COMP4522 Advanced Databases – Project (17%)

A. Summary

In this project you will have to put into practice several techniques you have learned during lectures' time. The project has 3 components:

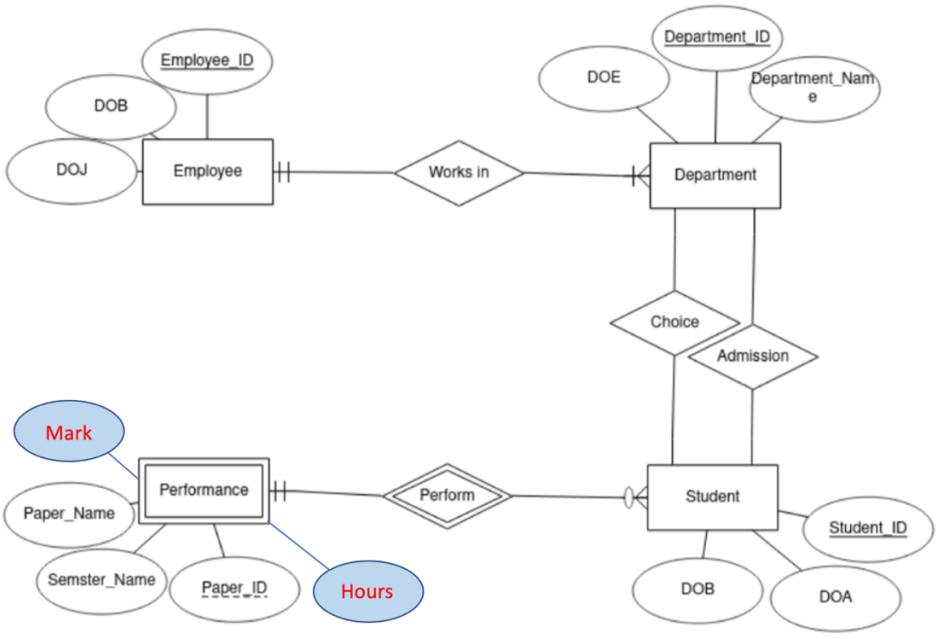
1. Four tables will be given to you, and you will have to load them into **MariaDB**, creating a small database that reflects the E-R Diagram provided below. You can assume that those tables are in .csv format.
2. The information in those tables must be re-arranged and reformatted, as needed, to fit the purpose of your *data mining* task at hand. You may think of this step as the process of mapping the original data format(s) to a new template that is convenient for the task at hand. Also, as we all know, data is not always clean, it rarely is, so you may have to inspect the data for several common quality problems -and fix the issues that may negatively impact your task No. 3 (below).
3. You will have to do some data mining on the data you have prepared in step 2. Descriptive data mining will be your first task, followed by predictive analytics for a specific request.

Yes, you guessed it right! Your project is a hybrid between *Data Warehousing* and *Data Mining* via ***ETL*** (Extract, Transform and Load).

B. Introduction

The background of this project is as follows. Here is the **ER-Diagram** for the Relational Database System in place.

ER Diagram

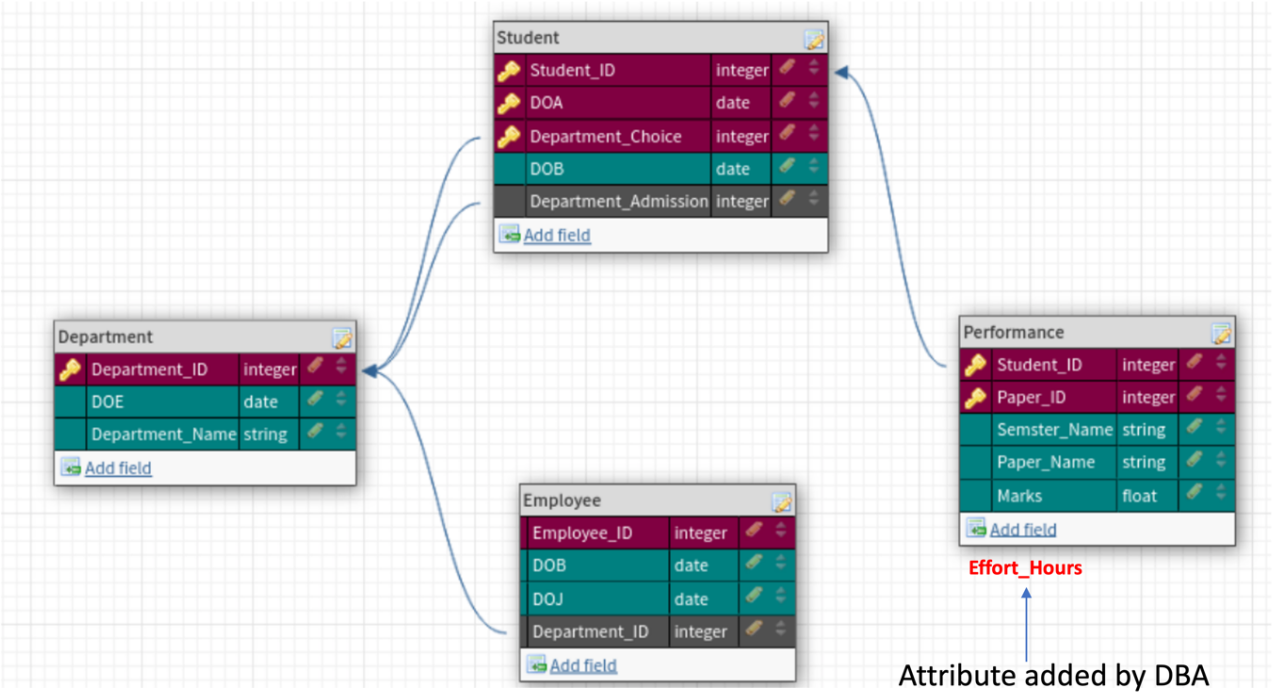


But what does this system is for? The ER-Diagram and the system in place have been designed with the purpose of recording the journey of students, right from his/her admission until the last of her/his degree. The data associated to the system contains information about an *imaginary educational institution*. It contains data about:

- Departments
- Professors (university employees)
- Student Counselling
- Coursers offered
- Courses that students selected to get admission to the institution
- Student performance in different examinations they sat through

A *possible* database **schema** for this system follows:

Schema



Keys, primary and foreign, have been defined as follows:

Relation/Table	**Primary Key**	**Foreign Key**
Department	Department_ID	
Employee	Employee_ID	Department_ID
Student	• Student ID • Department_Choices • DOA	• Department_Choices • Department_Admission

Relation/Table	Primary Key	Foreign Key
Performance	• Student_ID • Paper_ID • Semester_Name	Student_ID

C. Goals of this project

The goal of the project is to **load** the original data into a RDBMS (MariaDB) and **extract** data from the original RDBMS system into a format that fits your data mining needs (some form of ETL). Note that you will have to **pre-process** and **transform** data as needed to perform **statistical analysis** -including building a linear regression model on a subset of the transformed data-.

STEPS

1. Load the data into MariaDB. Each .csv file corresponds to an entity.
2. Extract, validate and clean the data, transforming it into a valid aggregated dataset, fit for purpose for further analysis.
   - There are multiple tables/files, and joins may be required
   - You must find and remove data inconsistencies, as needed
   - Aggregate data as needed for your analysis
   - Do intermediate computations, as needed
3. Apply visualization techniques to the data in order to extract some valuable insights (using Jupyter Notebook).
4. Find if you can "excerpt" more information from this data, than what is "factual."
   - Use statistical analysis (**descriptive** data mining) for the data structure containing the transformed data ONLY. At this point, there is no need to do analysis anything else.
5. Create a predictive model to forecast the grade that a given student will obtain in the next paper (assignment) assuming the existing history of **Effort vs. Mark** for *such* a student. You will build a linear regression model in *Jupyter*. Again, another data mining activity (predictive, this time).
6. Take note that the Professors members of the Student Council, whom are in control of assigning funding, asked all students to provide the number of hours they spent in each paper that they did write. The Council requested to the Database Department that the (new) information, related to Effort (hours invested in each paper), were added to the Database. As the request was urgent and everyone was busy, the Database Administrator (DBA), added an attribute (column) to the "Student_Performance_Data" table, as the last attribute to the right. Hence, now the aforementioned relation looks like this:

Student_ID	Semester Name	Paper_ID	Paper_Name	Marks	Effort_Hours
SID12345678	Sem_x	SEMI1234567	Paperx	XX	YY

Notes:
These activities are very common in the industry and are usually grouped together as ETL or ELT, combined with data warehousing and data mining. Data quality is an underlying layer to all the above. Think about the implications.

D. Data Inspection, Exception Reporting and Cleansing

Data cleansing is a complex task. In this project we want to expose you to some of these problems, but we have no time for a full-fledge data pre-processing activity. Hence, we will ask you to do quality check for the following *quality dimensions* only:

- *Completeness*: is the data complete for the predictive data mining task at hand? Do you have null values or out-of-range values that could **spoil** your predictions?
- *Validity*: here we have several aspects to consider:
  - Check for format, data type and range.
    - Dates must be in range for day, month, and year.
    - Check for validity related to key values used to link relations.
- *Consistency*: check you primary and foreign keys to ensure safe navigation among relations
- *Uniqueness*: you must check for duplicates and for non-existent values for important occurrences. For instance:
  - Are there any duplicate student IDs?
  - Is there duplicate Paper IDs?

Also, we will limit the number of relations (tables) we will ask you to inspect, and the focus will be in **producing exceptions** rather than fixing the issues -except when the exception values are part of the data you will use for *regression analysis*-. In the latter case, you must *either* remedy the data or completely ignore that data instance in your computations (either way, you must report the exception). As such, see the table below for the required actions:

Table / Relation	**Attribute**	**Validation Required**	**Action**
Department_Information	Department_ID	Uniqueness	Report Exception
Department_Information	Department_Name	Uniqueness	Report Exception
Department_Information	DOE	Year >= 1900	Report Exception
Department_Information	All	Missing values	Report Exception
Employee_Infomation	None	None	None
Student_Counceling_Information	Department_Admission	Missing Values	Report Exception

Table / Relation	Attribute	Validation Required	Action
Student_Counceling_Information	Department_Admission	Department_Admission does not exist	Report Exception
Student_Performance_Data	Marks	Range: 0 to 100	Discard entries with issues, and report them
Student_Performance_Data	Hours	Min = 0 Max = any positive integer	Discard entries with issues, and report them
Student_Performance_Data	Student_ID and Paper_ID	A given Student_ID cannot have more than 1 mark per each Paper_ID	Report Exception
Student_Performance_Data	All	Missing values	Discard entries with issues, and report them

E. Data Mining

Descriptive: once you create a *transformed dataset* -the data structure that you have created out of the original .csv files using aggregation and data cleansing principles- *then* use Descriptive Analytics to describe and understand the data at hand. You can use the Jupyter Notebook template provided in class as a starting point.

Predictive: Consider that the "Student_Performance_Data" relation provides you with invaluable historical info about the performance (scores per paper) that the students achieved in the past, based on the effort they put into each assignment. You will have to **predict the scores that a given student will most likely obtain in the next paper** (based on their historical performance). You have a wealth of historical data available! Answer the question for each of the students listed below, assuming that they will put an effort of 10 hours into the next paper. Then, list the results as follows:

Student	**Predicted Score in next paper**	**Department**
SIDyyyyy	Out of 100 predicted score	IDEPxxxx

Student IDs to be computed: 'SID20131151', 'SID20149500' and 'SID20182516'.

F. Software Requirements

It is possible to perform the statistical analysis portion using any programming language. However, **Python** has become the dominant force in this area, so we *strongly recommend* that you embrace Python for the *data mining* portion of the task. You will need:

- Scikit-learn (site and download available in the LINKS section in D2L)
- Python (downloadable for many platforms at https://www.python.org)
- Jupyter Notebook (https://docs.jupyter.org/en/latest/install/notebook-classic.html)
- MariaDB (https://mariadb.com/products/community-server/). An instance will be made available to you by MACO, but you are welcome to do your own install in your own hardware. There are Windows and Linux versions.

Jupyter Notebook is a web interface that makes data mining and analytics, convenient and straight forward. Check D2L for instructions on installing the software. We used it in class during the lectures.

G. Project phases / grading schema and due date

Due date: **13-November-2024** @5pm.

Phase	**Description**	**Your Task**	**Points**
Loading	*Loading* the data to the database. The original is given to you as four .csv files, representing each the existing relations/tables. Review the E-R Diagram provided.	Apply your knowledge of RDBMS to load the data properly.	**3**/17
Extracting, Transforming & Data Quality	*Apply* the required extraction and transformations to perform proper data mining and elicit details from the data.	Apply the required transformations, including cleaning data and dealing with outliers, to generate a data structure (format) that will enable you to perform data mining. You may have to write your own scripts to achieve an effective *transformation*.	**5**/17
Data Mining	*Analyze* and visualize the data, arrive to conclusions, and build a simple predictive model.	Explore the data using statistical concepts and use linear regression to predict students' next grade, as requested above.	**5**/17

Phase	**Description**	**Your Task**	**Points**
Report	*Write* a report covering your main design findings and including your outputs. Include all ETL tasks you have performed. The latter should be interpreted and analyzed.	Produce a report as specified under “Description” (on the left)	**4**/17
		TOTAL ……	**17**/17