



Data Base and Data Mining Group of Politecnico di Torino

## Data Warehousing and Data Mining

Politecnico di Torino  
School of Information Engineering

### Data warehouse design – XtraSport exercise

XtraSport is a company whose business is focused on wholesale. It has **agencies** all over Italy. Every agency has a number of **agents** who are assigned a sale district. Most of the **customers** are chains of stores with many branches in different provinces and regions all over the country.

A **sale** consists of a receipt with one or more rows, each representing units of goods sold. XtraSport sales strongly depend on the season, thus the company turnover changes over the year for different products. Every year mainly consists of two sale **seasons**: fall & winter and spring & summer. Each sale season starts and ends in different months with respect to the orders and to the billing.

The company aims at aggregating information from its agencies in a central data warehouse. The data warehouse users are interested in two main events: the turnover (i.e., the **sales**) and the **orders**.

The data warehouse must be designed to answer promptly to queries performed by company managers. Some of the queries the managers are interested in are the following.

a. Turnover (sales)

1. Monthly, 3-monthly, 6-monthly, and yearly sales for each province.
2. Monthly, 3-monthly, 6-monthly, and yearly sales for each region.
3. Monthly, 3-monthly, 6-monthly, and yearly sales for each customer.
4. Monthly, 3-monthly, 6-monthly, and yearly sales for each customer group.
5. Monthly, 3-monthly, 6-monthly, and yearly sales for each agent.
6. Monthly, 3-monthly, 6-monthly, and yearly sales for each agency.
7. Monthly, 3-monthly, 6-monthly, and yearly sales for each item category.
8. Sales of each agency for each month and item category.
9. Number of items sold for each item category, time period (month, 3-month, half-year, and year), customer, and agency.
10. Absolute discount value (not percentage) depending on the item category, time period (month, 3-month, half-year, and year), customer, and agency.

b. Orders

1. Total monthly, 3-monthly, 6-monthly, and yearly order price for each province.
2. Total order price for each item category according to the sale season (fall & winter or spring & summer) for each province.
3. Total monthly order price for each item category, customer, and agent.
4. Average delivery time of different item categories for each customer, province, agency, month.

**TASK-1.** Design the **conceptual** and **logic** models of the XtraSport data warehouse.

**TASK-2.** - Considering the characteristics and cardinalities of the designed data warehouse, decide whether and which **materialized views** can be useful to optimize the performance of the following queries.

- Q1. 3-monthly, 6-monthly, and yearly sales for each province.
- Q2. 3-monthly, 6-monthly, and yearly sales for each region.
- Q3. 3-monthly, 6-monthly, and yearly sales for each item category and province.
- Q4. Total order price for each item category according to the sale season for each province.
- Q5. Average delivery time of different item categories for each region and month.

To evaluate the materialized view performance gain, the following statistics are known.

- Customers
  - o Customers  $\approx$  1000
  - o Provinces  $\approx$  100
  - o Regions  $\approx$  20
  - o Customer groups  $\approx$  4
- Agents
  - o Agents  $\approx$  100
  - o Agencies  $\approx$  30
- Items
  - o Item categories  $\approx$  10
- Time
  - o Year = last year
  - o Days  $\approx$  360
  - o Months = 12
  - o Trimesters = 4
  - o Semesters = 2
  - o Sale seasons = 2

The worst case has the following **cardinalities** for the facts.

- Sales
  - o 1000 customers
  - o 12 months
  - o 10 item categories
  - o 100 agents
  - o Total  $\approx 12 \times 10^6$  (12 millions)
- Orders
  - o 1000 customers
  - o 360 days
  - o 10 item categories
  - o 100 agents
  - o Total  $\approx 360 \times 10^6$  (360 millions)