Exercises Data Warehousing
SQL Extensions & DB Explosion Problem

Solutions

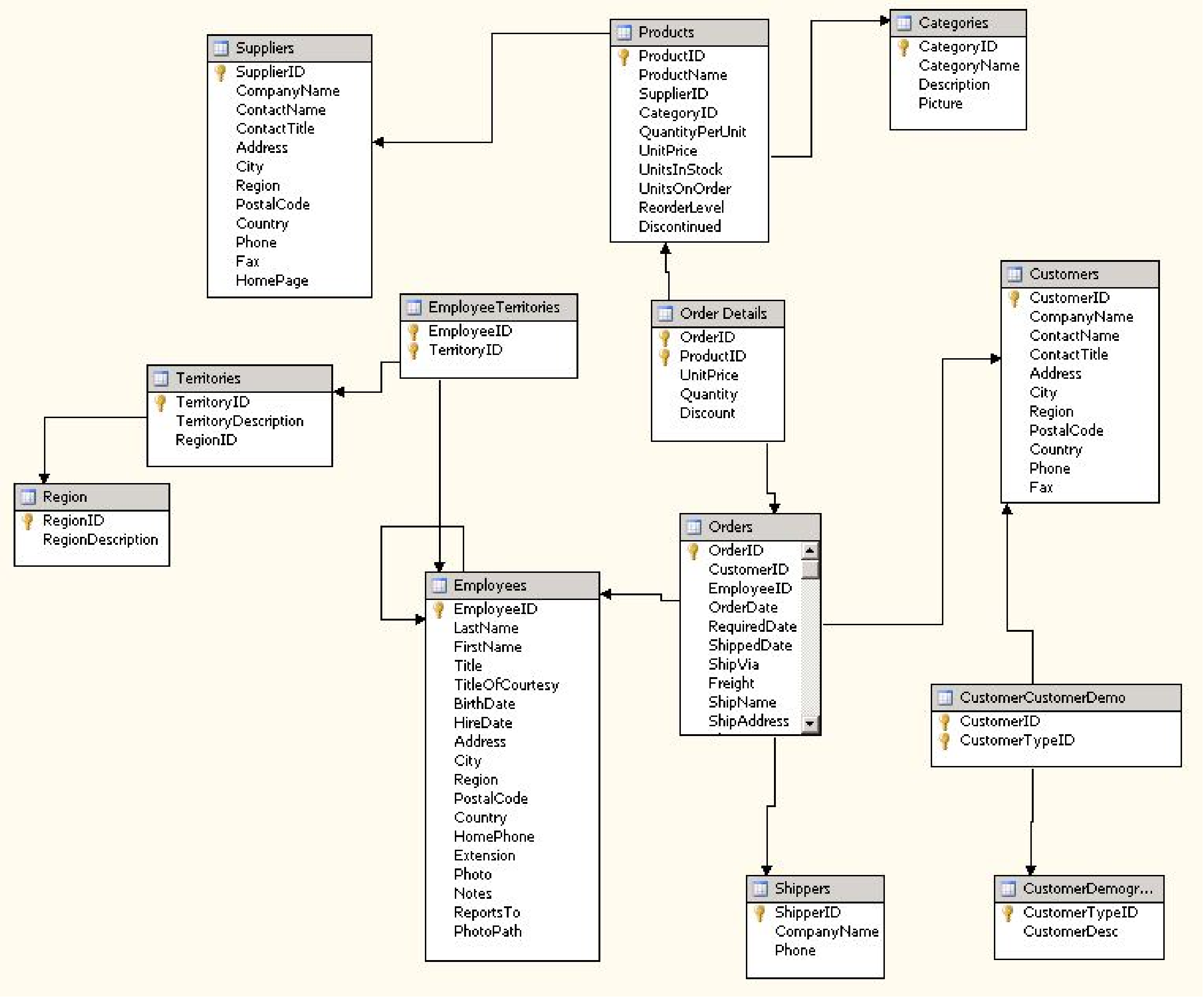

Figure 1: Schema of the Northwind database

1. Consider the Northwind database whose schema is given in Figure 1. This database contains information of orders placed by customers. For every order the detail is given of what products were sold, for what unit price and in what quantity. The employee that secured the order is recorded as well as the date in which the order was inserted. For customers the city they live in etc. is recorded, and for employees their salesdistrict. For this database, create queries to generate the following reports:

   (a) Select the number of sales per category and country.

   ```
   SELECT CategoryName, Country, COUNT(*) AS Count
   FROM "Order Details" O, Products P, Categories C, Suppliers S
   WHERE O.ProductID = P.ProductID AND P.CategoryID = C.CategoryID AND
         P.SupplierID = S.SupplierID
   GROUP BY CategoryName, Country
   ORDER BY CategoryName, Country
   ```

   (b) Select the 3 top-selling categories overall (hint: use “select top 3” construction).

   ```
   SELECT CategoryName, Country, COUNT(*) AS Count
   FROM "Order Details" O, Products P, Categories C, Suppliers S
   ```

```
WHERE O.ProductID = P.ProductID AND P.CategoryID = C.CategoryID AND
      P.SupplierID = S.SupplierID
GROUP BY CategoryName
ORDER BY CategoryName
```

(c) Produce an overview of sales by month for these categories (hint: get month and year with "month" and "year" functions). Are there countries and product categories for which the trend over time is increasing?

```
WITH Top3Categories AS (
      SELECT TOP 3 C.CategoryId, CategoryName, COUNT(*) AS Count
      FROM "Order Details" O, Products P, Categories C
      WHERE O.ProductID = P.ProductID AND P.CategoryID = C.CategoryID
      GROUP BY C.CategoryId, CategoryName
      ORDER BY Count DESC )
SELECT CategoryName, Country, year(OrderDate) AS Year,
      month(OrderDate) AS Month, COUNT(*) AS Count
FROM Orders O, "Order Details" OD, Products P, Top3Categories C, Suppliers S
WHERE O.OrderID = OD.OrderID AND OD.ProductID = P.ProductID AND
      P.CategoryID = C.CategoryID AND P.SupplierID = S.SupplierID
GROUP BY CategoryName, Country, year(OrderDate), month(OrderDate)
ORDER BY CategoryName, Country, year(OrderDate), month(OrderDate)
```

(d) List total amount of sales in $ by employee and year (discount in OrderDetails is at UnitPrice level). Which employees have an increase in sales over the three reported years?

```
SELECT FirstName, LastName, year(OrderDate) AS Year,
       FORMAT(SUM((1-Discount)*OD.UnitPrice*Quantity),'C', 'en-us') AS TotalAmount
FROM Orders O, "Order Details" OD, Employees E
WHERE O.OrderID = OD.OrderID AND O.EmployeeID = E.EmployeeID
GROUP BY FirstName, + LastName, year(OrderDate)
ORDER BY FirstName, LastName, year(OrderDate)
```

(e) Get an individual sales report by month for employee 9 (Dodsworth) in 1997.

```
SELECT month(OrderDate) AS Month,
       FORMAT(SUM((1-Discount)*OD.UnitPrice*Quantity),'C', 'en-us') AS TotalAmount
FROM Orders O, "Order Details" OD
WHERE O.OrderID = OD.OrderID AND O.EmployeeID = 9 AND year(OrderDate) = 1997
GROUP BY month(OrderDate)
ORDER BY month(OrderDate)
```

(f) Get a sales report by country and month.

```
SELECT Country, year(OrderDate) AS Year, month(OrderDate) AS Month,
       FORMAT(SUM((1-Discount)*OD.UnitPrice*Quantity),'C', 'en-us') AS TotalAmount
FROM Orders O, "Order Details" OD, Products P, Suppliers S
WHERE O.OrderID = OD.OrderID AND OD.ProductID = P.ProductID AND
      P.SupplierID = S.SupplierID
GROUP BY Country, year(OrderDate), month(OrderDate)
ORDER BY Country, year(OrderDate), month(OrderDate)
```

2. The sales department of a supermarket chain wants to have a system to support the strategic planning and evaluation of promotions. To this end, they need sales information over the different stores of the supermarket chain. For their analysis tasks they want to compute average sales and total sales, for different products, either at product level or brand level, for different stores at different levels of granularity: individual store, province where the store is located, and country, and for different time periods: per year, month, quarter, semester and also by day of the week.

   (a) How would you conceptually model the data needed by the sales department as a data cube? E.g., what are the measures, the dimensional attributes, the hierarchies, the aggregations that are needed?

   Solution: The dimensions are as follows

   - Product(Product, Brand, Type),
   - Store(Store, Province, Country), and

- Date(Month, Semester, Year, Weekday).

The measure is sales. The aggregation functions are sum (for total sales) and average (for average sales). The hierarchies are as follows:

Product	→	Brand				
Store	→	Province	→	Country		
Date	→	Month	→	Semester	→	Year
	→	Weekday				

(b) Given the cube of (a), explain how you would construct the answers to the following queries with the operations slice-and-dice, pivot, roll-up, and drill-down. If necessary, indicate in which cell(s) of the constructed cube the answer can be found:

i. Give the total overall sales per store.
Solution: Cells (Store, all, all) of the original cube. Slice on Product=all and Date=all. The measure is TotalSales.

ii. Give an overview of the average sales per month per province.
Solution: The measure is AverageSales. Slice on Product=all. Roll-up Store to Province, and Day to Month. Represent it using a pivot-table on dimensions Store and Day.

iii. Give the subcube with only dimensions store at level province and day at level month for the average and total sales for the period 1999 till 2005.
Solution: Slice: date must be in 1999 till 2005. Roll-up Store to Province, Date to Month.

(c) Give an SQL:1999 expression that produces the datacube (i.e., contains all aggregates of the cube using the null value in an attribute to represent aggregation on the corresponding dimension). How do you handle the multiple measures? The hierarchy?

We assume that the base data is stored in the following relational tables:

- Product(ProductID, Brand, Type)
- Store(StoreID, Province, Country)
- Date(Date, Weekday, Month, Semester, Year)
- Sales(ProductID, StoreID, Day, Amount)

```
SELECT P.ProductID, Brand, S.StoreID, Province, Country, D.Day,
      Weekday, Month, Semester, Year, SUM(Amount) AS Total, AVG(amount) AS Average
FROM Product P, Store S, Day D, Sales Sa
WHERE P.ProductID = SA.ProductID and S.StoreID = Sa.StoreID and D.day = Sa.day
GROUP BY ROLLUP(brand,P.ProductID), ROLLUP(country, province, S.StoreID),
         ROLLUP(year, semester, month, D.day), ROLLUP(weekday, D.day);
```

3. Give SQL:1999 expressions for the queries in 2(b).

Solution
Let Cube be the result of the query in 2(c).

- Give the total overall sales per store.

```
SELECT StoreID, Total
FROM Cube
WHERE Brand IS NULL AND Year IS NULL AND Weekday IS NULL AND StoreID IS NOT NULL
```

- Give an overview of the average sales per month per province.

```
SELECT Month, Year, Province, Average
FROM Cube
WHERE Brand IS NULL AND StoreID IS NULL AND Province IS NOT NULL AND
      Date IS NULL AND Month IS NOT NULL AND Weekday IS NULL
```

- Give the subcube with only dimensions store at level province and date at level month for the average and total sales for the period 1999 till 2005.

```
SELECT Month, Year, Province, Average, Total
FROM Cube
WHERE Brand IS NULL AND StoreID IS NULL AND Province IS NOT NULL AND
      Date IS NULL AND  Month IS NOT NULL AND Year >= 1999 AND Year <= 2015 AND
      Weekday IS NULL
```

4. Suppose that we have a relation Sales(Product, Month, Store, Amount). There are five products: P1, P2, P3, P4, P5, 12 months of data, and three stores: S1, S2, and S3.

   (a) (Dense setting) Suppose that every product has been sold in every month in every store; i.e., for every combination of a product p, a month m, and a store s, there is a tuple (p, m, s, a) with a non-zero amount.

      i. How many tuples does this relation contain?
      Solution: $5 \times 12 \times 3 = 180$

      ii. How many tuples does a data cube with dimensions Product, Month, Store, and measure Amount contain?
      Solution: $6 \times 13 \times 4 = 312$

   (b) (Sparse setting) Consider the following (sparse) relation:

Product	Month	Store	Amount
P1	Jan	S1	a1
P1	Jan	S2	a2
P2	Feb	S2	a3
P2	Feb	S3	a4
P3	Jan	S1	a5
P3	Feb	S1	a6
P4	Feb	S1	a7
P5	Jan	S3	a8

   How many non-empty cells does the data cube of this relation contain?

   Solution:

Group By	#	Group By	#
PMS	8	P	5
PM	6	M	2
PS	7	S	3
MS	6	()	1

   Hence, in total: 38 non-empty cells.