Data Analysis and Integration

1st semester

Project

In this project, you will be working with the Northwind database. For more information about this database, the course bibliography can be useful, especially the book *Data Warehouse Systems: Design and Implementation*, by A. Vaisman and E. Zimányi (Springer, 2014).

The provided script **northwind.sql** contains the SQL instructions to create the database in MySQL. You can run this script and explore the database to become familiar with it.

Creating a data warehouse

The first goal of this project is to create a data warehouse based on the Northwind database. For this purpose, you should:

1. Develop an SQL script (**northwind_dw.sql**) to create the data warehouse tables (this is similar to what **steelwheels_dw.sql** does).
2. Implement a complete ETL process in Pentaho Data Integration (PDI), including all the necessary transformations (***.ktr**) and a job (***.kjb**) to populate the data warehouse from the Northwind database.
3. Define the OLAP data cube (**northwind_dw.xml**) using Pentaho Schema Workbench.

The data warehouse should have a star schema with the following dimensions:

- A *customer* dimension with company name, city and country.
- A *product* dimension with product name and category name. This should be a slowly-changing dimension.
- A *supplier* dimension with company name, city and country.
- A *shipper* dimension with company name.
- A *time* dimension with day, month and year. Use 3-letter months, e.g. Jan, Feb, Mar, etc.

There should be two measures: *sales* and *quantity*.
The *sales* measure should be calculated as: UnitPrice * Quantity * (1 - Discount).

Analyzing the data

The second goal of this project is to perform multidimensional analysis over the data warehouse to answer a set of business questions.

4. Load the cube definition into Saiku and use this OLAP front-end (and its MDX capabilities) to carry the following analysis:
   a) Analyze sales by customer country and year to discover the country, the year, and the pair country-year with the most sales.
   b) Analyze sales by product category and year to discover the category, the year, and the pair category-year with the most sales.
   c) Analyze quantity by shipping company and year to discover the shipper, the year, and the pair shipper-year with the most quantity.
   d) Analyze sales by customer country and product category to identify the pairs of country-category with no sales at all.
   e) Analyze quantity by supplier country and customer country to identify the pairs of countries with no quantities being shipped between them.
   f) Analyze quantity by product category and shipping company to identify the pairs of category-shipper with no quantity at all.

Documenting the results

To submit the project, you should prepare a document with the following contents:

1. Present the SQL instructions to create the data warehouse tables. The code should be formatted and indented in a way that makes it easy to read for a human.

2. For each transformation/job that you develop in PDI, present:
   - a screenshot of the entire transformation/job,
   - screenshots of the configuration window and of the preview window for each step (for transformations only).

3. Present the XML code for the cube definition. The code should be formatted and indented in a way that makes it easy to read for a human.

4. For each analysis query that you develop in Saiku, present one of the following:
   - If you developed the query by drag-and-drop, present a screenshot of the Saiku user interface, showing the measures, columns, rows and filters used in the query, together with the query results.
   - If you developed the query in MDX mode, present the MDX code together with a screenshot of the query results.

Submitting the project

Prepare a PDF document with the contents above (code and screenshots). Please make sure that the text is properly formatted and the images have good quality. Submit the PDF document in Fénix until the deadline (Nov 11, 2021).