NoSQL Graph & Distributed Data Processing

Project Title: Hackaton

Group Name: Alone

Boulay Mathias

DO5 2025_Polytech

March 28, 2025

About the Class Activities or Exercises

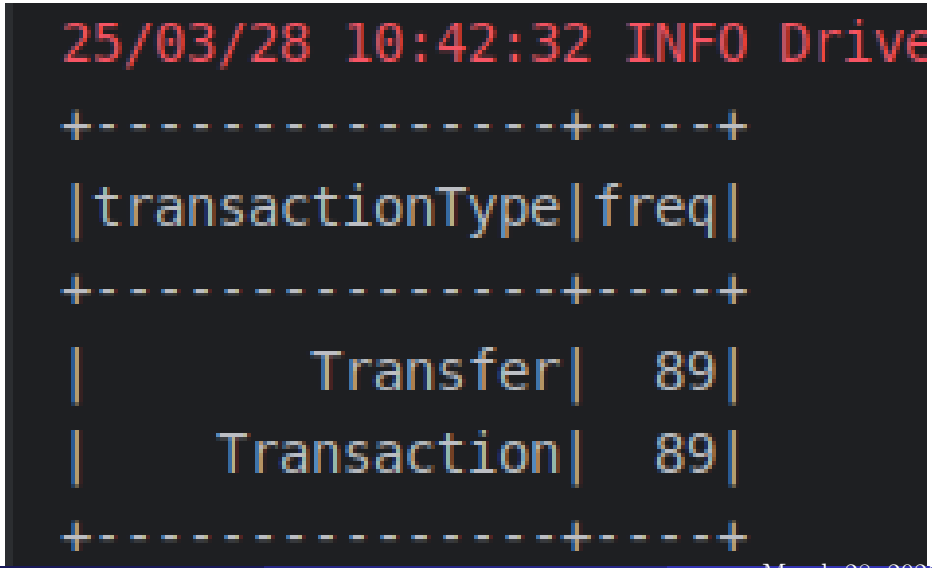
- **Which Activities were interesting**
  - I liked playing around with neo4j, and understanding the graph style databases
- **Which Activities were difficult and not useful in your understanding**
  - Small scale apache spark fails to show how much it can handle
  - The mongo db connector activity doesn't seem useful.
- **What would you have liked to see ?**
  - Large optimization guidelines for neo4j
  - neo4j isn't cheap, what about the rest of the ecosystem for graph databases ?
  - Efficient ways to split the data for spark ?

About the Project

- **How did you solve the problems/task:**
  - transaction types
  - criminal groups over 9
  - criminal groups over 10
- **Methodology or Approach:**
  - Everything was running on a k3d instance, aside from neo4j
  - For development speed, running Scala programs locally can be done with port forward and etc/hosts tweaks

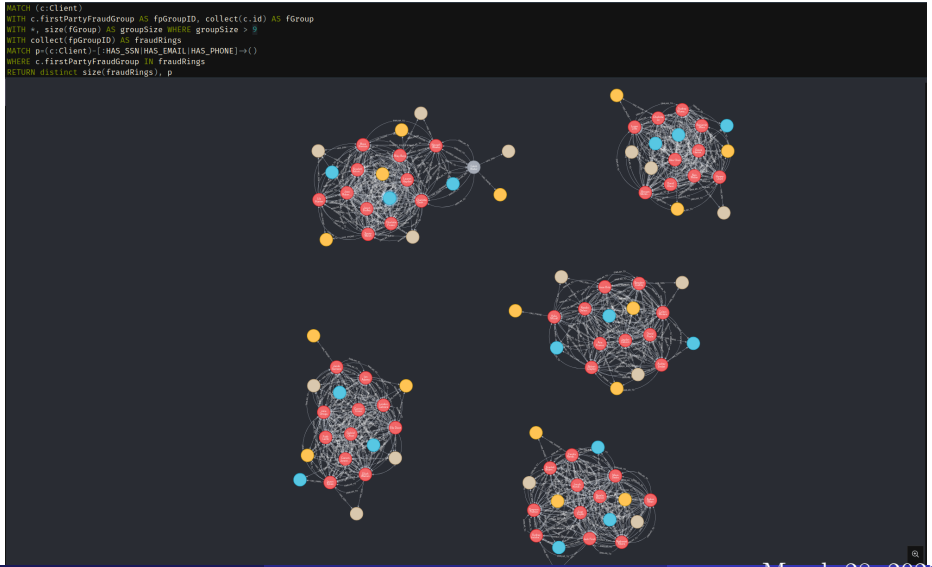
About the dataset

Transaction types



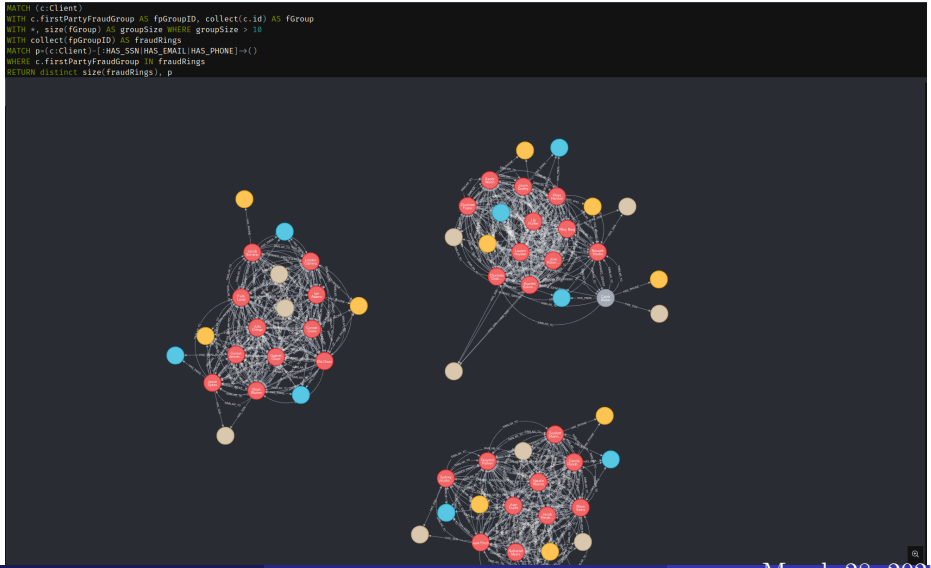
Groups (strictly) over 9

```
MATCH (c:Client)
WITH c.firstPartyFraudGroup AS fpGroupID, collect(c.id) AS fGroup
WITH *, size(fGroup) AS groupSize WHERE groupSize > 9
WITH collect(fpGroupID) AS fraudRings
MATCH p=(c:Client)-[:HAS_SSN|HAS_EMAIL|HAS_PHONE]->()
WHERE c.firstPartyFraudGroup IN fraudRings
RETURN distinct size(fraudRings), p
```



Groups (strictly) over 10

```
MATCH (c:Client)
WITH c.firstPartyFraudGroup AS fpGroupID, collect(c.id) AS fGroup
WITH *, size(fGroup) AS groupSize WHERE groupSize > 10
WITH collect(fpGroupID) AS fraudRings
MATCH p=(c:Client)-[:HAS_SSN|HAS_EMAIL|HAS_PHONE]->()
WHERE c.firstPartyFraudGroup IN fraudRings
RETURN distinct size(fraudRings), p
```



Challenges and Learning

- **Challenges Faced and Their Resolutions:**
  - It's been a while since I've done Scala (refreshed my memory after)
  - First time seeing cypher queries (had to learn it through examples)
- **Learning Gained:**
  - Learned how to apply algorithms to graph databases
  - basics of Apache spark, graph databases
- **Future Improvements:**
  - Learn to optimize graph databases