Data Engineer Task

Introduction

An e-commerce shop would like to onboard new suppliers efficiently. To enable the onboarding process, the customer needs us to integrate product data from suppliers in various formats and styles into the pre-defined data structure of their e-commerce shop application.

Input files

Supplier: supplier_car.json
Target: Target Data.xlsx (database dump file)

Tasks

Your goal is to write a Spark Pipeline that transforms the supplier data so that it could be directly loaded into the target dataset without any other changes. Your pipeline needs to include the code of steps 1 to 4 below.

1) Pre-processing

Here you need to load the data into a dataset and transform the supplier data to achieve the same granularity as the target data. (Hint: how many rows per product do you have in the target data?)
Be aware of character encodings when processing the data.

2) Normalisation

Normalisation is required in case an attribute value is different but actually is the same (different spelling, language, different unit used etc.).
Example: if the source encodes Booleans as "Yes"/"No", but the target as 1/0, then you would need to normalise "Yes" to 1 and "No" to 0.
Please normalise at least the following supplier attribute:

- "BodyColorText": needs to be translated into English and to match target values in target attribute "color"
- "MakeText": needs to be normalised to match target values in target attribute "make"

Input: pre-processed data
Output: normalised supplier data

3) Extraction

Some relevant features for the product can be extracted from supplier attributes and stored in new attributes.
Please extract at least:

- The value of the consumption from the supplier attribute "ConsumptionTotalText" into an attribute called: "extracted-value-ConsumptionTotalText"
- The unit of the consumption from the supplier attribute "ConsumptionTotalText" into an attribute called: "extracted-unit-ConsumptionTotalText"

Input: normalised supplier data
Output: extracted supplier data

4) Integration

Integration is to transform the supplier data with a specific data schema into a new dataset with target data schema, such as to:

- keep any attributes that can be mapped to the target schema
- discard attributes not mapped to the target schema
- keep the number of records from the supplier data as unchanged

Please integrate the dataset using at least the 5 attributes mapping as follows: (structure is given as "supplier attribute" ⬄ "target attribute")

- the normalised "BodyColorText" ⬄ "color"
- the normalised "MakeText" ⬄ "make"
- "ModelText" ⬄ "model"
- "TypeName" ⬄ "model_variant"
- "City" ⬄ "city"

Input: extracted supplier data
Output: integrated supplier data

5) Product matching, enriching existing products, adding new products

Product matching is to identify:

- products from the supplier file that are already existing in the target file / target database. In this case, we would need to enrich the target data with additional supplier information
- new products from the supplier file that are not yet in the target file / target database. In this case, all these new products should be added to the target database

Without implementing/coding any of this step, please come up with a proposal of how you would approach this task (ideas/ways to solve product matching, challenges, data flow and methodology/steps).

Deliverables

A Git project or a ZIP file containing:

- 4 CSV files, using encoding UTF-8, containing the results of steps 1 to 4 (i.e., pre-processing/normalisation/extraction /integration)
- A Spark Pipeline (in any language of your choice) that can be executed to provide the above CSV files
- A data-flow presentation to illustrate the above processing steps 1 to 5 as a data-flow, to show the logic applied in this onboarding