# Servoo Data Scraping Practical Test

Servoo is building a digital grocery platform that compares and consolidates supplier catalogs with online product listings (e.g., supplier sites, e-commerce stores, and B2B sources).   
You are tasked to design and demonstrate a small scraping & enrichment pipeline.

## 📁 Files Provided

You will be provided with a sample product list

## ⚙️ Tasks

### 1️⃣ Extract (Web Scraping)

Scrape or query public websites to find matching products by name.  
Example sources: Carrefour UAE, Lulu Hypermarket, Tradeling, Noon, Amazon.ae or manufacturer websites.  
For each product, extract:  
• Product Name, Brand name  
• Product Description  
• Price (AED preferred)  
• Image URL (main product image)  
• Manufacturer Barcode / GTIN / EAN   
• Source URL  
If a product is not found, mark it as “Not Available”.

### 3️⃣ Load (Database / Output)

Prepare a clean output file:  
Output\_File: Servoo\_Scraped\_Data.csv  
Columns:  
Scrape\_ID | Serial\_Number | Product\_Name | Matched\_Product\_Name | Description | Price (AED) | Image\_URL | Barcode | Source\_Website | Last\_Updated  
Load the data into a SQL table for queries.

### 4️⃣ Analysis & Reporting

Provide short reports or queries to answer:  
• Total products successfully matched.  
• Count of products with/without barcodes or images.  
• Common missing fields (e.g., description or barcode).

### 5️⃣ Visualization (Power BI or Python)

Create simple visualizations showing:  
• Products with missing barcodes or images.  
• Source websites contributing the most matches.

### 6️⃣ 🧩 Bonus — Automation & Scalability

Describe how you would:  
• Scale this scraping pipeline for hundreds of suppliers and thousands of products daily.  
• Automate scraping orchestration using Crawlee / Playwright / Selenium / Apify / n8n / Airflow.  
• Use Proxies / Captcha handling for anti-bot sites.  
• Schedule scrapes periodically and feed results into the master product catalog from the ETL test.