First MidTerm of *Decision Support Systems / Decision Support Databases*

It is forbidden to consult any material during the test. Duration of written exam is 1.5h.

1. Let us consider the following database, without null values:

Products

PkP	UnitPrice	...
10	5	...
20	10	...
30	20	...
...	...	...

Sales

FkP	Qty	...
10	50	...
20	10	...
30	20	...
10	30	...
20	100	...
30	10	...
10	30	...
...	...	...

(a) **(1 point)** Give an SQL query to find the total sales revenue by product. Revenue is quanity times unit-price.

(b) **(1 point)** Give the query logical tree, the type and the value of the result of the SQL query.

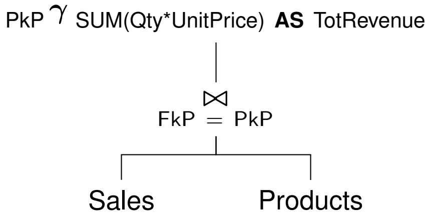
(c) **(2 points)** Explain the meaning of a "semi-additive" measure. Let F(D1, D2, M) be a table with the "semi-additive" measure M with respect to D1. Give a *wrong* query on F with the aggregation SUM(M).

Solution:

(a)

```
SELECT PkP, SUM(Qty*UnitPrice) AS TotRevenue
FROM Sales, Products
WHERE FkP = PkP
GROUP BY PkP
```

(b) The logical query plan is:



The type of the result is {{(PkP:int, TotRevenue:float)}}. The value of the result is:

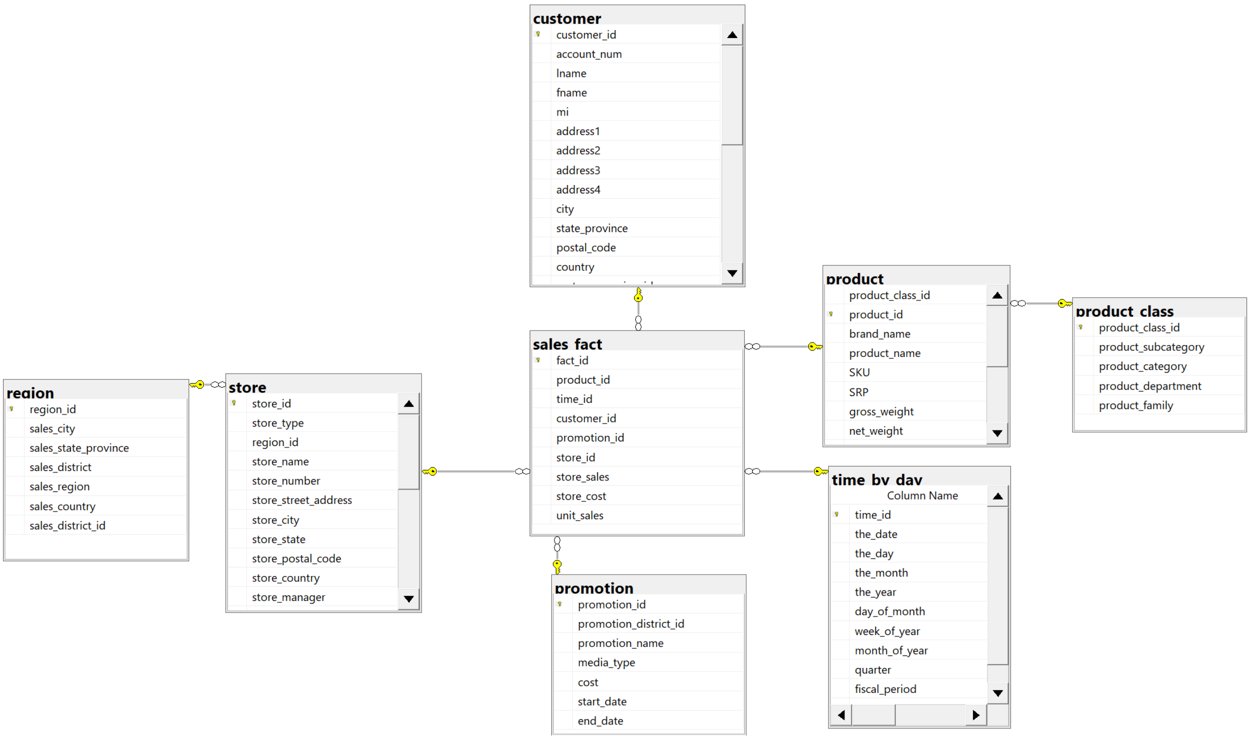
PkP	Revenue
10	550
20	1100
30	600
...	...

(c) A semi-additive measure with respect to a dimensional attribute (or sets of attributes) is a measure that cannot be aggregated over a set of tuples containing two or more distinct values of the dimensional attribute(s). A wrong query is:

```
SELECT SUM(M)
FROM F
```

unless D1 is constant over the table F.

2. **(4 points)** Consider the FoodMart datawarehouse.



Write **analytic** SQL queries to solve the following problems:

(a) for each brand_name, the total sales of the brand_name and the ratio of it over the total sales of its product_category.

(b) for each brand_name, how many distinct other brand names have higher total sales

Solution:

The following query solve (a) in the output attribute Ratio and (b) in the output attribute N,

```
SELECT brand_name,
  SUM(store_sales) /
   SUM(SUM(store_sales)) OVER (PARTITION BY product_category) AS Ratio,
  (-1+DENSE_RANK() OVER (ORDER BY SUM(store_sales) DESC)) AS N
FROM sales_fact S, product P, product_class C
WHERE S.product_id = P.product_id
      AND P.product_class_id = C.product_class_id
GROUP BY brand_name, product_category
ORDER BY brand_name
```

3. A farm produces several types of products (doors, windows, gates, etc.). Employees are distributed in work shifts (*turni di lavoro*), typically 6am-2pm or 2pm-10pm but sometimes employee may work less than 8 hours (unassigned time). An employee is paid by an hourly salary based on the employee's specialization (basic, intermediate, advanced). The farm is interested in the analysis of the worker's productivity and to related it to anomalies in products. Examples of business questions are:

(i) Number of hours of work by type of product.

(ii) Total cost of employees by specialization.

(iii) Number of anomalies by type of product (employee specialization).

With respect to the above business scenario, answer the following questions:

(a) **(4 points)** Design a conceptual schema for the data mart to support the business questions. Your schema should at least be able to satisfy the above mentioned analysis requirements. You may motivate other suitable attributes for the dimensions. Follow all the steps of the conceptual design methodology.

(b) **(2 points)** Give a logical data mart design, which includes the treatment of updates for dimensional attributes.

Solution:

(a), We follow the process of datawarehousing conceptual design. Tables 1 and 2 show the preliminary analysis.

ID	Dimensions	Measures	Metrics
(i)	Product (type)	Worked hours	Tot worked hours
(ii)	Employee (specialization)	Hourly salary	Total cost
(iii)	Product (type), Employee (specialization)	Anomaly	Num anonmalies

Table 1: Business process requirements

Description	Preliminary Dimensions	Preliminary Measures
A fact *WorkedShift* is a single work shift of a single worker working on a single product	Product (type), Employee (specialization)	Worked hours Hourly salary, Anomaly

Table 2: Fact description

Anomaly is a binary measure: 0 if the product had no anomaly, and 1 if it had some.

In addition to the preliminary dimensions, from the context description, we also consider a Date dimension with attributes Day, Month, Year; and a degenerate dimension Shift to record the work shift within the day.

Moreover, Employee and Product may be extended with other typical attributes of those two dimensions.

Figure 1 shows the conceptual DFM schema.

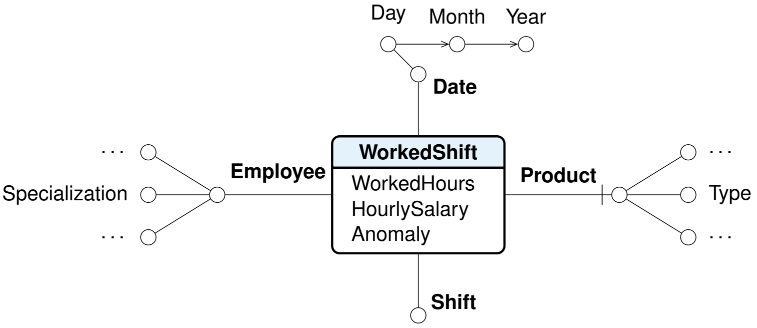


Figure 1: DFM Schema

Notice that the Product dimension is optional, as to model unassigned time to employees.

(b) Figure 2 shows the logical design assuming that only specialization changes as Type 2, while all other dimensional attributes changes are Type 1.

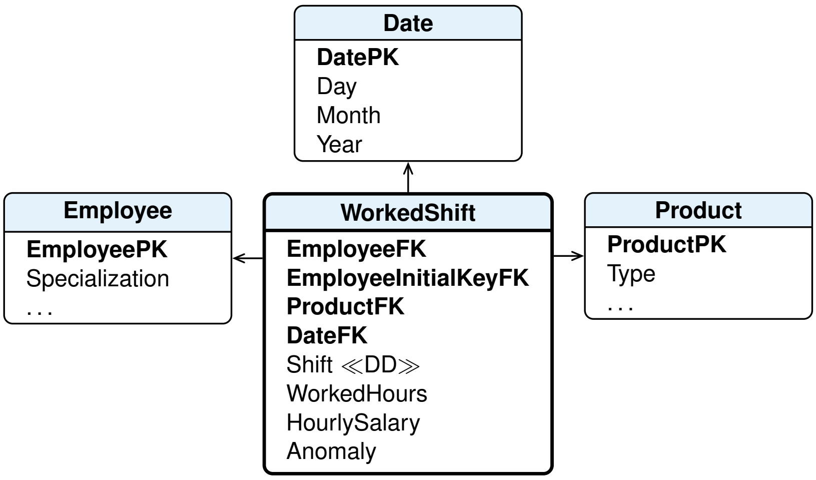


Figure 2: Logical schema

4. **(2 points)** A multidimensional cube has 4 dimensions:

   - product, with 20 different products
   - employee, with 5 different employees
   - date, with 100 different days
   - anomaly, with 2 possible values (yes, no)

   Answer the following quesitons:

   (a) what is the size of the multidimensional data cube, i.e., how many cells are in it?

   (b) assuming that each product is worked by a single employee, how can the number of cells be reduced and to what number?

   Solution:

   (a) The size of the (extended) multidimensional cube is $(20+1)\cdot(5+1)\cdot(100+1)\cdot(2+1) = 38178$. See Lesson 13.

   (b) Since a fixed product (functionally) determines a fixed employee, the cells where product is fixed are non-zero only for the determined employee or for $\star$, hence the part $(20+1)\cdot(5+1)$ reduces to $20\cdot(1+1)+1\cdot(5+1)$ and the total number of cells to $(20\cdot(1+1)+1\cdot(5+1))\cdot(100+1)\cdot(2+1) = 13938$.