Client Communications

Consider the work you just completed in the Data Engineer Exercise. Please write no more than 2 paragraphs explaining to a client how you approached this task. Remember that clients do not need to know every process detail, but do want to understand how and why your choices contribute to our overall strategy and any benefits of the end product.

Thank you for the opportunity to work with you. According to the request received, three datasets were received, one from Constituent Information with 700,000 records and 29 columns, another from Constituent Email Addresses with 1,400,000 records and 16 columns and finally another with Constituent Subscription Status with 350,000 records and 6 columns. To answer the first question, We started with a merge between Constituent Information and Constituent Email Addresses using the cons_id key and using an inner join to guarantee all people with email. We have also only selected the records from cons_email where is_primary equals 1 according to requirement. With the above, a first version of the people dataset was obtained with the columns source, create_dt, modified_dt and e_mail in addition to the keys that were used to make the joins. With this result, a second merge was performed with the Constituent Subscription Status data using the cons_email_id key and a left join with which the isunsub column was added. We also filtered the dataset where chapter_id is 1 since we were given a requirement to remove all other chapte_ids in Constituent Subscription Status.

Finally, the null values in the isunsub column of the result were changed to 0 according to the requirement (If an email is not present in this table, it is assumed to still be subscribed where chapter_id is 1). With the above, the result had the necessary columns, so We proceeded to eliminate the unnecessary columns (cons_id and cons_email_id). Next, the type of data and the name of the columns were also verified, and they were changed according to the schema provided. Finally, the result with 605,639 records and six columns was exported to a .CSV file. For the second exercise, the previous results were grouped by days using the create_dt column and count using the email column to obtain the number of acquisitions for each day. Then the columns were renamed according to the scheme. Finally, the result with 18,445 rows and two columns was exported to a .CSV file.