1 Synthetic dataset creation

Given the confidential and private nature of bank data, it was not possible to find any real bank datasets to perform our experiments on. In this regard, a synthetic property graph bank dataset was built based on the *Wisabi Bank Dataset*[1]. It is a fictional banking dataset that was designed and architected by Obinna Iheanachor, and that was made publicly available in the Kaggle platform.

In particular it contains 10 CSV tables. Five of them are of transaction records of five different states of Nigeria (Federal Capital Territory, Lagos, Kano, Enugu and Rivers State) that refers to transactions of cardholders in ATMs. In particular they contain 2143838 transactions records done during the year 2022, of which 350251 are in Enugu, 159652 in Federal Capital Territory, 458764 in Kano, 755073 in Lagos and 420098 in Rivers.

Then, the rest of the tables are: a customers table ('customers_lookup') where the data of 8819 different cardholders is gathered, an ATM table ('atm_location lookup') with information of each of the 50 different locations of the ATMs, and then three remaining tables as complement of the previous ones ('calendar lookup', 'hour lookup' and 'transaction_type lookup') (tables summary).

Based on the aforementioned dataset we created our own synthetic dataset. For simplicity and to do it in a more stepwise manner, we are going to first create all the CSV data tables for the nodes and for the relations in the corresponding format and then we will populate the Neo4j GDB with those. The description of the creation of the CSV data tables as well as the specific details are described in what follows:

Bank

Since a unique bank instance is considered, the values of the properties of the bank node are manually assigned, leaving them completely customisable. Bank node type properties consist on the bank *name*, its identifier *code* and the location of the bank headquarters, expressed in terms of *latitude* and *longitude* coordinates, as seen in Table 1. For the bank, we will generate `n` ATM and `m` Card entities. Note that apart from the generation of the ATM and Card node types we will also need to generate the relationships between the ATM and Bank entities (`belongs_to` and `external`) and the Card and Bank entities (`issued_by`).

[1] Wisabi bank dataset on kaggle

Name	Description and value
`name`	Bank name
`code`	Bank identifier code
`loc_latitude`	Bank headquarters GPS-location latitude
`loc_longitude`	Bank headquarters GPS-location longitude

Table 1: Bank node properties

ATM

Name	Description and value
`ATM_id`	ATM unique identifier
`loc_latitude`	ATM GPS-location latitude
`loc_longitude`	ATM GPS-location longitude
`city`	ATM city location
`country`	ATM country location

Table 2: ATM node properties

The bank operates `n` ATMs, categorized in:

- Internal ATMs: ATMs owned and operated by the bank. They are fully integrated within the bank's network.
- External ATMs: These ATMs, while not owned by the bank, are still accessible for the bank customers to perform transactions.

Both types of ATMs are considered to be of the same type of ATM node. Their difference is modeled as their relation with the bank instance: `belongs_to` for the internal ATMs and `external` for the external ATMs, having:

```
n = n_internal + n_external
```

where `n_internal` is the number of internal ATMs owned by the bank and `n_external` is the number of external ATMs that are accesible to the bank.

The ATM node type properties consist on the ATM unique identifier *ATM_id*, its location, expressed in terms of *latitude* and *longitude* coordinates, and the *city* and *country* in which it is located, as seen in Table 2. Note that the last two properties are somehow redundant, considering that location coordinates are already included. In any case both properties are left since their inclusion provide a more human-understandable way to easily realise about the location of the ATMs.

The generation of `n` ATMs for the bank is done following the geographical distribution of the locations of the ATMs in the *Wisabi Bank Dataset*. On this dataset there are 50 ATMs locations distributed along Nigerian cities. Note that for each of these ATMs locations, there can be more than one ATM. However,

this is not taken into account and only one ATM per location is assumed for the distribution.
⇒ Put a plot of the distribution of the ATM locations
This distribution of the ATMs matches the relevance of the location in terms of its population, since the number of ATM locations is larger in the most populated Nigerian cities (30% of the ATM locations are in the city of Lagos, then the 20% in Kano...). Therefore, for the generation of the location of each of the `n` ATMs, the location/city of an ATM selected uniformly at random from the *Wisabi Bank Dataset* is assigned as *city* and *country*. Then, new random geolocation coordinates inside a bounding box of this city location are set as the *loc_latitude* and *loc_longitude* exact coordinates of the ATM.
Finally, as the ATM unique identifier *ATM_id* it is assigned a different code depending on the ATM internal or external category:

$$ATM\_id = \begin{cases} bank\_code + " - " + i & 0 \leq i < \texttt{n\_internal} \text{ if internal ATM} \\ EXT + " - " + i & 0 \leq i < \texttt{n\_external} \text{ if external ATM} \end{cases}$$

Card

- Explicar las properties con la tabla y de la forma que se hizo descriptiva para ATM y Bank.

- number_id: Unique identifier of the card.
- client_id: Unique identifier of the client.
- expiration: Validity expiration date of the card.
- CVC: Card Verification Code.
- extract_limit: Limit amount of money extraction associated with the card.
- loc_latitude: Client address GPS-location latitude.
- loc_longitude: Client address GPS-location longitude.

Aspects to explain:

- Extended behavior fields: to not only withdrawal.
- Location: Explain the 2 options we have developed and the one used so far.
- Explain how the behavior of the card is generated based on a random selected wisabi customer.
- Extract_limit: explain how and why?

For the Card entity generation there are some different aspects that are worth to mention:

- First, in relation with the card and client identifiers (`number_id`, `client_id`), for the moment we define that each client has 1 card, later this can be modified.
  ⇒ For the moment 1 client : 1 card. So far, for the kind of frauds we are considering this is the easier approach. In a future, if needed, it could be generated the case that 1 client has n cards

- `Expiration` and `CVC` fields: they are not relevant, could be empty fields indeed or for all the Cards the same values. For simplicity and completeness we chose them to be the same values for all the Cards.

  Expiration: set for completeness the same date in all of them but in a far future!

- Behavior: For each Card object to be generated, we assign it a *behavior* based on the transaction behavior of a randomly selected wisabi customer. The behavior is gathered from the transaction history of the customer on the wisabi dataset. This will be particularly useful later for the generation of the synthetic transactions, so that for each of the cards their transactions can be simulated based on this gathered behavior. In particular the customer *behavior* refers to:

  - `extract_limit`: maximum normal amount that a card can extract. `amount_avg` $*$ 5.
  - `amount_avg`: extracted amount average of the customer on a transaction.
  - `amount_std`: extracted amount standard deviation of the customer on a transaction.
  - `withdrawal_day`: average number of transactions per day of the customer. Withdrawals only for the moment!.

  - New added behavior: interesting for the generation of transactions that are not only withdrawals: balance inquiries, deposits, transfers.

  Note that all this fields are additionally added to the Card CSV records. Some additional remarks:

  - For the moment we only consider the *withdrawal* type of transaction in the behavior. However *transfer* and *deposit* could be also considered.

  - This behavior is gathered from one random customer of the wisabi dataset per each of the Cards, so that we have more variability. However, we could also assign the same behavior to all the clients, and this behavior be like a summary of all the wisabi dataset clients behavior. Also the behavior could be assigned drawning it from taylored distributions selected by us, in a more customizable manner.

- Location (`loc_latitude`, `loc_longitude`): Two possible options in this case:
  - 1. Assign a random location of the usual ATM city/location of the random selected wisabi customer. This way, we maintain the geographical distribution of the wisabi customers.
  - 2. Assign a random location of the city/location of a random ATM of the newly generated ATMs objects. (* For the moment)

2 Indexing

Useful for ensuring efficient lookups and obtaining a better performance as the database scales.

→ indexes will be created on those properties of the entities on which the lookups are going to be mostly performed; specifically in our case:

- Bank: `code` ?
- ATM: `ATM_id`
- Card: `number_id`

Why on these ones?

→ Basically the volatile relations / transactions only contain this information, which is the minimal information to define the transaction. This is the only information that the engine recieves from a transaction, and it is the one used to retrieve additional information - the complete information details of the ATM and Card nodes on the complete stable bank database. Therefore these parameters/fields (look for the specific correct word on the PG world) are the ones used to retrieve / query the PG.

By indexing or applying a unique constraint on the node properties, queries related to these entities can be optimized, ensuring efficient lookups and better performance as the database scales.

From Neo4j documentation:

An index is a copy of specified primary data in a Neo4j database, such as nodes, relationships, or properties. The data stored in the index provides an access path to the data in the primary storage and allows users to evaluate query filters more efficiently (and, in some cases, semantically interpret query filters). In short, much like indexes in a book, their function in a Neo4j graph database is to make data retrieval more efficient.

Some references on indexing:

- Search-performance indexes
- The impact of indexes on query performance
- Create, show, and delete indexes

Okay... but before diving deeper...:
To Index or Not to Index?

When Neo4j creates an index, it creates a redundant copy of the data in the database. Therefore using an index will result in more disk space being utilized, plus slower writes to the disk.
Therefore, you need to weigh up these factors when deciding which data/properties to index.
Generally, it's a good idea to create an index when you know there's going to be a lot of data on certain nodes. Also, if you find queries are taking too long to return, adding an index may help.

From another tutorial on indexing in neo4j