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 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), and 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

Every month the quantity of lost deals is higher than the quantity of closed deals.

On average, 57% of the deals in one given month is lost and this percentage has been steady throughout the year of 2021.

SQL query available here.

deal_created_month	closed	lost
2021-01	49	68
2021-02	57	58
2021-03	66	86
2021-04	67	98
2021-05	69	98
2021-06	77	101
2021-07	87	110
2021-08	81	129
2021-09	88	107
2021-10	90	113
2021-11	96	121
2021-12	99	122

Data analysis - Monthly amount of deals closed per product

Every month the Data Aggregation product is responsible for more than half of the total amount generated by all the three products.

SQL query available here.

deal_created_month	data_aggregation	data_enrichment	payments
2021-01	21588	1473	2341
2021-02	20592	3878	5606
2021-03	26118	2164	15816
2021-04	28416	1336	13732
2021-05	27783	3595	5859
2021-06	33229	1000	9245
2021-07	33515	2581	12115
2021-08	36661	1501	5206
2021-09	37088	5362	4469
2021-10	34010	4845	13708
2021-11	39367	3056	11971
2021-12	34820	6529	19966

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 approximately 16 days to close a Data Aggregation deal, and it takes approximately 29 days to close a Data Enrichment or Payments deal.

deal_product	avg_days_to_close_deal
Data Aggregation	16
Data Enrichment	29
Payments	29

SQL query available here.

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

There are not many recurrent companies - i.e., once a deal was closed, it is not usual that the same company closes another deal in the future.

Out of the 17 recurrent companies, only 3 of them closed three deals, the remaining 14 closed two deals. All the other companies only closed one deal.

SQL query available here.

company_name	closed_deals_qty	closed_deal_amt
Brown Inc	3	1294
Collins Group	3	1283
Harris Group	3	1145
Alexander PLC	2	2964
Garcia LLC	2	2470
Meyer LLC	2	1481
Miller Group	2	1386
Smith PLC	2	1313
Jackson Ltd	2	1306
Allen Ltd	2	1262
Clark PLC	2	1108
Sullivan LLC	2	969
Moore Group	2	959
Howard Group	2	817
Sullivan Inc	2	741
Williams and Sons	2	718
Johnson and Sons	2	693

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

Google Ads has been the best acquisition channel throughout the year of 2021, with 59% of the closed deals being acquired through it.

SQL query available here.

contact_channel	closed_deals_qty	closed_deal_amt
Google Ads	542	288624
Website	83	41995
Partner	82	52012
Facebook Ads	76	38398
Blog	71	33577
Referral	44	49085
Prospecting	28	26850

Suggested improvements

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

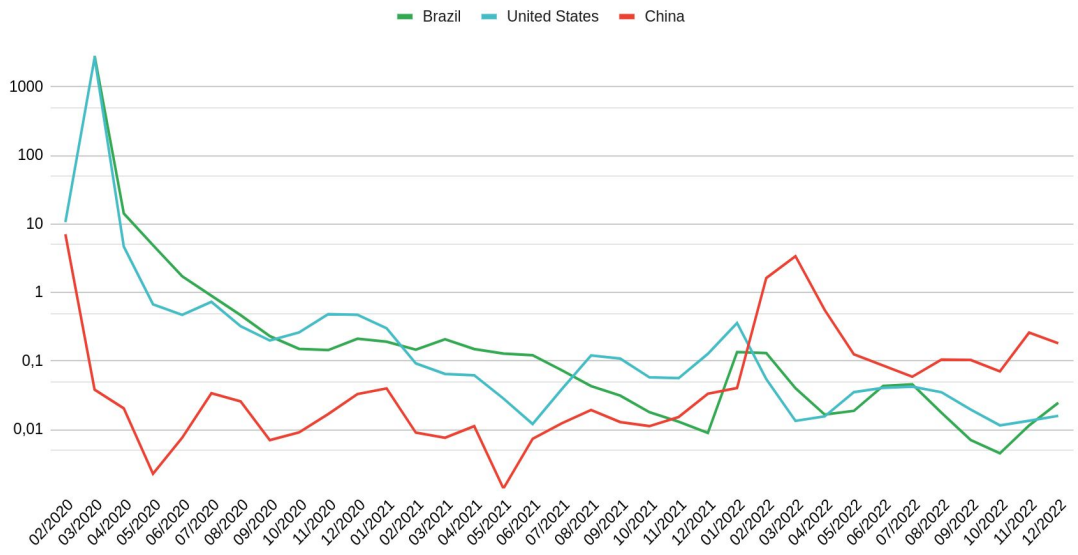
It would also be necessary to get a better **deals** data extraction and/or get more information with the responsible team to confirm if there can only be one contact and one company per deal. If not, then it will be necessary to change the current table structure.

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 for weekly growth rate per capita per country available here.

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



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

Thank you