Big Data Engineer Assignment Solution

The decisions I have made in order to join the 3 datasets are:

1. What column will you use to join?

Answer: I selected the **'domain'** column as the primary joining key. This choice is grounded in the notion that a domain is typically unique to a company.
Assuming all three datasets from Facebook, Google, and the Company Website consistently use the same domains for corresponding companies, the 'domain' column seems to be an apt primary key for merging.

2. If you have data conflicts once you join, which one do you believe?

Answer: To address data conflicts, it's essential to rank the data sources based on their perceived reliability. In this situation:

Company Website: Given that companies directly oversee their websites, the data from this source is presumed to be the most accurate and current, making it the top priority.

Google: Ranking second in reliability, Google's vast data scraping and validation systems offer a trustworthy data set.

Facebook: This platform is last in the trust hierarchy. The rationale is that its data might derive from user-generated content or business profiles that aren't updated as frequently.

I would use the following priority order to pick the first available non-null value:
Company Website > Google > Facebook.
This sequence is premised on the belief that data from a company's website is more dependable than data from third-party platforms such as Google or Facebook.

3. If you have very similar data, what information will you keep?

Answer: When confronting similar data points, the objective is to conserve the most precise and detailed information:

If the same data is presented in two datasets, but one offers a more comprehensive version, the richer dataset should be favored.

For minor discrepancies in similar data (like phone numbers or addresses with different formatting), it's essential to standardize this information to a unified format, retaining the more detailed variant.

If two datasets present entirely disparate data for a specific column (like two distinct phone numbers attributed to one company), both entries should be kept for subsequent manual review unless a clear priority hierarchy exists.

Data Cleaning: The data was further refined using the read_csv and lowercase_and_strip functions for better consistency and to handle potential discrepancies.

Performance: Although I've employed Pandas for this solution, using Apache Spark could offer swifter processing for larger datasets. Performance enhancements can further be explored for improved efficiency.