COMS W4111: Introduction to Databases
Spring 2024, Sections 002/V02

Homework 1
Introduction to Core Concepts, ER Modeling, Relational Algebra, SQL

Introduction

This notebook contains Homework 1. **Both Programming and Nonprogramming tracks should complete this homework.**

Submission Instructions

- You will submit **PDF and ZIP files** for this assignment. Gradescope will have two separate assignments for these.
- For the PDF:
  - The most reliable way to save as PDF is to go to your browser's menu bar and click `File -> Print`. Switch the orientation to landscape mode, and hit save.
  - **MAKE SURE ALL YOUR WORK (CODE AND SCREENSHOTS) IS VISIBLE ON THE PDF. YOU WILL NOT GET CREDIT IF ANYTHING IS CUT OFF.** Reach out for troubleshooting.
  - **MAKE SURE YOU DON'T SUBMIT A SINGLE PAGE PDF.** Your PDF should have multiple pages.
- For the ZIP:
  - Zip a folder containing this notebook and any screenshots.
  - You may delete any unnecessary files, such as caches.

Add Student Information

In [10]:
```
# Print your name, uni, and track below

name = "Naomi Jiang"
uni = "yj2747"
track = "Non-Programming Track"

print(name)
print(uni)
print(track)
```

```
Naomi Jiang
yj2747
Non-Programming Track
```

Setup

SQL Magic

The `sql` extension was installed in HW0. Double check that if this cell doesn't work.

In [11]:
```
%load_ext sql
```

```
The sql extension is already loaded. To reload it, use:
  %reload_ext sql
```

You may need to change the password below.

In [12]:
```
%sql mysql+pymysql://root:dbuserdbuser@localhost
```

In [13]:
```
%sql SELECT * FROM db_book.student WHERE ID = 12345
```

```
 * mysql+pymysql://root:***@localhost
1 rows affected.
```

Out[13]:

ID	name	dept_name	tot_cred
12345	Shankar	Comp. Sci.	32

Python Libraries

In [18]:
```
from IPython.display import Image
import pandas
```

Written Questions

Chapter 1 from the recommended textbook Database System Concepts, Seventh Edition (https://codex.cs.yale.edu/avi/db-book/) covers general information and concepts about databases and database management systems. Lecturing on the general and background information is not a good use of precious class time. To be more efficient with class time, the chapter 1 information is a reading assignment.

Answering the written questions in HW 1, Part 1 does not require purchasing the textbook and reading the chapter. The chapter 1 slides (https://codex.cs.yale.edu/avi/db-book/slides-dir/index.html) provided by the textbook authors provide the necessary information. In some cases, students may also have to search the web or other sources to "read" the necessary information.

When answering the written questions, do not "bloviate". The quantity of words does not correlate with the quality of the answer. We will deduct points if you are not succinct. The answers to the questions require less than five sentences or bullet points.

W1

What is a database management system and how do relational databases organize data?

- A database management system (or DBMS) is essentially a computerized data-keeping system. Users of the system are given facilities to perform several kinds of operations on such a system for either manipulation of the data in the database or the management of the database structure itself.
- Database Management Systems (DBMSs) are categorized according to their data structures or types. In relational databases, all the data is stored in various tables.
- A relational database organizes data into rows and columns, which collectively form a table. Data is typically structured across multiple tables, which can be joined together via a primary key or a foreign key.

https://www.ibm.com/docs/en/zos-basic-skills?topic=zos-what-is-database-management-system (https://www.ibm.com/docs/en/zos-basic-skills?topic=zos-what-is-database-management-system)
https://www.ibm.com/topics/relational-databases#:~:text=A%20relational%20database%20organizes%20data,key%20or%20a%20foreign%20key (https://www.ibm.com/topics/relational-databases#:~:text=A%20relational%20database%20organizes%20data,key%20or%20a%20foreign%20key)

W2

Columbia University uses several applications that use databases to run the university. Examples are SSOL and CourseWorks. An alternate approach could be letting students, faculty, administrators, etc. use shared Google Sheets to create, retrieve, update, and delete information. What are some problems with the shared spread sheet approach and what functions do DMBS implement to solve the problems?

- In a shared spreadsheet environment, there's a high risk of data redundancy since multiple copies of the same data can be created and maintained by different users. However, a DBMS controls redundancy by maintaining a single repository of data that is defined once and accessed by many users.
- For the shared spreadsheet, multiple people are accessing the same data at the same time, which can lead to accidental deletion or overwriting. However, the DBMS provides a locking system and comprehensive backup and recovery mechanisms to prevent such anomalies from occurring.
- Google Sheets are easy to share, but this can decrease the security of the data. In contrast, a DBMS has specialized features that help provide protection for its data.
- Compared to Google Sheets, a DBMS is more easily maintainable due to its centralized nature.

Reference: https://www.geeksforgeeks.org/advantages-of-dbms-over-file-system/ (https://www.geeksforgeeks.org/advantages-of-dbms-over-file-system/)

W3

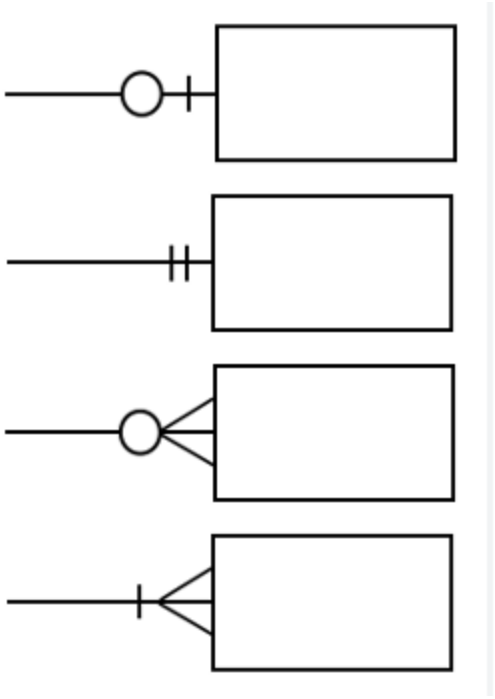
Explain the differences between SQL, MySQL Server and DataGrip.

- SQL is the language used for interacting with databases
- MySQL Server is a database management system that uses SQL for its operations ◦ Itstores and converts character sequences into tables and rows.
- DataGrip is an IDE that supports SQL and MySQL, providing a powerful environment for database professionals.

https://www.jetbrains.com/datagrip/ (https://www.jetbrains.com/datagrip/)

W4

Crow's Foot Notation has four endings for relationship lines. Briefly explain the meaning of each ending.



- Zero or One: The origin can have zero or one linked entity with the other entity set.
- One and Only One: The origin entity has only one entity linked with the other entity.
- Zero or Many: The origin can have zero or many linked entities with the other entity set.
- One or Many: The origin can have one or many linked entities with the other entity set.

https://www.edrawsoft.com/er-diagram-symbols.html (https://www.edrawsoft.com/er-diagram-symbols.html)

W5

What is a primary key and why is it important?

- A primary key is a unique identifier for each record in a database table.

- It ensures that no two rows can have the same key, which guarantees the 'uniqueness' property for

W6

The relational algebra is closed under the operators. Explain what this means and give an example.

- Closure under its operators means any operation applied to relations produces another relation. This allows for the chaining of operations, where the output of one can be the input for another.
- For example, selecting students with more than 90 credits yields a new relation of qualifying students, demonstrating closure with the selection operator.

https://www.w3schools.in/dbms/relational-algebra#:~:text=So%20the%20output%20from%20one,are%20closed%20under%20arithmetic%20operatio (https://www.w3schools.in/dbms/relational-algebra#:~:text=So%20the%20output%20from%20one,are%20closed%20under%20arithmetic%20operatio

W7

Some of the Columbia University databases/applications represent the year/semester attribute of a section in the form "2023_2". The first four characters are the academic year, and the last character is the semester (1, 2, or 3). The data type for this attribute might be CHAR(6). Using this example, explain the concepts of domain and atomic domain. How is domain different from type?

- A domain in database terminology refers to the set of allowable values that a database field (or attribute) can hold.The "2023_2" format used to represent the year/semester attribute, the domain refers to the specific set of values that this attribute can take, such as any valid year followed by a semester indicator (1, 2, or 3).
- An atomic domain means that the values within the domain are indivisible in the context of the database. ike "2023_2", are indivisible in the context they are used, meaning they represent a single, discrete piece of information (a specific year and semester) without further breakdown.
- The difference between domain and type is that while the domain specifies the permissible values for an attribute, such as the format "YYYY_S", the type (e.g., CHAR(6)) describes the kind of data (character string of length 6).

W8

Briefly explain the difference between a database schema and database instance.

- Schema is structure of the database
- Instance is the actual content of the database at a particular point in time

W9

Briefly explain the concepts of data definition language and data manipulation language.

- Data Definition Language (DDL) is a syntax for defining and modifying database structures, including tables, fields, and relationships; it results in updates to the database schema stored in a data dictionary without manipulating the data itself.

- Data Manipulation Language (DML) is a set of commands used to access, modify, and manage data within the database, allowing for operations such as insertion, deletion, and updating of data according

W10

What is physical data independence?

- physical data independence: the ability to modify the physical schema without changing the logical schema

---

Entity-Relationship Modeling

Overview

The ability to understand a general description of a requested data model and to transform into a more precise, specified *logical model* is one of the most important skills for using databases. SW and data engineers build applications and data models for end-users. The end-users, product managers and business managers are not SW or data modeling experts. They will express their *intent* in imprecise, text and words.

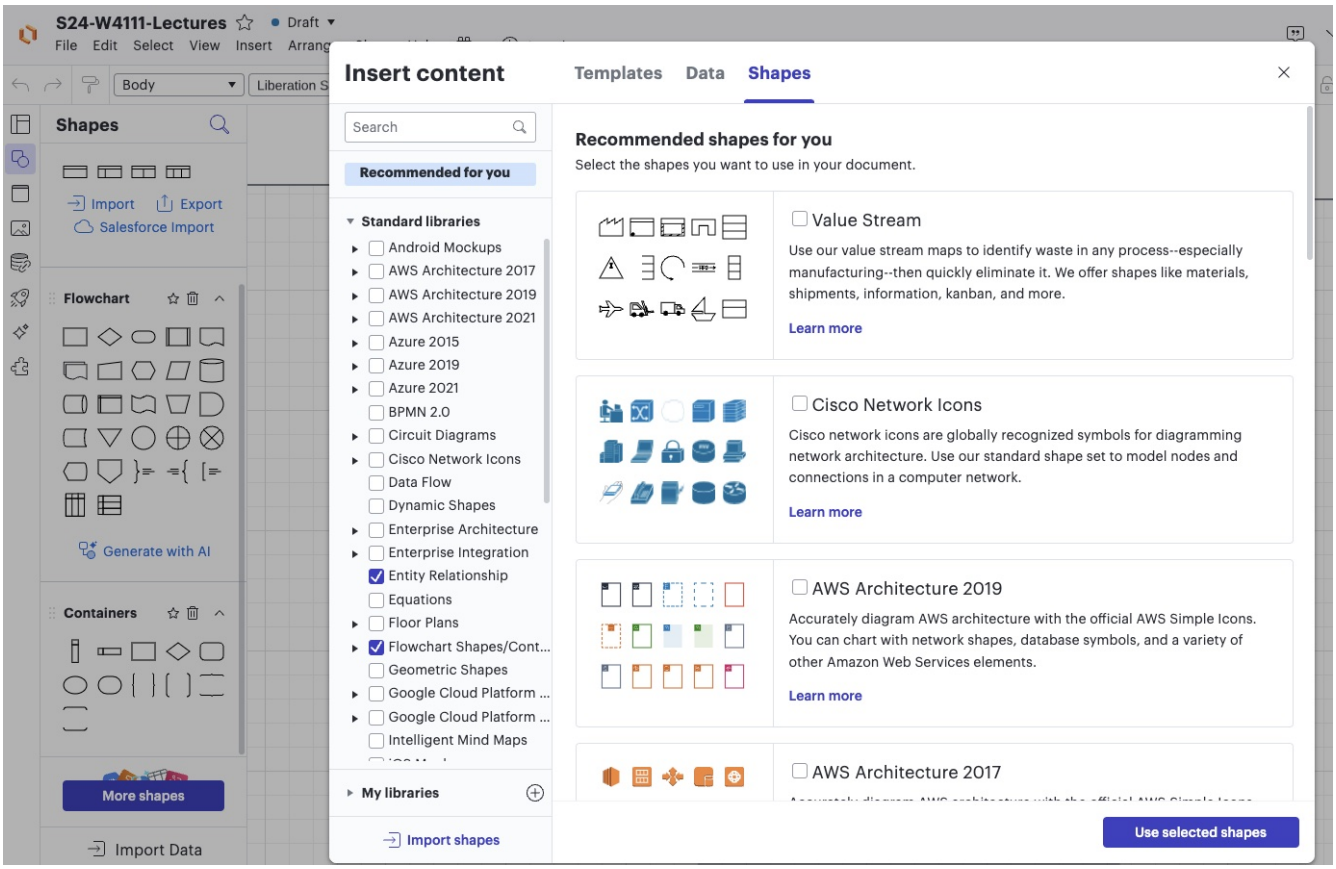
The users and business stakeholder often can understand and interact using a *conceptual model* but details like keys, foreign keys, ... are outside their scope.

In this problem, you will:

- Understand a short written description of a requested data model.
- Produce a *conceptual data model diagram* using Lucidchart.
- Produce a *logical data model diagram* using Lucidchart.

You can sign up for a free Lucidchart account. (https://www.lucidchart.com/pages/landing) The free account provides the capabilities you will need for this course.

To draw the diagrams, you need to add the *entity relationship* shapes. Lecture 2 demonstrated how to add the shapes.



Adding Entity Relationship Shapes

We provide a simple [Lucidchart document (https://lucid.app/lucidchart/828777b1-7b2d-4828-bedb-37b6d456c33e/edit?invitationId=inv_a142899a-7e60-44e9-b18e-335d7c9767fc)](https://lucid.app/lucidchart/828777b1-7b2d-4828-bedb-37b6d456c33e/edit?invitationId=inv_a142899a-7e60-44e9-b18e-335d7c9767fc) from Lecture 2 that helps you get started. You need a Lucidchart account to access the document and diagrams.

Data Model Description

The data model represents banks, customers. employees and accouts. The model has the following entity types/sets:

1. *Customer*
2. *Employee* of the banking company
3. *Branch,* which is a location of one of the banks offices
4. *Savings Account*
5. *Checking Account*
6. *Loan*
7. *Portfolio*

Customer has the following properties:

- *customerID*
- *lastName*
- *firstName*
- *email*
- *dateOfBirth*

Employee has the following properties:

- *employeeID*
- *lastName*
- *firstName*
- *jobTitle*

Branch has the following properties:

- *branchID*
- *zipCode*

Savings Account has the following properties:

- *accountID*
- *balance*
- *interestRate*

Checking Account has the following properties:

- *accountID*
- *balance*

Loan has the following properties.

- *loanID*
- *balance*
- *interestRate*

Portfolio has the following properties:

- *portfolioID*
- *createdDate*

The data model has the following relationships:

- *Customer Branch* connects a customer and a branch. A *Customer* is connected to exactly one *Branch*. A *Branch* may have 0, 1 or many customers.
- *Employee Branch* connects an employee and a branch. An *Employee* is connected to exactly one *Branch*. A *Branch* may have 0, 1 or many associated employees.
- *Savings Account Branch*, *Checking Account Branch*, and *Loan Branch* all have the same pattern.
  - An account/loan has exactly one branch.
  - A *Branch* many have 0, 1 or many accounts/loans.
- *Savings Customer*, *Checking Customer*, *Loan Customer*, and *Portfolio Customer* follow the same pattern.
  - The account/loan has exactly one customer.
  - The customer may have 0 or 1 of each type of account.
- A *Portfolio* is related to exactly one *Customer*, exactly one *Savings Account*, exactly one *Checking Account*, and exactly one *Loan*.
- *Portfolio Advisor* relates a *Portfolio* and *Employee*. An *Employee* may be the advisor for 0, 1 or many *Portfolios*. A *Portfolio* may have at most one *Employee* advisor.

Answer

1. Place your Logical Model diagram below.
2. You *may* have to add attributes to entities to implement the model.

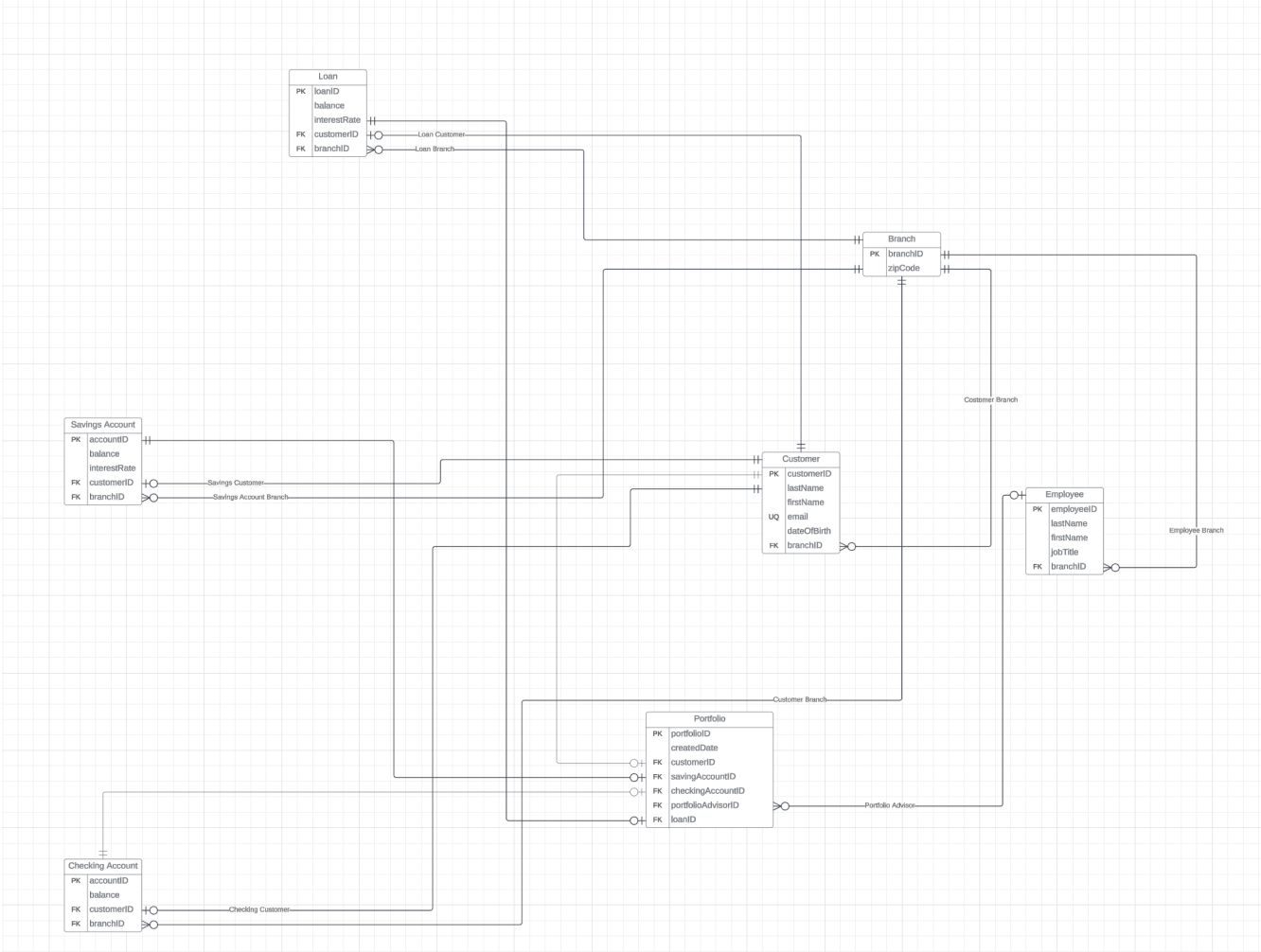
3. You *may* make reasonable assumptions. Please document your assumptions below. You may add comments/notes to your diagram for clarity.

Assumptions:

- A Branch many have 0, 1 or many accounts/loans. So I assume a branch have 0 or many accounts/loans.
- A Branch may have 0, 1 or many associated employees. So I assume a branch have 0 or many employees.
- savingAccountID, checkingAccountID, loanID, and customerID are all unique, primary keys.
- Email is the unique constraint that make sure no two users can register with the same email address.
- customerID are foreign keys to portfolio, checking account and saving account
- savingAccountID, checkingAccountID, loanID and portfolioAdvisorID are Portfolio's foreign keys

ER Diagram:

Save your diagram to an image, place in the same directory as your notebook and change the file name in the HTML `img` tag in this Markdown cell.



Logical ER Diagram

Relational Algebra

R-1

The following is the SQL DDL for the `db_book.classroom` table.

```
CREATE TABLE IF NOT EXISTS db_book.classroom
(
    building    VARCHAR(15) NOT NULL,
    room_number VARCHAR(7)  NOT NULL,
    capacity    DECIMAL(4)  NULL,
    PRIMARY KEY (building, room_number)
);
```

Using the notation from the lecture slides, provide the corresponding relation schema definition.

$classroom(\underline{building}, \underline{room\_number}, capacity)$

Answer Format

R0

Write a relational algebra statement that produces a table of the following form:

- ID is the instructor ID
- name is the instructor name
- course_id, sec_id, semester, year of a section
- building, room_number

Note:

1. You will have to use the instructor, teaches and section relations
2. Your answer should only include sections taught in `Comp. Sci.` in `2009`

Algebra statement:

```
π ID, name, course_id, sec_id, semester, year, building, room_number(
    (σ dept_name='Comp. Sci.' ∧ year=2009
        (teaches ⋈ instructor)
        )
⋈ section)
```

Execution:

RO Execution Result

R1

Write a relational algebra statement that produces a relation with the columns:

- `student.name`
- `student.dept_name`
- `student.tot_cred`
- `instructor.name` (the instructor that advises the student)
- `instructor.dept_name`

Only keep students who have earned more than 90 credits.

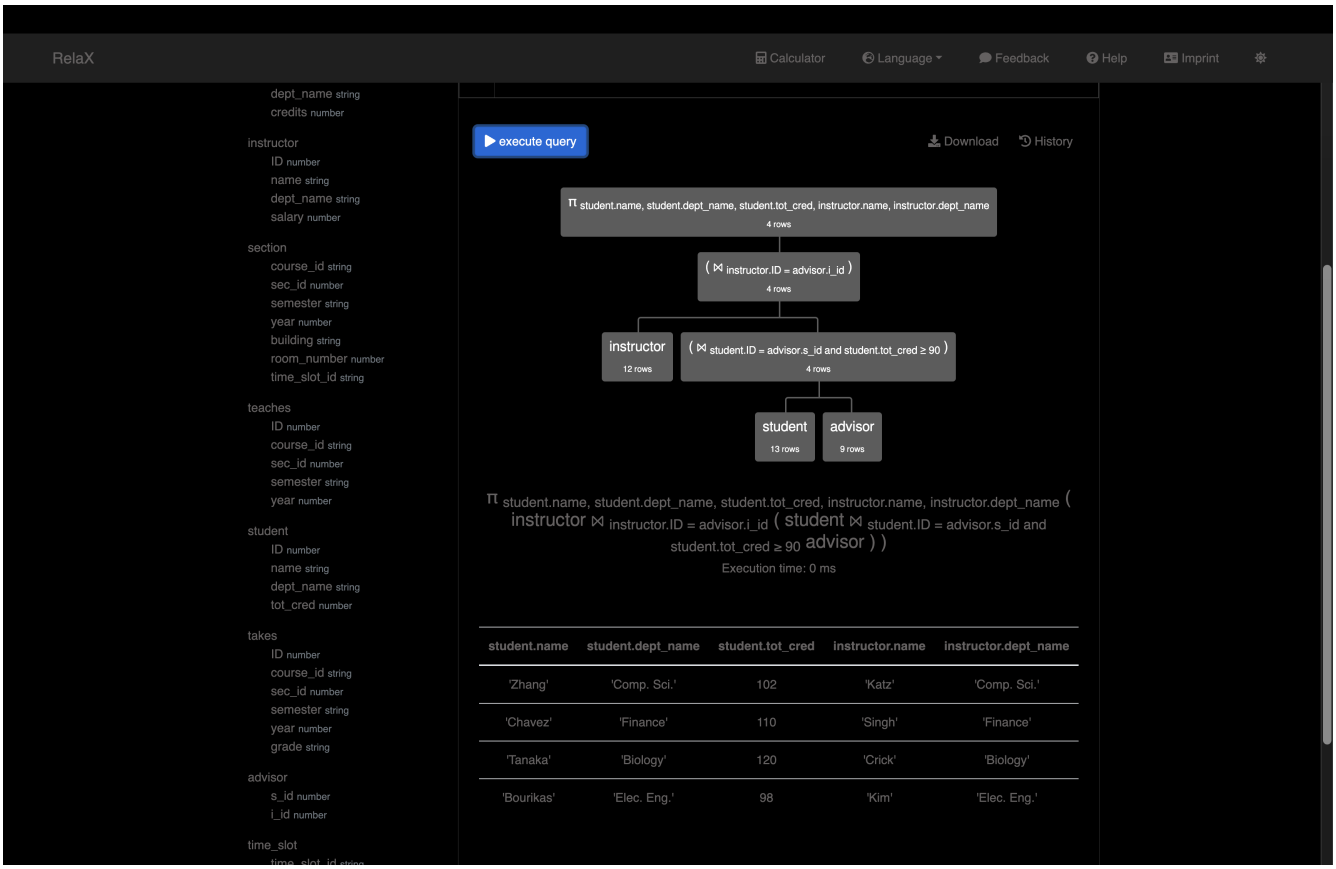
Note:

1. You will have to use the `student`, `instructor`, and `advisor` relations.
2. You should only include students that have an advisor, i.e., `instructor.name` and `instructor.dept_name` should be non-null for all rows.

Algebra statement:

```
π student.name , student.dept_name , student.tot_cred , instructor.name ,
instructor.dept_name
((instructor ⋈ instructor.ID=advisor.i_id
(student ⋈ student.ID=advisor.s_id ∧ student.tot_cred ≥90 advisor)))
```

Execution:



R2

Write a relational algebra statement that produces a relation with the columns:

- `course_id`
- `title`
- `prereq_course_id`
- `prereq_course_title`

This relation represents courses and their prereqs.

Note:

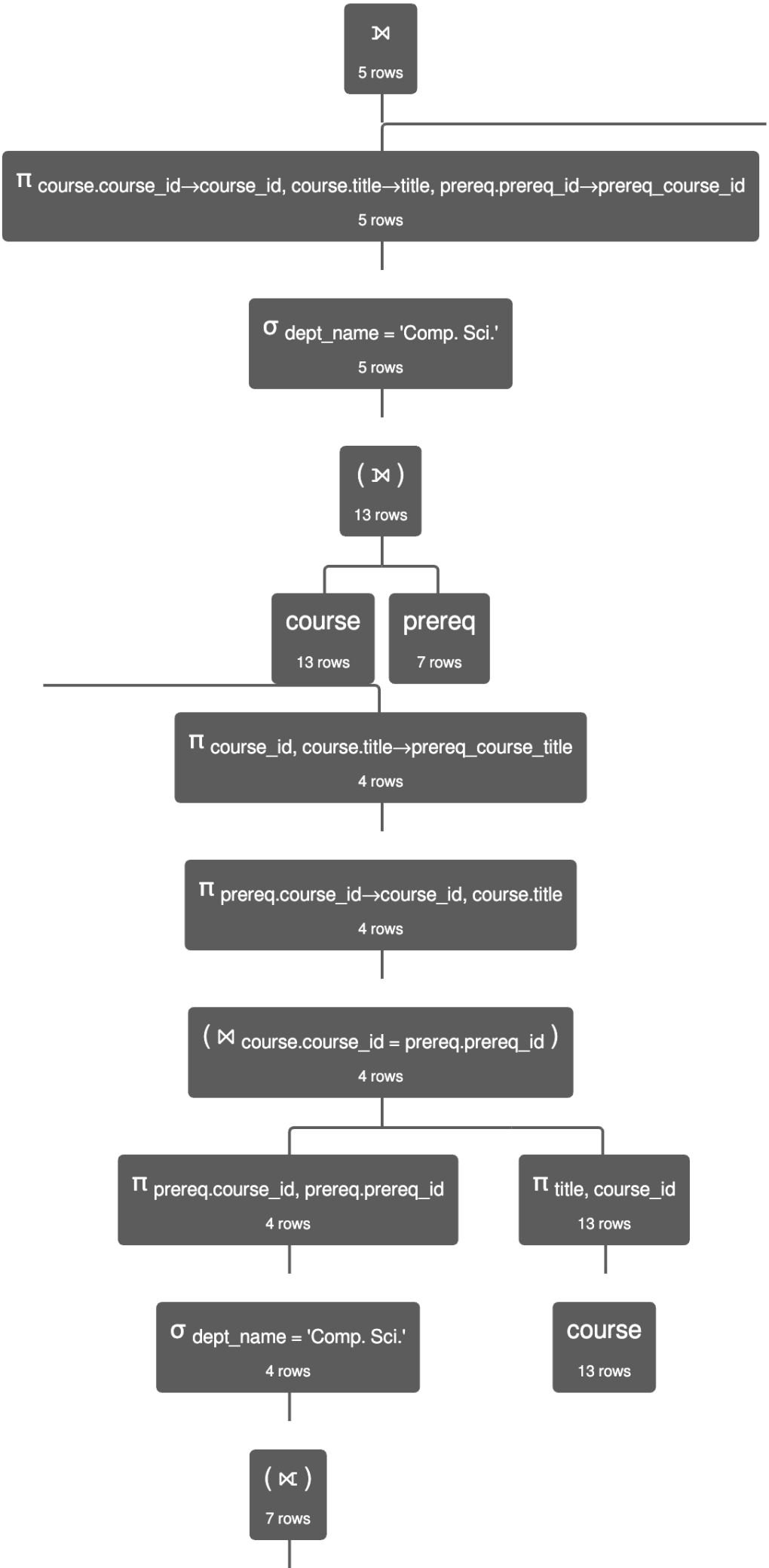
1. This query requires the `course` and `prereq` tables.
2. Your answer should only include courses in the `Comp. Sci.` department.
3. If a course has no prereqs, `prereq_course_id` and `prereq_course_title` should both be *null*.
4. You *may* have to use table and column renaming.

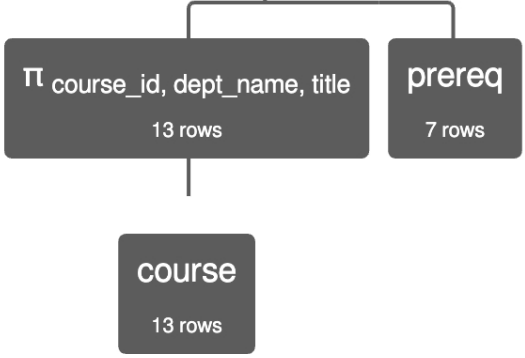
Algebra statement:

```
π course.course_id→course_id, course.title→title, prereq.prereq_id→prereq
_course_id(
σ dept_name='Comp. Sci.'
(course⋈prereq)
)
⟖
(π course_id, course.title→ prereq_course_title (π prereq.course_id→cours
e_id, course.title(
π prereq.course_id, prereq.prereq_id(
    (σ dept_name='Comp. Sci.'
        (π course_id,dept_name,title(course)
            ⋈
        prereq)
        )
)
⋈ course.course_id = prereq.prereq_id
π title, course_id(course))))
```

Execution:

|


⋈
5 rows
π course.course_id→course_id, course.title→title, prereq.prereq_id→prereq_course_id
5 rows
σ dept_name = 'Comp. Sci.'
5 rows
(⋈)
13 rows
course
13 rows
prereq
7 rows
π course_id, course.title→prereq_course_title
4 rows
π prereq.course_id→course_id, course.title
4 rows
(⋈ course.course_id = prereq.prereq_id)
4 rows
π prereq.course_id, prereq.prereq_id
4 rows
π title, course_id
13 rows
σ dept_name = 'Comp. Sci.'
4 rows
course
13 rows
(⋈)
7 rows



SQL

New Database

MySQL Tutorial (https://www.mysqltutorial.org/) is a good site with information that complements and extends the core material in our course. Much of the material the site covers is applicable to other SQL products. MySQL Tutorial uses an interesting dataset that is more complex than the simple "db_book" database. This is the Classic Models Dataset (https://www.mysqltutorial.org/getting-started-with-mysql/mysql-sample-database/). The complexity allows us to better appreciate more complex SQL concepts.

You learned how to run a SQL script/file as part of HW0. **Use the same approach to load and create the `Classic Models Database`.** The file is `classic_models_database.sql` and is in the HW1 folder.

To test loading the data, you can use the cell below.

In [19]:
```
%sql show tables;
```

```
 * mysql+pymysql://root:***@localhost
8 rows affected.
```

Out[19]:

Tables_in_classicmodels
customers
employees
offices
orderdetails
orders
payments
productlines
products

In [28]:
```
%sql USE classicmodels;
```

```
 * mysql+pymysql://root:***@localhost
0 rows affected.
```

Out[28]:
```
[]
```

SQL 1

This query uses `customers` and `employees`.

Write and execute a SQL query that produces a table with the following columns:

- `customerContactName`
- `customerPhone`
- `salesRepName`

Only keep customers from France. Order your output by `customerContactName`.

Notes:

- The names of your columns must match exactly with what is specified.
- `customerContactName` can be formed by combining `customers.contactFirstName` and `customers.contactLastName`.
- `salesRepName` can be formed by combining `employees.firstName` and `employees.lastName`.

In [26]:
```
%%sql
SELECT
    CONCAT(customers.contactFirstName, ' ', customers.contactLastName) AS custome
    customers.phone AS customerPhone,
    CONCAT(employees.firstName, ' ', employees.lastName) AS salesRepName
FROM
    customers
JOIN
    employees ON customers.salesRepEmployeeNumber = employees.employeeNumber
WHERE
    customers.country = 'France'
ORDER BY
    customerContactName;
```

```
 * mysql+pymysql://root:***@localhost
12 rows affected.
```

Out[26]:

customerContactName	customerPhone	salesRepName
Annette Roulet	61.77.6555	Gerard Hernandez
Carine Schmitt	40.32.2555	Gerard Hernandez
Daniel Tonini	30.59.8555	Gerard Hernandez
Daniel Da Silva	+33 1 46 62 7555	Loui Bondur
Dominique Perrier	(1) 47.55.6555	Loui Bondur
Frédérique Citeaux	88.60.1555	Gerard Hernandez
Janine Labrune	40.67.8555	Gerard Hernandez
Laurence Lebihan	91.24.4555	Loui Bondur
Marie Bertrand	(1) 42.34.2555	Loui Bondur
Martine Rancé	20.16.1555	Gerard Hernandez
Mary Saveley	78.32.5555	Loui Bondur
Paul Henriot	26.47.1555	Loui Bondur

SQL 2

This query uses `employees`, `customers`, `orders`, `orderdetails`.

Write and execute a SQL query that produces a table showing the amount of money each sales rep has generated.

Your table should have the following columns:

- `salesRepName`
- `moneyGenerated`

Order your output from greatest to least `moneyGenerated`.

Notes:

- The names of your columns must match exactly with what is specified.
- `salesRepName` can be formed by combining `employees.firstName` and `employees.lastName`.
- To calculate `moneyGenerated`:

    - Every order in `orders` is associated with multiple rows in `orderdetails`. The total amount of money spent on an order is the sum of `quantityOrdered * priceEach` for all the associated rows in `orderdetails`. **Only consider orders that are `Shipped`.**
    - A customer can have multiple orders. The total amount of money a customer has spent is the sum of the money spent on all that customer's orders.
    - A sales rep can have multiple customers. `moneyGenerated` is the sum of the money spent by all that sales rep's customers.
- You may find the [WITH keyword (https://www.tutorialspoint.com/mysql/mysql_with.htm)] to be useful for cleaner code.

In [27]:

```
%%sql
WITH OrderAmounts AS (
    SELECT
        o.orderNumber,
        SUM(od.quantityOrdered * od.priceEach) AS orderTotal
    FROM
        orders o
        JOIN orderdetails od ON o.orderNumber = od.orderNumber
    WHERE
        o.status = 'Shipped'
    GROUP BY
        o.orderNumber
),
CustomerTotals AS (
    SELECT
        c.salesRepEmployeeNumber,
        SUM(oa.orderTotal) AS customerTotal
    FROM
        OrderAmounts oa
        JOIN orders o ON oa.orderNumber = o.orderNumber
        JOIN customers c ON o.customerNumber = c.customerNumber
    GROUP BY
        c.salesRepEmployeeNumber
)
SELECT
    CONCAT(e.firstName, ' ', e.lastName) AS salesRepName,
    SUM(ct.customerTotal) AS moneyGenerated
FROM
    CustomerTotals ct
    JOIN employees e ON ct.salesRepEmployeeNumber = e.employeeNumber
GROUP BY
    salesRepEmployeeNumber
ORDER BY
    moneyGenerated DESC;
```

```
 * mysql+pymysql://root:***@localhost
15 rows affected.
```

Out[27]:

salesRepName	moneyGenerated
Gerard Hernandez	1065035.29
Leslie Jennings	1021661.89
Pamela Castillo	790297.44
Larry Bott	686653.25
Barry Jones	637672.65
George Vanauf	584406.80
Loui Bondur	569485.75
Peter Marsh	523860.78
Andy Fixter	509385.82
Foon Yue Tseng	488212.67
Mami Nishi	457110.07
Steve Patterson	449219.13
Martin Gerard	387477.47
Julie Firrelli	386663.20
Leslie Thompson	307952.43

In [ ]:

In [ ]: