Advanced Data Management Technologies
Data Warehouse Project

2017/2018

1 Objective

The aim of the project is to motivate, design, implement, and query a DW (data warehouse). The business domain of the project is a print and publishing group that owns various subsidiaries, such as printed newspapers (daily, weekly, and monthly), online newspapers, radio stations, book shops, print shops and so on.

Each subsidiary stores its own data in a relational database and the databases are neither synchronized nor linked. The group is in the process to develop a DW solution across subsidiaries to get a comprehensive overview of their business, such as customers, advertisement campaigns, and so on. You will start with creating an example DB and design, implement and populate a prototype DW using PostgreSQL that is capable of convincing the management of the group to implement your DW. The DW should be representative for the business, as such you have to carefully chose some business questions that are relevant for the management's and company's business needs. To implement your project, you need to think of and generate your own DB and data.

2 Tasks

The project is divided into a number of steps that are described below.

2.1 Task1: Domain Analysis and Description

- Describe the domain and business, provide motivations for the development of a DW, collect information and documentation, etc.
- What are the business processes that you want to model and what business questions should they help to answer. Consider minimum two business processes. Which granularity level is appropriate for the described processes?

2.2 Task2: Conceptual Design

- Write what do the facts represent, what are their dimensions and measures. The facts should have a least four dimensions and one or more measures. Indicate the properties of the measures.
- Draw the fact schemas using DFM for the chosen facts. Mark properly on the schema the following concepts: fact, dimensions with hierarchies, measures, descriptive attributes, non-additivity of the measures, convergence, shared hierarchies, multiple arcs, optional arcs, etc.

2.3 Taks3: Logical Design

- Draw the star schemas (and/or snowflake schema if it suits better) for your facts including primary/foreign key relations, attributes and (estimated) cardinalities. Describe all the non-trivial choices that you made, e.g., how did you model a multiple arc or a recursive hierarchy.
- Choose two of the business questions that you find important (one for each fact) and write SQL queries that solve them. Draw sample instances of the tables that are involved in the queries (several rows for the dimension tables and up to 15 rows for the fact tables). Show the results of the queries on the instances.

2.4 Task4: Implementation

- Write an SQL script that creates your data warehouse (fact and dimension tables).
- Populate the DW form your relational schema. Add instructions that populate your DW and describe any data cleaning steps that may be involved.
- Write two queries for your data warehouse: One query that uses the ROLLUP, CUBE or GROUPING SETS operator and one query that uses the GROUPING ID and/or GROUP ID function.

2.5 Task5: Querying

- Write queries for each of the following points: one ranking query using NTILE, RANK or DENSE RANK functions; one windowing query using the windowing clause; and one period-to-period comparison query (a query comparing values across time periods, e.g., compare sales for every week of the current year with the sales of the corresponding weeks in the past year). In your answer, include the queries in natural language, their SQL codes and the results.

2.6 Task6: Data Analysis Tool

- Use a data analysis tool, such as pentaho (`http://community.pentaho.com/`), to show your results from the previous tasks. Alternatively, you may manually create your own plots and show your results using any graph tool, such as gnuplot (`http://gnuplot.sourceforge.net`).

3 Deliverables

- A project report of approx. 20 pages targeting at non-technical and technical people.
- A presentation and demo targeting at non-technical and technical people.

4 Project Evaluation

The evaluation of the project will be based on the complexity, arguments, and the completeness of report and presentation. In particular, some important evaluation criteria are as follows:

- Complexity of your DW solution.
- Reasoning and completeness of the dimensions in your DW.
- Strength of your arguments for your DW in report and presentation.