UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2022/2023 : SEMESTER 1

WQD7005 : Data Mining

Name: Sadman Chowdhury ID: S2199546

Duration: From 14/01/2023, 8.00 PM to 14/01/2023, 9.30 PM

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (25 marks).

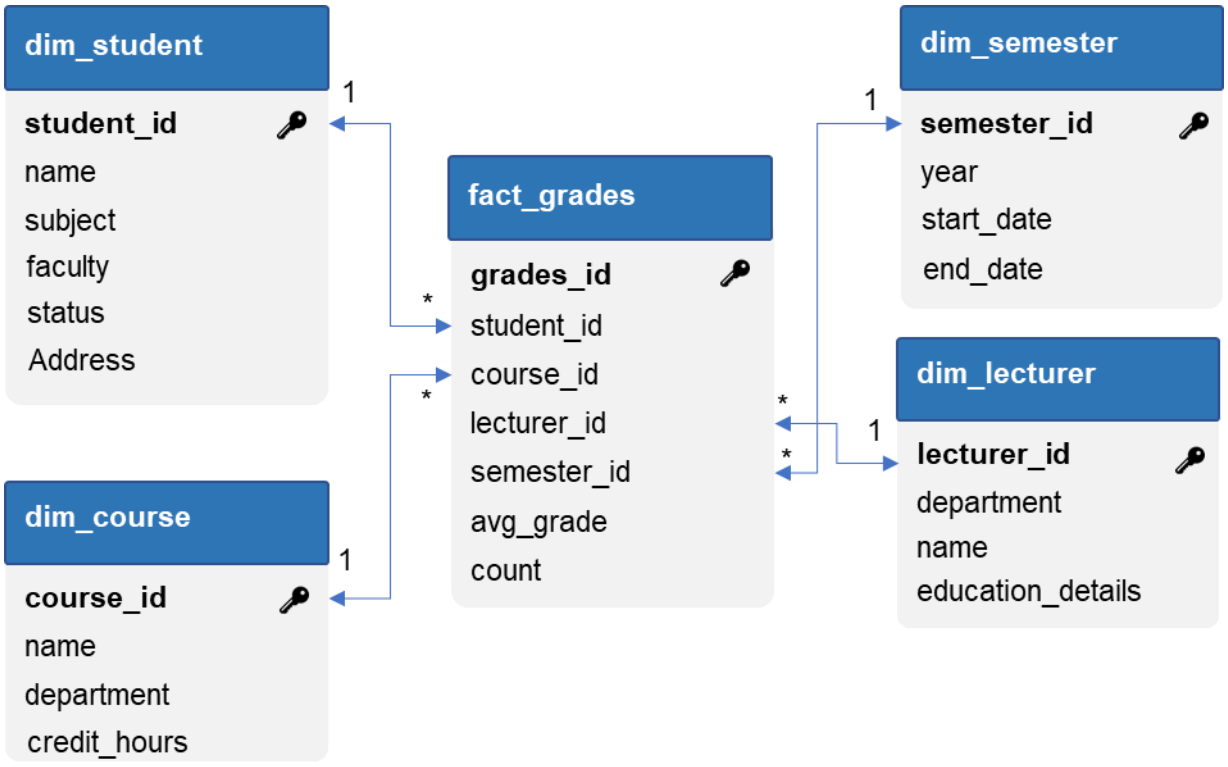
Github Link: *https://github.com/siam923/datamining_mid*

1 Suppose that a data warehouse for **Faculty of Computer Science, UM** consists of the following four dimensions: **student**, **course**, **semester**, and **instructor**, and two measures **count** and **avg grade**.

Draw a schema diagram for the above data warehouse using Star schema.

(10 marks)

Ans:



Justification:

To see the summary of a **student** based on his average grade, his name, id, subjects, address, and enrollment status will be necessary to evaluate his performance based on these factors.
For the **course** dimension, only the name and the department offering the course are enough, as it has no other information to be included for analytical purposes.
The **semester** dimension, is a **time dimension**. Generally, there are 2 or 3 semesters every year. So, only the semester and year are adequate for the analysis.
In the **lecturer** dimension, the name attribute can be used to compare how students of specific lecturers perform. The education_info & department attributes will also help to analyze his teaching efficiency based on his study area and level of education.
Besides, all dimensions have an id attribute as their **primary key**, which will be used to connect with the fact table and perform join operations if necessary.

2 **Define Data Warehouse in your own words.** (5 marks)

Ans:

A *Data Warehouse* is a system that acts as a centralized repository for storing and managing large amounts of data that is used for analysis and reporting. It includes tools, processes, and a way of representing data that makes it easy for people to analyze and understand. It is like a special database optimized for reading and analyzing large amounts of data rather than inserting, updating, or deleting data. Using a Data Warehouse makes it possible to quickly and easily access and analyze data from different sources, such as sales, customer, and financial data. Proper data warehouse development makes it a powerful tool for business intelligence, decision-making, and reporting.

3 Discuss the differences between database processing query and data mining processing query with your own three examples which are based on your group project dataset.

(6 marks)

Ans:

The difference is explained in the table below:

	Database Processing Query	Data Mining Processing Query
Def:	Database processing query, also known as Online Transaction Processing (OLTP), is focused on retrieving specific pieces of information from a database based on a specific need, such as finding the total sales of a specific product in a specific month.	Data mining processing query, also known as Online Analytical Processing (OLAP), is focused on finding patterns, trends, and relationships in the data, such as identifying the top-selling products by category and country or identifying the age group and gender of customers who buy the most expensive products. These queries are usually short and simple and focus on a small data subset.
Type	These queries are usually short and simple and focus on a small data subset.	Such queries are usually more complex, require more computation, and focus on a larger subset of the data.
EX 1	Total sales of a specific product category “Bike" in a specific month "December".	Total total sales by category of the product.
EX 2	Counting the number of unique customers by age who bought a bike in a specific state " Oregon"	Counting the number of similar customers by age group and gender to understand behaviour.
EX 3	Total sales of a specific product part such as " Tires and Tubes" in a specific year "2017"	Most Profitable product in a specific year "2017"

4 Data quality can be assessed in terms of several issues, including incomplete, noisy, inconsistent and intentional. How do you address these 4 issues in your group project dataset?

(4 marks)

Ans:

We have followed several strategies to assess the four issues for our project.

1. **Completeness:** To handle the completeness issue, we have checked if there are any missing values in the dataset. We used Talend to run a query that counts the number of null or empty values for each column. For the missing values, we decided whether to remove those rows or to fill in the missing values with a default value or a value that is estimated based on other data depending on the need.
2. **Noise:** To address noise issues, we checked if there were any outliers in the dataset. We created the graphs like box plots and histograms for each column. These charts showed us the distribution of the data and any outliers. After finding outliers, we removed some of them and kept some outliers intentionally, which seemed important to us.
3. **Consistency:** To address consistency issues, we checked for inconsistencies like duplicates, type mismatches, or wrong or unrelated values in the data. In Talend, we ran queries that checked for duplicate values or values that were not in the expected format. For example, if we expect the "Country" column to contain only uppercase characters, we can run a query that checks for any lowercase characters in that column.
4. **Intentional issues:** To address intentional issues, we have scan the dataset to see if any values are intentionally entered incorrectly or not logically. For example, we found few negative profit values, which was kept intentionally to indicate loss.

END