Tredence SQL Hackathon: Round 2 - Online University Learning Analytics

Welcome to round 2 of Tredence SQL Hackathon. Please read the problem statement and data description carefully.

Online University Learning Analytics Dataset contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses. Courses go live in February and October and they are marked by “B” and “J” respectively. The dataset consists of tables connected using unique identifiers. All tables are stored in the CSV format.

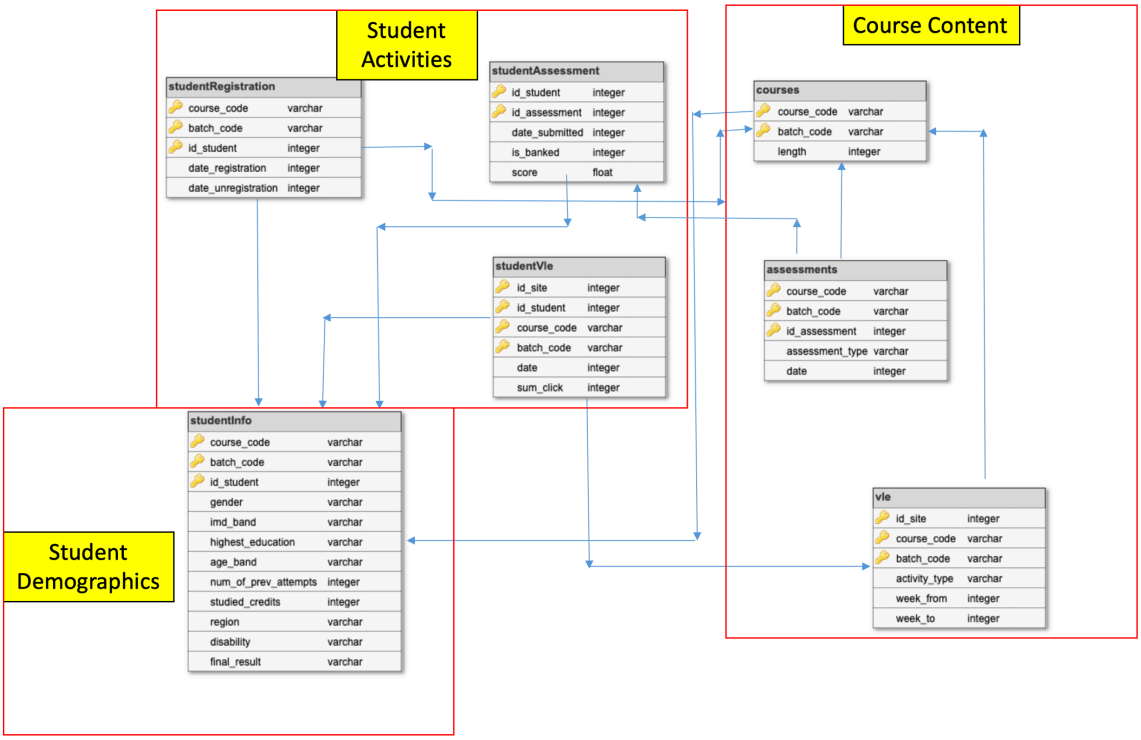
Data Description & Download

The database schema depicts information collected for students in different categories:

- Course Content
- Student Demographics
- Student Activities

Download the datasets from this link.

Database Schema



studentRegistration	
course_code	varchar
batch_code	varchar
id_student	integer
date_registration	integer
date_unregistration	integer

studentAssessment	
id_student	integer
id_assessment	integer
date_submitted	integer
is_banked	integer
score	float

courses	
course_code	varchar
batch_code	varchar
length	integer

studentVle	
id_site	integer
id_student	integer
course_code	varchar
batch_code	varchar
date	integer
sum_click	integer

assessments	
course_code	varchar
batch_code	varchar
id_assessment	integer
assessment_type	varchar
date	integer

studentInfo	
course_code	varchar
batch_code	varchar
id_student	integer
gender	varchar
imd_band	varchar
highest_education	varchar
age_band	varchar
num_of_prev_attempts	integer
studied_credits	integer
region	varchar
disability	varchar
final_result	varchar

vle	
id_site	integer
course_code	varchar
batch_code	varchar
activity_type	varchar
week_from	integer
week_to	integer

Course Content

courses.csv

This file contains the list of all available courses and their respective batches. The columns are:

- **course_code** – code name of the course, which serves as the identifier.
- **batch_code** – code name of the batch. It consists of the year and “B” for the course starting in February and “J” for the course starting in October.
- **length** - the length of the course-batch in days.

Course code and batch code together represent a unique course offering with given duration. The combination of both are present in all tables to refer to the exact offering and both should be used to join tables for analysis

assessments.csv

This file contains information about assessments in each course-batch. Usually, every course-batch has a number of assessments followed by the final exam. CSV contains columns:

- **course_code** – identification code of the course, to which the assessment belongs.
- **batch_code** - identification code of the batch
- **id_assessment** – identification number of the assessment.
- **assessment_type** – the type of assessment. Three types of assessments exist - Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
- **date** – information about the final submission date of the assessment calculated as the number of days since the start of the course-batch. The starting date of the batch has number 0 (zero).
- **weight** - the weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%.

If the information about the final exam date is missing, it is at the end of the last batch week.

vle.csv

This CSV file contains information about the available materials in the VLE. Typically these are HTML pages, pdf files, etc. Students have access to these materials online and their interactions with the materials are recorded. The vle.csv file contains the following columns:

- **id_site** – an identification number of the material.
- **course_code** – an identification code for the course.
- **batch_code** - the identification code of batch.
- **activity_type** – the role associated with the course material.
- **week_from** – the week from which the material is planned to be used.
- **week_to** – week until which the material is planned to be used.

Student Demographics

studentInfo.csv

This file contains demographic information about the students together with their results. The file contains the following columns:

- **course_code** – an identification code for a course on which the student is registered.
- **batch_code** - the identification code of the course batch in which the student is registered.
- **id_student** – a unique identification number for the student.
- **gender** – the student's gender.
- **region** – identifies the geographic region, where the student lived while taking the course-batch.
- **highest_education** – highest student education level on entry to the course batch.
- **imd_band** – specifies the Index of Multiple Deprivation band of the place where the student lived during the course-batch.
- **age_band** – the band of the student's age.
- **num_of_prev_attempts** – the number times the student has attempted this course.
- **studied_credits** – the total number of credits for the courses the student is currently studying.
- **disability** – indicates whether the student has declared a disability.
- **final_result** – student's final result in the course-batch.

Student Activities

studentRegistration.csv

This file contains information about the time when the student registered for the course-batch. For students who unregistered, the date of unregistration (withdrawn) is also recorded. The file contains five columns:

- **course_code** – an identification code for a course.
- **batch_code** - the identification code of the batch.
- **id_student** – a unique identification number for the student.
- **date_registration** – the date of student's registration on the course batch, this is the number of days measured relative to the start of the course-batch (e.g. the negative value -30 means that the student registered to course batch 30 days before it started).
- **date_unregistration** – date of student unregistration from the course batch, this is the number of days measured relative to the start of the course-batch. Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final_result column in the studentInfo.csv file.

studentAssessment.csv

This file contains the results of students' assessments. If the student does not submit the assessment, no result is recorded. The final exam submissions are missing if the result of the assessments is not stored in the system. This file contains the following columns:

- **id_assessment** – the identification number of the assessment.
- **id_student** – a unique identification number for the student.
- **date_submitted** – the date of student submission, measured as the number of days since the start of the course batch.
- **is_banked** – a status flag indicating that the assessment result has been transferred from a previous batch.
- **score** – the student's score in this assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail. The marks are in the range from 0 to 100.

studentVle.csv

The studentVle.csv file contains information about each student's interactions with the materials in the VLE. This file contains the following columns:

- **course_code** – an identification code for a course.
- **batch_code** - the identification code of the batch.
- **id_student** – a unique identification number for the student.
- **id_site** - an identification number for the VLE material.
- **date** – the date of the student's interaction with the material measured as the number of days since the start of the course-batch.
- **sum_click** – the number of times a student interacts with the material in that day.

Problem Statement

You have been hired by Online University to mine their extensive data and extract useful information. Following are some questions asked by the leadership team. Using MySQL queries, find solutions to the following queries:

1.
    1. Which region has the highest median score in the final exam?
    2. Which course has the highest pass rate? What is the percentage of each activity_type within this course?
    3. Which activity type is most exercised amongst the students who got the distinction?
    4. During what day (Number of days from course-batch start) for each course-batch do we see maximum activity (clicks) on VLE?
    5. What is the extent of activity (number of clicks) for course batch AAA 2013J above which 95% of students pass the course?
    6. Compare the average result of batches 2013J and 2014J only for females who engaged with Virtual Learning Environment in 28th day of the course-batches for 1st and 2nd Tutor Marked Assignment (TMA) using combined weighted score.

7. Print the id of students who achieved full scores (100) for more than one assessment. Order your output in descending order by the total number of assessments in which the student earned a full score. If more than one student received full scores in the same number of assessments, then sort them by decreasing average score (weighted) across all assessments
8. (A) Print total number of unique students who made at least 20 clicks each week for course-batch AAA 2013J (starting from the first day of the course), and (B) find the student id of the student who made the maximum number of clicks each week. In case more than one such student has a maximum number of clicks, print the lowest student id.
9. Print top 5 id_sites for material from VLE in each course-batch for which the total number of clicks has the highest correlation (Pearson's correlation) with pass rate in decreasing order of correlation within each course-batch.
10. What are the primary factors you think are leading students to withdraw or fail? Which of these can be checked from the datasets provided? Illustrate using query results. (For only this question, participants are allowed to use other open source tools as well but only for visualizations)

Submission Format

For a valid submission entry, the participant must adhere to the provided **SUBMISSION FORMAT** and submit the following requirements in a zipped file format below. All submissions must go in 2 folders namely:

1. **code_files:** Code files used by the participant for each question must be submitted in a separate folder named 'code_files' with file name as Question number - e.g. Qustion_1.txt and so on. An optional explanation for each query can be included within the txt file itself.
2. **Results:** Resultant table of each query must be pasted in the excel file - 'Question_1-9_query_results.xlsx' at the corresponding sheet for each question as provided in the submission format.

For Question 10, in the code_file folder provide all the MySQL queries used along with the visualisation code. Also, in the Results folder include your findings, queries and explanations in a docx file named 'Question_10.docx'.

Note: Last submission made by the participant would be considered as final submission

Evaluation

The submitted SQL queries would be evaluated on the following criteria:

- **Query Correctness:** Number of questions correctly answered
- **Query Optimization:** The run time of the SQL queries (How fast it can produce results from given data?)

- For last question, the solution will be evaluated by the Tredence Team on the basis of following criteria:
    - Problem Understanding & Hypothesis Generation
    - Visualization
    - Presentation of the solution