Data warehouse design exercise

Problem specifications

A computer assembler would like to analyze costs and income of the component assemblage and sale of its Personal Computers.

PCs are assembled by buying 4 main components from external suppliers: RAM, Hard Disk (HD), processor (CPU), and graphic board. Every component has its own alphanumeric code, depending on its manufacturer and its features. Each component is made by a specific manufacturer, which is also its supplier.
Once bought, components are assembled together into a pre-determined standard PC configuration, which is identified by the four alphanumeric codes of the components. Each standard configuration targets a specific market segment by addressing its most common requests. Groups of assembled PCs are then sold to computer retailers and supermarkets.

For the assembling phase, the number of components, their cost, and the cost of delivery must be analyzed according to:

- the supplier of the component
- the supplier location (region and geographical area: north, center, south Italy, or islands)
- the component type (RAM, CPU, HD, graphic board)
- the component features:
    - code of the component
    - for the RAM component: memory capacity, type, access time
    - for the HD component: memory capacity, physical dimensions
    - for the CPU: type and clock speed
    - for the graphic board: maximum supported resolution
- the month, two-month period, four-month period, semester, year when components have been bought.

Regarding the sales, the cost of production (i.e., the sum of the component costs and the assembling cost), the selling price, the number of sold PCs and the number of defective PCs returned under warranty must be analyzed according to:

- the PC configuration
- the component codes (RAM, HD, CPU, and graphic board)
- the component suppliers
- the final retailer/supermarket
- the number of shops of the final retailer/supermarket
- the month and two-month period of sale
- the three-month period and the year of the sale

The following are **some** of the frequent queries of interest:

a) For each four-month period in 2005 and 2006, select the number of CPUs bought from north-Italy suppliers, separately for each CPU type
b) For each year, select the configuration with the lowest "production cost / price" ratio, considering only configurations having both RAM and CPU from the same supplier.
c) For each component, select the number of pieces bought in March 2007 and the total number of pieces bought in the whole 2007.
d) **Considering only 2002, for each configuration whose RAM has been bought from the "IntelligenceDevice" supplier, select the percentage of defective products with respect to the total number of sold products.**
e) **Considering only 2007 and RAM components, for each four-month period select the total cost (component cost + delivery cost) and the cumulative cost since the beginning of the year, separately for each geographical area and RAM type.**
f) For each month of 2007, select the number of different configurations that have been sold to each retailer/supermarket.
g) For each geographical area and each year, select the total delivery cost of the components.
h) **Considering only suppliers from which, in 2003, more than 100 000 components have been bought in total, select for each supplier and each component type the number of pieces bought and the total cost in the second two-month period of 2003.**

Design

The data warehouse will store information of the 1998-2007 period. The following cardinalities are known (suppose data is uniformly distributed):

- Components: ~1000
- Suppliers: ~20
- Retailers/supermarkets: ~15
- Average number of shops per retailer/supermarket: ~500
- Configurations: ~30
- Products: ~100 000
- Ski tow types: ~5
- Ski tow altitude gaps: ~5
- Ski tow lengths: ~10
- Ticket types: ~3
- Payment types: ~4

Unknown cardinalities can be supposed and motivated by the student.

1) Design the data warehouse to address the described issues. In particular, the designed data warehouse must allow efficient execution of **all** the queries described in the specifications.
2) Write the frequent queries **(d)**, **(e)** and **(h)** of the "problem specifications" using the extended SQL language.
3) Considering the designed data warehouse and its cardinalities, decide whether and which materialized views are convenient to improve response time of the frequent queries (consider **all** the frequent queries). Explain reasons for your choices.
4) Possible changing dimensions must be properly handled: describe how to address this issue and explain reasons for your choices.