):** Jede Dimension wird zu einer eigenen Relation / Tabelle.

Kennzahlentabelle:

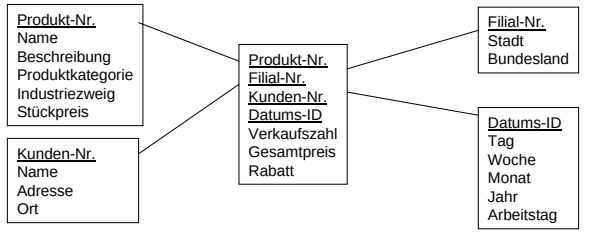
- Jedes **Tupel der Kennzahlentabelle** besteht aus
  - einem Zeiger für jede Dimensionstabelle (Fremdschlüssel), die den Kontext eindeutig definieren und
  - den numerischen Werten (**Daten**) für den jeweiligen Kontext.
- Sie enthält die eigentlichen Geschäftsdaten, die analysiert werden sollen.
- Die Kennzahlentabelle kann sehr viele Zeilen enthalten (Millionen).
- Der Schlüssel der Kennzahlentabelle wird durch die Gesamtheit der Dimensionszeiger gebildet

2.3.3 Mehrdimensionales Datenmodell

Dimensionstabelle:

- Jede **Dimensionstabelle** enthält
  - einen eindeutigen Schlüssel (z.B. Produktnummer) und
  - beschreibende Daten der Dimension (**Attribute**).
- Dimensionstabellen sind deutlich kleiner als die Kennzahlentabelle.
- Zusammenhang zur Kennzahlentabelle über Schlüssel/Fremdschlüssel-Relation

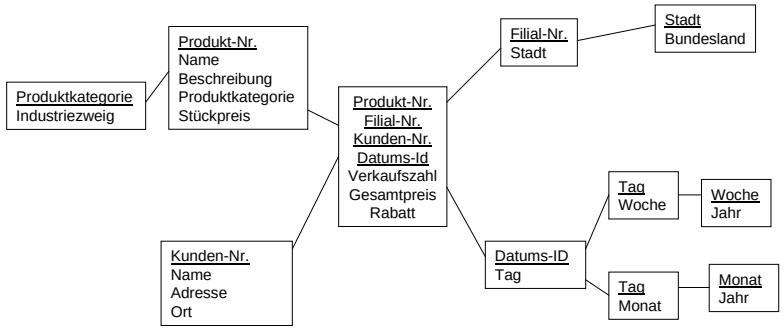
Beispiel: Tabellen abgeleitet aus einem Stern-Schema:



2.3.3 Mehrdimensionales Datenmodell

Schneeflocken-Schema:

- Stern-Schema repräsentiert die Attribut-Hierarchien in den Dimensionen nicht explizit.
- Explizite Hierarchie kann durch sog. **Schneeflocken-Schemata (Snowflake Schema)** erreicht werden.
- **Beispiel:** Schneeflocken-Schema



2.3.3 Mehrdimensionales Datenmodell

MOLAP: **Multidimensional On-Line Analytical Processing**

Spezifische Produkte für OLAP, die auf einer eigenen, proprietären mehrdimensionalen Datenbank beruhen.

Intern beruht die Datenbank auf einer Zell-Struktur, bei der jede Zelle entlang jeder Dimension identifiziert werden kann.

ROLAP: **Relational On-Line Analytical Processing**

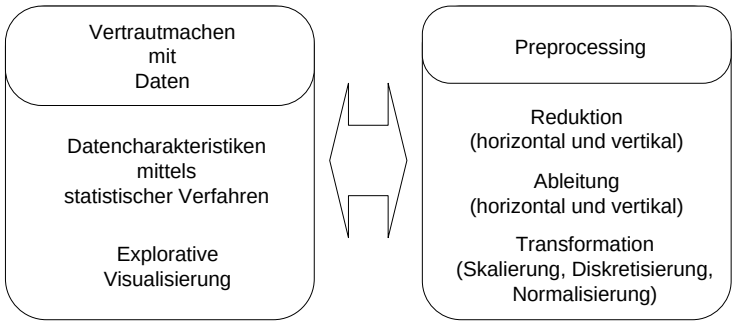
Produkte, die eine multidimensionale Analyse auf einer relationalen Datenbank ermöglichen.

Sie speichern eine Menge von Beziehungen, die logisch einen mehrdimensionalen Würfel darstellen, aber physikalisch als relationale Daten abgelegt werden.

2.4 Preprocessing

Preprocessing

Preprocessing bereitet sowohl die Daten als auch den Analysten auf die Aufgabe vor.



Mehr zum Preprocessing, wenn die Verfahren eingeführt sind!