Data Management for Data Science

Corso di laurea magistrale in Data Science
Sapienza Università di Roma
2015/2016

Exercise on OLAP

Riccardo Rosati

Dipartimento di Ingegneria Informatica Automatica e Gestionale A. Ruberti

Exercise

We want to store a multidimensional structure containing the following information about sales:

- quantity (number of items sold)
- customer (name of the customer)

over the following dimensions:

- Time (day, week, month, quarter, year)
- Product (type, brand, category, group)
- Location (city, region, country, continent)

Exercise (contd.)

1. Define a star schema to represent the above multidimensional structure;
2. Define a snowflake schema that reduces (at least on one dimension) the redundancy of the star schema defined at the previous point;
3. Write an SQL query over the star schema defined at point 1 that returns the names of the customers who bought a product from category "Car" in 2015 in Italy;
4. Write the SQL query over the snowflake schema defined at point 2 that returns the names of the customers who bought a product from category "Car" in 2015 in Italy.

Solution (point 1)

Star schema:

Sales(keyTime, keyProduct, keyLocation, quantity, customer)

Time(keyTime, day, week, month, quarter, year)

Product(keyProduct, type, brand, category, group)

Location(keyLocation, city, region, country, continent)

Solution (point 2)

To eliminate redundancy from the dimensions Product and Location, we identify the following functional dependencies:

category $\rightarrow$ group

region $\rightarrow$ country

country $\rightarrow$ continent

(Remark: the functional dependency brand $\rightarrow$ category does not hold, since the same brand can produce items from different categories)

Solution (point 2)

We obtain the following snowflake schema:

Sales(keyTime, keyProduct, keyLocation, quantity, customer)

Time(keyTime, day, week, month, quarter, year)

Product(keyProduct, type,brand, keyCategory)

Category(keyCategory, category, group)

Location(keyLocation, city, keyRegion)

Region(keyRegion, region, keyCountry)

Country(keyCountry, country, continent)

Solution (point 3)

SQL query over the star schema:

```
SELECT customer
FROM Sales, Product, Time, Location
WHERE Sales.keyTime=Time.keyTime AND
   Sales.keyProduct=Product.keyProduct AND
   Sales.keyLocation=Location.keyLocation AND
   Time.year="2015"  AND
   Product.category="Car" AND
   Location.country="Italy"
```

Solution (point 4)

SQL query over the snowflake schema:

```
SELECT customer
FROM Sales, Product, Time, Location, Category, Region, Country
WHERE Sales.keyTime=Time.keyTime AND
   Sales.keyProduct=Product.keyProduct AND
   Sales.keyLocation=Location.keyLocation AND
   Time.year="2015"  AND
   Product.keyCategory=Category.keyCategory AND
   Category.category="Car" AND
   Location.keyRegion=Region.keyRegion AND
   Region.keyCountry=Country.keyCountry AND
   Country.country="Italy"
```