Please follow the instructions in the Readme.md file before going through this file

Questions Asked During the Study

1. Churn over time: Users at month start – users at month end/users at month start
2. Revenue over time: Number of Subscriptions * 20 over time
3. Cumulative User Counts: Count of users grouped by months and years
4. Profit (Important for Stakeholders): Assuming each user makes a single appointment and gets a Diagnosis, 20*number of appointments - money spent during marketing
5. Return on Investment (Important for Stakeholders): Considering the total cost on Marketing as the only cost as ‘a’, and considering the change in profit calculated above over time as ‘b’, ROI can be calculated by (a-b)/b
6. Impact of Gender: Customers Lost by Gender
7. Impact of lead time: Average user retention time
8. Impact of Marketing Strategy on Profit

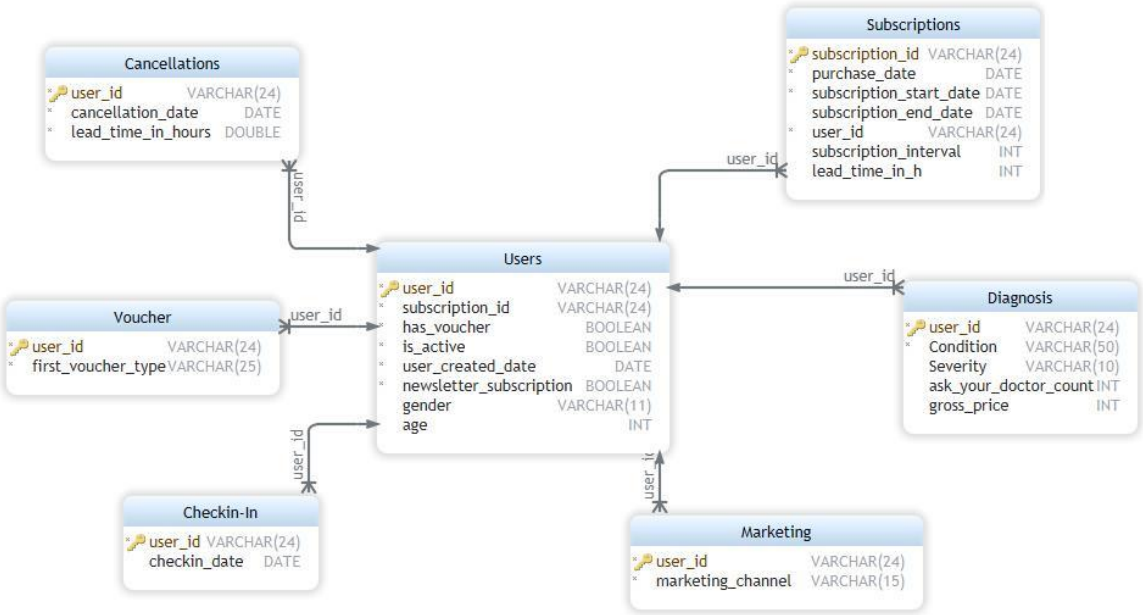
Key Performance Indicator:

Impact of Marketing Strategy on Profit

Technologies Used:

Big Query, Google Data Studio, Python

Schema:



ETL Pipeline Explained:

Extraction

1. The input files are read into a pandas Data frame.
2. The data Schema is then executed over Big Query from a schema file.
3. Then the data is cleaned up to minimize Null values, and then separated into their respective entities (See Diagram Above).
4. The data is then uploaded to Big Query after parsing data types.

Transformation

1. SQL Queries saved in the 'Data_Transformation/Queries' folder are executed over Big Query, and then the results are stored into CSVs in the 'Result_Sets' folder.

Loading

1. Data in the CSVs is read into pandas data frames.
2. The CSV is loaded into the respective worksheet on Google Sheets.
3. Data from Google Sheets is used in Google Data Studio in order to generate reports.

Whenever new data is uploaded to the sheets, the Reports in Data Studio would change as well according to the source sheets.

Head of BI considerations:

Resource Assignment:

3 People for setting up the Cloud Infrastructure using Big Query and Cloud Functions and Google Sheets, and for Future uses, given that we will have a data stream rather than a static data, the approach will change from Batch Processing to a Lambda Stream and then to Kappa over time, there setting-up data streams using Apache Spark or Kafka or GCloud Pub-Sub.

2 people for Data Cleanup and Quality Check, keeping the data updated and with as few Null values as possible.

2 people to write the Code and Queries, preferably using Python and SQL.

Since Big Query and Python are being used as the main technologies, for future considerations a pub-sub Cloud Function should be setup in order to create streams, rather than running multiple instructions.

The SQL Queries/Python Code written should not have the dates hard coded but should use SQL or Python constructs to generalize the groupings and partitions.

Future Considerations:

Setting up data streams.

Cost Estimations from GCP should be considered, when presenting monthly reports.

Time Management:

At least 2 weeks for Data Cleaning and Coding.

1 to 2 months for setting up the Cloud Infrastructure.

Therefore, a meeting with the Stakeholders should be planned accordingly, given that the team has enough resources available, and the Pipeline is functioning as a data stream.

Future Project for Formel:

As mentioned in the interview with Laura Nicholas, I used to work in Kassenärztliche Vereinigung Berlin, which is the statutory healthcare authority for Registered Doctors, KV is available all of Germany, and they are responsible for making the billings for the doctors and Insurance Companies, so that the Doctors get their payments directly from KV as compared to through the Insurance Companies, thus making the Process streamlined and well isolated.

During the Billing Process at KV, the incoming doctor's data is fitted into a Data Schema that is standardized for the Insurance Companies, this data schema is found in the "vda_Schnittstellenbeschreibung_V2.0.pdf" file, which I am providing alongside the study.

During the Covid Pandemic, doctors started offering Video Consultations, through a service offered by the University of Duisburg, KV had to implement a new Data Pipeline (which I was directly involved with) in order to make this new incoming data a part of their regular Billing Process.

Since Formel also offers Video Consultations and in case the Users decide to pay with insurance, the Doctors must prepare the Billing Certificates and send them over to KV themselves, to get paid.

If Formel decides to provide data to the Insurance companies themselves, that might change the Business Model of the Company, as Formel would then have to hire the doctors themselves, and pay them the money that Formel would receive from the Insurance Companies