



Data Warehousing (I)

Praktikum Data Warehousing und Mining

Aufgabe 1:

In einem Unternehmen existieren die Relationen *promotions* (Daten zu Werbekampagnen), *products* (Daten zu im Unternehmen erstellten Produkten), *customers* (Kundendaten), *times* (Hierarchie über die Zeit) und *sales* (Daten über Verkäufe).

Auf der Basis dieser Relationen soll nun ein Data Warehouse für den Vertrieb erstellt werden.

- Wählen Sie alle Relationen und Attribute aus, die für die Nutzung in einem Data Warehouse geeignet sind. Modellieren Sie das entsprechende Data Warehouse mittels ME/R.
- Vertriebsmitarbeiter liefern folgende zusätzliche Informationen: Das Attribut *CHANNEL_ID* der Relation SALES kann die Werte 2, 3, 4, 5 und 9 annehmen. Dabei haben die einzelnen Zahlenwerte folgende Bedeutung: 2: Verkauf an Partnerunternehmen, 3: Direktverkauf, 4: Verkauf über Internethandel, 5: Katalogbestellungen und 9: Teleshopping. Direktverkauf und Teleshopping werden dann zu Verkäufen mit direktem Kundenkontakt zusammengefasst, Internethandel und Katalogbestellungen kommen durch indirekten Kundenkontakt zu Stande. Der Verkauf an Partnerunternehmen lässt sich hierbei nicht klassifizieren. Alle Vertriebswege werden insgesamt unter dem Titel ‚Vertrieb gesamt‘ aggregiert. Berücksichtigen Sie diese Informationen soweit wie möglich im ME/R Diagramm.
- Setzen Sie nun das ME/R Diagramm sowohl in das Snowflake-Schema als auch in das Star-Schema um. Welche Vor- bzw. Nachteile bieten die einzelnen Repräsentationsweisen? Welche Repräsentation ist in der Praxis geeigneter?
- Welche Umsetzung ist ausgehend von den gegebenen Relationen am schnellsten zu realisieren? Legen Sie die hierfür benötigten Relationen an und füllen Sie diese mit den entsprechenden Daten.
Hinweise: 1. Beachten Sie bei der Erstellung neuer Relationen Fremd- und Primärschlüsselbeziehungen!
2. Es ist möglich diese Aufgabe durch Erstellung genau einer Relation zu lösen.

Aufgabe 2:

Auf den gegebenen Daten sollen nun Anfragen gestellt werden. Geben Sie die Anfragen der einzelnen Teilaufgaben in **genau** einer SQL Anweisung wieder.

- Wie viele Y-Box Geräte (*Y Box*) sind zwischen 08.01.2001 und 03.12.2001 über das Internet verkauft worden. (*Hinweis: time_ID* hat das Format ‘31-JAN-03’.)
Ergebnis: 869
- Wie hoch sind die Umsätze durch den Verkauf von 18" Flat-Panel Monitore (*18" Flat Panel Graphics Monitor*) über indirekte Vertriebswege?
Ergebnis: 1148972.72
- Geben Sie für jeden Monat des Jahrs 2000 und jede Produktkategorie an, wie viele Produkte verkauft wurden. Ordnen Sie das Ergebnis nach Monaten.
Beispiel tupel: {April, Electronics, 2906}, {April, Hardware, 230}
- Ist es möglich das Ergebnis aus c) so zu modifizieren, dass es für jeden Monat genau das Tupel ausgibt, in dem die meisten Produkte verkauft wurden? Wenn ja, wie sieht die entsprechende Anfrage aus?
Beispiel tupel: {April, Software/Other, 7965}, {August, Software/Other, 9356}

- e) Wie viele Artikel wurden in den einzelnen Jahren abgesetzt. Schlüsseln Sie die Produkte nach Produktkategorien auf.
Beispieltupel: {2000, Electronics, 36161}
- f) Bestimmen Sie für jeden Vertriebsweg und jedes Jahr die Höhe des Umsatzes.
Beispieltupel: {Direkt, 1998, 15569726.2}, {Internet, 1998, 2862770.55}
- g) In welchem Monat des Jahres 1999 wurden die meisten Produkte abgesetzt? Wie viele waren es?
Ergebnis: February, 24122
- h) Wie viele Y-Box-Geräte (Y Box) wurden im Jahr 2000 und wie viele davon im Monat Dezember abgesetzt? (*Hinweis: to_char()* erlaubt das Umwandeln von Zahlenwerten in Characterwerte.)
Ergebnis: {2000, 2010}, {December, 147}
- i) Wie hoch waren die Einnahmen durch den Verkauf von Produkten aus der Produktkategorie Hardware (*Hardware*) in den einzelnen Jahren, Quartalen und Monaten? Ordnen Sie Ihre Ausgabe absteigende chronologisch und blenden sie alle Einträge aus, in welchen weniger Umsatz als 300000 gemacht wurden. Ihr Ergebnis sollte ähnlich dem Folgenden sein:

| Jahr | Quartal | Monat | Umsatz |
|------|---------|---------|------------|
| 2001 | null | null | 5684370.01 |
| 2001 | 4 | null | 1174512.68 |
| 2001 | 4 | October | 346370.41 |
| ... | ... | ... | ... |

Beispieltupel: {2001, null, null, 5684370.01}, {2001, 4, null, 1174512.68}, {2001, 4, October, 346370.41}

- j) Die hier gestellten Anfragen entsprechen typischen Anfragen an ein Data Warehouse. Worin liegen nach Ihrer Ansicht die Schwächen von SQL in dieser Anwendungsdomäne? Wie könnte diesen Schwächen begegnet werden?

Aufgabe 3:

Die bisher genutzten Daten sollen in einem nächsten Schritt in einen Data Cube übertragen werden.

- a) Erstellen Sie zunächst die Dimensionen wie Sie sie im ME/R Modell modelliert haben.
- b) Erstellen Sie den Würfel *VERTRIEB_CUBE* unter Verwendung Ihrer Faktentabelle und Ihren Dimensionen.
- c) Für welche Art von Anfragen hat die Darstellung in Form eines Data Cubes Vorteile gegenüber der Darstellung in Relationen?