# Servoo Data Engineering Practical Test

Servoo is building a digital grocery platform. We receive raw product catalogs from many suppliers in different formats. Your task is to clean, standardize, and consolidate the four CSV catalogs we provide, and build a small ETL + reporting pipeline.

## Files Provided

* Amal Trading - Sheet1 (2).csv
* CATALOG - Future.csv
* CATALOG - RED FROZEN.csv
* CATALOG - CHETTIOT.csv

Each file may use different column names and have inconsistent or missing data.

## Tasks

### 1️⃣ Extract

Import all four CSV files.  
Standardize them into one clean table with the columns:  
  
Product\_ID, Product\_Name | Serial\_Number | Supplier | Weight\_Quantity | Packaging\_Type (CTN / NON-CTN) | Units\_Per\_Carton

Products will have weight/quantity in their name and maybe carton pieces. For example, PRIME CHICKEN BURGER 1KG X10, the weight/quantity = 1KG while units per carton = 10.  
  
(Map, create and rename columns as needed — some files might not have all fields or might use different names, Product\_ID is to be created by you, use any)

### 2️⃣ Transform (Data Cleaning & Quality Checks)

Identify and report:

• Products with missing Serial\_Number.

• Products with Duplicate Serial\_Numbers.

• Correct Weight/Quantity inconsistencies (e.g., “1kg”, “1000 g”).

Interpret packaging info:

- Identify which products are Carton (CTN) and which are Non-Carton.

- Extract Units\_Per\_Carton where available (e.g., “+30” = 30 pieces per carton).

### 3️⃣ Load (Master Database)

Create an SQL master product table to store the cleaned data, for example:  
  
Product\_ID (PK)  
Product\_Name  
Serial\_Number  
Supplier  
Weight\_Quantity  
Packaging\_Type (CTN / NON-CTN)  
Units\_Per\_Carton  
Source\_File  
Last\_Updated  
  
Load your cleaned and standardized data into this table.

### 4️⃣ Analysis & Reporting

Use SQL or Power BI to answer:

• Total number of products.

• Number of CTN vs NON-CTN products.

• Total units available (cartons × units per carton + individual).

• Count of missing / duplicate Serial\_Numbers.

• Supplier-wise product counts.

• Products appearing in multiple catalogs vs unique ones.

### 5️⃣ Visualization

Create a Power BI dashboard showing:  
• CTN vs NON-CTN distribution.  
• Missing/duplicate serial number counts.  
• Supplier-wise product counts.  
• Overlap of products across the four catalogs.  
• Weight/quantity distribution (e.g., most common pack sizes).

### 6️⃣ Bonus (Optional)

Explain how you would automate & scale this ETL process when Servoo has thousands of supplier files (e.g., Python pipelines, SQL, Airbyte/n8n, Azure Data Factory, or scheduled Power Query refresh).