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.

**Answer:** I identified that this task involves merging multiple spreadsheets and performing query operations. A brief survey of available python libraries led me to PySpark, third-party library which helps achieve data manipulation and SQL-like operations with ease.

The end product is a result of merging, querying and formatting the three files: *cons\_email\_chapter\_subscription.csv, cons\_email.csv and cons.csv*. This is achieved by identifying the unique field across the files and combining them based on the given conditions to generate aggregate datasets: *people.csv* and *acquisition\_facts.csv.* Of note is the custom function *save\_spark\_df,* which saves the result into a spreadsheet file. This replaces the available save function from PySpark because it generates temporary files without deleting them. This custom function, therefore, deletes temporary files and keeps the working directory uncluttered.