Professor: Ana Escobar Llamazares

NEO4J IN-CLASS EXERCISES

DATASET

Node person. Properties:
- Person ID (person_id)
- Person Name (person_name)
- Person Health Status (health_status: infected/healthy)
- Timestamp representing the moment a person is notified about the PCR test result, whether they are infected or not (test_result_time)
- Latitude: latitude coordinate of the person's residence (residence_latitude)
- Longitude: longitude coordinate of the person's residence (residence_longitude)

Node Location. Properties:
- Location ID (location_id)
- Location Name (location_name)
- Location Type (location_type)

Node visit (from people to different places (hospital, school, bar, restaurant, etc.). Properties:
- Visit Identifier (visit_id)
- Location Identifier (location_id)
- Person Identifier (person_id)
- Start time of the visit (visit_start)
- End time of the visit (visit_end)

There are three types of relationships between nodes:
- (Person) [VISITS_LOCATION] (Location)
- (Person) [MAKES_VISIT] (Visit) [TO_ESTABLISHMENT] (Location)

Load the datasets (csv files) in Neo4j and answer the following questions:

QUESTIONS

1. Identify the number of infected and healthy people in the sample of 40 individuals. Return the result in table or text format.

2. Find healthy individuals who have been in contact with a person who tested positive. Due to being in the early stages of studying the disease, it is assumed that COVID can infect healthy individuals who have been in the same place as an infected person, even on different days. Identify all healthy individuals who have been in the same place (regardless of date or time) where an infected person has been. Return the result in table or text format.

3. Show a graph with healthy individuals who have coincided with an infected person, specifically with "Maxwell Ramirez." Display this person's node along with all the places visited and nodes of healthy individuals who have also visited that place afterward. Based on the graph results, comment on six individuals who have been in a location where Marcelino has been afterward; they have less risk than the rest of being infected.

4. Build the same query as above but show the result as a table and not as a graph (table or text format) displaying the fields:
   - Virus_spreader
   - Start_virus_spreading
   - Establishment
   - Person_at_risk
   - Start_visit_person_at_risk

5. Build a table (text and table format) that identifies for each infected person (first column), the healthy individuals with whom they have coincided in an establishment at the same time. Build as the second column an array of JSON elements called "Contacts" with keys:
   - Person_in_contact
   - Establishment
   - Start_date_overlap
   - End_date_overlap

6. Once the previous query is obtained, by adding three statements to it, obtain a table that has the name of the infected person and another column with the number of healthy individuals they have had contact with, obtaining that number from the elements of the JSON array. Order the results by "Number_of_healthy_contacts" in descending order.

7. Find those individuals (if any) who visited an establishment even after knowing they had tested positive.

8. Now that all healthy individuals who coincided in any establishment with an infected person have been obtained, it is desired to find out the exact time (duration) that each healthy person coincided with person p1. Express the duration in hours and rounded to four decimals. Return the result in table or text format.

9. A person has been in two different places with infected individuals; in one, they were in contact for an hour and a half, and in the other, for two hours. The total exposure of that person will have been three and a half hours. The duration of each contact between a healthy person and an infected person will be the result obtained in the previous question. Therefore, you can use the previous query as a base and add something more to get the expected result. If a healthy person coincided with two infected individuals on the same day in the same establishment, the time spent in contact with each infected person will also be added, understanding that being surrounded by more infected individuals implies a higher risk of contracting the disease. Only the top five healthy individuals with the most exposure time will be shown in the table (table or text format). For those five individuals, an immediate call will be made to start quarantine. The total time will be displayed in hours rounded to four decimals (for example: 9.4972 hours, which will be nine hours and 30 minutes).

10. It is intended to try to reduce attendance and implement even more precautionary measures in those establishments where infected individuals have spent more time. Please return a table that contains each establishment, visited by at least one infected person, the total visits of infected individuals in each establishment, the total visits in each establishment, the percentage of visits by infected individuals relative to the total visits to each establishment, and the city to which the establishment belongs. Express the percentage rounded to two decimals. Comment on the two establishments with the highest and the two with the lowest percentage of visits by infected individuals relative to the total for each establishment.