Université Paris-Saclay
Master 2 Data and Knowledge

Professor: Ioana Manolescu
Teaching assistant: Paweł Guzewicz[1]

Big Data Architectures Lab 3

October 14, 2019

Submission deadline: Sunday, October 27, 2019 at 23:59:59.

1 Lab organization[2]

1.1 Neo4j setup

Neo4j is supported under several Linux distributions, macOS and Windows platforms.

1. Download and install Neo4J Community Edition following instructions at `https://neo4j.com/download-center/#community`.
2. Launch Neo4j and check if it works by visiting `http://localhost:7474`.

1.2 Neo4j tutorials

Get familiar with the Neo4j data model and Cypher query language by going through the tutorials at `http://localhost:7474`. To start them use the commands `:play concepts`, `:play cypher`, `:play movie-graph`, and `:play stackoverflow`, respectively.

1.3 Assignment and questions

Please read the assignment description in the next section. If you have any questions about the lab or on the course material, don't hesitate to ask them during the lab session or via email (don't wait until the last moment).

2 Assignment

Download this archive file: `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching/2019_2020_Big_Data_Architectures/lab_3_dataset.zip`[3]. It contains two CSV files (where CSV stands for comma-separated values file format). Using this dataset, you need to solve the following tasks.

Task 1: data import

1. Import the files `boston-crime-incident-reports-10k.csv` and `boston-offense-codes-lookup.csv` into Neo4j, pay attention to their headers; see this page for details: `https://neo4j.com/developer/guide-import-csv/`. Before importing, make sure your `conf/neo4j.conf` file allows importing files from the directory where you store them[4]; recommended solution in case of problems: comment with a `#` the line `dbms.directories.import=import` in the configuration file.
2. Upon loading, add Cypher commands to model of your data as a property graph; see examples here: `https://neo4j.com/developer/guide-importing-data-and-etl/`.

[1] pawel.guzewicz@lix.polytechnique.fr; `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching`

[2] All lab materials based on previous course editions courtesy of Ioana Manolescu and Silviu Maniu (with some changes)

[3] The lab dataset is a fragment of a publicly available dataset that can be found at `https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/resource/12cb3883-56f5-47de-afa5-3b1cf61b257b`

[4] More information in the replies to this question: `https://stackoverflow.com/questions/28398778/cypher-neo4j-couldnt-load-the-external-resource/42094102`

Task 2: data querying and analysis

1. Find the number of incidents by `Drug Violation` offense group.
2. Find the names of offense codes for incidents of `Investigate Person` offense group.
3. PROFILE and EXPLAIN two above queries.
4. Add indexes on your graph. What are the changes in the plans?
5. Explore a graph, and write a query of your choice using ORDER BY clause.

Task 3: results visualization
Using built-in Neo4j functionalities and Cypher commands (see `https://neo4j.com/developer/graph-visualization/`) visualize the results of your last query in Task 2 and export it to a PNG file. What are your conclusions about readability?

2.1 Report

Write a report on your solutions for the tasks. It should include the following elements.

1. Answers: for each solution write down **the complete list of commands or queries** you used, as well as **all the results/output**.
2. Explanation of the commands or queries, i.e., what is a logical operation that you perform.

The report should not include the dataset nor any output that comprises the whole dataset.
You can separate the output of commands and queries, and/or screenshots of terminal(s) into some files and then refer to those files in the report. (Please include any external files in the submission archive.)

2.2 Submission guidelines

Please follow submission rules and guidelines: `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching/2019_2020_Big_Data_Architectures/submission_rules_and_guidelines.pdf`.

Moreover, I encourage you to read my advice on lab sessions and submissions: `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching/2019_2020_Big_Data_Architectures/advice_on_lab_sessions_and_submissions.pdf`