**Elementor - Data Analyst Team Lead Home Assignment – Doron Hadar**

Preliminary assumptions on Sales data:

1. Data is a scrambled set from real data.
2. Data represents sales only and not license management.
3. The column min\_created\_ts represents the record creation date for the assignment. All other date columns were not adjusted for the task, therefore cannot be used for actual analysis on this set. However, they can be used for reporting and analysis ideas (e.g. distribution of how many days to ask for the 30-days-money-back).
4. There are cases in which certain fields values repeat in rows which don't seem to have any relation to each other. These repetitions include (not a finite list):
   * Same Payment\_ID for different transactions (different customers, etc.)
   * Same Customer\_ID for different transactions, carrying different User\_IDs, countries, names, email address, etc.
   * Same User\_ID for different Customer\_IDs, names, etc.
   * Same License\_ID for different transactions, carrying different Customer\_IDs, etc.

These repetitions may be caused by the following reasons:

1. Data is flowing from multiple transactional systems; each uses its own ID system. However, this explanation is not supported by the data in the provided file. There is a field called "data\_source" which carries the same value for all rows and there is no other field which might indicate such structure nor explain how a duplicate license ID can be verified against a license engine.
2. Damaged transactional system. This is highly improbable due to the extent of repetition in the file (or someone should lose his job…)
3. Some repetitions may represent real changes (e.g. a customer may change his country of residence, so we may see the same customer with multiple countries as a "slowly changing dimension"). This is also highly improbable, as the level of different details per repetition is high (e.g. different email addresses and name for same customer ID), and all the records are in the range of 2 months.
4. Repetition is caused by data scrambling for the assignment. This is the most probable assumption. Under this assumption, repetitions will be ignored and each row will be considered as a unique customer, license ID, etc. However, this does not suggest that reporting on returning customers should be ignored.
5. The "domain" field – The Elementor checkout page doesn't have this input, and based on the business model of Elementor, I assume this is the domain in which the license was activated. However, since there are plans for multiple sites, I assume this piece of data was merged to the sales and in actuality there is a one-to-many relationship between a license ID and a domain.
6. The sales file doesn't include explicit renewal transactions (e.g. payment for a 2nd year). However, fields of such an action do exist. The assumption is that renewal transactions can be distinguished.
7. The sales file seem to have the last state of a purchase (again, assuming payment ID or license ID repetition were created by mistake) and not a change log of a license (assuming the "action\_x" is a count of the actions – how many failed attemtpes, etc.). It is unclear if a history is required, as there is no mention of other history data (timestamps for failed payment attempts, etc.)
8. It is unclear why there seems to be a discount in the final total field when there is no indicator of any other action in a row. Example: customer\_ID 22780 purchased plan 3 which cost $199, but paid only $99.5. There is no discount or refund indicator, nor another line for this customer. Assumption is that this is another mistake in the file preparations.

**Tasks:**

1. Schema building  
   Assumptions:
   1. The schema is designed for reporting and analytics purposes and not the company's data warehouse. This assumption rules out Data Vault model, as performance for any reporting system will be hindered.
   2. Elementor is using Looker as its BI/Reporting suite. Looker's best practices suggest a snowflake model.
   3. Business model – The Elementor business model comprised of yearly plans. It is unclear if Elementor will ever sell individual items (e.g. only a template pack or create its own store for the 3rd party add-on developers). Such a change may require a change in the purchase data structure, as a transaction may include more than one item. Current assumption is one line of product per purchase.
   4. Date dimension and multiple date fields in a fact table – There is one date dimension, but there could be multiple date fields in a fact table. The schema itself will not solve this. The work on reporting platform will connect the same dimension to multiple fields per the options and best practices of the reporting platform
   5. Currency – the sales file contains only USD transactions and the media spent file is also in USD. Currently, the checkout page doesn't have a currency change functionality. It is unclear if Elemntor will ever allow payment in other currencies. The current assumption that this will not change and there is no need for currency exchange logging.
   6. Breakdown of totals per action in the sales file (e.g. net\_new, net\_renew) – although seems to be redundant, as there is only one action per line, for reporting performance purposes, it seems that the breakdown should remain.
   7. Changelogs and Slowly Changing Dimensions. While it is clear that business plans may change over time (e.g. pricing) and some customer data may change (e.g. email address) and there is a need to log these changes, it is unclear if there is a requirement to log changes in a transactions, as it is already noted that the file shows the last status of the transaction. Therefore, plans and customer data will be built using SCD2 method (The same ID, but with a calculation of what is the current data).
   8. Nitpicking – it is unclear why there is an Israeli term "maam" (VAT) in some of the sales file fields, with a duplication of net fields (why does the "net\_total" field already include tax? Either it's a net or not). I assume Elementor is an Israeli taxable entity and doesn't have any legal taxable entities in the rest of the world, and as such need to collect only the Israeli VAT, so some kind of a patch was laid on the transactions system. Fields will be consolidated to tax fields, net, etc

Explanation for Schema:

1. Dim\_date table was built with additional attributes for date, like month and year, and a few examples for flags that will improve performance for reporting, such as if the date is in the past 7 days, etc. These flags will need to be re-calculated every day

II. Field name changes:

min\_created\_ts is now created\_date in the purchases table

action is now transaction\_last\_status

The Israeli "maam" will be consolidated into the tax columns

III. No need to the revenue by plan in the media table

2 – SQL queries built on Snowflake

2a: assumption – profitability doesn't include affiliate costs.

SELECT country\_name

FROM (

SELECT country\_name, RANK() OVER (ORDER BY SUM(net\_total) DESC) AS rank\_order

FROM countries AS ct

INNER JOIN customers AS co

INNER JOIN purchases AS pr

ON co.customer\_id=pr.customer\_id

ON ct.country\_iso2=co=country\_iso2

GROUP BY 1) dt

WHERE rank\_order IN (2,3)

2b:

As per my conversation with Menachem, the wording of this question is a bit misleading. According to Menachem, the purpose of this task is to provide suggestion for future distribution of media. The wording of the question should have been: "Build a media spend distribution suggestion for each day based on performance of the following parameters:"

The logic of the below query is as follows:

* Performance is captured for the past 7 days before running the query.
* Query provides only top 3 results
* Both the above numbers are arbitrary and can be changed. However, IMHO, I believe these are optimal settings
  1. By license type:  
     WITH top\_counts AS   
      (SELECT plan\_id,COUNT(1) AS transactions  
      FROM Purchases   
      WHERE transaction\_last\_status='new'   
      AND created\_date BETWEEN   
      DATEADD(DAY,-7,CURRENT\_DATE()) AND   
      DATEADD(DAY,-1,CURRENT\_DATE())  
      GROUP BY 1   
      ORDER BY 2 DESC  
      LIMIT 3)  
     SELECT l.plan\_id,r.transactions\*1.0/l.tot\_trans AS pct   
     FROM top\_counts AS l   
      JOIN (SELECT sum(transactions) tot\_trans FROM top\_counts) r

ii. By Country:  
 WITH top\_counts AS   
 (SELECT country\_name, COUNT(1) AS transactions  
 FROM Purchases pr

INNER JOIN Customers co

INNER JOIN Countries ct

ON co.country\_iso2=ct.country\_iso2

ON pr.customer\_id=co.customer\_id  
 WHERE transaction\_last\_status ='new'   
 AND created\_date BETWEEN   
 DATEADD(DAY,-7,CURRENT\_DATE()) AND   
 DATEADD(DAY,-1,CURRENT\_DATE())  
 GROUP BY 1   
 ORDER BY 2 DESC  
 LIMIT 3)  
SELECT l. country\_name,r.transactions\*1.0/l.tot\_trans AS pct   
FROM top\_counts AS l   
 JOIN (SELECT sum(transactions) tot\_trans FROM top\_counts) r

iii. Theoretically, one can look into the best ranking sites

4. Additional directions:

* Emphasis on renewal. Although not present in the data (and thus no way to create dashboards for it in this exercise), measuring renewals (and churned customers) is extremely important, as a healthy business should have a thick layer of returning customers
* Active sites per license (is 1000 sites really the correct number)?
* USA as a country is missing from the transactions. Is it in a different system or simply was out because of the scramble?
* Comparing periods of time. Not feasible for this exercise (data for less than 2 months), but comparing to previous week, month, quarter, year, etc. are important (this is the reason for the flags in the dim\_date table)
* Transaction history. The data is the last status, but it could be that a changelof of the purchase is in order to understand processes within the ordering system