Big Graph Processing Systems 2025
Warm-up: Schema Discovery

General Information

The goal of this component is to obtain a schema for the given property graph. It consists of seven tasks.

Important

- Do not use `CALL ... YIELD` clauses to call procedures. They are Neo4j specific and are not part of openCypher/GQL.
- Make sure that all queries are free of syntax errors by running them, even if you make only small changes.
- A good explanation helps someone familiar with basic knowledge on openCypher to understand the (key) idea(s) of a query. "Reading" or repeating the query in natural language is, for example, not very helpful.
- Solve the tasks yourself, on your own.

Task 1

a) Install Neo4j.
- Make sure you are using **version 5.26 or newer**.
- Make sure you have enough memory, at least 20GB.
- A simple option is to install Neo4j Desktop by following the official instructions: `https://neo4j.com/docs/desktop-manual/current/installation/`.

b) Download and import the database.
- Download the database dump file from `https://partage.liris.cnrs.fr/index.php/s/LoEtp24fk38P6n5`.
- Import the dump file by following the official instructions: `https://neo4j.com/docs/desktop-manual/current/operations/create-from-dump/`.

Task 2

a) Write a Cypher query that returns the set of all distinct node labels.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

Task 3

a) Write an Cypher query that returns the set of all distinct relationship types.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

Task 4

    a) Write a Cypher query that returns each distinct node label together with its corresponding set of properties.

    b) How many answers does your query return?

    c) Explain concisely how you came up with your query.

Task 5

    a) Write a Cypher query that returns each distinct relationship type together with its corresponding set of properties.

    b) How many answers does your query return?

    c) Explain concisely how you came up with your query.

Task 6

Write at least three additional Cypher queries that are needed for you to understand the underlying structure of the dataset.

Hint: What relationship types have outgoing edges of nodes with a specific label? Given a node n with label L, what labels can nodes reachable from n have?

Task 7

Draw the schema of the pangenomic property graph, and explain all parts and properties of the schema, which are not illustrated in the picture.