

Advanced Data Management - Assignment N°01

The assignment sheet contains one exercise with seven tasks. Last chance of uploading the assignments will be 30.10.2023. For each task, there is a maximum allotted time. Please do not work longer than the specified time on a task! If you think that you will not be able to finish the task in the given maximum time, stop working on it 15 minutes before the end, and provide an explanation containing the following information:

- Whether you think that the task is solvable with the current system at all, and why?
- If you think that is solvable with more time: which approach, would you try out next?

Exercise 1: Relational Databases

DATA IN RELATIONAL DATABASE

In relational databases data is organized in tables. The data model for this task is shown in figure 1.

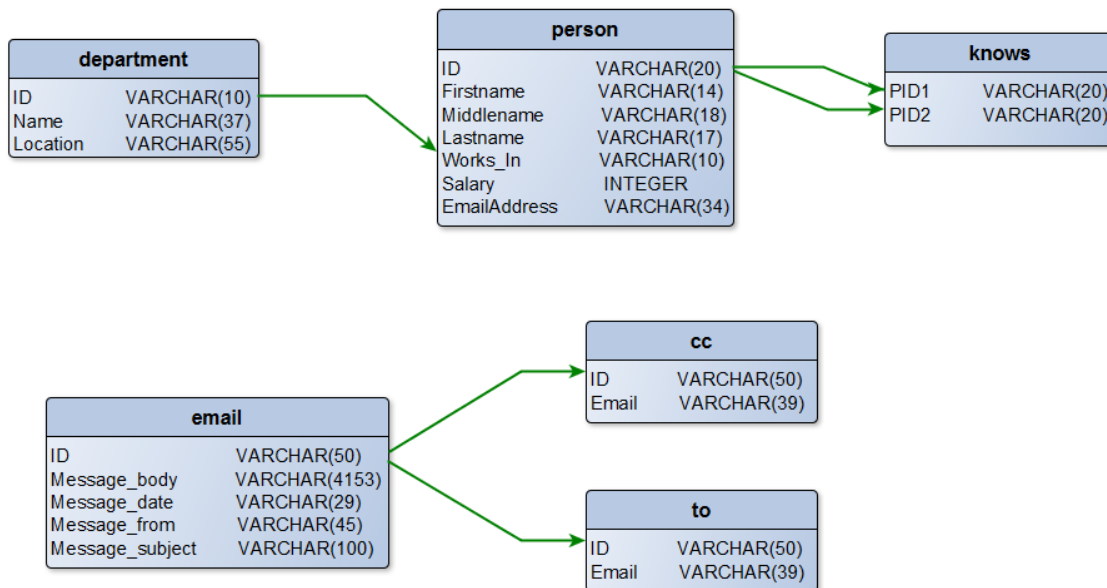


Figure1. Tables and attributes in relational database

SYSTEM PREPARATION

For relational database we use PostgreSQL server. To connect to the server, follow the following steps:

1- You should have received your user name and password via Moodle

2- Use the user name and password to log in to our PgAdmin on :

`http://162.55.237.41:8080/`

3- Once you have logged in, create a new sever connection by right clicking on “**Sever**” in the right side menu, then select “**create-> server**”. Enter the information for the connection tab:

Host address: 162.55.237.41. port: 5432

Username: your user name password: your password

4- Follow the path “Databases -> enron -> schemas. You will find three schemas:

a- **email**: this is where the data and tables of the enron data set reside. You have read-only access to those tables.

b- **public**: this is a public schema where everyone can write, read and alter. Please always keep this schema empty by dropping your created tables as soon as you leave the site.

c- **private**: you don't have access to this schema.

You can try some simple queries like:

— `SELECT * FROM email.department;`

— `SELECT * FROM email.emails;`

Exercises of relational database

Exercise1 TASK 1 Use Case: Equi Join

1.1 List of people with their department Max Time: 0,5h

For each person you want to know in which department she or he works. Therefore, you have to make an output that contains a person's first name and last name and the name of the department she or he is working at.

1.2 Number of emails sent out per department Max Time: 0,5h

For each department: Find out how many emails in total were sent out from employees working there. The output per department shall contain the corresponding number of emails.

1.3 Number of emails received per department Max Time: 0,5h

For each department: Find out how many emails in total were sent to employees working there (hint: carbon copies included). The output shall have the same structure as the output of Task 1.2.

Exercise1 TASK 2 Use Case: Theta Join

2.1 Correlation between salary and number of emails Max Time: 2h

Do people that earn more than the average salary in their department write more emails than those who don't?

Query for people that earn more than the average salary at their department and find out whether they write more emails than the other employees that earn less than the average salary at their department (equal is not considered). Check that for each department. First compute the result for the average salary (avg. S.) per department that contains the brief-name of all the departments and the average salary for that department. Then produce the output of all the people that earn more than the avg. Salary and accordingly produce the output for all the people who earn less than the avg. Salary.

Produce a query result per department that contains the number of emails written by the people earning more and the people earning less than the average.

Exercise1 TASK 3 Use Case: Schema Evolution

3.1 Add information for entity set Max Time: 0,1h

Hint: First create a copy of email table (in the “**public**” schema) and name it with “email_yourname” (e.g. email_elmamooz) and execute the queries of task 3 (3.1, 3.2) on this new table. Add/delete data and attributes just in your own copy of email table.

You have to introduce a new element (attribute) to the email's entity set (to your own copy of email table you created in task 3.1). Find the general syntax to do that.

3.2 Add information for entity set with default value Max Time: 0,4h

Now, use the syntax from *task 3.2* and add a new element “priority” to the email's entity set with a default value of 1 for each entry. Then take a single entry of your choice (with a certain id) and set its priority to a value of 3.

Remember : Drop the table you created in task 3.1.

Exercise1 TASK 4 Use Case: Missing values

4.1 Find missing values Max Time: 0,5h

Find missing values for each attribute of the e-mails. Which attribute has the most missing values?

Exercise1 TASK 5 Use Case: Range queries

5.1 Emails between two dates Max Time: 0,2h

Select all emails that have been written between the 01.09.2001 and the 31.10.2001. First, find out which date and time format is used in email!

5.2 Emails between two dates for Larry John May Max Time: 1h

Larry May is an employee of Enron. Find all emails he received between the 01.09.2001 and the 31.10.2001.

Exercise1 TASK 6 Use Case: Network analysis

Network analysis can be done to investigate social structures. In the related field of social network analysis a network is characterized by *nodes* (individual actors, people, or things within the network) and the *edges* (relationships or interactions) that connect them. To find out, how far from each other two nodes of the network are, we can count hops. To calculate the number of hops, we calculate $m+1$, where m is the number of intermediate nodes between the two nodes we're looking at. In Enron example, each employee can be seen as a node in the network and edges are represented by the emails sent from one to another employee or by the "knows" relationship between two employees. A relational Database provides us only records; therefore solutions to network analysis must be derived somehow different.

Hint: you can use a UDF (User defined function) to solve the following two tasks.

Remember: Delete any created functions before you leave the site.

6.1 Network size by e-mail Max Time: 2h

If the network nodes (the persons) are fully connected, how many hops are needed to reach everyone in Enron from Larry May by email? Consider the "from" and "to" fields to compute the amount of hops that is needed to reach everyone in Enron.

6.2 Network size by "knows" relation Max Time: 0,5h

How many hops are needed to reach everyone from Larry May by their "*knows*" relationship (similar to task 6.1)?

6.3 Two-hop email network Max Time: 1h

Which people are in the 2-hop email network? Again, consider the "knows" relationship, but only for people that are reachable with two hops.

6.4 Count outgoing edges Max Time: 1h

Find out who sent emails to exact 7 TO-recipients. The output shall contain the name(s) of the sender(s).