Data Warehousing

Übungen zu DIS, Sommersemester 2010

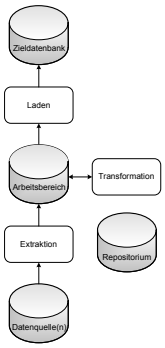
Data-Warehouses

- Beispielanwendung
  - Ein Manager möchte diverse Kennzahlen über seine Abteilungen erfahren:
    - Durchschnittliche Gehälter, Anzahl der Mitarbeiter pro Organisationseinheiten und die Zielerfüllung - jeweils nach Jahr, Alter und Geschlecht aufgeschlüsselt.
  - Die Daten sind in verschiedenen Systemen für Personalverwaltung, Projektverwaltung und Unternehmenssteuerung gespeichert.

Übersicht	Anzahl Beschäftige nach Geschlecht und Alter					Gehalt		Zielerfüllung
	Gesamt	bis 39		ab 40		Gesamt	Durchschnitt	Gesamt
		männlich	weiblich	männlich	weiblich			
2009								
Gesamt	68942	22337	15443	21544	9618	126107773,82	1770,97	...
Fertigung	18614	8376	3257	5584	1397	32384781,88	1689,42	...
Nahrung	2329	466	583	682	598	3436462,79	1475,51	...
Maschinenbau	6523	2962	1231	1957	373	12333492,71	1890,77	...
andere	9762	4948	1443	2945	426	16614826,38	1701,99	...
Dienstleistung	37918	10051	9562	12438	5867	72475706,84	1911,38	...
andere	12410	3910	2624	3522	2354	21247285,10	1712,11	...
2010								
Gesamt	69037	21859	15672	22738	8768	1231036676	1699,23	...
...	...	...	...	...	...	...	...	...

Data-Warehouses (2)

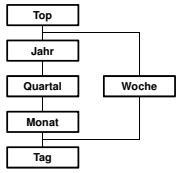
- ETL-Prozess
  - **E**xtraktion
    - ... der relevanten Daten aus verschiedenen Datenquellen
    - Erfolgt periodisch, ereignisgesteuert oder anfragegesteuert
  - **T**ransformation
    - ... der Daten in das Schema und Format der Zieldatenbank
    - Syntaktisch (Datumsformate, Zeichenketten, ...)
    - Semantisch (Duplikate, Werte umrechnen, ...)
  - **L**aden
    - ... der Daten in die Zieldatenbank
- ETL-Prozess durch Repositorien unterstützt:
  - Enthalten Regeln für Extraktion und Transformation



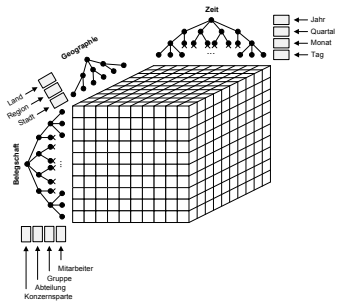
Multidimensionale Sichtweise

- Datenwürfel
- Datenstrukturen
  - Qualifizierende Daten (Kategorieattribute)
  - Quantifizierende Daten (Summenattribute)

Dimensionen
(Kategoriehierarchien)



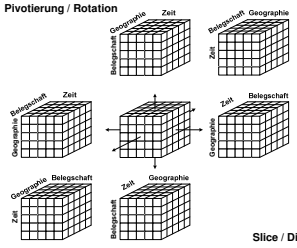
Datenwürfel mit mehreren Dimensionen



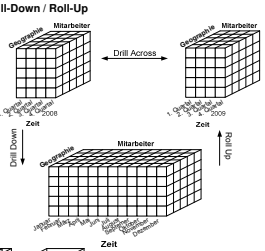
Multidimensionale Sichtweise (2)

- Operationen während der Datenanalyse

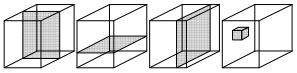
Pivotierung / Rotation



Drill-Down / Roll-Up



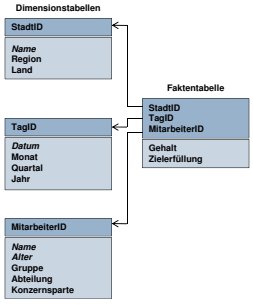
Slice / Dice



Relationale Abbildung multidimensionaler Daten

- Trennung von Struktur und Inhalt:
  - Zentrale **Faktentabelle**
    - Wenige Spalten, viele Tupel
    - Teilweise numerische quantifizierende Attribute für Performanz
  - **Dimensionstabellen**
    - Merkmals- und Kategorisierungsattribute (Strukturdaten)
    - Viele Spalten da Strukturdaten
    - Relativ wenige Tupel
    - Fremdschlüsseleigenschaft verbindet die Faktentabelle mit Dimensionstabellen

Stern-Schema

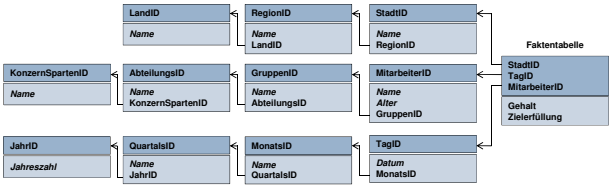


Relationale Abbildung multidimensionaler Daten (2)

- Problem beim Stern-Schema:
  - Redundanz

➔ Dimensionstabellen normalisieren

➔ Snowflake-Schema

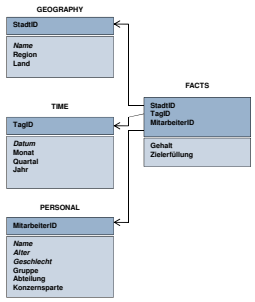


Star Query

- Gruppierungsanfragen auf Stern- (bzw. Snowflake-) Schema
- Beispiel:

  Gesamtgehalt der Mitarbeiter pro Region und Geschlecht, sofern mehr als 1400€ Gehalt

```
SELECT      G.Region, P.Geschlecht,
            SUM(Gehalt)
FROM        facts       NATURAL JOIN
            geography G NATURAL JOIN
            personal P
WHERE       …
GROUP BY    (G.Region, P.Geschlecht)
```



Star Query (2) – Probleme Kreuztabelle

- Kreuztabelle
  - Typische Ergebnisausgabe
  - Zeilen und Spaltensummen
- Gruppierungskombinationen
  - Mehrere Anfragen nötig, da jede Granularität eine eigene Gruppierungsanfrage benötigt (z.B. beim Roll-Up)
  - Bei n Attributen gibt es 2^n mögliche Kombinationen
  - Kombinationen im Beispiel (3 Attribute = 2^3 = 8 Kombinationen):
    - 3er: (Konzernsparte, Jahr, Geschlecht)
    - 2er: (Konzernsparte, Jahr), (Jahr, Geschlecht), (Konzernsparte, Geschlecht)
    - 1er / 0er: (Geschlecht), (Jahr), (Konzernsparte), ()
- Idee: Erweiterung des GROUP BY auf Mengen von Granularitäten

Übersicht		Anzahl Beschäftige nach Geschlecht		
		Gesamt	männlich	weiblich
2008	Gesamt	37780	22337	15443
	Fertigung	11633	8376	3257
	Dienstl.	19613	10051	9562
	andere	6534	3910	2624
2009	Gesamt	37688	22388	15300
	Fertigung	11590	8379	3211
	Dienstl.	19522	10199	9323
	andere	6576	3810	2766
Gesamt		75468	44725	30743

SQL-Erweiterungen: Mehrfachgruppierung

- Erweiterung der GROUP-By-Klausel:
  - Explizite Aufzählung der gewünschten Gruppierungskombinationen

```
SELECT Jahr, Konzernsparte, Geschlecht, SUM(Gehalt) FROM …
GROUP BY GROUPING SETS ((Konzernsparte, Jahr), (Jahr, Geschlecht), (Jahr), ())
```

```
JAHR  KONZ.SPARTE  GESCHL  SUM(Gehalt)
----- ------------ ------- ------------
2008  Fertigung    -       1200000
2008  Dienstl.     -       2500000
2009  Fertigung    -       1500000
2009  Dienstl.     -       3000000
2008  -            m       3000000
2008  -            w       7000000
2009  -            m       3000000
2009  -            w       1500000
2008  -            -       3700000
2009  -            -       4500000
-     -            -       8200000
```

  - Jedes GROUPING SET erzeugt eine eigene Tupelmenge!

SQL-Erweiterungen (2)

- Erweiterung der GROUP-By-Klausel:
  - Abkürzung für die Aufzählung aller 2^n möglichen Gruppierungskombinationen

```
SELECT Jahr, Konzernsparte, SUM(Gehalt) FROM …
GROUP BY CUBE(Jahr, Konzernsparte)
```

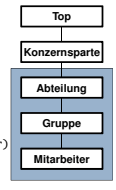
ist äquivalent zu

```
SELECT Jahr, Konzernsparte, SUM(Gehalt) FROM …
GROUP BY GROUPING SETS ((Jahr, Konzernsparte), (Jahr), (Konzernsparte), ())
```

```
JAHR  KONZ.SPARTE  SUM(Gehalt)
----- ------------ ------------------
2008  Fertigung    1200000
2008  Dienstl.     2500000
2009  Fertigung    1500000
2009  Dienstl.     3000000
2008  -            3700000
2009  -            4500000
-     Fertigung    2700000
-     Dienstl.     5500000
-     -            8200000
```


SQL-Erweiterungen (3)

- Mehrfachgruppierung mit abhängigen Attributen
  - Erinnerung: Schema einer Dimension
    - Halbgeordnete Menge von Attributen
  - Berechnung aller Gruppierungen mit Abteilung, Gruppe, Mitarbeiter
    - Aufgrund der funktionalen Abhängigkeiten gibt es äquivalente Gruppierungskombinationen, beispielsweise




```
(Abteilung, Gruppe, Mitarbeiter)   ■ (Gruppe, Mitarbeiter)
                                   ■ (Abteilung, Mitarbeiter)
                                   ■ (Mitarbeiter)

(Abteilung, Gruppe)                ■ (Gruppe)
```

  - **ROLLUP**-Operator
    - Nicht wie bei CUBE 2^n, sondern nur (n+1) Gruppierungskombinationen

```
... GROUP BY ROLLUP(Abteilung, Gruppe, Mitarbeiter)
```

ist äquivalent zu

```
... GROUP BY GROUPING SETS ((Abteilung, Gruppe,
Mitarbeiter), (Abteilung, Gruppe), (Abteilung), ())
```