COMP40370 Practical 3

Data & Data Warehouses

Prof. Tahar KECHADI

Academic year 2021-2022

The aim of this practical is to develop data warehouses for various data-driven applications. To do this, you need to use and apply some of the concepts and techniques introduced on the lectures so far. Use Python and its libraries to define and setup a data warehouse for one data-driven application. In this practical you need to create a data warehouse where you can store datasets, arrays and records into. It is required to use PostgreSQL to implement your Data warehouse.

The first thing to do is to install PostgreSQL on your computer, if you have not done so.

Once, we complete the design of a DW for a given application, we will continue with the implementation details.

Data files

• `./practical3.pdf`	// assignment questions (this file)
• `./specs/DW_dataset.csv`	// data file for Question 2
• `./specs/`*`input_DW_data.csv`*	// data file for Question 3 you need to create

Programming requirements and tools

The assignment should be solved in Python, version 3.8 or above (3.9 is recommended). You can use the following packages for this assignment:

- SQLAlchemy 1.4+ → will be used to connect to your database
- You need to install and import all the necessary libraries (e. g. psycopg2 drivers)
- Pandas 1.3+

The documentation of **SQLAlchemy** can be found here: https://docs.sqlalchemy.org/en/14/. In previous practical assignments, we have introduced some packages. There are very interesting tutorials you can go through to help you understand how to connect to a DB/DW, how to interact with it, etc. There is no need to mention them here again. Please use those packages as you need them.

Question 1:

Consider a dataset **D** that consists of customers purchasing tour packages to various places at different prices. We can perform different kinds of data operations on the dataset **D**. Try to categorise the following operations into a) **simple query of data retrieval**, b) **Online Analytical Processing**, or c) **Data Mining**.

1- Find names of customers who have purchased tours that cost less than €500.
2- List the names of the customers, the number of tour packages that the customers have purchased, and the total cost for the tours.
3- Calculate the difference in quarterly sales of tours between this year and the previous two years.
4- Find a rule such as "**IF** customers purchase a tour package to France, **THEN** it is 80% likely that the same customers also purchase a tour package to Spain.
5- From the customer purchase history, build a model for predicting the kinds of customer who are likely to purchase tours to a certain country.

Question 2: Data Warehouse & Data

Consider the dataset given in "*DW_Dataset.csv*", representing the employee data of a company.

1- If the attribute "***Salary***" needs to be discretised into three pay bands, suggest a simple yet sensible solution for the discretisation based with a valid argument.
2- Miss Davis's salary is unknown, and the unknown value needs to be imputed, what is a sensible replacement value and why?
3- Among the employee records, which record can be considered as an outlier? What harm can an outlier cause to the understanding of the dataset?

Online Analytical Processing (OLAP) perceives a dataset in a multi-dimensional space. For the dataset given in "*DW_Dataset.csv*", perform the following tasks/operations:

4- Draw a diagram of a 3D view using the following attributes: ***Year of Birth***, ***Status***, and ***Salary***.
5- What do the data points inside the cube represent? Use the cube as an example to discuss the meaning of OLAP operations such as pivoting, slicing and dicing, rolling up and drilling down.

Question 3: Data Warehouse - Implementation

Suppose that a data warehouse for Big University consists of the four dimensions ***student***, ***course, semester***, and ***instructor***, and two measures ***count*** and ***avg_grade***. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the ***avg_grade*** measure stores the actual course grade of the student. At higher conceptual levels, ***avg_grade*** stores the average grade for the given combination.

1- Draw a snowflake schema diagram for the data warehouse.

2- Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should you perform to list the average grade of CS courses for each Big University student.

3- If each dimension has five levels (including all), such as “*student* < *major* < *status* < *university* < *all*”, how many cuboids will this cube contain (including the base and apex cuboids)?

The following questions will guide you to implement a data warehouse using PostgreSQL. At this stage, we assume that you have already defined the DW schema, with subjects, dimensions, and measures.

4- First, you need to establish a connection with the database to create a table where you can store records and arrays of data. Make sure you follow the PostgreSQL naming convention.

5- Second, you need to create another connection to the DW where you will store your datasets in.

6- Create a data file that contains 5 entries, called “*input_DW_data.csv*”, which you need to store in your DW.

7- Finally, you need to define the following functions to read, write, update, and list your data to / from the data warehouse.

```
def read_record (field, name, engine): …
def write_record (name, details, engine): …
def update_record (field name, new value, engine):
def read_dataset (name, engine):
def write_dataset (name, dataset, engine):
def list_datasets (engine):
```

The final deadline for the submission is Thursday, 7th of October at 23:00. All submissions must be done in Brightspace.