Kapitel 6
Einführung in Data Warehouses

Vorlesung: PD Dr. Peer Kröger

Dieses Skript basiert auf den Skripten zur Vorlesung Datenbanksysteme II an der LMU München von Prof. Dr. Christian Böhm (Sommersemester 2007), PD Dr. Peer Kröger (Sommersemester 2008) und PD Dr. Matthias Schubert (Sommersemester 2009)

http://www.dbs.ifi.lmu.de/cms/Datenbanksysteme_II_13

6 Einführung in Data Warehouses

Übersicht

6 Einführung in Data Warehouses

Übersicht

6.1 Einleitung

Zwei Arten von DB-Anwendungen

- Online Transaction Processing (OLTP)
  - Routinetransaktionsverarbeitung
  - Realisierung des operationalen Tagesgeschäfts wie

    “Buchen eines Flugs”

    “Verarbeitung einer Bestellung”

    “Ein- und Verkauf von Waren”

    …
  - Charakteristik:
    - Arbeitet auf dem jüngsten, aktuellsten Zustand der Daten
    - Änderungstransaktionen (kurze Lese-/Schreibzugriffe)
    - Zugriff auf sehr begrenzte Datenmenge
    - Sehr kurze Antwortzeiten erwünscht (ms-s)
  - OLTP-Datenbanken optimieren typischerweise den logischen und physischen DB-Entwurf hinsichtlich dieser Charakteristik

6.1 Einleitung

Zwei Arten von DB-Anwendungen (cont.)

- Online Analytical Processing (OLAP)
  - Bilden Grundlage für strategische Unternehmensplanung (Decision Support)
  - Anwendungen wie

    „Entwicklung der Auslastung der Transatlantik-Flüge über die letzten 2 Jahre?“

    „Auswirkungen spezieller Marketingaktionen auf Verkaufszahlen der Produkte?“

    „Voraussichtliche Verkaufszahl eines Produkts im nächsten Monat?“

    …
  - Charakteristik:
    - Arbeitet mit „historischen“ Daten (lange Lesetransaktionen)
    - Zugriff auf sehr große Datenmengen
    - Meist Integration, Konsolidierung und Aggregation der Daten
    - Mittlere Antwortzeiten akzeptabel (s-min)
  - OLAP-Datenbanken optimieren typischerweise den logischen und physischen DB-Entwurf hinsichtlich dieser Charakteristik

6.1 Einleitung

Zwei Arten von DB-Anwendungen (cont.)

- OLTP- und OLAP-Anwendungen sollten nicht auf demselben Datenbstand ausgeführt werden
  - Unterschiedliche Optimierungsziele beim Entwurf
  - Komplexe OLAP-Anfragen könnten die Leistungsfähigkeit der OLTP-Anwendungen beeinträchtigen
- Data Warehouse
  - Datenbanksystem, indem alle Daten für OLAP-Anwendungen in konsolidierter Form gesammelt werden
  - Integration von Daten aus operationalen DBs aber auch aus Dateien (Excel, …), …
  - Daten werden dabei oft in aggregierter Form gehalten
  - Enthält historische Daten
  - Regelmäßige Updates (periodisch)

6.1 Einleitung

Operationales DBS vs. Data Warehouse

	operationales DBS	**Data Warehouse**
Ziel	Abwicklung des Geschäfts	Analyse des Geschäfts
Focus auf	Detail-Daten	aggregierten Daten
Versionen	nur aktuelle Daten	gesamte Historie der Daten
DB-Größe	~ 1 GB	~ 1 TB
DB-Operationen	Updates und Anfragen	nur Anfragen
Zugriffe pro Op.	~ 10 Datensätze	~ 1.000.000 Datensätze
Leistungsmaß	Durchsatz	Antwortzeit

6.1 Einleitung

Data Warehouses

- Begriff:

  A Data Warehouse is a subject-oriented, integrated, non-volatile, and time variant collection of data to support management decisions

  [W.H. Inmon, 1996]

6.1 Einleitung

Data Warehouses (cont.)

- Begriff: *A Data Warehouse is a **subject-oriented**, integrated, non-volatile, and time variant collection of data to support management decisions*

  [W.H. Inmon, 1996]

- Fachorientierung (subject-oriented)
  - System dient der Modellierung eines spezifischen Anwendungsziel (meist Entscheidungsfindung in Unternehmen)
  - System enthält nur Daten, die für das Anwendungsziel nötig sind. Für das Anwendungsziel irrelevante Daten werden weggelassen.

6.1 Einleitung

Data Warehouses (cont.)

- Begriff: *A Data Warehouse is a subject-oriented, **integrated**, non-volatile, and time variant collection of data to support management decisions.*

  [W.H. Inmon, 1996]

- Fachorientierung (subject-oriented)
  - System dient nicht der Erfüllung einer Aufgabe (z.B. Verwlatung von Personaldaten)
  - System dient der Modellierung eines spezifischen Anwendungsziel

- Integrierte Datenbasis (integrated)
  - Verarbeitung der Daten aus unterschiedlichen Datenquellen

6.1 Einleitung

Data Warehouses (cont.)

- Begriff: *A Data Warehouse is a subject-oriented, integrated, **non-volatile**, and time variant collection of data to support management decisions.* [W.H. Inmon, 1996]

- Nicht-flüchtige Datenbasis (non-volatile)
  - Stabile, persistente Datenbasis
  - Daten im Data Warehouse werden nicht mehr entfernt oder geändert

6.1 Einleitung

Data Warehouses (cont.)

- Begriff: *A Data Warehouse is a subject-oriented, integrated, non-volatile, and **time variant** collection of data to support management decisions*

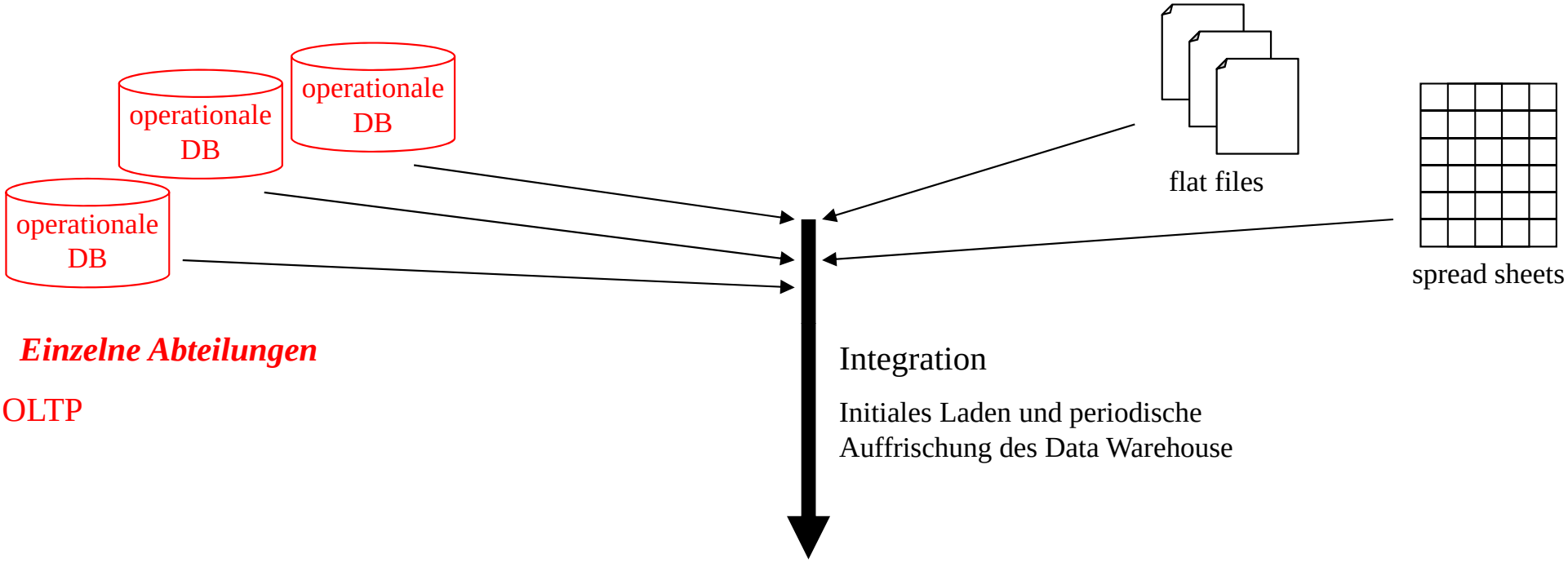
  [W.H. Inmon, 1996]

- Nicht-flüchtige Datenbasis (non-volatile)
  - Stabile, persistente Datenbasis
  - Daten im Data Warehouse werden nicht mehr entfernt oder geändert

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

6.1 Einleitung

Architektur eines Data Warehouse



Einzelne Abteilungen

- OLTP

Integration

Initiales Laden und periodische Auffrischung des Data Warehouse

Data Warehouse

Management

- OLAP
- Decision Support Systems (DSS)
- Executive Information Systems (EIS)
- Data Mining

6.1 Einleitung

Data Warehouses und Data Marts

- Manchmal kann es sinnvoll sein, nur eine inhaltlich beschränkte Sicht auf das Data Warehouse bereitzustellen (z.B. für eine Abteilung)

  => Data Mart

  - Gründe:

    Eigenständigkeit, Datenschutz, Lastenverteilung, …

  - Realisierung:

    Verteilung der DW-Datenbasis

  - Klassen:

    - Abhängige Data Marts: Verteilung eines bestehenden DWs

      => Analysen auf DM konsistent zu Analysen auf gesamten DW

    - Unabhängige Data Marts: unabhängig voneinander entstandene „kleine“ DWs, nachträgliche Integration zum globalen DW

      => unterschiedliche Analysesichten

6 Einführung in Data Warehouses

Übersicht

6.2 Datenmodellierung

Motivation

- Datenmodell sollte bzgl. Analyseprozess optimiert werden
- Datenanalyse im Entscheidungsprozess
  - Betriebswirtschaftliche Kennzahlen stehen im Mittelpunkte (z.B. Erlöse, Gewinne, Verluste, Umsätze, …)

    => ***Fakten***
  - Betrachtung dieser Kennzahlen aus unterschiedlichen Perspektiven (z.B. zeitlich, regional, produktbezogen, …)

    => ***Dimensionen***
  - Unterteilung der Auswertungsdimensionen möglich (z.B. zeitlich: Jahr, Quartal, Monat; regional: Bundesländer, Bezirke, Städte/Gemeinden; …)

    => ***Hierarchien***, ***Konsolidierungsebenen***

6.2 Datenmodellierung

Kennzahlen/Fakten

- Kennzahlen/Fakten
  - Numerische Messgrößen
  - Beschreiben betriebswirtschaftliche Sachverhalte
- Beispiele: Umsatz, Gewinn, Verlust, …
- Typen
  - Additiv: (additive) Berechnung zwischen sämtlichen Dimensionen möglich (z.B. Bestellmenge eines Artikels)
  - Semi-additiv: (additive) Berechnung möglich mit Ausnahme temporaler Dimension(z.B. Lagerbestand, Einwohnerzahl)
  - Nicht-Additiv: keine additive Berechnung möglich (z.B. Durchschnittswerte, prozentuale Werte, …)

6.2 Datenmodellierung

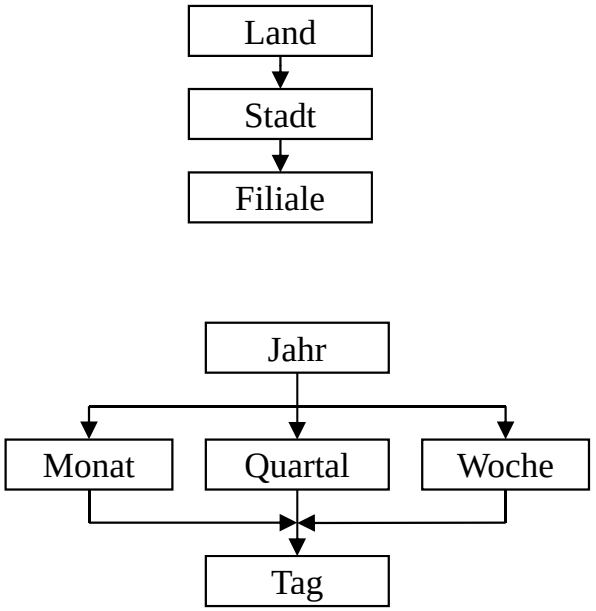
Dimensionen

- Dimension
  - Beschreibt mögliche Sicht auf die assoziierte Kennzahl
  - Endliche Menge von $d \geq 2$ Dimensionselementen (Hierarchieobjekten), die eine semantische Beziehung aufweisen
  - Dient der orthogonalen Strukturierung des Datenraums
- Beispiele: Produkt, Geographie, Zeit

6.2 Datenmodellierung

Hierarchien in Dimensionen

- Dimensionselemente sind Knoten einer Klassifikationshierarchie
- Klassifikationsstufe beschreibt Verdichtungsgrad
- Darstellung von Hierarchien in Dimensionen über Klassifikationsschema
- Formen
  - Einfache Hierarchien: höhere Ebene
    - enthält die aggregierten Werte genau
    - einer niedrigeren Hierarchiestufe
  - Parallele Hierarchien: innerhalb einer
    - Dimension sind mehrere verschiedene
    - Arten der Gruppierung möglich



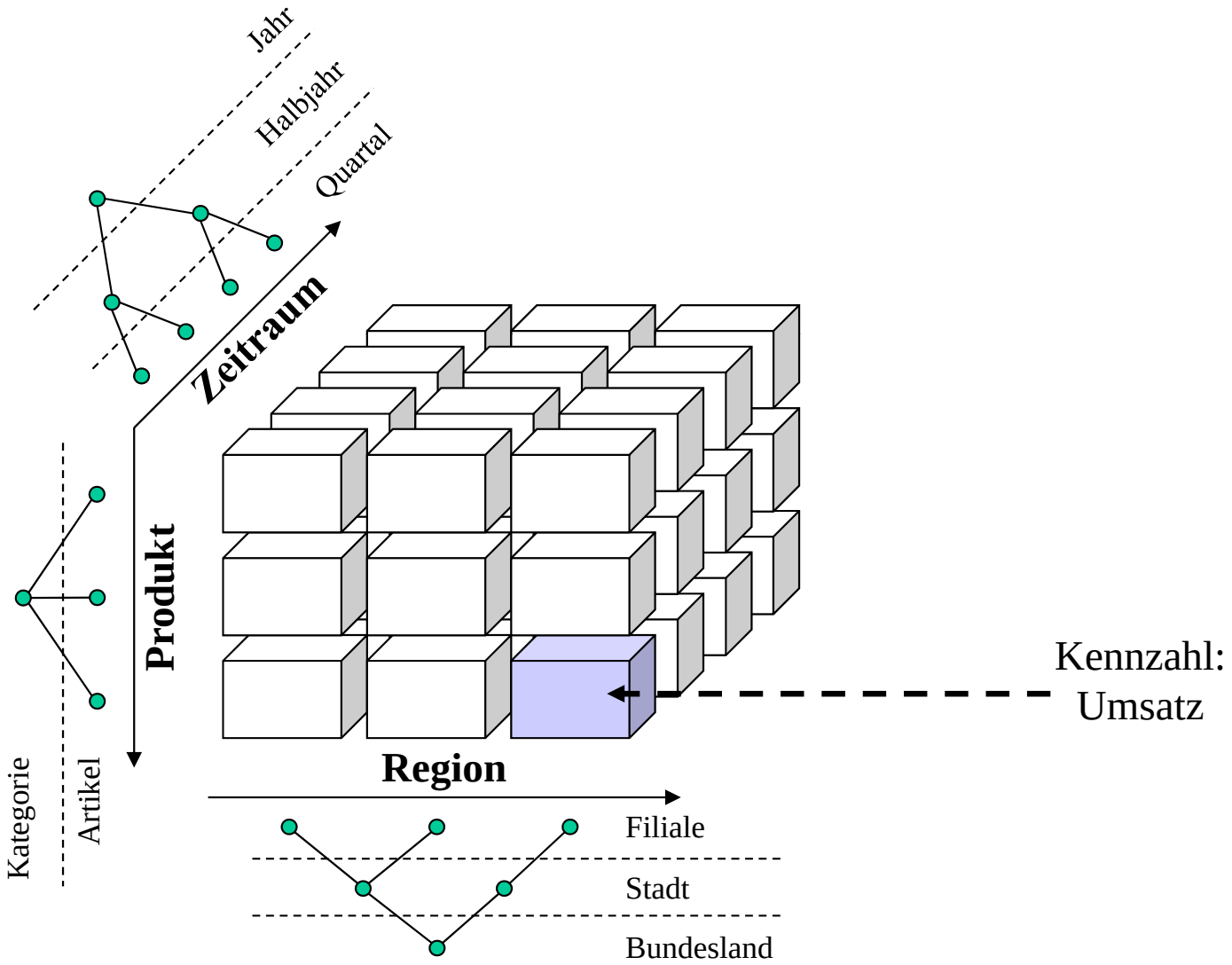
6.2 Datenmodellierung

Data-Cubes

- Grundlage der multidimensionalen Datenanalyse:
  Datenwürfel (***Data-Cube***)
- Kanten des Cubes: Dimensionen
- Zellen des Cubes: ein oder mehrere Kennzahlen (als Funktion der Dimension)
- Anzahl der Dimensionen: Dimensionalität des Cubes
- Visualisierung
  - 2 Dimensionen: Tabelle
  - 3 Dimensionen: Würfel
  - >3 Dimensionen: Multidimensionale Domänenstruktur

6.2 Datenmodellierung

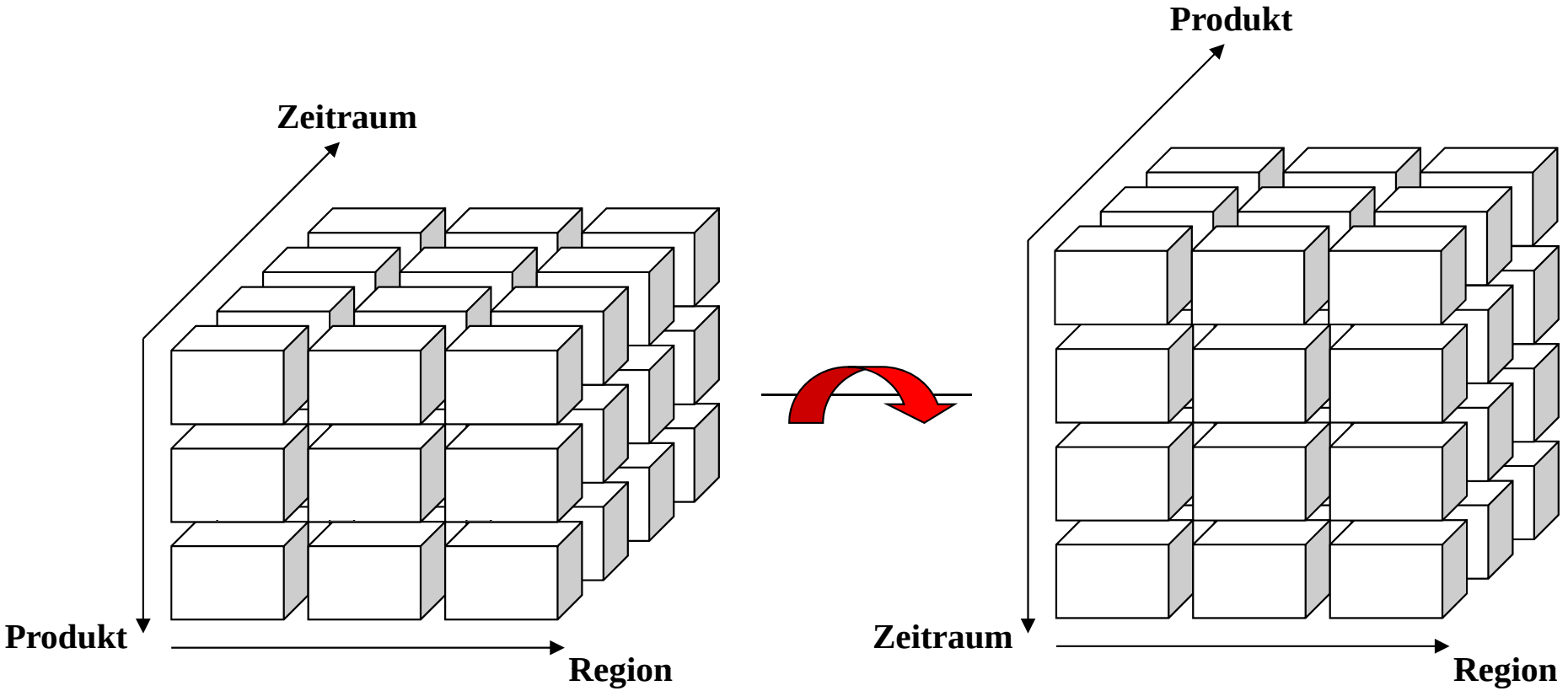
Beispiel: 3D Data-Cube



6.2 Datenmodellierung

Standardoperationen zur Datenanalyse

- Pivotisierung/Rotation
  - Drehen des Data-Cube durch Vertauschen der Dimensionen
  - Datenanalyse aus verschiedenen Perspektiven
  - Beispiel:



6.2 Datenmodellierung

Standardoperationen zur Datenanalyse (cont.)

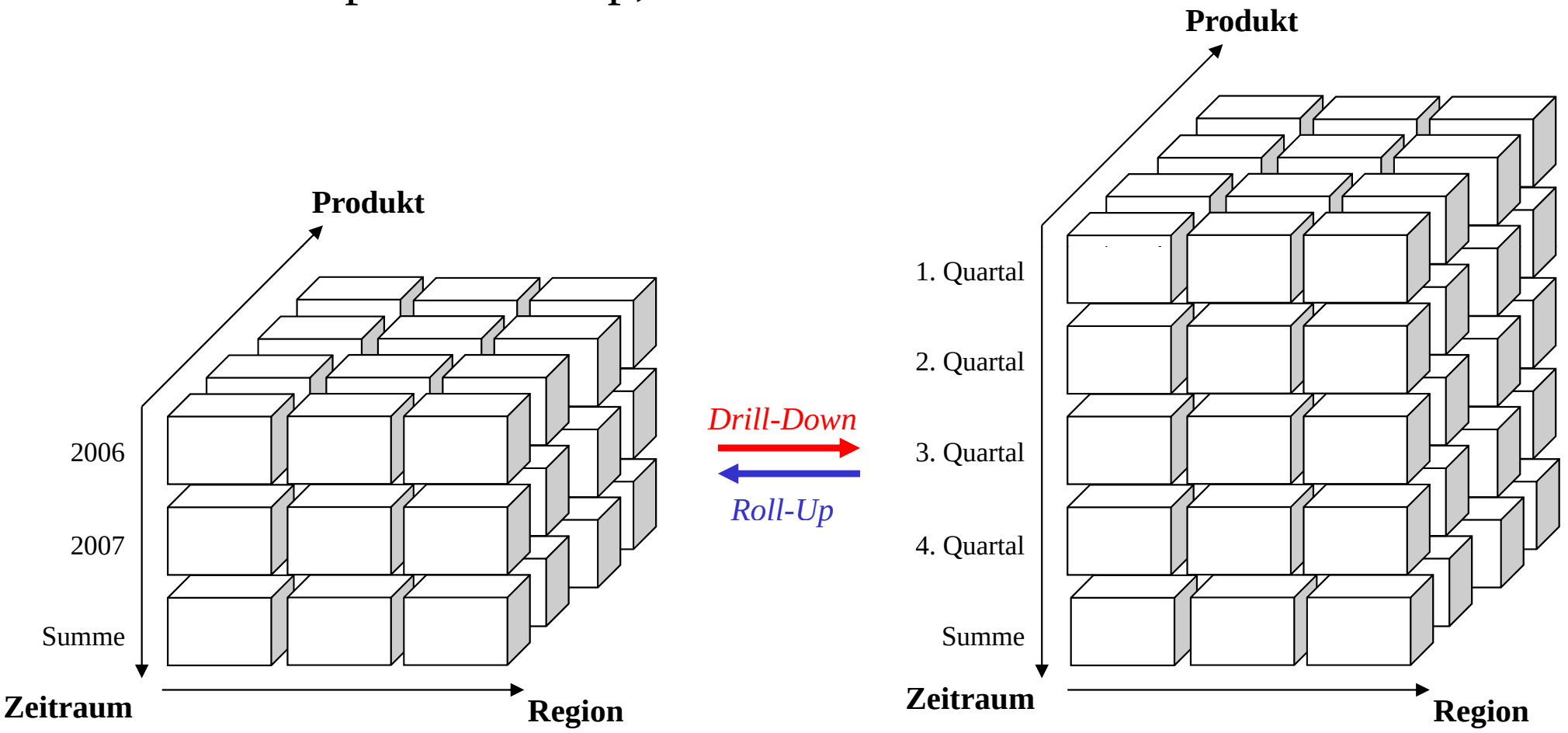
- Roll-Up
  - Erzeugen neuer Informationen durch Aggregierung der Daten entlang der Klassifikationshierarchie in einer Dimension

    (z.B. Tag => Monat => Quartal => Jahr)
  - Dimensionalität bleibt erhalten
- Drill-Down
  - Komplementär zu Roll-Up
  - Navigation von aggregierten Daten zu Detail-Daten entlang der Klassifikationshierarchie
- Drill-Across
  - Wechsel von einem Cube zu einem anderen

6.2 Datenmodellierung

Standardoperationen zur Datenanalyse (cont.)

- Beispiel: Roll-Up, Drill-Down



6.2 Datenmodellierung

Standardoperationen zur Datenanalyse (cont.)

- Slice und Dice
  - Erzeugen individueller Sichten
  - Slice:
    - Herausschneiden von „Scheiben“ aus dem Cube (z.B. alle Werte eines Quartals)
    - Verringerung der Dimensionalität
  - Dice:
    - Herausschneiden eines „Teil-Cubes“ (z.B. Werte bestimmter Produkte und Regionen)
    - Erhaltung der Dimensionalität
    - Veränderung der Hierarchieobjekte

6.2 Datenmodellierung

Standardoperationen zur Datenanalyse (cont.)

- Beispiel: Slice