Informationssysteme und Datenanalyse

Tutorium: Data Warehousing

Tutoren

Fachgebiet Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

http://www.dima.tu-berlin.de/

Überblick

- Heute:
  - OLAP vs. OLTP
  - Data Warehouse
  - Multidimensionale Modellierung

OLAP vs. OLTP (I)

- Datenbankanfragen können grob in zwei Gruppen eingeordnet werden: **OLTP** und **OLAP**.

- OLTP („Online Transactional Processing“):
  - *Anfragen zur operativen Verwaltung von Daten.*
  - *Hohes Volumen an kurzen, transaktionalen Anfragen, die (in der Regel) nur wenige Datensätze berühren.*
  - *Anfragen verändern einzelne Einträge in der Datenbank.*
  - *Fokus auf schnelle Transaktionen & Datenintegrität.*
  - *Typische Anwendungen: Kontenverwaltung, Bestellabwicklung, Rechnungssystem, Finanztransaktionen, …*

- OLAP („Online Analytical Processing“):
  - *Anfragen zur Analyse von Daten.*
  - *Niedriges Volumen an teuren Anfragen, die die gesamte (oder zumindest große Teile) der Datenbank berühren.*
  - *Anfragen verändern die Datenbank nicht.*
  - *Fokus auf effizienten Lesezugriff & schnelle analytische Anfragen.*
  - *Typische Anwendungen: Budgetverwaltung, Reporting, Verkaufsanalyse, Marketingstrategien, …*

OLAP vs. OLTP (II)

	OLTP	OLAP
Typische Operationen	Insert, Update, Delete, Select	Select, Bulk-Inserts
Transaktionen	viele, kurze	Lange Lesetransaktionen
Typische Anfragen	Einfache Anfragen, Primärschlüsselzugriff, Schnelle Abfolgen von Selects/inserts/updates/deletes	Komplexe Anfragen: Aggregate, Gruppierung, Subselects, etc. Bereichsanfragen über mehrere Attribute
Daten pro Operation	Wenige Tupel	Mega-/ Gigabyte
Datenmenge in DB	Gigabyte	Terabyte
Eigenschaften der Daten	Rohdaten, häufige Änderungen	Abgeleitete Daten, historisch & stabil
Erwartete Antwortzeiten	Echtzeit bis wenige Sek.	Minuten
Modellierung	Anwendungsorientiert	Themenorientiert
Typische Benutzer	Sachbearbeiter, Kunde	Management

Aufgabe 1: OLAP vs. OLTP

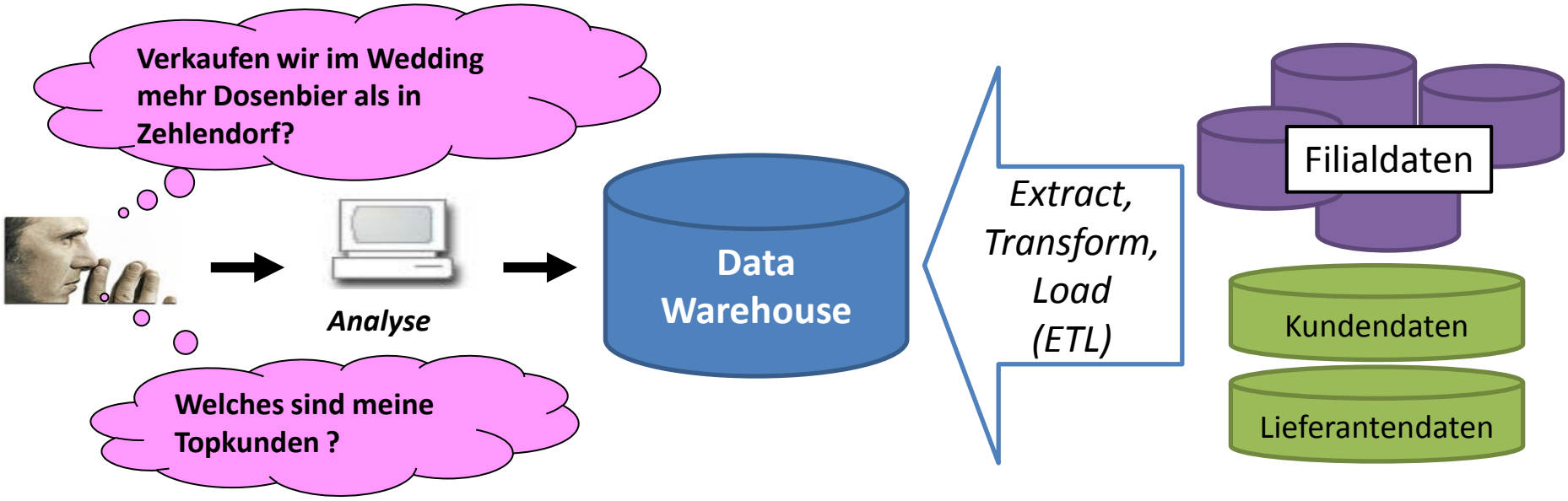
- Klassifizieren sie die folgenden Anwendungen in OLAP und OLTP:
  - Kassenverwaltung im Supermarkt.
  - Auswirkung von Werbekampagnen auf Verkaufszahlen bestimmen.
  - Ticketwebseite für Konzerte.
  - Überwachung des Flugraums (Fluglotsen).
  - „Wird oft zusammen gekauft“ (Amazon).
  - Identifizieren der wichtigsten Kunden.

Aufgabe 1: Lösung

- Klassifizieren sie die folgenden Anwendungen in OLAP und OLTP:
  - Kassenverwaltung im Supermarkt.
    - *OLTP: Verkäufe sind Transaktionen.*
  - Auswirkung von Werbekampagnen auf Verkaufszahlen bestimmen.
    - *OLAP: Vergleich von Verkaufsdaten vor & nach der Kampagne.*
  - Ticketwebseite für Konzerte.
    - *OLTP: Ticketverkäufe sind Transaktionen.*
  - Überwachung des Flugraums (Fluglotsen).
    - *OLTP: Viele transaktionale Updates (Flugzeugpositionen).*
  - „Wird oft zusammen gekauft“ (Amazon).
    - *OLAP: Untersuchung von Verkaufsdaten.*
  - Identifizieren der wichtigsten Kunden.
    - *OLAP: Untersuchung von Verkaufsdaten.*

Data Warehouse

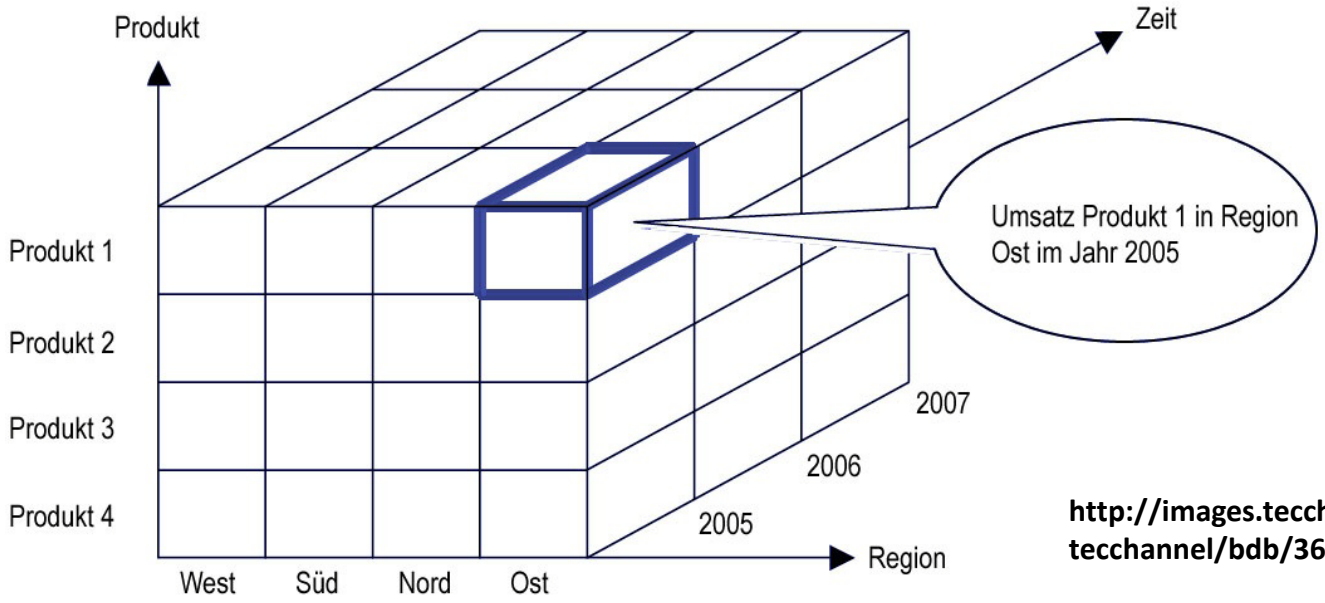
- Ein Data Warehouse ist eine zentrale Datenbank in der Daten aus verschiedenen Quellen einheitlich zusammengefasst und für analytische Anfragen bereitgestellt werden.
  - *Quellen sind typischerweise OLTP Systeme.*
  - *Daten werden in regelmäßigen Abständen aus den Quellen extrahiert und dem Data Warehouse hinzugefügt (ETL).*
  - *Analysten können (lesend!) auf die Datenbasis im Data Warehouse zugreifen um Fragen zu beantworten.*



Der OLAP-Würfel

- Die typische Darstellung von Daten in einem Data Warehouse ist der „OLAP-Würfel“ (OLAP cube).
  - *Einzelne Datenpunkte sind Elemente eines mehrdimensionalen Würfels.*
  - *Dimensionen sind häufig hierarchisch unterteilt.*

- Beispiel: Data Warehouse für Verkaufsdaten.
  - *Dimensionen: Verkaufsdatum, Region und Produkt.*
  - *Zellen enthalten Verkaufspreise eines bestimmten Produktes das zu einem bestimmten Zeitpunkt in einer bestimmten Region verkauft wurde.*
  - ***Beachte:** Logische Repräsentation, die meisten Zellen sind leer!*

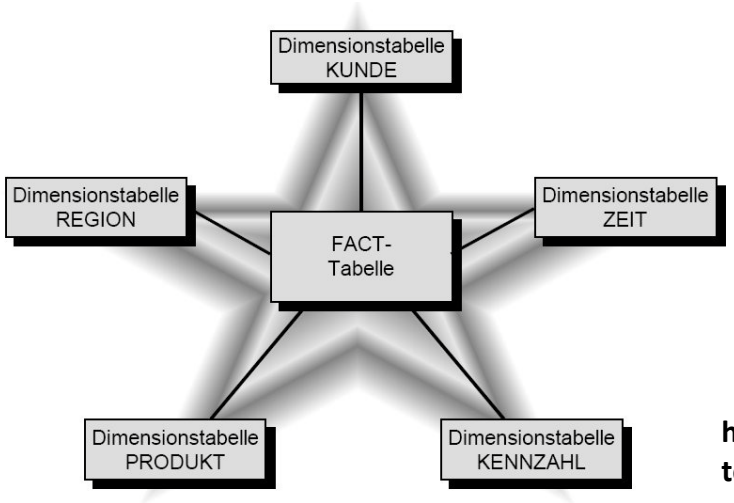


http://images.tecchannel.de/images/tecchannel/bdb/362924/890.jpg

Mehrdimensionale Modellierung

- Der OLAP Würfel ist lediglich eine logische Repräsentation der Daten in einem Data Warehouse.
  - *In der Datenbank: Darstellung durch Relationen.*

- Das Sternschema („Star Schema“) ermöglichte die Modellierung von OLAP Würfeln mittels relationaler Tabellen:
  - *Jede Zelle im OLAP Würfel entspricht einem Eintrag in einer zentralen Faktentabelle.*
  - *Informationen über Dimensionen sind in Dimensionstabellen referenziert*
  - *Diese sind sternförmig um die Faktentabelle angeordnet, und per PK-FK mit dieser verknüpft.*



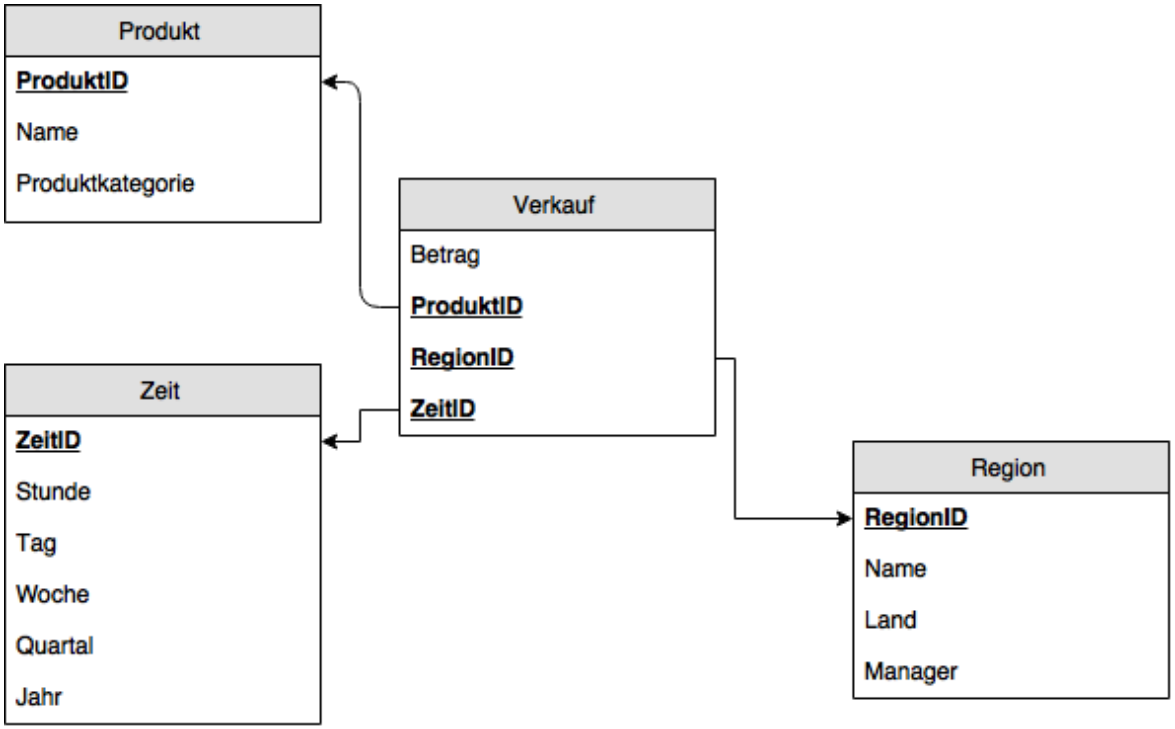
http://images.tecchannel.de/images/tecchannel/bdb/364608/890.jpg

Aufgabe 2: Mehrdimensionale Modellierung

- Modellieren Sie den OLAP Würfel von Folie 8 (Verkaufsdaten nach Produkt, Region, Zeit) als Sternschema.
  - *Erstellen Sie ein ER Schema, achten Sie auf korrekt gesetzte Primär- und Fremdschlüssel.*

- Beachten Sie die folgenden Informationen über die Dimensionen:
  - *Jedes Produkt hat einen Namen und gehört einer Produktkategorie an.*
  - *Jede Region hat einen Namen, ist einem Land zugeordnet und hat einen zuständigen Manager.*
  - *Zeit wird unterteilt in Jahre, Quartale, Wochen, Tage & Stunden.*

Aufgabe 2: Lösung

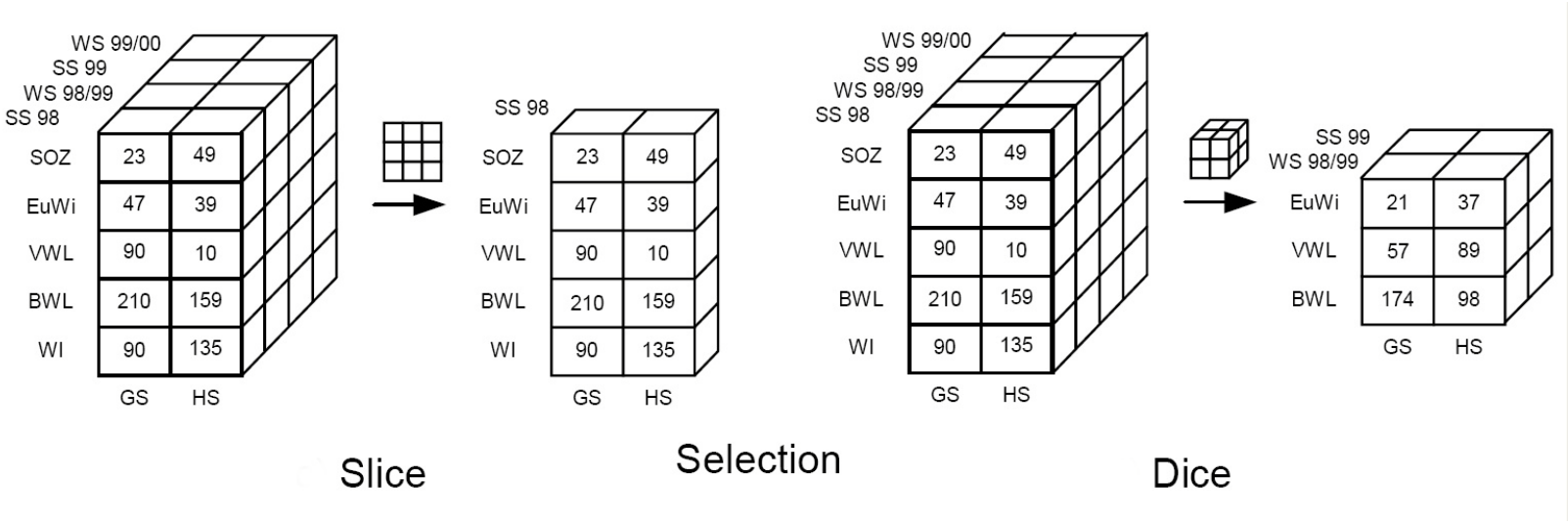


- **Hinweis:** Zeit kann auch als Timestamp modelliert werden. In diesem Fall müssen Funktionen verwendet um die benötigten hierarischen Informationen (Woche/Quartal/Jahr/…) zu erhalten.

Operationen auf dem OLAP Würfel

- Analytische Anfragen auf einem OLAP Würfel sind häufig gruppierte Aggregationen nach den Dimensionen:
  - Gesamtverkäufe nach Quartalen und Abteilungen.
  - Wie oft wurde Produkt X in den verschiedenen Regionen verkauft?
  - Was ist das am häufigsten gekaufte Produkt in Deutschland?

- Logisch werden diese Anfragen auf dem OLAP-Würfel durch Operationen dargestellt, die die Form des Würfels verändern.

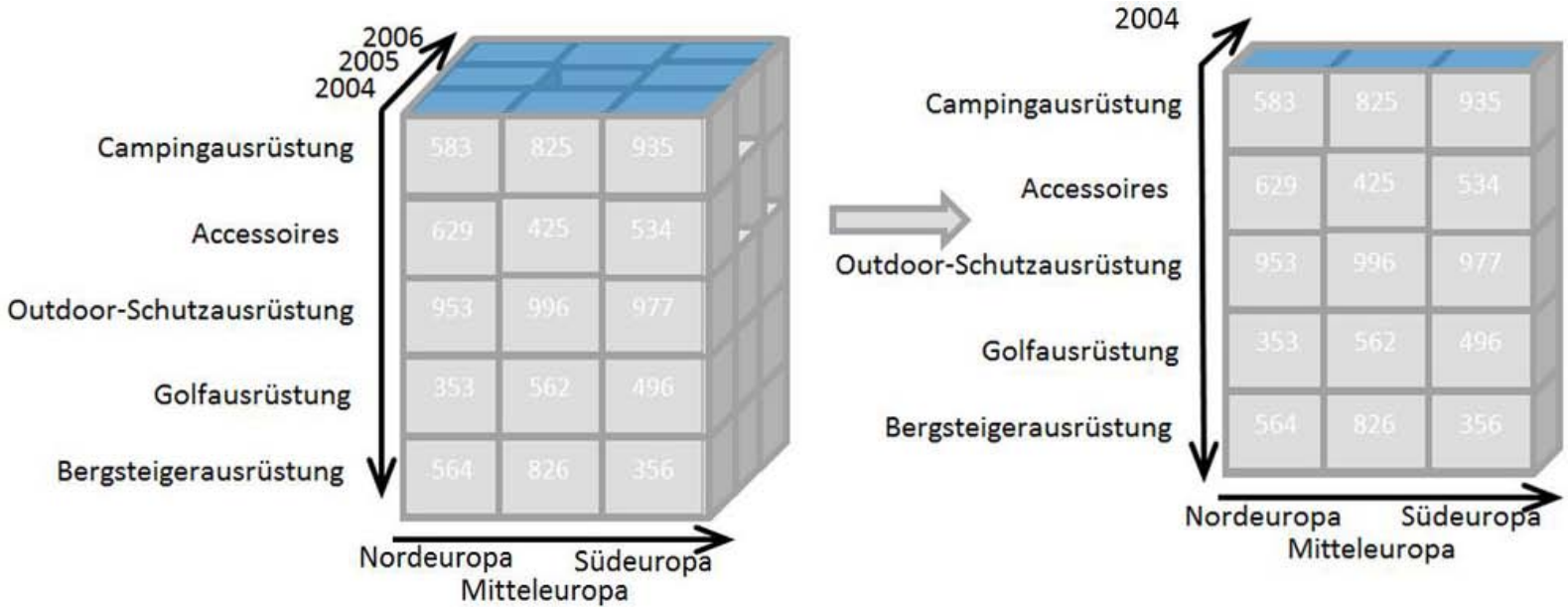


http://images.tecchannel.de/images/tecchannel/bdb/366351/890.jpg

OLAP Operationen: Slice.

- Schneidet eine Scheibe aus dem Würfel heraus.
  - *Erlaubt eine gezielte Analyse für einzelne Werte.*

- Beispielanfragen:
  - *Untersuchung der Verkaufszahlen für das 2. Quartal 2004.*
  - *Wie oft wurde Produkt X in den verschiedene Regionen verkauft?*

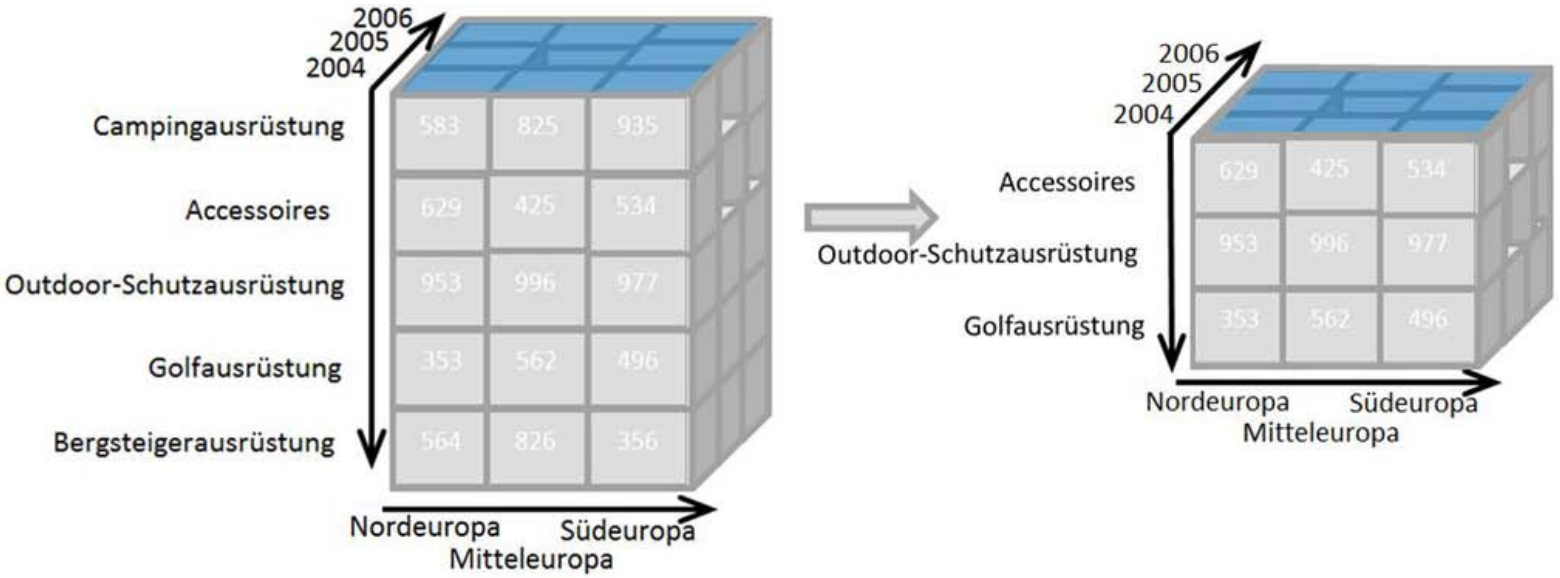


https://upload.wikimedia.org/wikipedia/commons/f/ff/OLAP_slicing.png

OLAP Operationen: Dice.

- Erzeugen eines kleineren Würfels, der einen Teilbereich enthält.
  - *„Generalisierte“ Form der Slice Operation.*

- Beispielanfragen:
  - *Untersuchung der Verkaufszahlen für alle Quartale des Jahres 2004.*
  - *Wie oft wurden die Produkte der Kategorie „Accessoires“ in den verschiedene Regionen verkauft?*

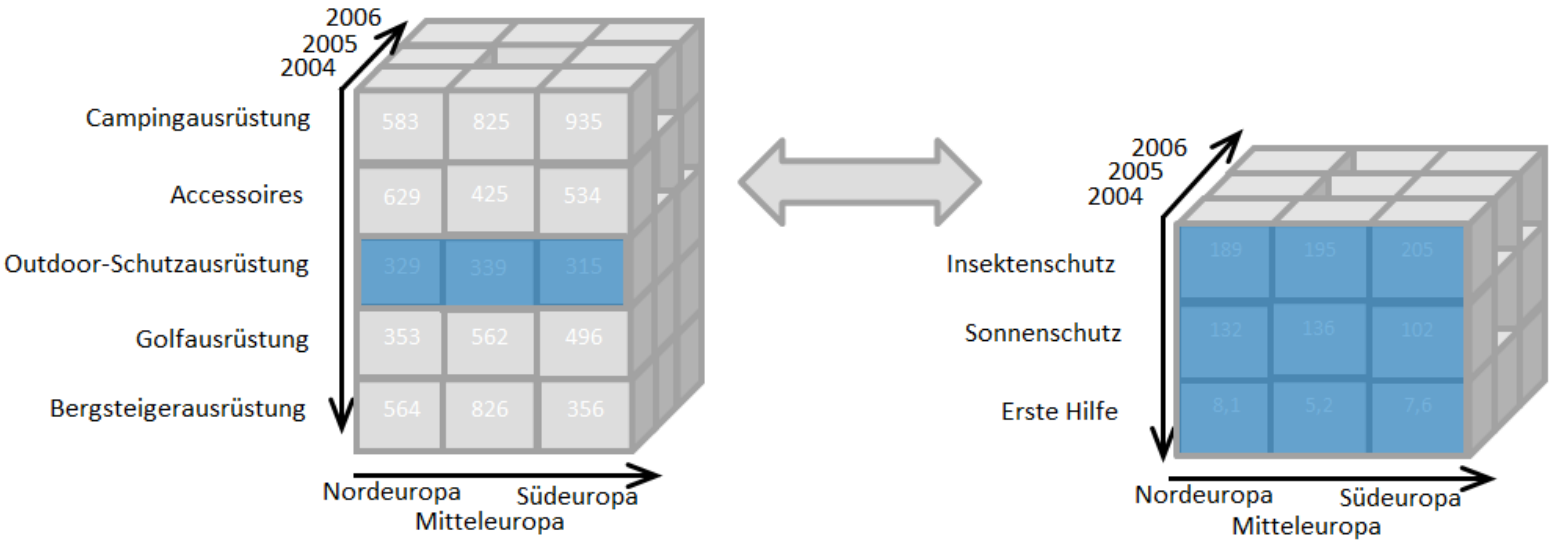


https://de.wikipedia.org/wiki/OLAP-W%C3%BCrfel#/media/File:OLAP_dicing.png

OLAP Operationen: Drill-Up / Drill-Down.

- Verfeinern des Würfels, indem Dimensionen in ihrer Hierarchie genauer dargestellt werden.
  - *Erlaubt es allgemeine Aggregate genauer zu untersuchen.*

- Beispielanfragen:
  - *Wie oft wurden die verschiedenen Unterprodukte der Kategorie „Accessoires“ verkauft?*
  - *In welchem der Länder Europas haben wir den größten Umsatz?*



https://de.wikipedia.org/wiki/OLAP-W%C3%BCrfel#/media/File:OLAP_drill_up%26down.png

Aufgabe 3: OLAP Operationen.

- Geben Sie – für das Schema aus Aufgabe 2 – Beispielanfragen in SQL für die folgenden Operationen an:
  - Slice:
    - Wie oft wurde das Produkt „Samsung Galaxy 6“ in den verschiedenen Regionen und Quartalen verkauft.
  - Dice:
    - Was war der durchschnittliche Umsatz für Produkte der Kategorie „Elektronik“ in den deutschen Regionen im letzten Jahr?
  - Drill-Down:
    - Wie hat sich der Gesamtumsatz für das Produkt „Samsung Galaxy 6“ im letzten Quartal entwickelt, heruntergebrochen nach:
      - … Tagen.
      - … Wochen.
      - … Monaten.

Aufgabe 3: Lösung

- Wie oft wurde das Produkt „Samsung Galaxy 6“ in den verschiedenen Regionen und Quartalen verkauft.
  - ```
    SELECT COUNT(*), R.Name, Z.Quartal
    FROM Verkauf V
      JOIN Region R ON (V.RegionID = R.RegionID)
      JOIN Zeit Z ON (V.ZeitID = Z.ZeitID)
      JOIN Produkt P ON (V.ProduktID = P.ProduktID)
    WHERE P.Name=„Samsung Galaxy 6“
    GROUP BY R.Name, Z.Quartal
    ```

- Was war der durchschnittliche Umsatz für Produkte der Kategorie „Elektronik“ in den deutschen Regionen im letzten Jahr?
  - ```
    SELECT AVG(V.Betrag), R.Name
    FROM Verkauf V
      JOIN Region R ON (V.RegionID = R.RegionID)
      JOIN Zeit Z ON (V.ZeitID = Z.ZeitID)
      JOIN Produkt P ON (V.ProduktID = P.ProduktID)
    WHERE P.Produktkategorie=„Elektronik“
          AND R.Land=„Deutschland“
          AND Z.Jahr=2014
    GROUP BY R.Name
    ```

Aufgabe 3: Lösung

- Wie hat sich der Gesamtumsatz für das Produkt „Samsung Galaxy 6“ im letzten Quartal entwickelt, heruntergebrochen nach [Tag/Woche/Monat]:
  - ```
    SELECT SUM(V.Betrag), Z.Tag
    FROM Verkauf V
      JOIN Zeit Z ON (V.ZeitID = Z.ZeitID)
      JOIN Produkt P ON (V.ProduktID = P.ProduktID)
    WHERE P.Name = „Samsung Galaxy 6“
          AND Z.Jahr=2015 AND Z.Quartal=2
    GROUP BY Z.Tag
    ```
  - ```
    SELECT SUM(V.Betrag), Z.Woche
    FROM Verkauf V
      JOIN Zeit Z ON (V.ZeitID = Z.ZeitID)
      JOIN Produkt P ON (V.ProduktID = P.ProduktID)
    WHERE P.Name = „Samsung Galaxy 6“
          AND Z.Jahr=2015 AND Z.Quartal=2
    GROUP BY Z.Woche
    ```
  - ```
    SELECT SUM(V.Betrag), Z.Monat
    FROM Verkauf V
      JOIN Zeit Z ON (V.ZeitID = Z.ZeitID)
      JOIN Produkt P ON (V.ProduktID = P.ProduktID)
    WHERE P.Name = „Samsung Galaxy 6“
          AND Z.Jahr=2015 AND Z.Quartal=2
    GROUP BY Z.Monat
    ```
    ```