Analytics Engineer Case

Case structure and source code available at github.com/izabel-ferrari/analytics-engineer-case

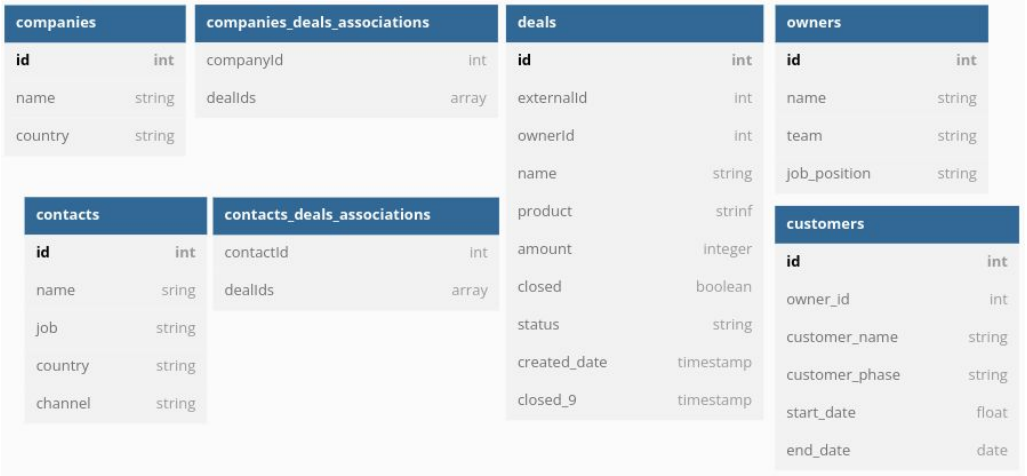
Part 1
Data ingestion, modeling and analysis

Data ingestion

The source csv files were uploaded into a dataset called **raw_data** in **BigQuery** using a Python script.

Each csv file became a table, as presented in the following diagram.

Python script created to perform the ingestion is available here.



Tables and its attributes after csv files ingestion in the **raw_data** dataset in BigQuery

Data modeling

Data transformations such as removing personal data from **owners** and **contacts** tables were performed and the resultant tables were landed in the **transformed_data_staging** dataset.

After that, further transformation was done, such as: removing intermediate tables found in **transformed_data_staging**, and adding references to **company_id** and **contact_id** directly in the **deals** table. The final tables were landed in the **transformed_data_final** dataset.

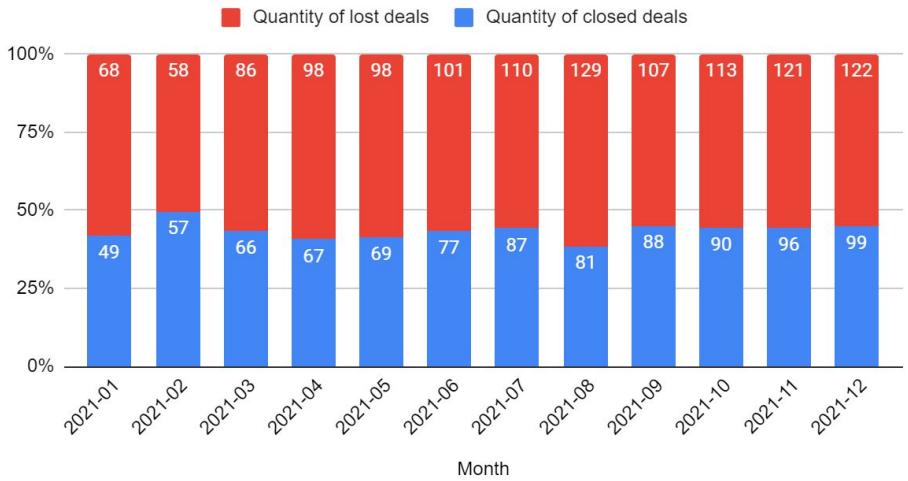
Source code of dbt project available here.

Tables relationship and attributes after data modeling

Data analysis - Quantity of closed and lost deals per month

Based on the information available, it is possible to see that every month the quantity of lost deals is higher than the quantity of closed deals.

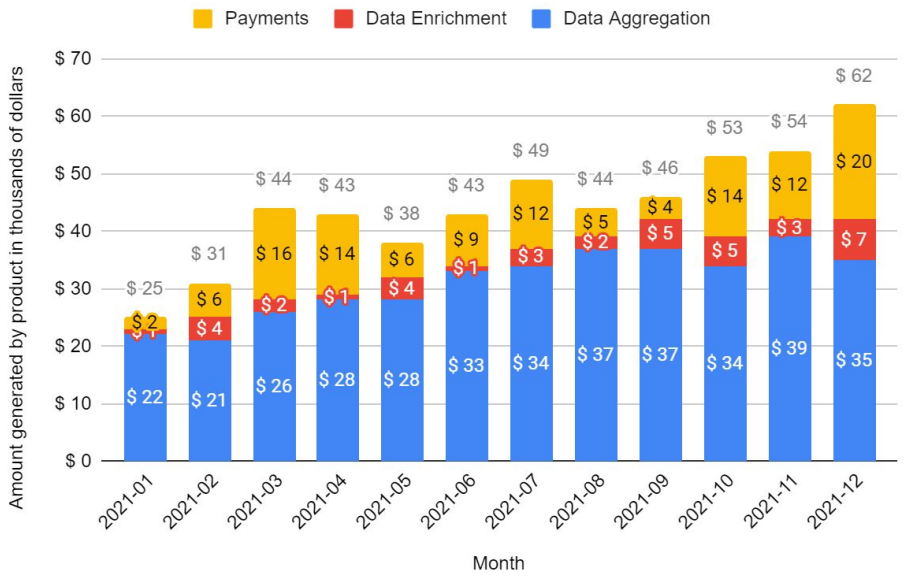
SQL query created to get this insight is available here.



Data analysis - Monthly amount of deals closed per product

The Data Aggregation product is responsible for more than half of the monthly total amount generated by all products.

SQL query created to get this insight is available here.

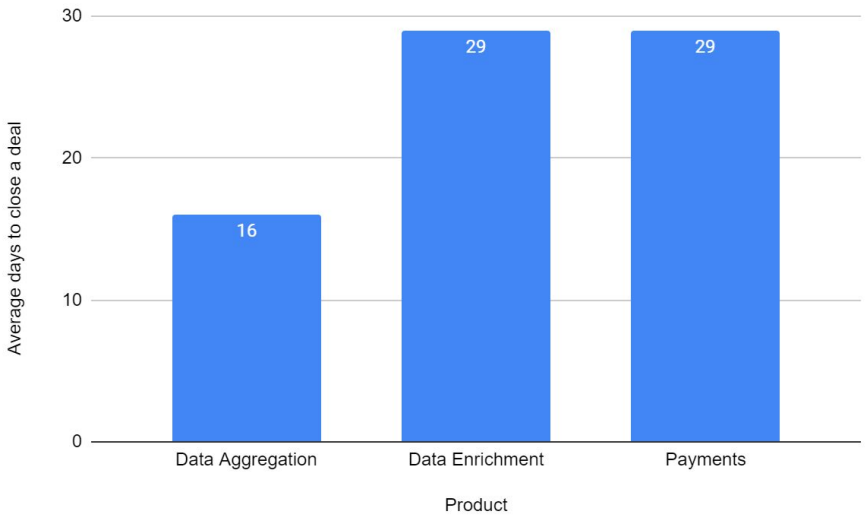


Data analysis - Average days to close a deal for each product

The Data Aggregation is the product that takes the least amount of time to close a deal.

It takes on average 16 days to close a Data Aggregation deal, whereas it takes approximately 29 days to close a Data Enrichment or Payments deal.

SQL query created to get this insight is available here.



Data analysis - Quantity and amount of closed deals per recurrent company

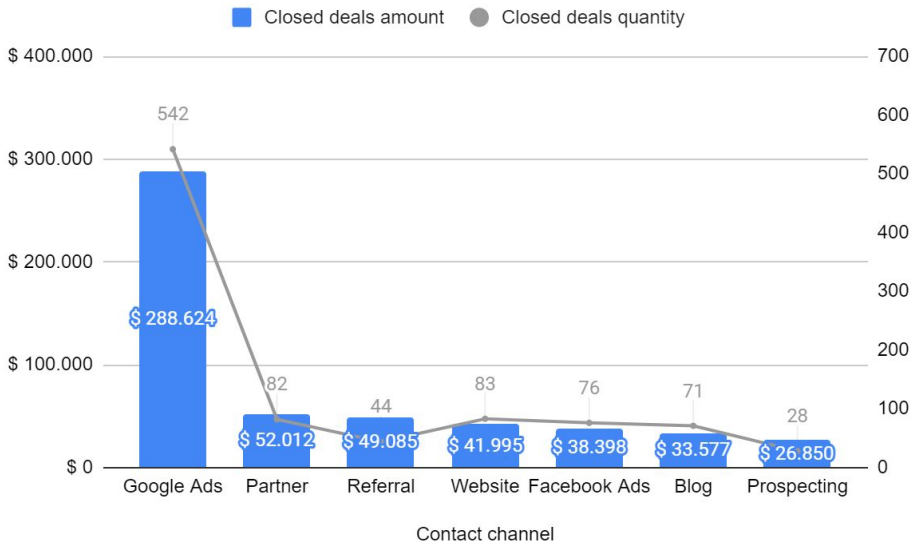
Out of the 17 recurrent companies (i.e., companies that close more than one deal), only 3 of them closed 3 deals, and the remaining 14 closed only 2 deals.

SQL query created to get this insight is available here.

Data analysis - Quantity and amount of closed deals per acquisition channel

Google Ads was the best channel of 2021, most of the closed deals were acquired through it.

SQL query created to get this insight is available here.



Suggested modifications

As a suggested improvement, it would be interesting to unify **customers** and **companies** tables to generate one dimension with all the companies information.

It would also be necessary to get a better **deals** data extraction and/or get more information with the responsible team to confirm if there can only be one contact and one company per deal. If not, then it will be necessary to change the current table structure.

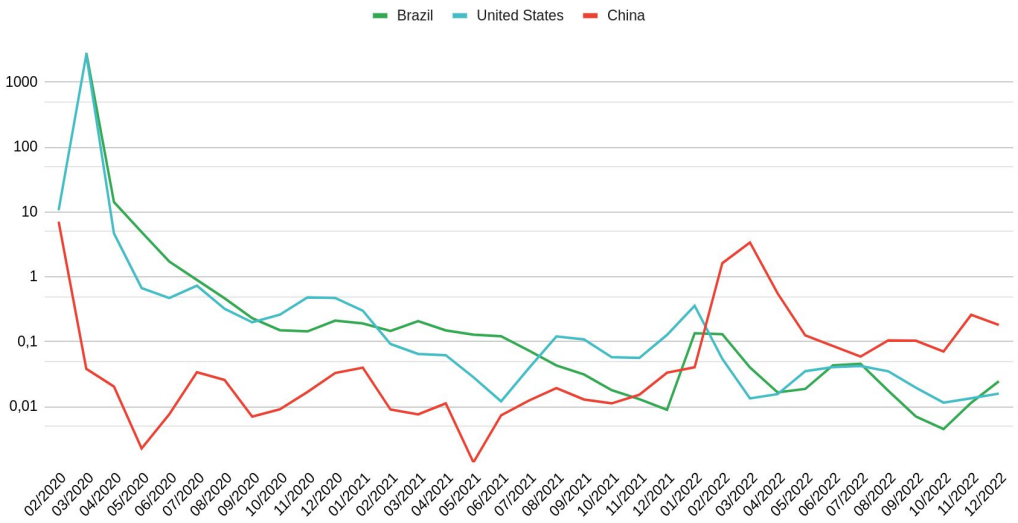
Finally, it would be also advisable to remove the data type inference in the data ingestion so we can have all the data in the raw_data schema exactly as they are in the source files.

Part 2
Query implementation

SQL query that generates per capita COVID total case growth rate per country per week is available here.

Interesting findings - China's spike of COVID-19 cases in 2022

In the following chart it is possible to see an increase of COVID-19 cases in China during the year of 2022, as opposed to an overall decline of new cases in countries as United States and Brazil.



Monthly growth rate of confirmed COVID-19 cases per capita in Brazil, United States and China (logarithmic scale)

Thank you