# Soleadify Test Project

# Intro

At the beginning, we have 3 datasets to be used as a source.
The given datasets are:

- facebook_dataset.csv
- website_dataset.csv
- google_dataset.csv

These data contain information about companies from different sources like Facebook, Google and websites. The main task of this project is to combine data from 3 given input datasets in a most accurate way, resulting in the 4th dataset with cleaned and transformed data.

# RAW part

Reading source data to create well-formed CSV files. This part is needed as Facebook and Google datasets are stored in a way that quoted fields has separators inside its data and this can lead to fields’ number explosion if to read in a wrong way.

# TRANSFORM part

Applying needed transformations and data cleaning such as:

- Columns renaming
- Lowercase columns
- Escape special characters, spaces
- Removing suffixes like (ltd, co ...)  in company names
- Filtering based on regular expressions
- Split and Explode fb_categories column into multiple rows
- Dropping NaN
- Cast column types
- Dropping duplicates

# TRANSFORMED DATASETS’ STRUCTURE

# WEB

web_domain
web_domain_suffix
web_language
web_company_name
web_city
web_country
web_region
web_phone
web_site_name
web_tld
web_category

# FACEBOOK

fb_domain
fb_address
fb_city
fb_country_code
fb_country
fb_description
fb_email
fb_link
fb_company_name
fb_page_type
fb_phone
fb_phone_country_code
fb_region_code
fb_region
fb_zip_code
fb_category

# GOOGLE

gg_address

gg_category

gg_city

gg_country_code

gg_country

gg_company_name

gg_phone

gg_phone_country_code

gg_raw_address

gg_raw_phone

gg_region_code

gg_region

gg_text

gg_zip_code

gg_domain

# PROCESS part

Joining 3 datasets into one based on domain, company name and phone columns to get final dataset with data from all sources.

# Results

- Cleaned and transformed data
- Joined data based on a custom strategy
- Got final dataset containing all the sources data
- Got data about 27288 companies from different sources

Tap to see full report about the project

# Soleadify Test Project