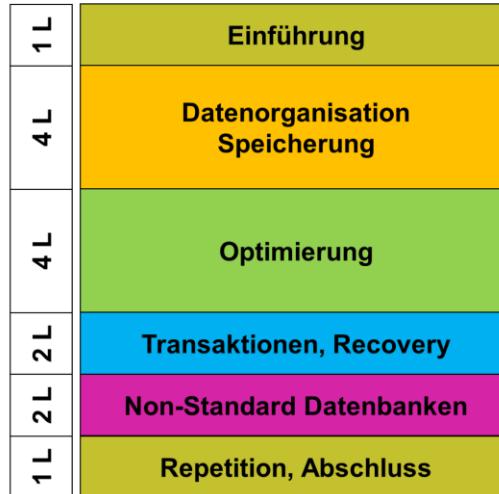


Lektion 12: Dispositive Systeme, data warehouses

Gliederung



← "You are here"

- Konzept der Sperren
- Begriff Blocking, Livelock und Deadlock
- Grundlagen von Recovery kennen:
 - Recovery-Komponenten eines DBMSs
 - Fehlerklassen
 1. Transaktionsfehler
 2. Systemfehler
 3. Mediafehler
 - Logging und Sicherungspunkte

Lernziele heute

Zürcher Hochschule
für Angewandte Wissenschaften



- Unterschiede zwischen **operativen** und **dispositiven** Systemen kennen.
- Begriff **Data warehouse** verstehen.
- Grundlagen der **mehrdimensionalen Modellierung** kennen.

Operative versus dispositive Daten

- **Operative Daten**

Sie werden von Administrations-, Planungs- und Abrechnungssystemen generiert und/oder verarbeitet. Grosse Teile der operativen Daten werden hierbei von sog. OLTP-Systemen erzeugt, bei denen mehrere Benutzer sich derselben Systeme und Datenbestände bedienen, wie beispielsweise bei Auskunfts-, Buchungs- und Bestellsystemen.

- **Dispositive Daten**

In Abgrenzung zu den operativen Daten werden die für management-unterstützende Systeme, sog. OLAP-Systeme, erforderlichen Daten als dispositive Daten bezeichnet. Diese Daten unterscheiden sich erheblich von dem operativen Datenmaterial, so dass ein direkter Durchgriff von management-unterstützenden Systemen auf operative Daten häufig nicht zielführend ist. Dispositive Daten werden aus operativen Daten erzeugt.

- **OLTP = Online Transaction Processing («Tagesgeschäft»):**

- Viele kurze Transaktionen (Datenänderungen und -einfügungen)
 - Normalisiertes Datenschema: Keine Redundanz in den Daten
 - Konsistente Zugriffe auf aktuelle Daten

- Beispiele:

- Hotelreservierungssystem
 - Bankomatabbuchungen
 - ERP-Systeme (Enterprise Resource Planning, Warenwirtschaft)
 - ...

- Ziel: So viele Transaktionen wie möglich pro Zeiteinheit verarbeiten



- **OLAP = Online Analytical Processing («Entscheidungsunterstützung»):**
 - Komplexe Queries mit vielen Joins: keine Updates
 - Redundanz in den Daten, um Abfragen zu optimieren:
 - Materialized views (Ergebnisse werden persistent gespeichert)
 - Spezielle Indizes (Bitmap Indizes)
 - De-normalisierung
- Beispiele:
 - Management-Informations-Systeme von Grossfirmen
 - Wissenschaftliche Datenbanken (CERN, Bioinformatik)
 - Wetterdatenauswertung
 - ...
- Ziel: Abfragen mit niedrigen Antwortzeiten (Sekunden bis Minuten)



OLTP versus OLAP

	Charakteristika operativer Daten	Charakteristika dispositiver Daten
Ziel	Abwicklung der Geschäftsprozesse	Informationen für das Management; Entscheidungsunterstützung
Ausrichtung	Detaillierte, granulare Geschäftsvorfalldaten	Verdichtete, transformierte Daten; umfassendes Metadatenangebot
Zeitbezug	Aktuell; zeitpunktbezogen; auf die Transaktion ausgerichtet	Unterschiedliche, aufgabenabhängige Aktualität; Historienbetrachtung
Modellierung	Altbestände oft nicht modelliert (funktionsorientiert)	Sachgebiets- o. themenbezogen, standardisiert u. endbenutzerfreundlich
Zustand	Häufig redundant; inkonsistent	Konsistent modelliert; kontrollierte Redundanz
Update	Laufend und konkurrierend	Ergänzend; Fortschreibung abgeleiteter, aggregierter Daten
Queries	Strukturiert; meist statisch im Programmcode	Ad-hoc für komplexe, ständig wechselnde Fragestellungen und vorgefertigte Standardauswertungen

8

Warum braucht es denn eine solches DWH eine Betrachtung aus einem anderen Blickwinkel? Wieso kann man nicht einfach VIEWS auf die bestehenden Systeme setzen? -> siehe Folie

Rahmenbedingungen in einer Grossbank

Zürcher Hochschule
für Angewandte Wissenschaften



- Filialen in Zürich, New York, Singapur, ...
- Aufgaben (Beispiele):
 - Monatsendabrechnung: Net New Assets (Neugeldzufluss), Assets under Management (verwaltete Bankprodukte).
 - Performance Analyse: Welche Produkte sind am rentabelsten?
 - Marketing: Welchen Kunden soll man das neue Produkt «High-Interest-Yielder» anbieten?
 - ...
- Ausgangslage:
 - Alle Filialen haben Daten in mehreren Excel-Sheets gespeichert.
 - Daten kommen von unterschiedlichen Systemen mit unterschiedlichem Aktualitätsgrad.
 - Die Berechnungen laufen zu unterschiedlichen Zeiten.
 - Buchhaltungssysteme gemäss lokalen Vorschriften.

- Unterschiedliche Excel-Sheets:
 - Möglicherweise keine gemeinsame «goldene» Quellen (z. B. Produkte)
 - Erkennt man einen Berechnungsfehler, so müssen alle Excel-Sheets ausgebessert werden, Synchronisation?
 - ...
- Unterschiedliche Systeme:
 - Unterschiedliche Berechnungsmethoden
 - Möglicherweise unterschiedliche Quellsysteme
 - Vergleich aggrigerter Werte unterschiedlicher Periodenlänge
 - ...
- Lokale Buchhaltungs- und Rechtsvorschriften:
 - Unterschiedliches Verständnis zu Sachverhalten / Positionen
 - Unterschiedliche Berechnungsmethoden
 - ...

● Gefahr man vergleicht «Äpfel mit Birnen»

10

Lösung: Data Warehouse(s) – DWH

- **Integration** unterschiedlicher Daten in ein System («goldene Quelle»)
- Daten werden regelmässig in das DWH geladen und **historisiert**
- Alle dispositiven Applikationen haben eine **gemeinsame Datenbasis**
- Die Berechnung der Kennzahlen geschieht an einem Ort (**zentral**)



11

Entscheidungsunterstützungssysteme sind Softwaresysteme, die für menschliche **Entscheidungsträger** relevante Informationen für **operative und strategische Aufgaben** ermitteln, aufbereiten, übersichtlich zusammenstellen und bei der Auswertung helfen.

[Wikipedia](#)

System	Einführung	Fokus
Decision Support Systems (DSS)	Anfang 1970er	Modellfokus
Executive Information Systems (EIS)	Ende 1980er	Präsentationsfokus
Data Warehouse (DWH)	Anfang 1990er	Datenfokus
OLAP (Online Analytical Processing)	Anfang 1990er	Modellfokus
Business Intelligence	Anfang 1990er	Präsentationsfokus
Advanced Analytics	Ab ca. 2012	Analysefokus

12

Definition: Data Warehouse

A **data warehouse** is a

- subject-oriented, – getrennt nach Fachbereichen (sachorientiert)
- integrated, – integriert mit einheitlichem Datenmodell
- time-variant, – Datenzustände zu unterschiedlichen Zeitpunkten
- nonvolatile – dauerhaft gespeichert und nicht mehr geändert

collection of data in support of management's decision-making process.

- Die Definition stammt von Bill Inmon:



- Der Begriff Data Warehouse wurde in den 1980ern durch IBM geprägt.

13

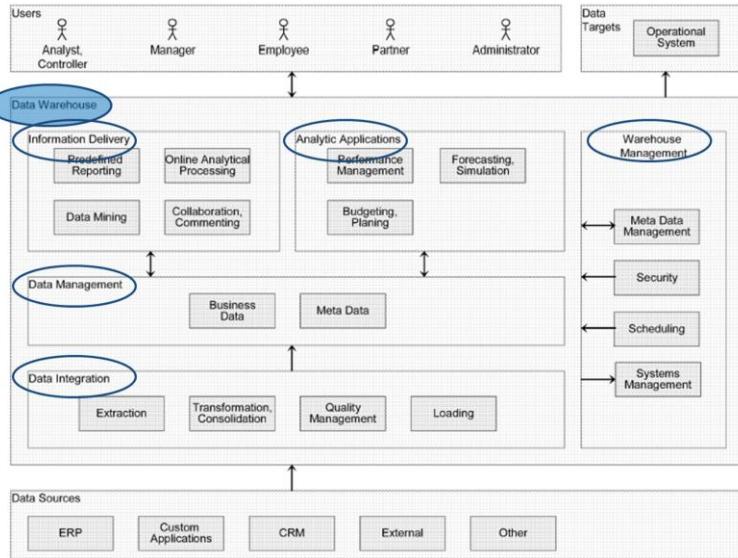
Zürcher Fachhochschule

Themenorientierte, zeitorientierte, integrierte und unveränderliche Datensammlung, deren Daten sich für Managemententscheidungen auswerten lassen. [Inmon 1994]

Merkmale:

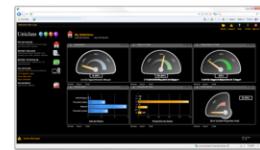
- Strikte **Trennung** von operativen und **dispositiven** Daten.
- **Integration** vielfältiger Datenquellen (auch externe).
- **Langfristige** Datenaufbewahrung, Daten historisieren.
- Auswertungen auf „**mehrdimensionalen**“ Daten.
- **Detaillierte** und **verdichtete** Daten.
- Hohe Datenvolumen (nur Wachstum, keine Löschungen).
- Immer **unternehmens-/anwendungsspezifisch** aufgebaut.
- ...

Data Warehouse – Referenzarchitektur



14

Definition: Business Intelligence



- Verfahren und Prozesse zur **systematischen Analyse** (Sammlung, Auswertung und Darstellung) von Daten in elektronischer Form.
 - Ziel ist die **Gewinnung von Erkenntnissen**, die in Hinsicht auf die Unternehmensziele **bessere** operative oder strategische **Entscheidungen** ermöglichen.
-
- Erstmalige Begriffsverwendung in IBM Journal 1958
 - Popularität durch Gartner 1989
 - Heute wird oft der Begriff „Analytics“ verwendet
-
- Achtung: der Begriff «Intelligence» hat im Englischen in dem Zusammenhang die Bedeutung von «Aufklärung, Einsicht, Information»

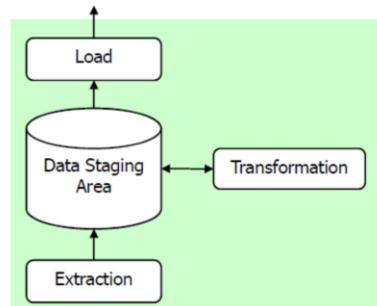
15

Begriff: Data Mart

- Ein Data-Mart ist ein langfristig gehaltener Datenbestand **innerhalb** eines Data-Warehouse-Systems oder die Kopie eines Teilbereichs des Data-Warehouse, der für einen **bestimmten Organisationsbereich** oder eine **bestimmte Anwendung** geschaffen wird. Hierdurch entsteht eine **Teilsicht** auf das Data-Warehouse.
- Viele Architekturvarianten im Einsatz. Übergänge sind fliessend und oft nicht scharf abgrenzbar.
- Datamarts sind «Analysedatenbanken»
(können auch Files sein wie z.B. Excel-Tabellen etc.)

Begriffe: Staging-Area, ETL-Prozess

- Die Daten werden zunächst aus den Datenquellen **extrahiert (E)** und in der Staging-Area **gesammelt und zwischengespeichert**. Dort werden die Daten bereinigt und **transformiert (T)**, ohne die Datenquellen weiter zu belasten. Die so aufbereiteten Daten werden anschliessend aus der Staging-Area in die Zieldatenbank **geladen (L)**.
- Eigenschaften:
 - Temporärer Speicher
 - Quellnahes Schema
 - Einzelne Arbeitsschritte sind effizient implementierbar (Mengenoperationen)
 - „Filterfunktion“: nur einwandfreie Daten gelangen ins Data Warehouse



«Multidimensionalität»

Zürcher Hochschule
für Angewandte Wissenschaften

- Auswertungen basieren i. d. R. auf den Data Marts
- Typische Fragestellungen einer Bank:
 - Entsprechen die Monatsberichte den Basel II / III Anforderungen?
 - Welche Kunden haben die letzten X Monate in die strukturierten Produkte Y investiert?
 - Welches sind die Risiken aller Vermögen, die von Zürich aus verwaltet werden?
 - Wie hoch ist der Neugeldzufluss im 3. Quartal?
 - Was sind die wichtigsten Kennzahlen des Unternehmens und wie ist deren Entwicklung der letzten 4 Quartale?
 - ...



Solche Fragen sollen «ad-hoc» formuliert werden können.

«Multidimensionalität»

Zürcher Hochschule
für Angewandte Wissenschaften

zhaw School of
Engineering
InIT Institut für angewandte
Informationstechnologie

Berechnete
Kennzahlen

Weitere
Kennzahlen

Gewinn pro Produkt und Monat?
Kosten pro Produkt und Monat?

Anteil des Gewinns der einzelnen
Produkte am Gesamtgewinn?
Veränderung des Umsatzes im
Vergleich zum Vormonat?

Weitere Aspekte
und Hierarchien

Umsatz pro
Distributionskanal?
Gewinn pro
Kundengruppe?

Grafische
Darstellungen

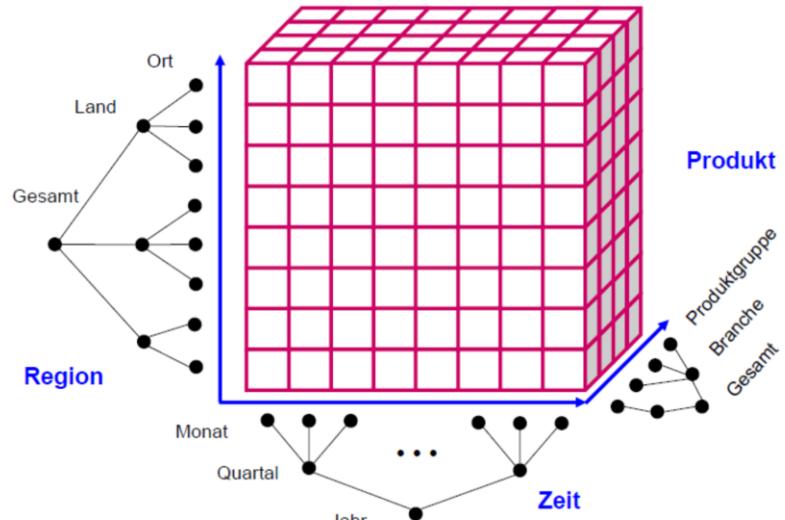
Geeignete graphische
Darstellungen?



...

19

«Multidimensionalität» – Grundideen



20

- Unterscheidung von:
 - **Fakten** (Measures): Gemessene Werte.
 - **Dimensionen**: Beschreibung der Fakten in Raum, Zeit, Organisation, ...
 - **Klassifikationshierarchien**: hierarchische Struktur von Dimensionen.
- Metapher: «**Würfel**» (Cube) bzw. Hypercube
 - Fakten: Punkte im multidimensionalen Raum.
 - Klassifikationshierarchien: Achsenbeschriftung in unterschiedlichem Verfeinerungsgrad.
- Analyse durch Operationen auf dem «Cube»:
 - Dimensionen ausblenden / einblenden.
 - Auswahl von Subwürfeln (Flächen, Punkten, ...).
 - Hierarchiestufe vergröbern/verfeinern.

«Multidimensionalität» – Fakten

Fakten, Measures:

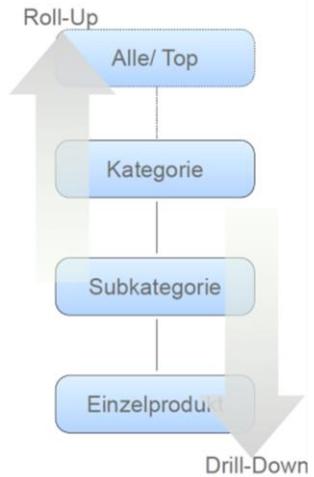
- Eigentliche Objekte der Analyse.
- Ansammlung von Einzelmessungen / Kennzahlen, verknüpft mit Kontextdaten (Dimensionen)
- Jedes Faktum repräsentiert einen Businesswert, eine Transaktion oder ein Ereignis
- In der Regel numerische Werte.
- Beispiele: Verkaufszahlen, Umsatz, Gewinn, Lagerbestand, ...
- Die Analyse der Fakten soll es dem Management ermöglichen, die Leistung des Unternehmens oder bestimmter Bereiche zu überwachen zu bewerten und zu steuern.
- Diese numerischen Werte lassen sich auf verschiedene Weise verdichten bzw. aggregieren.

Dimensionen:

- Kontext der Fakten.
- Eindeutige Strukturierung des Datenraums.
- Hoffentlich orthogonal, Abhängigkeiten zwischen Dimensionen bereiten aber an vielen Stellen Probleme.
- Definieren einen sog. Hyper-Würfel (hyper cube)
- Dimensionen haben Attribute. Diese können hierarchisch geordnet sein.
 - Zeit: Jahr – Quartal – Monat – Woche – Tag ...
- Die Anzahl von Dimensionen ist nicht beschränkt (in Lehrbüchern aber meist auf drei festgelegt. Warum?).
- Fast immer ist eine [Zeitdimension](#) vorhanden.

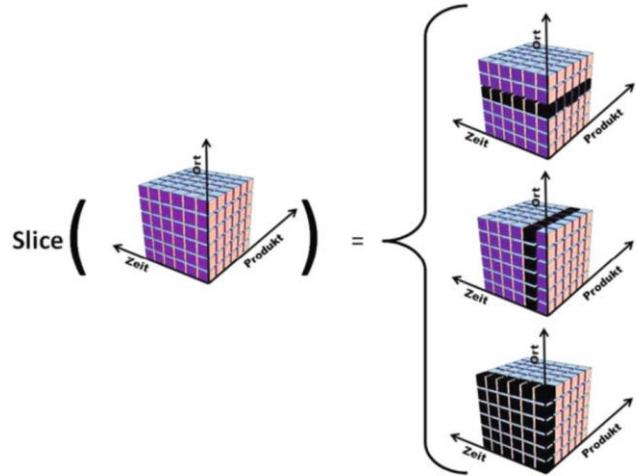
«Multidimensionalität» – Operationen

- **Drill-Down** (mehr Details zeigen)
 - Hierarchie von oben nach unten traversieren
 - Alle Produkte → Kategorie → Subkategorie → Einzelprodukt
 - «Verfeinerung» entlang einer Dimension
- **Roll-Up** (weniger Details zeigen):
 - Hierarchie von unten nach oben traversieren
 - Einzelprodukt → Subkategorie → Kategorie → alle Produkte
 - «Aggregation» entlang einer Dimension



«Multidimensionalität» – Operationen

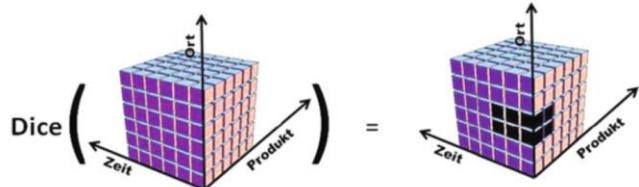
- **Slicing** («Scheiben» ausschneiden): Aggregation über mehrere Dimensionen



25

«Multidimensionalität» – Operationen

- **Dice** («Subwürfel» ausschneiden):

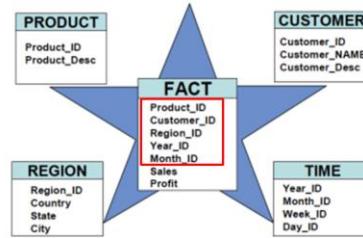


26

«Multidimensionalität» – Implementation

Zürcher Hochschule
für Angewandte Wissenschaften

- **Sternschema:** Kombination von Fakten (inkl. Kennzahlen) und Dimensionen

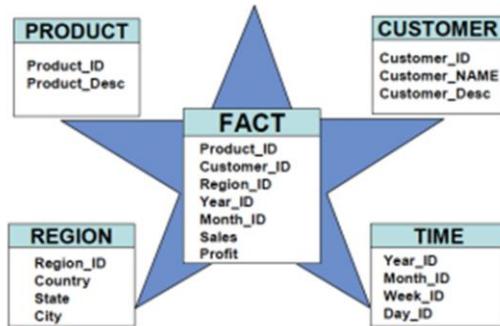


- **Faktentabelle:** Enthält Fakten und Fremdschlüsse (Foreign-Keys) auf Dimensionstabellen.
- **Dimensionstabellen:** Enthalten Dimensionsdaten und Primärschlüsse (Primary-Keys), die mit Faktentabelle «verknüpft» sind.

Eigenschaften des Sternschemas

- Faktendaten werden typischer Weise **sehr gross**:
 - Bis zu **Milliarden** von Records
 - **Partitionierung** der Faktentabellen und **Auslagerung** alter Records nötig
- Fakten sollten sich auf fachlichen Prozess beschränken: z.B. Verkauf
- Ein data mart kann aus unterschiedlichen Faktentabellen bestehen
- Dimensionstabellen werden typischer Weise **nicht sehr gross**, haben dafür aber oft viele Attribute (Spalten) für die hierarchische Gliederung
- Sternschemas sind nicht normalisiert
(es gibt keine updates → es gibt auch keine update-Anomalien)

Sternschemaabfrage – Beispiele



- Liste aller Verkäufe von iPhones in Zürich?

Sternschemaabfrage – Beispiele

Zürcher Hochschule
für Angewandte Wissenschaften



Product

Product_ID	Product_Desc
P1	Motorola
P2	iPhone
P3	Windows

Fact

Product_ID	Cust_ID	Reg_ID	Time_ID	Sales	Profit
P2	C1	R2	T1	1000	200
P2	C2	R3	T3	3000	500

Region

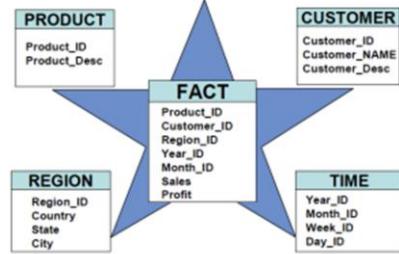
Region_ID	Country	State	City
R1	CH	ZH	Zurich
R2	CH	VD	Lausanne
R3	CH	ZH	Winti

Liste aller Verkäufe von iPhones in Zürich?

30

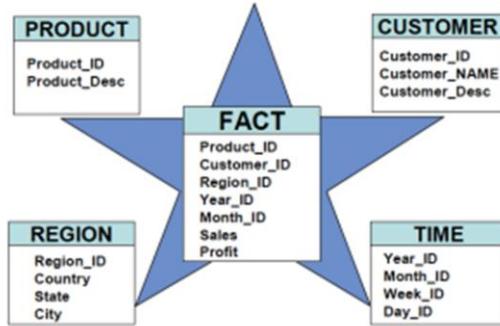
Sternschemaabfrage – Beispiele

```
SELECT r.city, f.sales
FROM Fact f, Product p, Region r
WHERE
    f.product_ID = p.product_ID AND
    f.region_ID = r.region_ID AND
    p.product_desc = 'iPhone' AND
    r.city = 'Zurich'
```



Liste aller Verkäufe von iPhones in Zürich?

Sternschemaabfrage – Beispiele

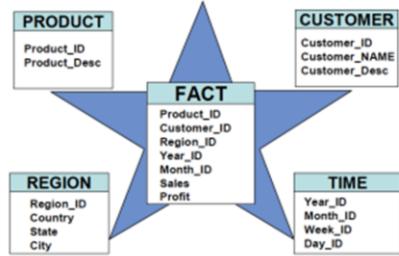


- Anzahl der iPhone-Verkäufe pro Kanton?

Sternschemaabfrage – Beispiele

```
SELECT r.state, sum(f.sales)
FROM Fact f, Product p, Region r
WHERE
    f.product_ID = p.product_ID AND
    f.region_ID = r.region_ID AND
    p.product_desc = 'iPhone' AND
    r.country = 'CH'
GROUP BY
    state
```

Anzahl der iPhone-Verkäufe pro Kanton?



Sternschemaabfrage – Bemerkungen

- SQL für Abfragen auf mehreren Dimensionen gleichzeitig, sehr mühsam:

emp_id	dept_id	first_name	last_name	salary	job	year
1111	10	Martha	White	4400.00	IT_PROG	2001
1112	10	John	Black	8800.00	IT_PROG	2003
1113	20	Bill	Austin	7600.00	MK_REP	2001
1114	20	Diana	Kimes	4300.00	MK_MAN	2003
1115	20	David	Peters	7600.00	IT_PROG	2004
1116	30	Sibile	Peterson	12000.00	AX_ASST	2001
1117	30	Jack	Klein	9900.00	MK_REP	2003
1118	30	Alex	Armstrong	8500.00	MK_REP	2004
1119	30	Jennifer	May	6700.00	AX_ASST	2005
1120	40	Roy	Hunt	9900.00	IT_PROG	2005
1121	40	Wendy	Blunt	8800.00	AX_ASST	2004
1122	50	Valli	Begg	7900.00	MK_MAN	2001
1123	50	Pat	Donaldson	4900.00	MK_MAN	2001

- SQL-Abfrage, die gleichzeitig
 - die Lohnsumme pro Abteilung und Jahr
 - die Gesamtlohnsumme pro Abteilung
 - die Gesamtlohnsumme pro Jahr
 - die Gesamtlohnsumme über alles zurückgibt?

Sternschemaabfrage – Bemerkungen

- SQL:1999, Gruppierung mit speziellen Operatoren:
 - CUBE
 - Berechnung aller 2^n Aggregationen für n Dimensionen
 - ROLLUP
 - Verdichtung entlang von Dimensionen
 - Beispiele:
 - SELECT Artikel, Ort, Datum SUM(Anzahl) FROM ...
GROUP BY CUBE (Artikel, Ort, Datum)
 - SELECT Artikel, Ort, Datum SUM(Anzahl) FROM ...
GROUP BY ROLLUP (Artikel, Ort, Datum)
- Grundsätzlich: Viele Joins nötig
- Queries werden i.d.R. «transparent» durch Tools (z.B. Excel) generiert!

35

Zürcher Fachhochschule

GROUP BY ROLLUP (Artikel, Ort, Datum) liefert die Gruppierungen nach:

- Artikel, Ort, Datum
- Artikel, Ort
- Artikel

GROUP BY CUBE (Artikel, Ort, Datum) liefert die Gruppierungen nach:

- Artikel, Ort, Datum
- Artikel, Ort
- Artikel, Datum
- Artikel
- Ort, Datum
- Ort
- Datum

Ausblick

Zürcher Hochschule
für Angewandte Wissenschaften



- Das nächste Mal: big data & NoSQL
- Lesen: Nichts!

36