Nubank Analytics Engineer Case

Introduction

You are in the next stage of the Nubank Analytics Engineer hiring process and now it is time to work in a hands-on project case.

The idea here is to show your knowledge based on previous experiences in fields that are relevant to this position applied to a very similar context to what happens at our company.

Data Architecture Overview

Here at Nubank, we have three different environments: Production, Data Warehouse and Reporting (shown in Figure 1).

The Production Environment is where Nubank's services live. In this environment, services are responsible for getting customers' information and storing them in their respective databases.

Then, the datasets generated by the services are maintained in the Data Warehouse Environment, where Business Analysts, Analytics Engineers and any other Nubanker create new tables, improve data models and work to turn raw data into easily consumable tables for analysis.

Finally, these tables feed into the Reporting Environment, where it is possible to create all sorts of dashboards and visualizations for monitoring, decision-making, processes and so on.

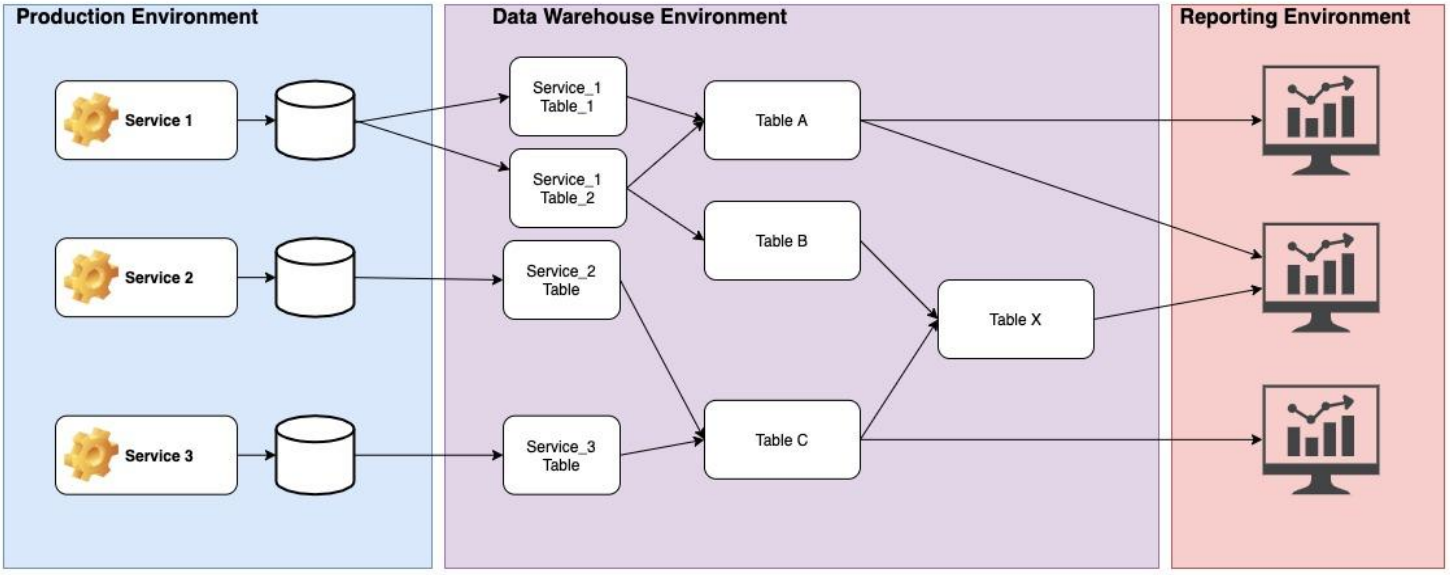

Figure 1 - Nubank's environments: Production, Data Warehouse and Reporting.

A slice of the table structure from the Data Warehouse Environment is depicted in Figure 2 below.

Apart from time (d_time, d_year, d_month, d_week, d_weekday), location (city, state, country), accounts, and customers tables, three tables store the financial movements of the accounts:

- **transfer_ins:** non PIX transfers made to an account (money arriving)
- **transfer_outs:** non PIX transfers made from an account (money leaving)
- **pix_movements:** transfers that are either received by or sent from an account using PIX

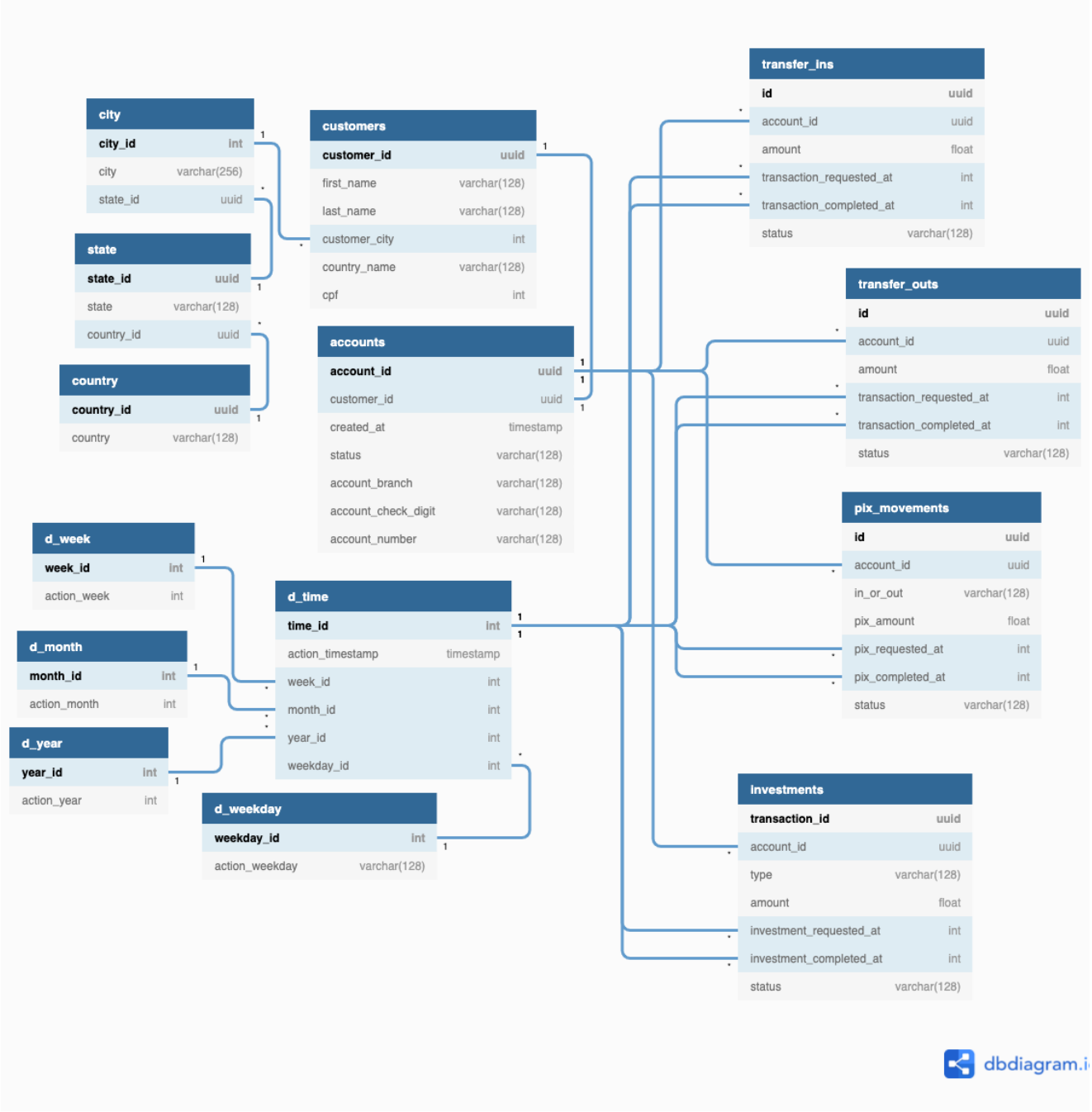

*Figure 2 - A slice of the table structure from the Data Warehouse Environment (diagram/table_diagram.png). You can check the code used to generate these diagrams using **dbdiagram.io** on file (diagram/table_diagram.txt).*

Business Context

To solve this case you need to be familiar with the concept of "Account Monthly Balance". Account Monthly Balance is the amount of money a customer had in their account at the end of a given month. This information can be calculated by adding all the movements inwards and subtracting all the movements outwards from the account balance of the previous month. Remember, there are two types of movements which you should consider in your calculations: PIX (the newest Brazilian transfer method which you can read more in the glossary) and non-PIX. You can see an example below:

Month	Customer	Total Transfer In	Total Transfer Out	Account Monthly Balance
1	A	1000	200	800
1	B	2000	0	2000
1	C	500	100	400
2	A	0	0	800
2	B	100	500	1600
2	C	0	100	300

Table 1 - An example of account monthly balance data.

Problem Statement

Your colleague Jane Hopper, the Business Analyst in charge of analyzing customer behavior, who directly consumes data from the Data Warehouse Environment, needs to get all the account's monthly balances between Jan/2020 and Dec/2020. She wasn't able to do it alone, and asked for your help. Add to your solution the SQL query (.sql file) used to retrieve the data needed (the necessary tables were sent in csv format along with this pdf, on folder tables/). Feel free to use the dialect of your choice, but please specify the SQL engine.

Now that you were able to work with these tables, you might have noticed that we could improve the data model somehow and want to suggest some changes. You may also consider that Nubank is always evolving with new products and it is also expanding to new countries, so our data warehouse model needs to accommodate all these incoming changes. Keep in mind that the new products sometimes are not related to peer-to-peer transactions – for example: life insurance, lending, rewards and other products – and some of them might be available only in some countries. Knowing all of that, which modifications would you propose to the current data model and why? Remember that other analysts will be using the same structure, so it should be as clear as possible. Feel free to change, remove and/or add tables and fields to generate a better data model design.

Since many people are already consuming data from the current model, we need to come up with a migration plan in order to change our data warehouse above with your suggestions. Which strategy would you propose in order to implement those changes?

On another note, Jane's friend, Pepino, wants to know how well our PIX product is doing inside Nubank. For that, he wants your help to come up with indicators that can be used to track the technical and business performance of the product. Which metrics would you suggest to track and why?

Additional Scenario

As mentioned before, Nubank is always evolving with new products and features. With this in mind, our product NuConta, which is the Nubank's customer banking account, has a feature that allows you to invest on a fixed rate income product. This product provides customers with a daily return of 0.01% (based on a governmental institution rate) according to their daily invested balance amount. Every day generates some return, including weekends. In order for the fixed income product to generate return, the customer has to voluntarily invest in this product at any time. **It is important to emphasize that the return is calculated on a daily basis after all withdrawals and/or deposits made in a given day.**

The transaction logs for this fixed income product is obtained through an API that provides this data in a JSON format. We need to integrate this data into the tables provided in order to calculate the total return generated by this product per customer.

An example scenario for customer A is given below, starting to invest in this fixed income product on day 16 of month 1. This is the first time this customer is depositing money into this investment, so the previous balance was null. His first deposit was 1000 (one thousand) and at the end of the day this amount has generated a 0.01% **income rate** to his balance. This customer continues to invest other times throughout the month in this same product. Keep in mind that this is a mock sample of the transaction log with calculations applied on a daily basis. Keep in mind that, in case of negative Movements, the income for that respective day should be set to zero.

$$Movements = Previous\ Day\ Balance + Deposit - Withdrawal$$

$$End\ of\ Day\ Income = Movements * Income\ Rate$$

$$Account\ Daily\ Balance = Movements + End\ of\ Day\ Income$$

Day	Month	Account ID	Deposit	Withdrawal	End of Day Income	Account Daily Balance
16	1	A	1000	0	0.1	1000.10
20	1	A	500	0	0.15	1500.55
2	2	A	0	200	0.13	1302.48
19	2	A	1000	200	0.21	2104.78

Table 2 - An example of investment transactions data.

Another business analyst, Sophia, would like your help to analyze how much money Nubank's customers' have on their investment account on a daily basis.

Summary

Imagine that we would present these solutions to either a business audience or a technical audience and people might access your solution in a moment that you are not available afterwards as well, so your solution should be self-explanatory as best as possible :)

Provide all required files mentioned for each Problem Statement below:

- **Problem Statement 1:** The SQL query used to help Jane retrieving the data needed using the dialect of your choice, but please specify the SQL engine
    a. Query in [_.sql] file extension
    b. Necessary inputs: transfer_in / transfer_out / pix_movements + dimensions if necessary
    c. Output of your query in [_.csv] file extension

- **Problem Statement 2:** Data model modification proposal with a visual representation and trade-off analysis
    a. Visual representation of the proposed changes
    b. Written explanation of the proposed changes or no changes. If you decide to not change some part of the data warehouse, we also expect a written explanation of why you decided to maintain it.

- **Problem Statement 3:** Migration plan and strategy in order to implement the data model modification proposal mentioned above
    a. Written or/and visual explanation of your plan and strategy

- **Problem Statement 4:** Suggestion of key performance indicators to track the technical and business performance of the PIX product:
    a. Written or/and visual explanation of the key performance indicators suggested

- **Problem Statement 5:** Calculate the total return generated by the investment product per account id:
    a. Script file with your code used to generate the output. You may use any programming language but DO NOT use SQL.
    b. Necessary inputs: investments + dimensions if necessary
    c. Output of your script in [_.csv] file extension
        i. filtered with:
            1. Dates from "2020-01-01" to "2020-12-31"
            2. Accounts listed in the available "investment_accounts_to_send.csv" file
        ii. following the schema below:

Columns	Account ID, Month, Day, Deposit, Withdrawal, End of Day Income, Account Daily Balance
Granularity	Account ID per Month per Day

Here are some tips that might help you create your solution:

- Keep in mind that if your case gets accepted, we'll schedule a talk so we can debate about your solution;
- If you want to send additional files, feel free to do so;
- Analytics Engineers are very worried about data governance;
- Analytics Engineers are the bridge between the Business and Software worlds;
- An image may add value for a solution to become self-explanatory;
- Analytics Engineers are very keen on documentation;

Suggestions for presentation

- The presentation must translate your thinking method to solve the case
- You will be required to explain your code
- Many questions will be made throughout the presentation, so make sure to account time for discussions
- The presentation should be easy to understand for non-tech people

Glossary

- **Account info (branch, number and check-digit)**
  In Brazil, a bank account can be uniquely identified by three numbers. Firstly the **branch** code that identifies in which bank branch the accounts were created. Secondly, the **account** number that identifies an account inside a branch. Finally, the **check-digit**, that exists only for error detection.

- **CPF**
  It is the Brazilian individual taxpayer registry identification.

- **PIX**
  This is the newest way to transfer money in Brazil. It's free. It is instantaneous, and one only needs to know the Pix-Key related to the account in order to make a transaction.

- **Non PIX transfers**
  Those are the traditional ways to move money from one bank account to another. In order to make this kind of transaction, one needs to inform the CPF, the branch code, the account-number and the check-digit of the account that will receive the money. The transaction usually takes several hours up to days to be confirmed, and most of the banks charge a fee in these transactions.