Solution to Big Data Engineer Assignment

The Datasets

Dataset from 3 different sources are given as below:

1. Facebook (facebook dataset.csv)

2. Google (google dataset.csv)

3. Company Website (website dataset.csv)

Data Observation and Cleansing

It has been observed that in google dataset, the field address has comma within col values and also comma is file delimiter, this ends up in data cleansing to get better results for joining.

Solution

1. What column will you use to join?
   *Most appropriate column for joining would be which is having **unique values, minimum nulls and duplicates, avoiding data skewness** and that column is identified as domain column- domain column from google data set == root_domain col in website dataset == domain col in Facebook dataset to join.*

2. If you have data conflicts once you join, which one do you believe?
   I would prefer defining rules for resolving the conflicts, below are the factors by which I shall be able to single out such records.
   - *The origin of data and its degree of truth will play an important role here*
   - *Will maintain all versions of data from there sources and will analyse which source is creating more conflict and that source can be flagged after analysis.*
   - *The majority which shows correct data out of three data sources will help in singling out the conflicting record.*

3. If you have very similar data, what information will you keep?
   - *Will create golden data record out of all three data sets, I will consider deduplication and merging of records based on the rules defined like keeping columns from datasets which are mostly populated with full values or maximum length or most frequently occurring and avoiding nulls.*
   - *To keep the phone or address information correct, overwriting will be considered based on the rules which fields to be overwritten from and to like select fields to overwrite from (longest/shortest, maximum/minimum, etc.). For example, we can choose to overwrite the Address of the golden record from the longest address in duplicate records.*
   - *The website data can be considered as base data or main source of data but to furnish more information against each record, google dataset seems to contain detailed information and Facebook dataset here can be used to enrich overall data by adding more columns from it to the main data.*