Data Analysis and Integration

Project

In this project, we will be using the Price Paid Data by the UK Government, which can be found here: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

We will start working with the data for 2025, and later we will try to add more data. Our main goal is to develop a star schema that we can use to perform multidimensional analysis on those data.

Price Paid Data is a dataset comprising a single table, with records of property sales. Each property sale has a price, a date, a location (comprising several fields), among other attributes.

For our purposes, we are interested in analyzing certain facts, namely, how much has been sold at each location and on each date. Here, location is defined at the granularity of town or city (there is no need to go to a more detailed level, such as street). For example, we could analyze the amount of sales by city and date, with the possibility of aggregating up to county and year, if necessary.

The following tasks describe what should be done.

Tasks

1. Access the Price Paid Data, and download the yearly CSV file for 2025:
   https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

2. Write an SQL script to create a database with a single table in MySQL, to store the data. Also, in that SQL script, write an instruction to load the CSV file into the table.
   Note: You may use online resources and the MySQL Reference Manual (for the MySQL version that you are using) to decide about column names, data types, and the appropriate instructions to load the data into the table. Present the SQL script in text format

3. From the database created in the previous step, we can try to get the data at the desired level of granularity. For this purpose, write an SQL query to get the amount of sales by location and date, where location does not need to go below the level of city.
   Note: Present the query and (possibly only an excerpt of) the results, sorted by amount, with the largest amount on top.

4. Write an SQL script to create a data warehouse (star schema) as a separate database in MySQL. The requirements are the following:
   - The star schema should have one fact table, and two-dimension tables for location and time.
   - Location should include the levels of city and above and should use an integer surrogate key.
   - Time should have day, month and year.
   - Facts should have a measure, and a primary key composed of foreign keys.

   Note: The purpose of this script is to define the structure for the data warehouse. The actual data will be populated into the tables through a series of transformations in the next tasks.

5. Using Pentaho Data Integration (PDI), develop a transformation to populate the location dimension.
   Note: Present a screenshot of the entire transformation, and screenshots of the configuration window and of the preview window for each step.

6. Using PDI, develop a transformation to populate the time dimension.

Note: Present a screenshot of the entire transformation, and screenshots of the configuration window and of the preview window for each step.

7. Using PDI, develop a transformation to populate the fact table.
   Note: Present a screenshot of the entire transformation, and screenshots of the configuration window and of the preview window for each step.

8. Using Pentaho Schema Workbench (PSW), develop an XML cube definition based on your star schema.
   Note: Present a screenshot of tree structure in PSW, and the contents of the XML file.

9. Configure Pentaho Server to use that cube and, using Saiku Analytics, present an analysis of the sales amount by month and district, but only for the county of Greater London.
   Note: Take a screenshot of the analysis and the results in Saiku Analytics. Also, export the results in Excel (XLS) format.

10. Using Pentaho Report Designer (PRD), create a report with the total amount per county in 2025. However, show only the counties with a total above one billion pounds (i.e. £ 1,000,000,000).
    Note: Present the underlying MDX query. Take screenshots of the report in design mode and in preview mode. Export the report in PDF format.

11. Download the CSV file for 2024, and write an SQL instruction to append those data to the database table, without erasing the data for 2025. Run your transformations again and, as you try to do this, what would you say are the main bottlenecks? Discuss using your own words.
    Note: Present the instructions to load the 2024 data and discuss the observed bottlenecks when running the transformations with the additional data.

12. Restart Pentaho Server and, using Saiku Analytics, write an MDX query that shows, for each district of Greater London: the amount of sales for January 2024; the amount of sales for January 2025; and the percent change from January 2024 to January 2025.
    Note: Present the underlying MDX query. Take a screenshot of the query and the results in Saiku Analytics. Also, export the results in Excel (XLS) format.

Submission

The project should be submitted in a zip file named **group_XX.zip** (where **XX** is a two-digit group number) and submitted via Fénix until 23:59 of the deadline date.

At the root of the zip file, there should be a slide presentation (prepared with PowerPoint or similar) in PDF format, named **slides_XX.pdf** (where **XX** is a two-digit group number). The slides should present the outputs (code, screenshots, results, discussion, etc.) requested for each task. You may organize the slides in a way to make it easier to locate a specific task.

In addition, the zip file should contain a folder for each task (**task_02/**, **task_03/**, ..., **task12/**) with each folder containing the files produced for each task, namely scripts (*.sql), transformations (*.ktr), cube (*.xml), exported results (*.xlsx), generated reports (*.prpt; *.pdf), etc.

Bear in mind that you should be able to explain all the SQL code and options. This means also that AI generated superfluous instructions may result in penalties on your final grade.

Note: Please follow the submission instructions rigorously. Evaluation elements that are not found as prescribed above may end up not being considered for grading. Do not wait until the last minute to submit, as late submissions will not be accepted.