Big Data Management and Analytics

Assignment 3

TU Clausthal, Institut für Informatik

!!! Due date: 27 May 2020, 1pm !!!

In this assignment, you will give you first experience with the NoSQL database management system *Neo4j*.

Here are useful links for learning more about Neo4j:

A brief introduction to Neo4j is given here:
`https://neo4j.com/developer/get-started/`

The query language Cypher to query Neo4j databases is introduced here:
`https://neo4j.com/developer/cypher-basics-i/`

An overview of Cypher queries:
`https://neo4j.com/docs/cypher-refcard/current/`

Read the following notes in detail for your submission of solutions.

- You can work in teams of up to 2.
- You can send us your solutions via the GATE-System `https://si.in.tu-clausthal.de/`
- Solutions sent after the deadline will automatically be marked with 0 points and treated as not participated in the respective assignment!
- Copying solutions from other students will be treated as cheating and will lead to exclusion from the course!

!! Submit your solutions via the GATE-System !!

There is guide in Moodle on how to register and how to upload your solutions.

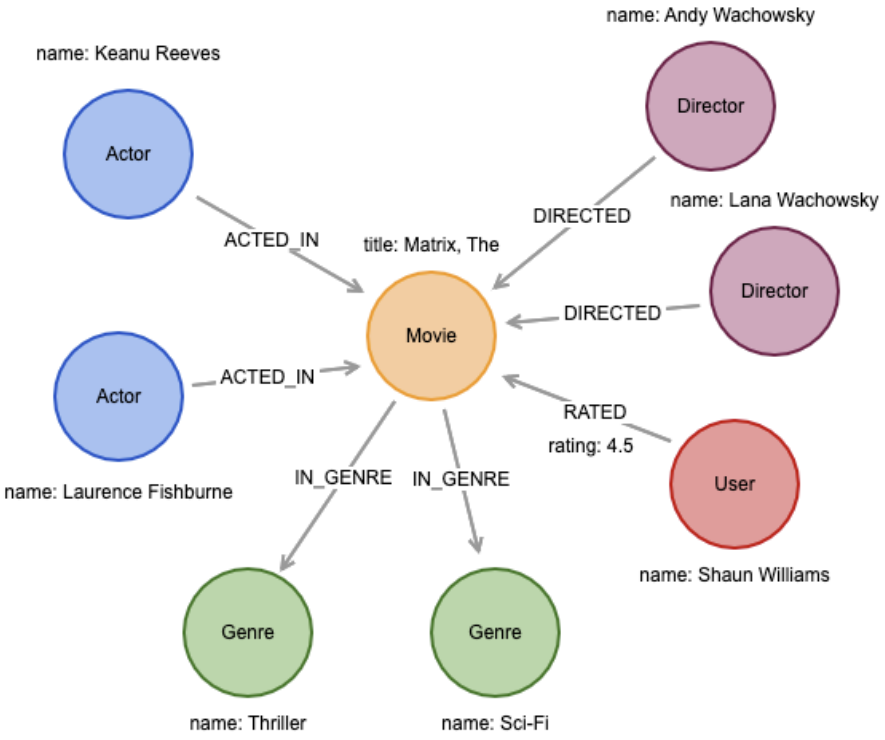
Task 1 **5 Marks**

In this task, we will get familiar with the NoSQL graph database Neo4j. You can read about Neo4j under this URL: `https://neo4j.com/developer/graph-db-vs-nosql/`.

(a) Explain in your own words what the major similarities and differences between graph databases and other NoSQL databases are.

Neo4j uses the query language Cypher. You can learn more about Cypher here: `https://neo4j.com/developer/cypher-basics-i/`

Consider the following outline of a Neo4j database:



(b) Write a Cypher query to retrieve the actor whose name is *Laurence Fishburne*.

(c) Write a Cypher query to find all directors whose name contains *Wachowsky*.

(d) Write a single Cypher query to retrieve the actors with name *Laurence Fishburne* and *Keanu Reeves* at the same time.

(e) Write a Cypher query to retrieve 100 movies. Afterwards alter your query, to sort your results descending by their title.

Note: If you like you can install and run Neo4j yourself on your own computer. Then you can execute your queries directly in Neo4j. You will find guides in Moodle on how to setup Neo4j and how to import the Movies database.

Task 2 **5 Marks**

In this task we will get familiar with the MapReduce programming model. For this we will use Python again. Recall the previous assignments, in which Python was introduced.

Suppose you are given the matrices $A \in \mathbb{R}^{i \times j}$ and $B \in \mathbb{R}^{j \times k}$.

Let there be $A := \begin{pmatrix} 2 & 3 & 5 \\ 7 & 11 & 13 \end{pmatrix}$ and $B := \begin{pmatrix} 17 & 19 \\ 23 & 29 \\ 31 & 37 \end{pmatrix}$

Through multiplying of both matrices we get a third (quadratic) matrix C with 2 rows and 2 columns.

(a) State how a matrix multiplication is done in general. Multiply both matrices with each other and write down the resulting matrix C. Could you achieve other solutions for C with the matrices A and B?

(b) Explain how the matrices can be multiplied by MapReduce. Begin with rewriting the matrices as tuples (i, j, a_{ij}) and (j, k, b_{jk}).
Explain all steps in general and give at least one numerical example for each step. Verify the results with your solution in (a).

(c) Implement a function *matrixMultiplication()* with python, which takes 2 matrices as arguments. The function returns the product of both matrices. Within the function use the type of functions, which were introduced in Assignment 2 Task 1 !! Again, verify your result with the solutions from (a) and (b).

+++++