Document details

Document:	Public Summary of CEU Capstone Project
Author:	Kristof Rabay
Title:	Exploration, utilization, and monetization of online search database
Company:	Hiflylabs, Hungary
Client:	*<anonymized>*

1. Project scope

I was working with Hiflylabs on a project with a Client in the online advertising industry. The Client had a search database that recorded every search event that happened on their website, with attributes such as whether or not the person who searched was logged in, what platform (i.e.: phone) they used, what parameters they used and what values they submitted to certain parameters. My job was to explore this database, analyze user behavior to let the Client know what their (possible) customers search for, what parameters they use and what values they submit.

We had outlined 4 main parts to the project:

(1) Exploring attribute information (such as user types, platforms, etc…) and basic parameter usage data (such as which are most frequent parameters)

(2) Exploring combination of parameters (there is an 'infinite' amount of possible combinations of parameters, Client was interested in the usual 'structure' of the searches)

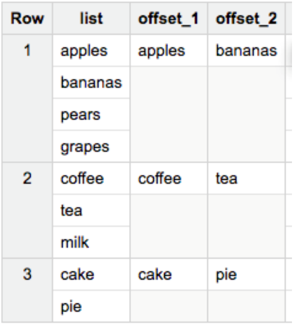
(3) Distribution analysis of continuous quantitative parameters such as price range. Create filter for user type and geographic location

(4) Geographic hierarchy analysis – when a user searches for a product, they may submit a geographic value, such as filter on city, filter on district, etc… Goal was to find the most commonly used level of detail when it came to geographic parameter usages

2. Data quality and availability

As a first step we needed to establish what data I was to use, where I could find it and how I could access it. Data was stored and collected in Google' Cloud Platform due to its size (> 500 Gb, >1 billion rows). The actual engine that supported the database was GCP's own BigQuery SQL, so most of the analytical work included submitting BigQuery-specific SQL commands to the raw database. What was different with regards to this SQL language, is that BigQuery supported a 'new' type of data, called 'repeated field', that are basically arrays / structs contained in a single cell. When needing to access data within these arrays, I first needed to unnest the field:

Figure 1: BigQuery's repeatable fields in 'list' column



Row	list	offset_1	offset_2
1	apples bananas pears grapes	apples	bananas
2	coffee tea milk	coffee	tea
3	cake pie	cake	pie

Source: https://stackoverflow.com/questions/48640838/select-first-n-items-in-a-google-bigquery-repeated-field/48643970#48643970

In the above example, the first row's first column is a repeatable field with 4 elements in the array. If we are to unnest it, row # 1 would be turned into 4 rows. With regards to my project, the only attribute to this size increase was the need to control costs (BigQuery is priced in a by-query way, so every ran query has a specific cost associated with it based on the size of the data that needs to be read and manipulated to complete the query). I've agreed with the Client to set a daily maximum limit in terms of cost and data usage, a limit that couldn't be breached, as if I were to reach that usage, I simply couldn't submit more queries that day.

3. Cooperation with project manager and Client

The project team consisted of 2 people: (1) my project manager and (2) me as the only analyst. We had weekly / semi-weekly VCs with the Client's operative team, where we discussed our previous and future deliverables, asked our questions and held some brainstorming sessions. While working on our deliverables, my PM and I held phone calls twice a week, where I presented my findings, we agreed upon my next tasks and deadlines and thought about possible next steps together.

4. Outputs for Client to support sales and marketing

As stated, during analysis I used Google's BigQuery SQL engine (95% of analysis), for distribution analysis I leveraged RStudio. For quick visualizations I used Google's Data Studio that could be directly connected to a specific BigQuery query and hence made it possible for me to visualize each query to the Client.

For the project closing meeting, the final outputs were dashboards in Tableau (in a story). The data sources that the visuals were built from were all live-connected to BigQuery, so one major takeaway for the Client was that the whole project could be done 'in the Cloud', meaning no data exports or extracts were created throughout the process.

For the final deliverable, I had created multiple dashboards that let the Client explore their search database with user and platform-specific filters to better familiarize themselves with their customers' behaviors. These dashboards included:

(1) Tool to explore mostly used parameters and parameter-combinations of different user types and different locations: Client may target different users in different regions with more data-driven parameter placement, Client may build on existing differentiating strategies (when it comes to geographic location of customers)

(2) Tool to explore continuous quantitative parameters such as price range: Client may explore different price expectations of user groups in locations. This may lead to more personalized targeting from the marketing team.

(3) Tool to explore advertisement-success with regards to geographic parameter usages (how likely is an ad to show up in results when users search for given city / zone / street, etc…). Client may better communicate with their advertisers how to check whether or not their ad-bidding was successful.

5. Future project ideas between Client and Hiflylabs

The project outlined numerous ideas about what sales increasing projects can take place after enhancing the current search database. Some of these are:

- adding user information and predicting which users will go from searchers to buyers
- time series analysis: cyclical / seasonal results may drive marketing in different times of year
- time series forecasting: which ads will have more visitors / clicks, which parameters / values will be used more frequently

6. What I've learned as a data scientist on this project

In terms of technology, I leveraged the BigQuery-specific SQL language for 95% of my analysis, when it came to distribution exploration and visualization I used RStudio with data.table and ggplot, otherwise I turned to Google's Data Studio for quick visuals. The final deliverable was a collection of interactive dashboards in Tableau, so overall I used 4 different software.

In terms of data scientist skills, having been the only analyst, I mostly gained communication skills, as I had to present my findings to my PM and Client, overall holding 4-5 meetings per week. I also needed to pay attention to time management, efficient analysis (cost and time were factors) and strategic thinking, as exploration of a database may lead to numerous possible pathways, of which I always needed to follow the ones leading to the bigger business results (finding the 80-20 ratio, keeping hypotheses in mind, driving work by Client's B2B, marketing and sales teams' preferences).