Advanced Databases UFCFU3-15-3: Task 3

Arthur Milner (21035478)

Graph Databases

The case study could benefit from graph databases as the data is very interconnected, with neighbours, addresses, and favourites, being shared amongst multiple people.

A graph database allows efficient performance when traversing connections, such as finding shared favourites between persons. This is especially useful if the data is required to provide recommendations, for example, recommending a favourite book to a Person. Graph databases also maintain the flexibility of alternative NoSQL databases as they are additive **(Decipher Zone, 2022)**.

A drawback of graph databases is the lack of a universally accepted query language. Existing languages, such as GOQL and SPARQL, lack the generalization of SQL for relational databases **(Sakr and Al-Naymat, 2010)**. In the case study, it would be important to select the most appropriate query language, which can require a deep understanding of graph algorithms/data structures. Another consideration is maintaining consistency in the data, graph databases lack data validation features. Consequently, validation logic is placed elsewhere, typically at the application level, to prevent situations such as a person having a drink as a neighbour. This would take considerably more effort than defining a validation schema or adding SQL constraints **(Medium, 2016)**.

Fig. 1 presents a draft of a possible graph database modelled after the case study.

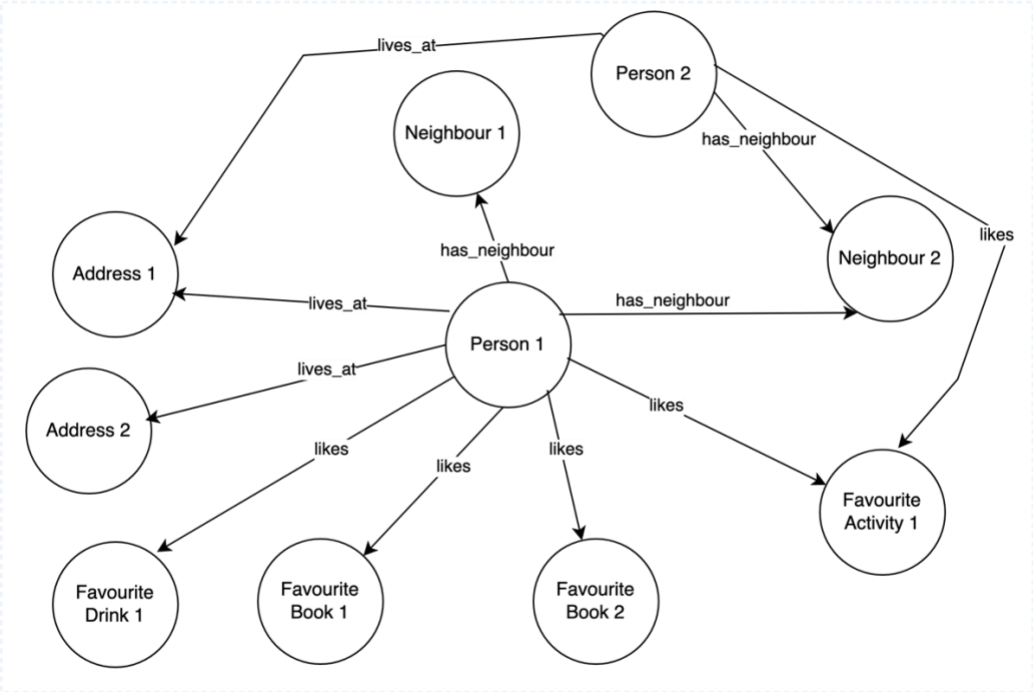

Figure 1 Draft of a graph database implementation of the case study

Personal Data Security & Legislation

The case study requires storage of names, birth dates, addresses, and emails, all of which are identifiable personal data. Consequently, concerns of data storage and legislation compliance can be said to apply.

Personal Data Security

Securing personal data is essential to protect data subjects and to align with strict legislation.

Hashing email addresses can add encryption to the data, but requires significant computation if indexing is unoptimized, and reversing the hash introduces complexity. Pseudonymization, by replacing names and addresses with IDs and storing the mapping separately, reduces risk of identification, while maintaining referential integrity **(Imperva, 2024)**, though query simplicity and performance are sacrificed as additional joins are required. Anti-virus software on the database host can protect against attacks but adds computational overhead. However, advanced attacks, such as the EasyJet data breach **(Wakefield, 2020)**, often bypass the software, causing a false sense of security.

A combination of these methods enhances security, but it would be important to manage the additional overhead and complexity.

Legislation

As identifiable personal data will be stored, **General Data Protection Regulation, 2016** compliance is essential. Key factors of GDPR and their relevance to the case study are:

- **Secure Storage:** Previously discussed methods ensure both secure storage and minimise the risk should the data be compromised.
- **Consent:** Consent would be required to store data on both a person and their neighbours. This could get difficult to manage, and should a neighbour withdraw consent, data on a person's neighbours would become inaccurate.
- **Specificity:** Each data field would require justification for its storage, for example, email is stored for communication. This greatly affects the flexibility of the system, new data fields would require approval and additional documentation.
- **Retention:** The database would require removal of redundant records, for example neighbours not attached to a person should be removed from the database, reducing data redundancy whilst adding the overhead to track and manage data.

Big Data and Data Lakes

Data Lakes

Using a data lake for storage in the case study can be justified in the large volume of records that are expected to be stored, alongside being a cost-effective and scalable solution for self-service access for users/analysts **(Microsoft, 2024)**. A major consideration when using a data lake is the risk of poor data governance resulting in a data swamp. For example, if users are mainly interested in a person's favourites and not their neighbours, it is important to ensure data can be flexibly filtered. In the case study example metadata to prevent this could be popularity on favourites or a confidence score regarding an address, or even an addresses neighbours, being valid.

Big Data

Much like graph databases help provide data insights, such as recommendations, big data can discover advanced patterns and is commonly paired with AI for further depth of analysis. If using AI it would be imperative to consider any potential bias to keep insights valuable and ethical, particularly since the case study contains personal data from which models might discriminate.

Analytics on the case study should be effective as it matches the "Five V's of Big Data Analytics" **(IBM, 2024)** in the following ways:

- **Volume:** There will be many records kept.
- **Velocity:** New records and updates to records, such as favourites, should be frequent.
- **Variety:** Both unstructured data, such as favourites, and structured, such as emails and zip codes.
- **Veracity:** Stored data would be accurate and validated.
- **Value:** Examples of valuable insights could include:
    - Informing businesses of locations they are likely to succeed based on the interests of the persons living there.
    - Detecting inaccurate addresses based on a person's neighbours.

References

- Decipher Zone (2022). ***When to Use Graph Databases: A Comprehensive Guide***. Available from: https://www.decipherzone.com/blog-detail/when-to-use-graph-database [Accessed 7 March 2025]
- General Data Protection Regulation (2016). Official Journal. L119, 4 May. EUR-Lex. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 [Accessed March 5 2025]
- IBM (2024) ***What is big data analytics?***. Available from: https://www.ibm.com/think/topics/big-data-analytics [Accessed 20 March 2025]
- Imperva (2024). ***Pseudonymization***. Available from: https://www.imperva.com/learn/data-security/pseudonymization/ [Accessed 13 March 2025]
- Medium (2016). ***The Challenges of Working with a Graph Database***. Available from: https://medium.com/vaticle/the-challenges-of-working-with-a-graph-database-2a5f9a7c903b [Accessed 12 March 2025]
- Microsoft (2024) ***What is a Data Lake?***. Available from: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-data-lake [Accessed March 19 2025]
- Sakr, S. and Al-Naymat, G. (2010). Graph Indexing and Querying: A Review. ***International Journal of Web Information Systems*** [online]. 6(2), pp. 101-120. [Accessed 10 March 2025]
- Wakefield, J. (2020) EasyJet Admits Data of Nine Million Hacked. ***BBC*** [online]. 19 May. Available from: https://www.bbc.co.uk/news/technology-52722626 [Accessed 15 March 2025]