Data Warehousing Data Mining - Exercises

Mateusz Pawlik

October 24, 2011

Contact

Mateusz Pawlik
mateusz.pawlik@unibz.it
Office hours: Mondays, 13:00-14:00, POS 2.12 (with possible changes)

Objectives

The exercises consider the data warehousing part of the course. Throughout the semester you will work on a small data warehousing project. You will design and implement an example data warehouse. At the end of the semester you will deliver a report and give a short presentation about the work done. Every labs meeting consists of a set of warming-up exercises (solved in class) and project's milestones tasks (individual work).

Project Description

Working in groups you will design and implement an example data warehouse. Each group will choose a project domain and go through several stages of the design process. The project is divided into several milestones. Each milestone contains a set of tasks to be solved by each group. Each milestone has to be submitted in the form of a report.

Project's Milestones

- Milestone 1 - Data warehouse design
  start 03.10.2011 14:00
  due 06.11.2011 23:59
  Delivery consists of a report, addressing all Milestone 1 tasks, and an SQL script (and any additional files) for creating and populating your data warehouse. If the process is different than running a script, remember to include a README file with the detailed instructions.

- Milestone 2 - Data warehouse querying
- Milestone 3 - Data warehouse performance optimization

Project Evaluation

The evaluation will be based on the milestones assignments and the final presentation given during the last exercises. Remember, that the project's mark makes 60% of the final course mark.

Schedule

Exercises 1 - Business domain - (03.10.2011, 14:00-16:00)

Warming-up Exercises

1. Write three examples of business intelligence queries.
2. Define and describe differences between OLTP and OLAP.
3. How can You define a data warehouse and data marts.
4. Name characteristics or features of a data warehouse.
5. Explain data granularity and how it is applicable to the data warehouse. Give an example of different granularities.
6. How are the top-down and bottom-up approaches for building a data warehouse different?
7. A data warehouse is subject-oriented. What would be the major critical business subjects for the following companies?
   - manufacturing company
   - bank
   - hotel chain
   - airlines
   - hospital
   - retailer
8. You are a Senior Analyst in the IT department of a company manufacturing automobile parts. The marketing VP is complaining about the poor response by IT in providing strategic information. Try to explain the reasons for the problems and why a data warehouse would be the only viable solution.

Milestone 1 Tasks

1. Split into groups of maximum 2 students. Choose the domain of your data warehouse project. Describe shortly an example company, reason the advantages of a data warehouse in your company and the data warehouse objectives. Send the group details (names, project domain) by e-mail to Mateusz Pawlik.

2. Choose and describe business processes significant to the chosen domain from the analyst's viewpoint.

3. Decide and argue the granularity of the chosen business processes.

Exercises 2 - Conceptual design - (10.10.2011, 14:00-16:00)

Warming-up Exercises

1. Define fact, measure, dimension, dimensional attribute and hierarchy. How are they represented on a fact schema in the Dimensional Fact Model (DFM)?

2. You are the Vice President of Marketing for a nation-wide appliance manufacturer with three production plants. Describe any three different ways you will tend to analyze your sales. Name facts, dimensions, attributes and measures. Draw a simple fact schema in the DFM.

3. Describe the following modeling concepts, give an example and how does the fact schema notation look like in the DFM. As a reference you can use `http://www-db.deis.unibo.it/~srizzi/PDF/isr08-1.pdf`.

   - Descriptive attributes.
   - Cross-dimensional attributes.
   - Convergence.
   - Shared hierarchies.
   - Multiple arcs.
   - Optional arcs.
   - Incomplete hierarchies.
   - Recursive hierarchies.
   - Additivity.

Milestone 1 Tasks

1. Define facts in your data warehouse (minimum 2).

2. Define dimensions.

3. Define measures.

4. Define attributes and their hierarchies.

5. Create a fact schema for each of the facts.
6. Describe the problems which you faced while creating the fact schema and argue the solutions. Remember about the modeling concepts discussed earlier.

Exercises 3 - Logical design - (17.10.2011, 14:00-16:00)

Warming-up Exercises

1. Define star schema and snowflake schema. Describe the representation of facts and dimensions. Describe pros and cons of each schema.
2. An online order wine company requires the designing of a data warehouse to record the quantity and sales of its wines to its customers. Part of the original database is composed by the following tables:
   CUSTOMER (Code, Name, Address, Phone, BDay, Gender)
   WINE (Code, Name, Type, Vintage, BottlePrice, CasePrice, Class)
   CLASS (Code, Name, Region)
   TIME (TimeStamp, Date, Year)
   ORDER (Customer, Wine, Time, nrBottles, nrCases)
   Note that the tables represent the main entities of the ER schema, thus it is necessary to derive the significant relationships among them in order to correctly design the data warehouse. Draw a star schema and a snowflake schema.

Milestone 1 Tasks

1. Draw a star and snowflake schema for every fact from your data warehouse.
2. Draw an example instance of created star and snowflake schema (several rows for each table).
3. Write what schema type suits your data warehouse most. Describe the design choices that you made. Draw the ”best” schema.

Exercises 4 - Physical design - (24.10.2011, 14:00-16:00)

Warming-up Exercises

1. Remind SQL language details for creating and populating databases.

Milestone 1 Tasks

1. Write an SQL script for creating your data warehouse in a chosen DBMS.
2. Populate your data warehouse with test data. Use real world data or consider using some data generator, e.g. ”generatedata.com”. Generating the data try to meet the real world scenarios, e.g. an online book retailer may offer 20000 books and sell 500 books a day.