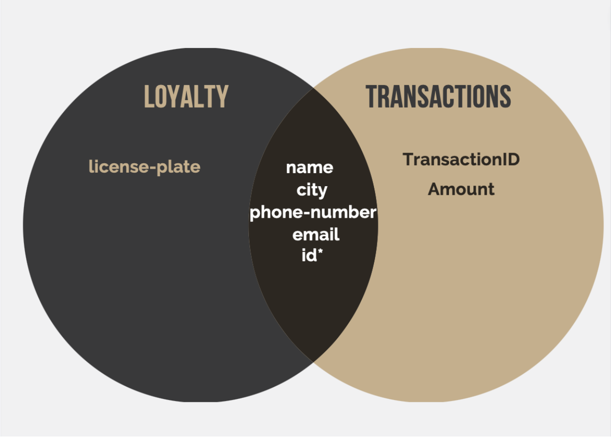
Lilian Pung Hui Ling - August 2022 - *lilianpunghuiling@gmail.com*

Challenge 1: Data Cleaning, Transformations and ETL pipeline architecture

Code can be found here:
https://github.com/LilianPung/TakeHomeTest---LilianPungHuiLing

1. **Explain 5 reasons, why did you join the datasets in that way?**



- From the Venn Diagram above, we can conclude that:
Loyalty = {“name”, “city”, “phone-number”, “license-plate”, “email”, “id”}
Transactions = {“name”, “city”, “phone-number”, “email”, “id”, “TransactionID”, “Amount}
Loyalty ∩ *Transactions* = {“name”, “city”, “phone-number”, “email”, “id”}
Both datasets have the same column “name”, “city,”, “phone-number”, “email” and “id”

- Both datasets have unique “**id”** as the **key** to identify each data entry, hence I’m using the column “**id”** to join both datasets.

- I noticed the columns “name”, “city”, “phone-number”, “license-plate”, “email” in dataset *Loyalty* were dirty (contained junk, special characters, space, etc), while the same columns in dataset *Transactions* were clean (except for column “license-plate”) hence I joined both dataset using column “id”.

- I dropped the columns “name”, “city”, “phone-number”, “email” in dataset *Loyalty*, and instead used the cleaned data in dataset *Transactions*. Now the joined dataset looks like this:
name (from *Transactions)*
city (from *Transactions)*
phone-number (from *Transactions)*
license-plate (*Loyalty)*
email (from *Transactions)*
transactionID (*Transactions*)
amount (*Transactions*)

- I renamed the columns “license-plate” and “phone-number” to “licensePlate” and “phoneNumber” to avoid syntax errors in Jupyter Notebook. Then, I removed special characters in column “licensePlate”

2. **Business Use Case**

Title: Automated Car Service System

Goal: To increase efficiency of client identification at the car service centre by minimising manual labour needed.

Description: A client goes into a car service centre for car maintenance. The client identification process will be hassle-free, the staff can identify the client by “license plate recognition” in which the information of the client would be shown on screen. After maintenance is done, the client can leave right away with the invoice being sent to their email. There will be minimal human interaction in which most processes are automated. The client’s information is shared amongst all car service centres (same brand), hence the vehicle’s records can still be traced even when the client is travelling out of the current city.

Pre-Condition: The client has to first register their information (name, city, phone number, license plate, email, etc) into the system before the automation process can be implemented.

Post-Condition: Client identification process is enhanced, overall client satisfaction rate is increased while decreasing overall manual labour cost.

Exceptions: It’s suggested for the centres to have stable internet connection to ensure all clients’ information are updated and stored properly.

3. **Cloud-based pipeline architecture:**

Goal: Processing and storing client's data to enhance business products and services, ensuring data quality, reliablity and data consistency for efficient data access across car service centres.

Data Sources:
The user historical data (eg: vehicle's last service date, vehicle component status, etc) will be stored in operational databases like SQL, NoSQL.
As client's data has key-value pairs, we would store their information as JSON format - this way the data is more readable, and with given permission, the data can be amended when needed.

Data Ingestion Strategy:
I would apply *Streaming Ingestion* - passing data along to its destination as it arrives in the system. However this step is relative, it's actually simply "micro-batching" the data, just sending it more frequently in much smaller groups.

Data Processing:
The data will then go through ETL processes in order to clean and format the data before it's being stored. We would use SQL queries to enhance data quality by removing redundant data and correcting the flaws.

Data Storage:
The data will be stored in a cloud-based data warehouse, i.e Google Cloud.

Data Workflow:
I will implement a work-flow that allows clients data to be read from different centres/locations, joining their data using a unique key (id) and then storing the transformed data at the data warehouse.

Data Monitoring & Governance:
The client's data will be encrypted via Customer Managed Keys (CMK) - in which the clients can independently monitor usage of their data and revoke access if needed.

How do you validate your analytics?
- Uniqueness: I would make sure all client's IDs are unique.
- Logic: make sure data's been entered in a logically consistent way
- Format: make sure data is stored in proper format to ensure consistency across data
- Data Type: make sure new data entered is following the predefined data format

Challenge 2 : Customer engagement

Proposal: Crisis monitoring for Nestle (Gerber Products Company)

Context: On 28 Sept 2021, a new congress report had been published regarding heavy metal being detected in Gerber baby food.

How I would present:
I would first describe the *total mentions and engagements* during the crisis monitoring period, followed by the *sentiment breakdown* (NSS score, positive - neutral - negative) and *channel breakdown* (which social media platforms - Facebook / Twitter / Instagram, etc).

Then, I would go into detail of the overview - including the elaboration of observation and analysis. I would demonstrate which social media mention contributes to the highest engagements - in this case, it's CNN's Facebook Page. I would then go into detail of the sentiment breakdowns: neutral mentions (netizens questioning FDA's method of approval, vaccine discussion, etc.), negative mentions (how Gerber do not care about their consumers, boycott Gerber Products, demands to have legal actions against Gerber, corruption, etc.), positive mentions (in this case, there is none).

Next, I would demonstrate the threat level imposed by this crisis and the recommendation I have for my client, and assure the client that we would continue to monitor the issue closely.

Before wrapping up, I would display the top 3 posts with highest engagement along with their NSS score.

**Glossary:
- Mention = A message which refers to or includes a monitored brand keyword.
- Engagement = An interaction (like, comment, share, retweet, view) received by the post.
- Sentiment = A measure of the message's tone (positive/neutral/negative)

Challenge 3: Scrapping , Datasourcing and Enriching data

1. Data Enrichment

- Data Enrichment is the process of enhancing/improving the data for a better real-time data analysis. With the additional resources being merged to the raw data, we can answer questions with deep/better insights as now we have a higher perspective of the business case.