INFS3200 Sem 1, 2025

Assignment 2: Data Warehouse

Guidelines for Assignment Submission:

1. Submission Deadline: **May 23, 2025, 15:00 (AEST).**
2. For questions requiring queries, kindly include both a screenshot of the query and a screenshot of the execution results.
3. When addressing discussion-based questions, provide comprehensive textual responses, supplemented with suitable visual aids if necessary.
4. The submission should be in PDF format, named as 'A2_s1234567.pdf' (replace '1234567' with your student ID). The submitted file should not exceed 10 MB.
5. All implementations and project components should be finalized within the UQZones environment. Evaluators might assess the assignment based on the established checkpoints within UQZones.
6. All assignment submissions must be submitted exclusively through the UQ Blackboard. Alternative methods of submission will not be acknowledged. Please be mindful that submissions via email will not be entertained under any circumstances.
7. It is imperative to adhere to the stipulated submission deadline to avoid penalties, as explained in the Educational Course Policies (ECP).
8. Ensuring the successful submission of your assignment within the designated timeline is your responsibility.

Task 1: Data Linkage (3 points)

In this part, you are provided with the restaurant dataset. The dataset is identical with the one in Prac3. Import data into PostgreSQL database and complete the following tasks:

1. (1 point) Link two two restaurant records by using edit-distance as the similarity measure. Report the hyper-parameter choice and the total number of similar records.
2. (1 point) Link two two restaurant records by using tri-grams (Jaccard Coefficient) as the similarity measure. Report the hyper-parameter choice and total number of similar records.
3. (1 point) Which similarity measure is better for the restaurant dataset? Provide the justifications.

Task 2: Data Warehouse (7 points)

You are provided with an electronic device wholesale dataset in CSV format. Download the dataset by using your UQZone terminal:

```
curl -O https://stluc.manta.uqcloud.net/infs3200/public/Sales.csv
```

The dataset is exported from an OLTP system. Each record (row) in the dataset indicates a sales transaction. You are required to construct a data warehouse to analyze the sales

performance. The following table describes the detail of attributes in the dataset. Here, 'Type' indicates the data type to be stored in database.

Attribute	Type	Description
SID	Int	The **unique** ID of a staff member.
FNAME	Varchar(20)	The first name of a staff member.
LNAME	Varchar(20)	The last name of a staff member.
STATE	Varchar(10)	The state that a staff member works in. Each state can have several stores.
STORE	Varchar(20)	The store that a staff member works in. Each store belongs to one of the states.
DATE	Date	The date of the transaction record.
PID	Int	The **unique** ID of a product.
PRODUCT	Varchar(40)	The name of a product. Each product belongs to one of the brands.
BRAND	Varchar(40)	The brand of a product. Each brand can have several products.
UNIT_COST	Decimal(10,2)	The purchase price of a product.
QUANTITY	Int	The number of products sold in this transaction.
PRICE	Decimal(10,2)	The sale price of a product in this transaction. Note that this value fluctuates case by case.

Table 1. A detailed description of attributes in the sales dataset.

1. (1 point) Design and construct the data warehouse under star schema that contains three dimension tables, including "Staff", "Product", and "Time_Period". Show the conceptual model of your design.
    a. Note: "Time_Period" must contain time of different granularity, for example, date, month, quarter, and year.

2. (1 point) Query the constructed data warehouse to provide the following basic statistics of the data.
    a. How many unique staff members?
    b. How many transactions have been made in 2022 Qtr3?

    Provide the query and query result.

3. (1 point) Construct a cube that contains the time, staff and sales information. Store the cube in a **materialized view** called 'Sales_Time_Staff'. The cube should summarize sales information by different levels of staff and time. Provide your query.

4. (2 points) Design a view to obtain profits from 'Sales_Time_Staff" cube. You should **only** use the cube to answer the question. Provide the query and fill in the following tables. Your solution will lose marks if it includes any unnecessary operations such as joins and group by. GROUP BY operation can only be used for adjusting the views of outputs.

The state sales profits in each quarter of 2021				
	QLD	NSW	WA	SA
2021 Q1				
2021 Q2				
2021 Q3				
2021 Q4				

The state sales profits in each year				
	QLD	NSW	WA	SA
2021				
2022				
2023				

5. (2 points) Construct a cube that contains the store (in staff dimension), product, and sales information. Store the cube in a **materialized view** called 'Sales_Product_Staff'. The cube should summarize sales information by different levels of staff and products. Based on the **materialized view** 'Sales_Product_Staff',
    a. Create a **view** that select top-3 stores with the highest gross profit. Gross profit of selling an item is calculated by Quantity * (Sale Price – Unit Cost). Provide your query and query results.
    b. Create a **view** that show the most profitable item for each store.

    Note: You should **only** use the cube 'Sales_Product_Staff' to create the views. You should not use any join operations for question a, and you can use only one join for question b. Your solution will lose marks if it includes any unnecessary operations such as joins and group by.