Data Analysis and Integration

1st semester

Project

In this project, you will be working with the *airports* database, which stores information about bookings, flights, airlines, airplanes, airports, and passengers.

The goal is to create a data warehouse where each fact can be described as follows:
"A certain flight took place between a certain airport([1]) and another airport([2]) from a certain time([3]) to a another time([4]) by a certain airline([5]) using a certain airplane([6]) carrying a certain number of passengers([7]) and with a certain revenue([8])."

Explanatory notes/details:
([1]) The airport of origin for the flight.
([2]) The airport of destination for the flight.
([3]) The date and time of departure for the flight.
([4]) The date and time of arrival for the flight.
([5]) The airline of the flight.
([6]) The airplane used in the flight.
([7]) The total number of passengers according to the bookings for that flight.
([8]) The total amount of revenue for the flight based on the price of bookings for that flight.

The data warehouse should have the following dimensions:

- An airport dimension for the airport of origin.
- An airport dimension for the airport of destination.
- A time dimension for the date/time of departure.
- A time dimension for the date/time of arrival.
- An airline dimension.
- An airplane dimension.

These dimensions should have the following levels:

- Airport dimension: name, city, country.
- Time dimension: day, month, year.
- Airline dimension: name.
- Airplane dimension: type.

Developing the data warehouse

In this project, you are expected to perform the following tasks:

1. Use the script **airports.sql** to create the airports database.

2. Develop a script **airports_dw.sql** to create the data warehouse.

3. Develop a set of PDI transformations to populate the data warehouse.

4. Use PSW to define the data cube (**airports_dw.xml**).

5. Use Pentaho Server & Saiku Analytics to perform three analysis queries over the data warehouse.
   - One of those queries is: passengers and revenue by airline and month.
   - For the other two queries, be creative, use different dimensions across the queries, and, if possible, avoid very large result tables that greatly exceed the size of the screen.

Testing on larger datasets

While the previous steps describe the essential tasks in this project, the following steps are aimed at testing the proposed solution (i.e. the data warehouse, the ETL process, and the analysis queries) on increasingly larger datasets.

6. Before proceeding, stop Pentaho Server. This is important for two reasons:
   - To free up memory on the VM for subsequent tasks.
   - To clear the cache of the OLAP server, so that the analysis queries will retrieve fresh results from the data warehouse, instead of returning previous results from cache.

7. Use **airports-large.sql** to create a larger version of the airports database.

8. Run you PDI transformations again to populate the data warehouse.

9. Start Pentaho Server, run your analysis queries again, and take note of the new results.

10. Try to repeat the steps above using **airports-large-extra.sql**. This may or may not be feasible.
    - If the attempt is successful, present your results by comparing to the previous results on smaller datasets.
    - If the attempt is unsuccessful, provide a detailed and convincing explanation of where and why it failed.

What should be delivered

To submit the project, prepare a single PDF document with the following contents:

a) 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.

b) For each transformation that you develop in PDI, present:
   - a screenshot of the entire transformation,
   - a screenshot of the configuration window and of the preview window for each step in the transformation.

c) 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.

d) For each analysis query, present a full screenshot of the user interface, showing the measures, columns, rows, filters used in the query, together with the query results.

e) For the larger datasets, present the results of your analysis queries in a similar way, or describe your difficulties while attempting to run them.

Submitting the project

Prepare a single PDF document with the contents a)-e) above. Make sure that the content is readable, the text is properly formatted, and the images have good quality. Submit the PDF document in Fénix until the project deadline.