DB - Study Point Assignment 2

Robert Pallesen

March 2023

Provide in writing, each student individually, answers to the following questions

1. What are the advantages and disadvantages of using graph databases and which are the best and worse scenarios for it?

   Graph databases is a type of database, that uses nodes and relations, to present visually and store data. An advantage of using graph databases includes it's ability to handle complex relationships/data and flexibility. Also, so far they seem way better to scale, than e.g. tabular databases. A disadvantage I've come to find, is that queries and such are complex than tabular databases (which I've been only been working with prior to this course).

2. How would you code in SQL the Cypher statements you developed for your graph algorithms-based query, if the same data was stored in a relational database?

   I will be using the a degree centrality algorithm in this example:

```
CALL gds.degree.stream('GoT')
YIELD nodeId, score
WITH gds.util.asNode(nodeId) AS node, score
RETURN node.id AS id, node.name AS name, score
ORDER BY score DESC
LIMIT 5
```

   For the rewrite for relational databases, it will look something like the following:

```
SELECT c.id, c.name, degree.score
FROM nodes c
INNER JOIN (
    SELECT source_id, COUNT(*) AS score
    FROM edges e
    INNER JOIN nodes n ON e.target_id = n.id AND n.name = 'GoT'
    GROUP BY source_id
) AS degree ON c.id = degree.source_id
ORDER BY degree.score DESC
LIMIT 5;
```

   This may be more pseudo code than actual code, but it is my best shot. We'll have to assume, that the graph will be stored in 2 different tables - a *node* table and an *edges* table.

   The *node* contains the nodeID and name, while the *edges* contains columns for the relationship (source node), which is nodeID and targetNodeId.

   Now, what I would like the query to do is, to select all nodes, that have a name that is exists in the GoT (Game of Thrones) graph.

   Then it finds the number of edges that comes from each of the nodes, by joining the *edges* table to itself, and then grouping by the nodeID.

   It then joins the *node* table to this result, and selects the top 5 nodes (highest to lowest).

3. How does the DBMS you work with organizes the data storage and the execution of the queries?

   DBMS stands for Database Management System, which is a system that's used for accessing, maintaining and administrating a database.

   The DBMS we're working with stores and organizes the data in the database, so it's quick to access and easy to handle.

   Execution of queries means the process, of executing requests on the database, which is commands to find and manipulate data. The DBMS will organize these requests and find the relevant data according to to the requests requirements and return the result to the user.

4. Which methods for scaling and clustering of databases you are familiar with so far?

   Personally I am not sure that I know any methods of scaling databases. Besides that, when we're speaking of clustering of databases, I am only familiar with what we've been introduced to (on the 28th of march), which is called "causal clustering".