Neobank Data Pipeline

Consultants: Alex, Enrico, Ikechi, Luiggi, Marlin

Business Problem

Neobank is a online business delivering banking services all around the globe

The CEO wants use analytics to improve the company’s performance. She asks the DE team to support the Finance and Marketing departments with self serving data insights to help their decision making:

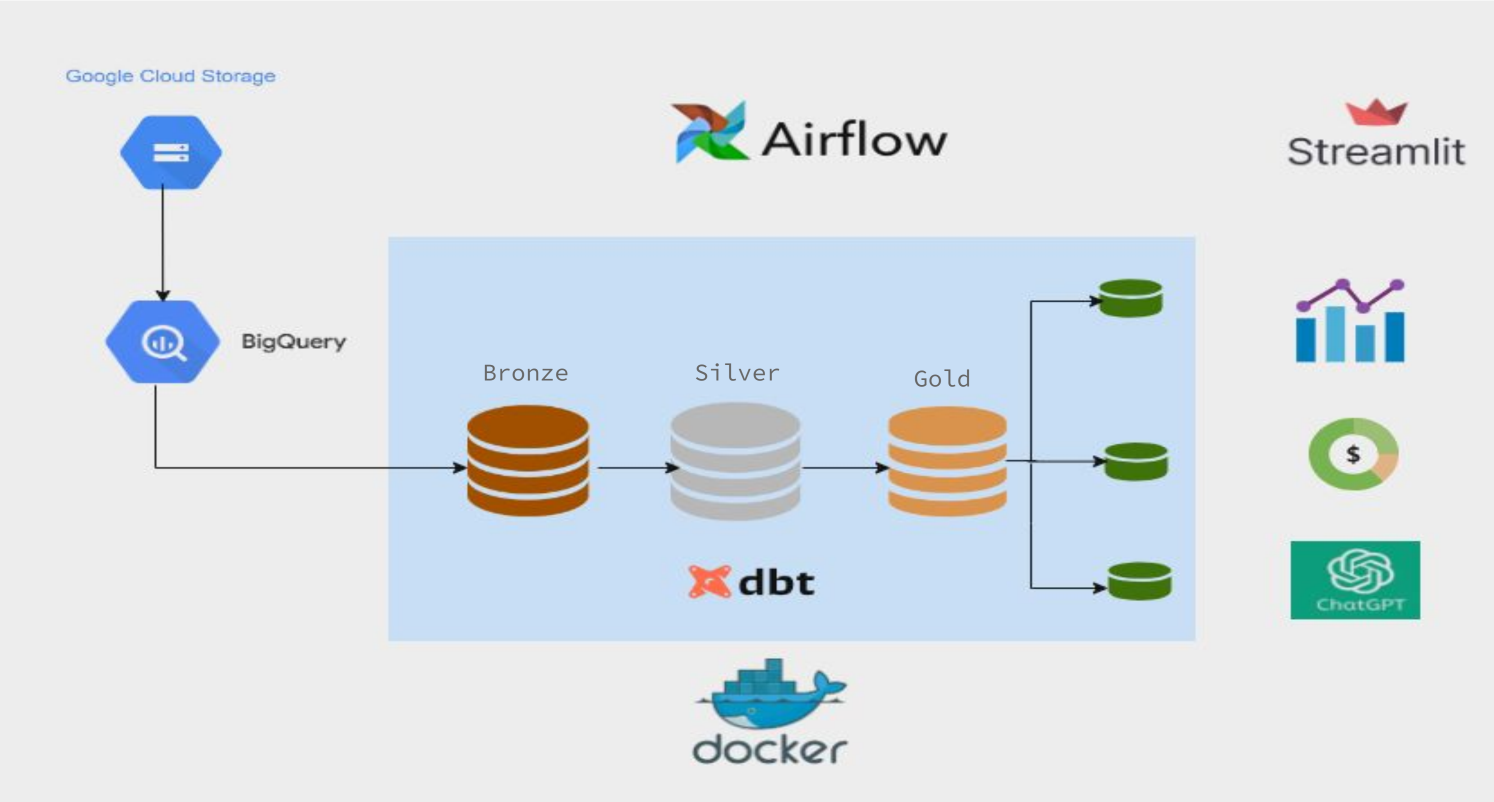
- What is the total amount of transactions by month/year?
- How much was charged in fees? Is it going up?
- Who are our users/customers?
- What are are the best communication channels?
- Are there any transactional patterns that could be of interest?

Engineering Challenge

Neobank gathered unorganized data of 2.7 million transactions across the globe

- Extracting the data from the different data sources (csv files)
- Storing the data in data warehouse
- Data modelling according to the business requirements
- Connecting to an interactive Dashboard
- Connecting a LLM to the database
- Making the process scalable and auto scheduled

Architecture



NeoBank

Extraction and Load

- Extraction: Created data lake buckets. Raw data was uploaded to these GCP buckets

Google Cloud Storage

- Load (to Google Bigquery): Created data tables, schema definition and uploaded data to Google Bigquery bronze layer

Google
Big Query

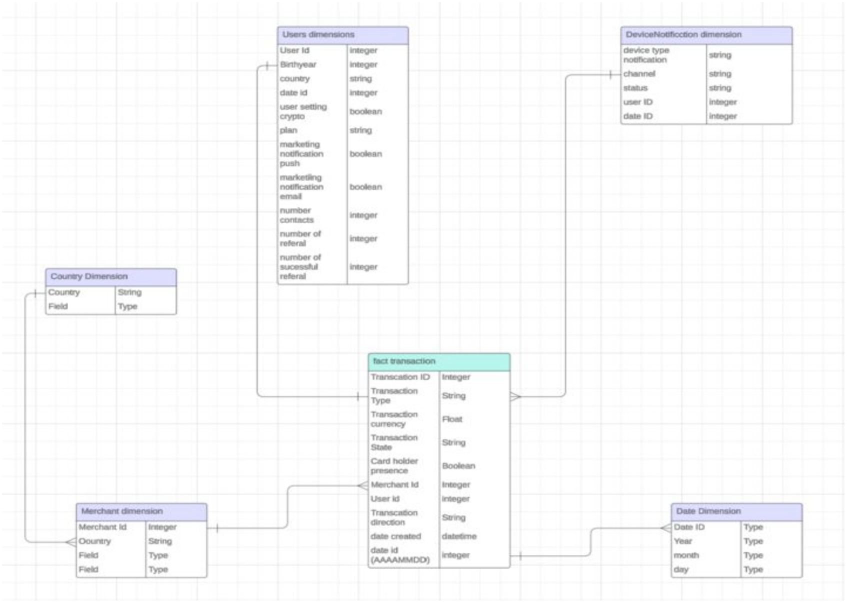
dbt Transformation

Bronze Layer: Raw Data uploaded from Data Lake (GCP bucket)

Silver Layer: Data Cleansing - Deleting unnecessary columns, formatting, Separating columns

Gold Layer: Building fact table and all dimension tables linking back to the fact table, creating a final star schema

Final Star Schema



Users dimensions

User Id	integer
Birthyear	integer
country	string
date id	integer
user setting crypto	boolean
plan	string
marketing notification push	boolean
marketiing notification email	boolean
number contacts	integer
number of referal	integer
number of sucessful referal	integer

DeviceNotifiocation dimension

device type notification	string
channel	string
status	string
user ID	integer
date ID	integer

Country Dimension

Country	String
Field	Type

fact transaction

Transcation ID	Integer
Transaction Type	String
Transaction currency	Float
Transaction State	String
Card holder presence	Boolean
Merchant Id	Integer
User id	integer
Transcation direction	String
date created	datetime
date id (AAAAMMDD)	integer

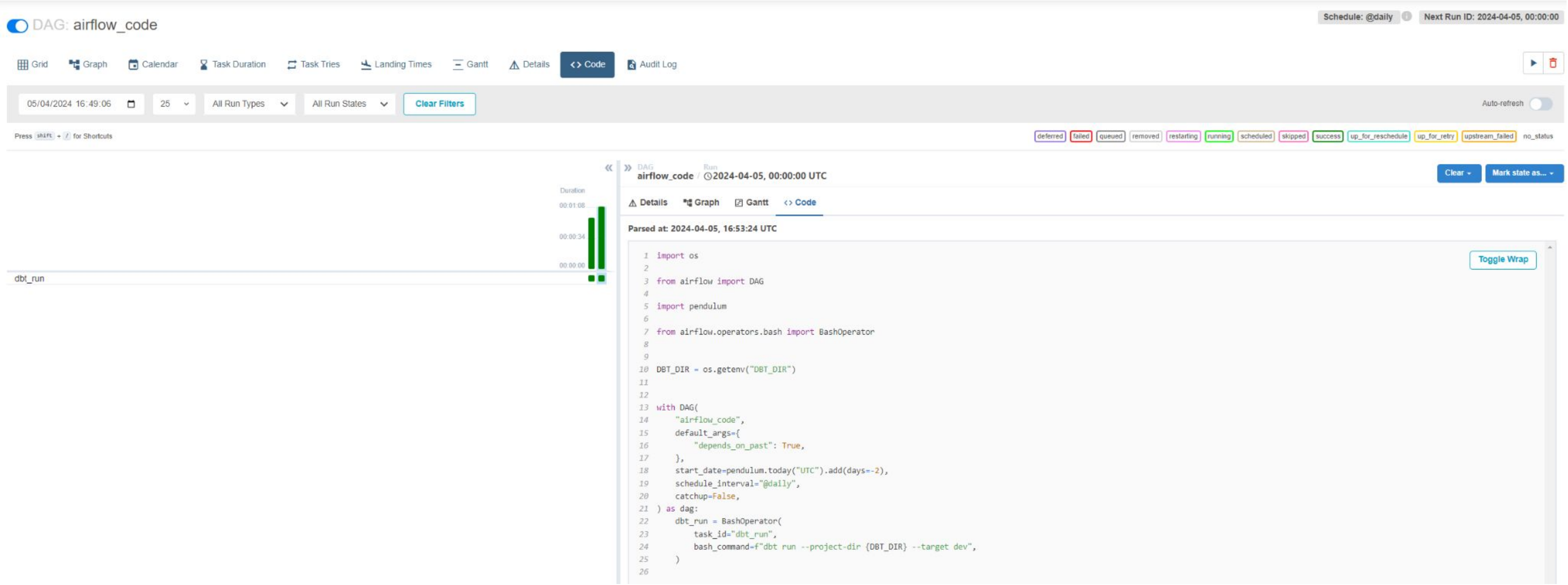
Merchant dimension

Merchant Id	Integer
Oountry	String
Field	Type
Field	Type

Date Dimension

Date ID	Type
Year	Type
month	Type
day	Type

Apache Airflow



DAG: airflow_code

Schedule: @daily

Next Run ID: 2024-04-05, 00:00:00

05/04/2024 16:49:06 | 25 | All Run Types | All Run States

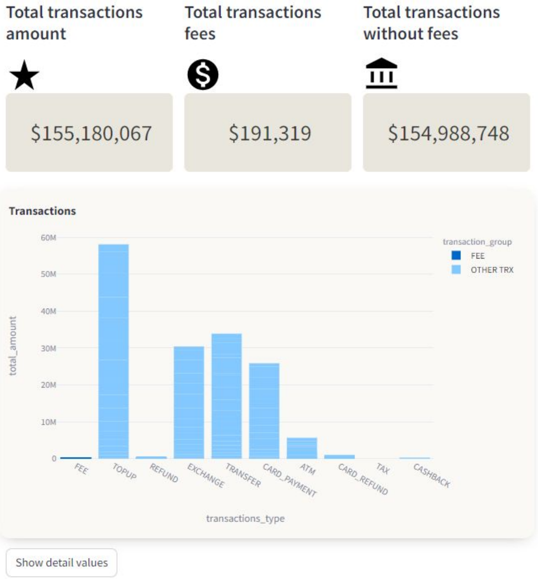
airflow_code / 2024-04-05, 00:00:00 UTC

Parsed at: 2024-04-05, 16:53:24 UTC

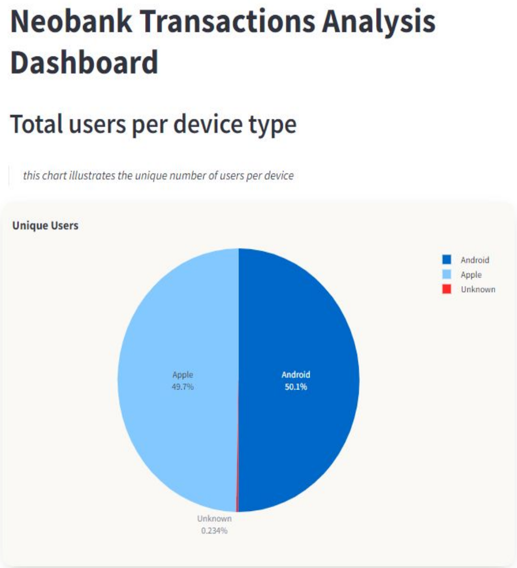
```python
import os

from airflow import DAG

import pendulum

from airflow.operators.bash import BashOperator


DBT_DIR = os.getenv("DBT_DIR")


with DAG(
    "airflow_code",
    default_args={
        "depends_on_past": True,
    },
    start_date=pendulum.today("UTC").add(days=-2),
    schedule_interval="@daily",
    catchup=False,
) as dag:
    dbt_run = BashOperator(
        task_id="dbt_run",
        bash_command=f"dbt run --project-dir {DBT_DIR} --target dev",
    )

```

Streamlit + Demo

https://dedemo-neobank.streamlit.app/



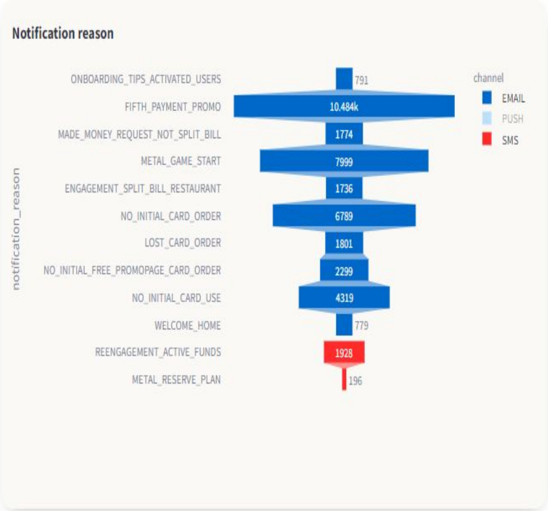
Total transactions amount	Total transactions fees	Total transactions without fees
★	$	🏛
$155,180,067	$191,319	$154,988,748



Neobank Transactions Analysis Dashboard

Total users per device type

this chart illustrates the unique number of users per device



NeoBank

QR Code

Challenges

- Github collaboration
- Package dependencies
- Connecting our Silver layer to a gpt3.5 model
- Securely deploying Streamlit (python versions, Streamlit secrets)

Future Improvements

- Auto orchestration of the extract and load process with Airflow
- Auto creation of table schema when uploading into BigQuery
- Create a CI/CD pipeline for seamless integration
- For own workflow:
  - Adding test files and documentation
  - Makefile to build, compile and organise source code
- Improving the LLM. We have a working variant, but we need to finetune it for more complex questions

Thank You

Tech Stack

Apache Airflow

ChatGPT

Google Big Query

docker

GitHub

dbt

Streamlit