Data Integration and Cleaning Problem

A) Relevance

- **Ability to handle real-world data integration challenges.**
- **Testing analytical skills in making decisions dealing with data conflicts and inconsistencies.**
- **Managing data and producing clean, reliable datasets.**

B) Outcome

- **Enhanced data quality**, which will improve decision-making.
- **Unified view of the companies**, consolidating analytics and insights.
- **Understanding one's market dynamics and customer behavior.**

C) Initial Steps

- Checking the number of rows and columns.
- Identifying common and unique column names.
- Verifying data types to identify potential inconsistencies.

D) Key Areas to Watch

- **Missing values analysis.**
- **Uniqueness and duplicates.**
- **Value counts.**
- **Data distribution** (e.g., anomalies in phone numbers).

E) Data Cleaning

- **Handling missing or inconsistent data.**
- **Standardizing data formats.**
    - Ensuring a consistent format for websites and domains.
    - *Note:* Extract domain names if needed for joining.
- **Fields to be standardized:**
    - Company names
    - Websites and domains
    - Addresses
    - Phone numbers
    - Categories
- **Removing errors or noise.**

    - Investigate data points that deviate significantly from the norm.
- **Performing consistency checks:**
    - Ensuring a match between phone number area codes and the regions in the address.
    - Verifying that there is a correspondence between the website domain and the company name.

F) Data Joining

- **Choosing the columns for joining.**
- **Selecting Primary Keys (PK) and Foreign Keys (FK).**
    - Use a combination of the website domain and the company name as PK to increase match accuracy and mitigate incorrect joins due to data inconsistencies.
    - *Scenario:* If the domain data is incomplete, this combination should fill the gaps.
- **Handling data conflicts by establishing source reliability and a conflict resolution strategy.**

i) Source Reliability – Interpretation

- **Facebook Dataset (+):**
    - Updates
    - Customer interaction
    - Social presence
- **Google Dataset (+):**
    - Reliable and frequently updated
    - Good for categorizing
- **Website Dataset (+):**
    - Reliable for official information (e.g., address)

ii) Conflict Resolution – Strategies

- Prioritize by reliability.
- Apply majority voting.
- *Note:* Document decisions by keeping a record of how conflicts were resolved for future reference.

G) Final Dataset – What to Keep

i) Important Fields

- **Company name**, for identification.
- **Website**, for online presence.
- **Address components**, for location-based analysis.
- **Phone number**, for contact.
- **Category**, for industry classification.

ii) Additional Fields – Enhancing Dataset Utility

- **Operational hours.**
- **Reviews** (from Google), if available.
- **Social links** (from Facebook).

H) Data Matching

- **Exact matching** using website domain and company name.
- **Fuzzy matching.**
- **Blocking techniques.**
    - *Note:* Reduce computational load by grouping by regions or categories.

I) Merging the Datasets

- Start by merging datasets with exact matches of the website domain and company name.
- For unmatched records, apply fuzzy matching on company names and addresses (second merge).
- Resolve conflicting data by applying different strategies.
- Include unique records to preserve available information.
- Ensure data integrity by verifying logical consistency.
- Check for duplicates or omissions.
- Use cross-validation with external data if possible.

J) Final Dataset

- Use clear column names with correct data types for future processing.

Coding – Quick Guide

- **Identifying common fields.**
- **Assessing data quality.**
- **Standardizing domain fields.**
    - *Note:* Reconstruct the website domain by combining multiple columns from the website_dataset.
    - Normalize domains across datasets by:
        - Converting domains to lowercase.
        - Removing the prefix (e.g., www).
        - Ensuring consistent formatting.
- **Cleaning company names** – standardization.
- **Standardizing addresses.**
- **Normalizing phone numbers.**
- **Standardizing categories.**
- **Primary Key (PK):** Domain.
- **Foreign Key (FK):** Company name (apply fuzzy string matching if differences occur).
    - *Note:* Apply tertiary join keys using address and phone for more accuracy.
- **Handling data conflicts by:**
    - **Establishing source reliability:**
        - **Category:** Google > Facebook > Website
        - **Contact:** Website > Google > Facebook
        - **Descriptions:** Facebook > Google > Website
    - **Setting up a conflict resolution strategy:**
        - Defining a source priority for each field.
- **Performing an inner join** on the standardized domain field.
- **Handling missing domains.**
- **Applying fuzzy matching** on company names.
- **Merging records.**
- **Constructing the final dataset.**
- **Converting columns' data types.**
- **Ensuring quality** by inspecting the records and checking for duplicates.