INFO20003 Semester 1, 2023

Assignment 2: SQL
Due: 6:00pm Friday, 28 April 2023
Weighting: 10% of your total assessment

Research Papers Database

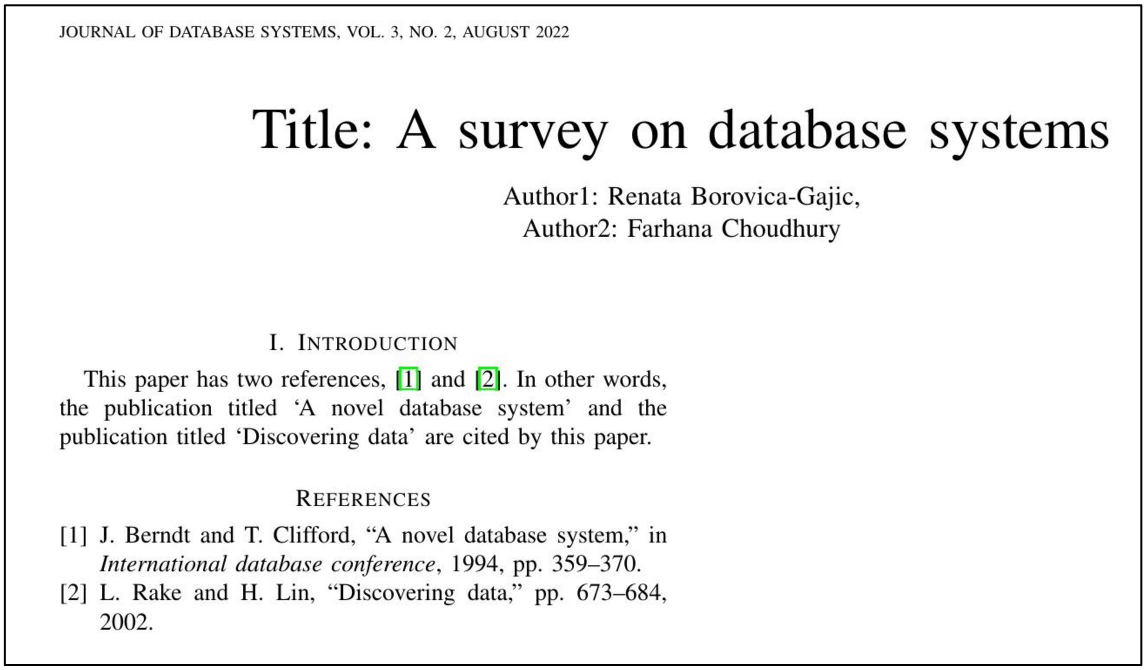
Description

You and a group of fellow undergrads have created a start-up like Google Scholar called 'newScholar'. newScholar is a free accessible web search engine that provides a broad range of information on scholarly publications. It also provides information on researchers and relations among scholarly publications and researchers.

For each researcher, newScholar records the researcher's details such as first name, last name, and one email address. Each researcher can also be associated with a few keywords representing their 'research area', such as Databases, Machine learning, Psychology, Medicine, etc. For each researcher, newScholar maintains their list of publications. The researchers who author a publication together are called 'co-authors'.

For each publication, newScholar stores its title, date of publication, start page number (e.g., 475), end page number (e.g., 500), and a list of authors (there can be multiple authors of a publication, where each author is a researcher). Each publication can also be associated with a few keywords representing its 'research area', such as Databases, Machine learning, Psychology, Medicine, etc. NewScholar database will not store the actual publications, but rather a link to the document objects. Each publication is linked to one document object. For each document object, newScholar stores the URL link and the document size in KB. Each publication has a list of references (i.e., it "cites" other publications), where each reference is another publication. If publication A is in the reference list of another publication B, then A is "cited by" B. Figure 1 shows an example publication with its basic information and its list of references.

NewScholar manually curates a list of top 10 publications every fortnight depending on the number of citations of the publications. A publication can make it to top 10 more than once over time. For each paper that is in the top 10 for a particular fortnight at a particular position, newScholar keeps a record of the start date when a publication reached that position (beginning of the fortnight) and end date (end of the fortnight). If a publication is again in the next fortnight's top 10 (whether in the same position, or different), a new row is recorded, with the start/end dates being the start/end dates of this new fortnight. For example, the publication entitled "Learning to index" can be number 1 for the fortnight from January 1st - January 14th, 2022 but the rank drops to number 3 from January 15th - January 28th, 2022. This would result in 2x rows, one with position 1 and one with position 3. Note that we do not need to know the rationale why this is the current ranking.



JOURNAL OF DATABASE SYSTEMS, VOL. 3, NO. 2, AUGUST 2022

Title: A survey on database systems

Author1: Renata Borovica-Gajic,
Author2: Farhana Choudhury

I. Introduction

This paper has two references, [1] and [2]. In other words, the publication titled ‘A novel database system’ and the publication titled ‘Discovering data’ are cited by this paper.

References

[1] J. Berndt and T. Clifford, “A novel database system,” in *International database conference*, 1994, pp. 359–370.
[2] L. Rake and H. Lin, “Discovering data,” pp. 673–684, 2002.

Fig 1: An example of publication with title, authors, and the list of references

The Data Model

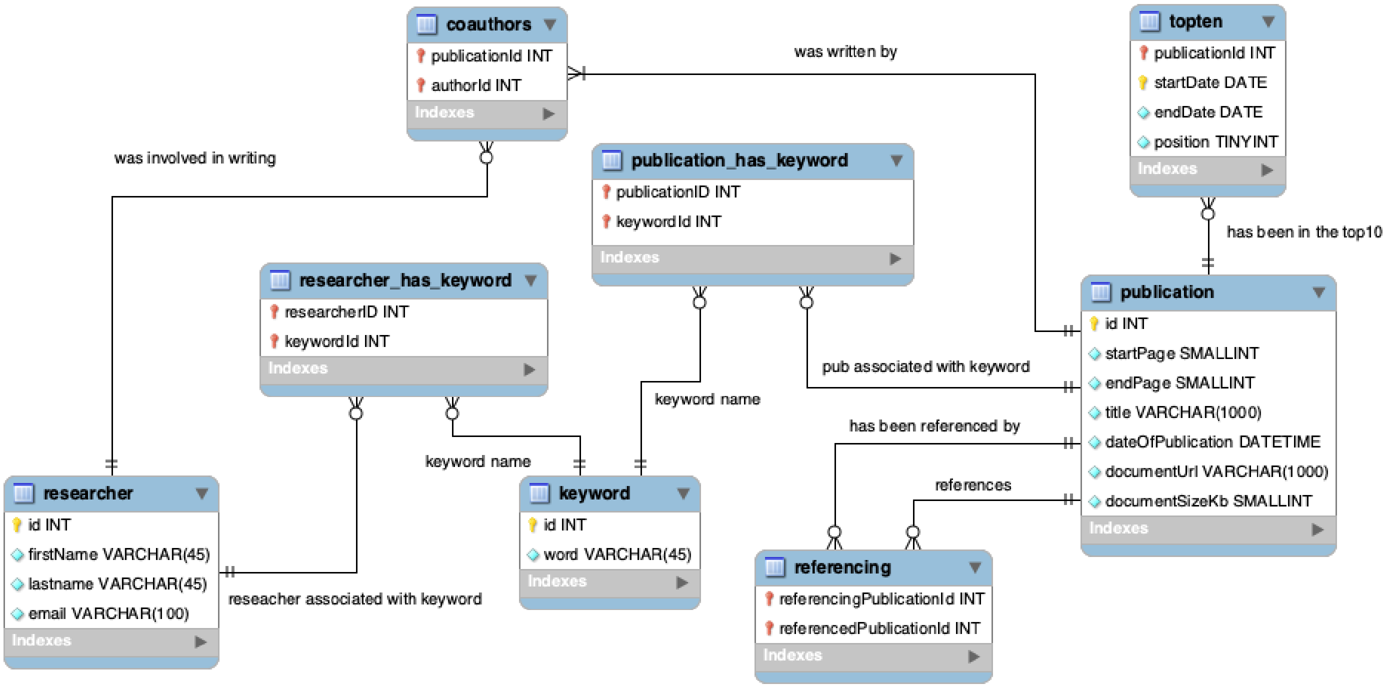


Fig 2: The physical ER model of newScholar startup database.

Assignment 2 Setup

A dataset is provided which you can use when developing your solutions. To set up the dataset, download the file **newScholar_2023.sql** from the Assignment link on Canvas and run it in Workbench. This script creates the database tables and populates them with data. Note that this dataset is provided for you to experiment with: but it is not the same dataset as what your queries will be tested against (the schema will stay the same, but the data itself may be different). *This means when designing your queries you must consider edge cases even if they are not represented in this particular data set.*

The script is designed to run against your account on the Engineering IT server (info20003db.eng.unimelb.edu.au). If you want to install the schema on your own MySQL Server installation, *uncomment the lines at the beginning of the script.*

Note: Do NOT disable only_full_group_by mode when completing this assignment. *This mode is the default, and is turned on in all default installs of MySQL workbench. You can check whether it is turned on using the command "SELECT @@sql_mode;". The command should return a string containing "ONLY_FULL_GROUP_BY" or "ANSI". When testing, our test server WILL have this mode turned on, and if your query fails due to this, you will lose marks.*

The SQL tasks

In this section are listed 10 questions for you to answer. Write one (single) SQL statement per question. Subqueries and nesting are allowed within a *single* SQL statement – however, you may be penalized for writing *overly* complicated SQL statements.

DO NOT USE VIEWS (or 'WITH' statements/common table expressions) to answer questions.

Some clarifications:

1. If publication A has 5 citations, that means publication A has been cited by 5 *other* publications (i.e., it does not mean that publication A cited 5 other publications).
2. A researcher's total number of citations refers to the count of citations of *all* the publications they've authored. If a publication has multiple citations, all the citations count towards the total count. For instance, if a researcher has 3 publications; the first one is cited 2 times, the second one 3 times, and the third one 5 times, we say that the researcher's total number of citations is 10.

Questions:

1. List all publications with no references (i.e., find publications that *do not* cite any other publications). Your query should return results of the form (publicationID, title). **(1 mark)**
2. Find the most recent publication. Assume there are no ties (only one publication is the most recent). Your query should return results of the form (publicationID, title, dateOfPublication). **(1 mark)**
3. List all papers from the author “Renata Borovica-Gajic” that are at least 10 pages in length. Your query should return results of the form (publicationID, title). **(1 mark)**
4. Find the publication with the highest number of citations. Note that, if a publication A is in the reference list of publication B and publication C, then the number of citations of A is 2. If there are ties, then you must return all publications with the highest number of citations. Your query should return results of the form (publicationID, title, citationCount), with one row per publication in case of a tie. **(2 marks)**
5. List the URLs of the papers that have ‘Databases’ as one of their research areas and have been cited at least once. Your query should return results of the form (publicationID, documentUrl). **(2 marks)**
6. Find the researchers who have at least 2 citations overall, and who have had publications in the top ten list at least once. Your query should return results of the form (firstName, lastName). **(2 marks)**
7. Find which research area has had the highest number of publications in the top-ten publication list over time. If a publication has appeared in the top 10 several times, count such publications just once in the calculation. If there are ties, then you must return all results. Your query should return results of the form (keywordID, word), with one row per research area in case of a tie. **(2 marks)**
8. List the researcher with the highest total citations. If there are ties, then you must return all such results. Your query should return results of the form (firstName, lastName, totalCitationCount), with one row per researcher in case of a tie. **(3 marks)**
9. Find the researchers in ‘Databases’ research area who have co-authored at least one publication with at least one researcher from ‘Machine learning’ research area (regardless of the research area of the publication itself). Note that we only want to consider co-authorships where the ‘Databases’ researcher is distinct from the ‘Machine Learning’ researcher: if there is a ‘Databases’ researcher who *also* is in the ‘Machine learning’ area, they must have co-authored with a *different* researcher who is in the ‘Machine learning’ area to be included. Your query should return results of the form (firstName, lastName). **(3 marks)**
10. Find the researchers who have not co-authored a publication on or after 01/01/2023 (where 01/01/2023 is the date of the publication) with the researcher named “Renata Borovica-Gajic” but have had at least one publication before that date co-authored with “Renata Borovica-Gajic” (that means they have had joint publications in the past, but not on or after 01/01/2023). Your query should return results of the form (firstName, lastName) for all such researchers. **(3 marks)**

SQL Response Formatting Requirements

To help us mark your assignment queries as quickly/accurately as possible, please ensure that:

- Your query returns the projected attributes in the same order as given in the question, and does **not** include additional columns. E.g., if the question asks 'return as (userId, name)', please write "SELECT userId, name …" instead of "SELECT name, userId…" *(you can name the columns using `AS` however you'd like, only the* ***order*** *matters).*
- Please do NOT use "databaseName.tableName" format. E.g., please write "SELECT userId FROM users…" instead of "SELECT userId FROM coltonc.users …".
- Ensure that you are using single quotes(') for strings (e.g. …WHERE name = 'bob'…) and double quotes (") only for table names (e.g. SELECT name FROM "some table name with spaces"…) . Do **NOT** use double quotes for strings "…WHERE name = "bob"…".

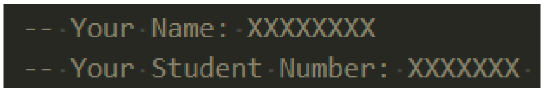
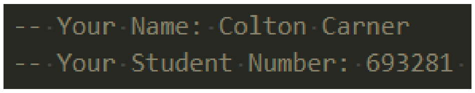
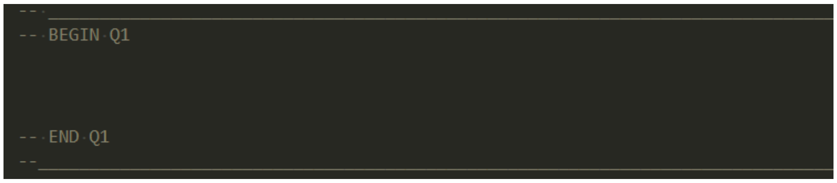
Submission Instructions

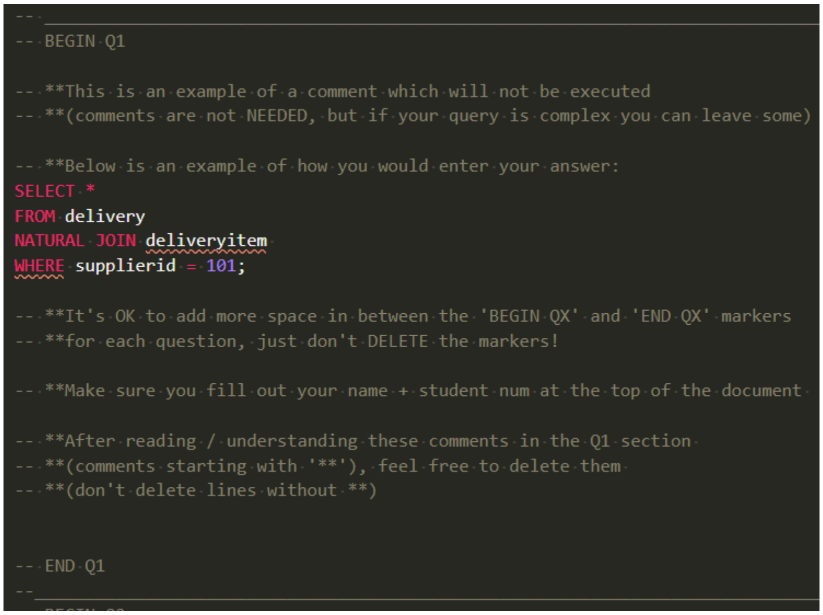
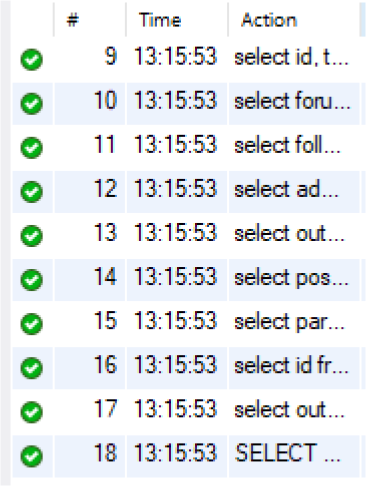
Your submission will be in the form of an SQL script. There is a template file on the LMS, into which you will paste your solutions and fill in your student details (more information below).

This .sql file should be submitted on Canvas by **6pm** on the due date of **Friday, 28 April 2023**. Name your submission as 987654.sql, where 987654 corresponds to YOUR student id.

Filling in the template file:

The template file on the LMS has spaces for you to fill in your student details and your answers to the questions. There is also an example prefilled script available on the LMS as well. Below are screenshots from those two documents explaining the steps you need to take to submit your solutions:

Step	Example
1. At the top of the template, you'll need to replace "XXXXXXXX" with your student number and name	*Template*  -- Your Name: XXXXXXXX -- Your Student Number: XXXXXXX
	Example Filled in  -- Your Name: Colton Carner -- Your Student Number: 693281
2. For each question 1-10, place your SQL solution in between the "BEGIN QX" and "END QX" markers. **Ensure each query is**	*Template*  -- -- BEGIN Q1 -- END Q1 --

terminated with a semicolon “;”	*Example Filled in*  -- -- BEGIN Q1 -- **This is an example of a comment which will not be executed -- **(comments are not NEEDED, but if your query is complex you can leave some) -- **Below is an example of how you would enter your answer: SELECT * FROM delivery NATURAL JOIN deliveryitem WHERE supplierid = 101; -- **It's OK to add more space in between the 'BEGIN QX' and 'END QX' markers -- **for each question, just don't DELETE the markers! -- **Make sure you fill out your name + student num at the top of the document -- **After reading / understanding these comments in the Q1 section -- **(comments starting with '**'), feel free to delete them -- **(don't delete lines without **) -- END Q1 --
3. Test that your script is valid SQL by running it from MySQL Workbench. Run the entire script by copy-pasting this entire file into a new workbench tab, placing your cursor at the start of the file (without selecting anything), and pressing the lightning bolt to run the entire file. All queries should run successfully one after another. If not, check to make sure you added semicolons ‘;’ after each query.	*All 10 queries ran sequentially and were successful.*  # Time Action ✓ 9 13:15:53 select id, t... ✓ 10 13:15:53 select foru... ✓ 11 13:15:53 select foll... ✓ 12 13:15:53 select ad... ✓ 13 13:15:53 select out... ✓ 14 13:15:53 select pos... ✓ 15 13:15:53 select par... ✓ 16 13:15:53 select id fr... ✓ 17 13:15:53 select out... ✓ 18 13:15:53 SELECT ...

Late submission

Unless you have an approved extension (see below), you will be penalised -10% of the total number of marks in the assignment per day that your submission is late. For instance, if you received a 78% raw score, but submitted 2 days late, you'd receive a 58% score for the assignment.

Requesting a Submission Deadline Extension

If you need an extension due to a valid (medical) reason, you will need to provide evidence to support your request by **5pm on Thursday 27 April**. Extensions received after this time may be rejected. Medical certificates need to be at least two days in length.

To request an extension:

- Email Timothy Hermanto (timothy.hermanto@unimelb.edu.au) from your university email address, supplying your student ID, the extension length that you require and supporting evidence. Please add INFO20003 in the subject title.
- If your submission deadline extension is granted you will receive an email reply granting the new submission date. Do not lose this email!

Reminder: INFO20003 Hurdle Requirements

To pass INFO20003, you must pass two hurdles:

- **Hurdle 1:** Obtain at least 50% (15/30) for the three assignments (each worth 10%)
- **Hurdle 2:** Obtain at least 50% (35/70) for the combination of the quizzes and final exam

Therefore, it is our recommendation that you attempt every assignment and question in the exam.

GOOD LUCK!