**Data Analysis and Integration**

**Data Investigation**

The first step is understanding the data. This involves:

1. **Exploring Each Dataset**: I identified the key columns in each file, understood the structure of the data (common columns: company\_name, domain, phone, address), and inspected for missing or insignificant values.
2. **Finding Common Points**: I examined the overlapping columns and assessed their strength and importance for the dataset. For example, ‘domain’ is a crucial factor for this dataset.
3. **Identifying Potential Conflicts**: I considered what conflicts might arise. The next step is to address missing values or highlight them, as they need to be taken into account. We can eliminate duplicates, and if there are essential columns containing missing data, a thorough check is required. Additionally, I noted any data with incorrect formats that need to be standardized (e.g., addresses and phone numbers).

I will use the ‘domain’ column as the primary merging key because it is a unique identifier for companies and is quite consistent. It directly corresponds to each company and allows us to combine information from all three sources.

I decided to keep the final dataset’s information for category, address, phone, and company\_name. In case of conflicts, I established a hierarchy of trust for each site to obtain a clearer result.

The third step involves analyzing the input data and identifying available information. Specifically, I performed descriptive statistics for each dataset, examined the distribution of values for patterns, and identified conflicts.

**Code Analysis:**

* The on\_bad\_lines parameter is used to ignore defective lines.
* I resolved the difference in delimiters between the datasets (using delimiter=';').
* I started with an overview of the data.
* I standardized the column names to maintain a uniform format across all DataFrames, facilitating easier merging.
* I used the phone number as an example for cleaning this column, eliminating missing values and non-numeric characters from the phone numbers to achieve a standardized numeric format.
* Company names are converted to lowercase to avoid duplication or confusion.
* I retained all records from the three DataFrames after merging using the domain\_name column.
* I combined the categories from the three sources, prioritizing them (Facebook > Google > Website).
* The same approach was applied to phone numbers and addresses.
* I eliminated columns that were used in the process and are no longer needed, such as ‘category\_facebook’, ‘category\_google’, etc.
* Finally, I exported the data into a final CSV file.