

Advanced Data Management - Assignment N°03

Assignment sheet N°03 contains Neo4j exercises. Deadline for uploading the assignment is 04.12.2023. For each task, there is a maximum time. Please do not work longer than that 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 contain 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?

You should have basic knowledge in graphs and Cypher language before solving the tasks. Thus, the maximum task's time does not count the time that you may spend on learning the graph and Cypher concepts.

Note: Because of the fact that all the students use the same data that is already available on the server, please don't change the structure and the data.

Graph Database

DATA IN GRAPH DATABASE

In graph databases data is structured to a set of nodes and edges between nodes. The small part of Enron database is shown in figure 1. In Enron example, each employee, department, and email is a node in the network and edges are represented by "EMAIL_FROM", "EMAIL_TO", "KNOWS", and "WORKS_IN". There are some emails which sent to or received from the people outside the Enron company.

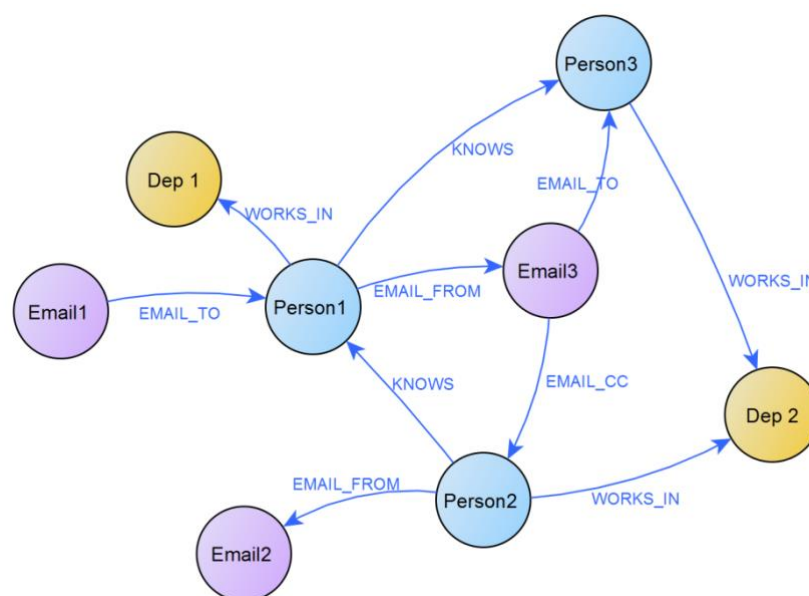


Figure1. nodes and links in graph database

SYSTEM PREPARATION

For graph database we use Neo4j server. To connect to the server please write this address to your browser: <http://162.55.237.41:7474/browser/> . Before getting to the data, please make sure to use the information provided in the screenshot. In the Host, please enter: neo4j://162.55.237.41:7687, username: neo4j and password: **Modyt6B*vsV**. You can see the node labels and relationship types in the dataset.

Connect to Neo4j
Database access might require an authenticated connection

Connect URL
neo4j:// 162.55.237.41:7687

Database - leave empty for default

Authentication type
Username / Password

Username
neo4j

Password
.....

Connect

Then you connect to the neo4j database. On the left panel you could see the database information

Database Information

Use database
neo4j

Node labels
(5,161) Department Email
Person

Relationship types
(32,882) EMAIL_FROM
EMAIL_TO KNOWS WORKS_IN

Property keys
deplD depld depname
emailID emailld email_address

neo4j\$

\$:play start

Getting started with Neo4j Browser
Neo4j Browser user interface guide

Get started

Sign t

Try a simple query like:

```
MATCH (n:Person) RETURN n LIMIT 50  
MATCH (n:Email) RETURN n LIMIT 50  
MATCH (n:Department) RETURN n
```

Exercises of Graph database

Exercise1 TASK 1 Use Case: Equi Join

1.1 List of people with their department Max Time: 1h

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: Missing values

2.1 Find missing values Max Time: 1h

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

Exercise1 TASK 3 Use Case: Range queries

3.1 Emails between two dates Max Time: 0,5h

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!

3.2 Emails between two dates for Larry John May Max Time: 0,2h

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

Exercise1 TASK 4 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.

4.1 Network size by e-mail Max Time: 1h

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 “EMAIL_FROM” and “EMAIL_TO” links to compute the amount of hops that is needed to reach everyone in Enron.

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

How many hops are needed to reach everyone by their “KNOWS” relationship **starting from Larry May** (similar to task 5.1)?

4.3 2-hop email network Max Time: 0,5h

Which people are in the 2-hop email network of **Larry May**? Again, consider the “KNOWS” relationship, but only for people that are reachable with two hops.

4.4 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).