### **1. Objective**

Develop an automated ETL pipeline to collect, process, and merge Egyptian tariff data from official government sources into a structured dataset for trade analysis, including:

* HS code classification
* Customs duty rates
* VAT specifications
* Trade agreement terms

### **2. Data Sources**

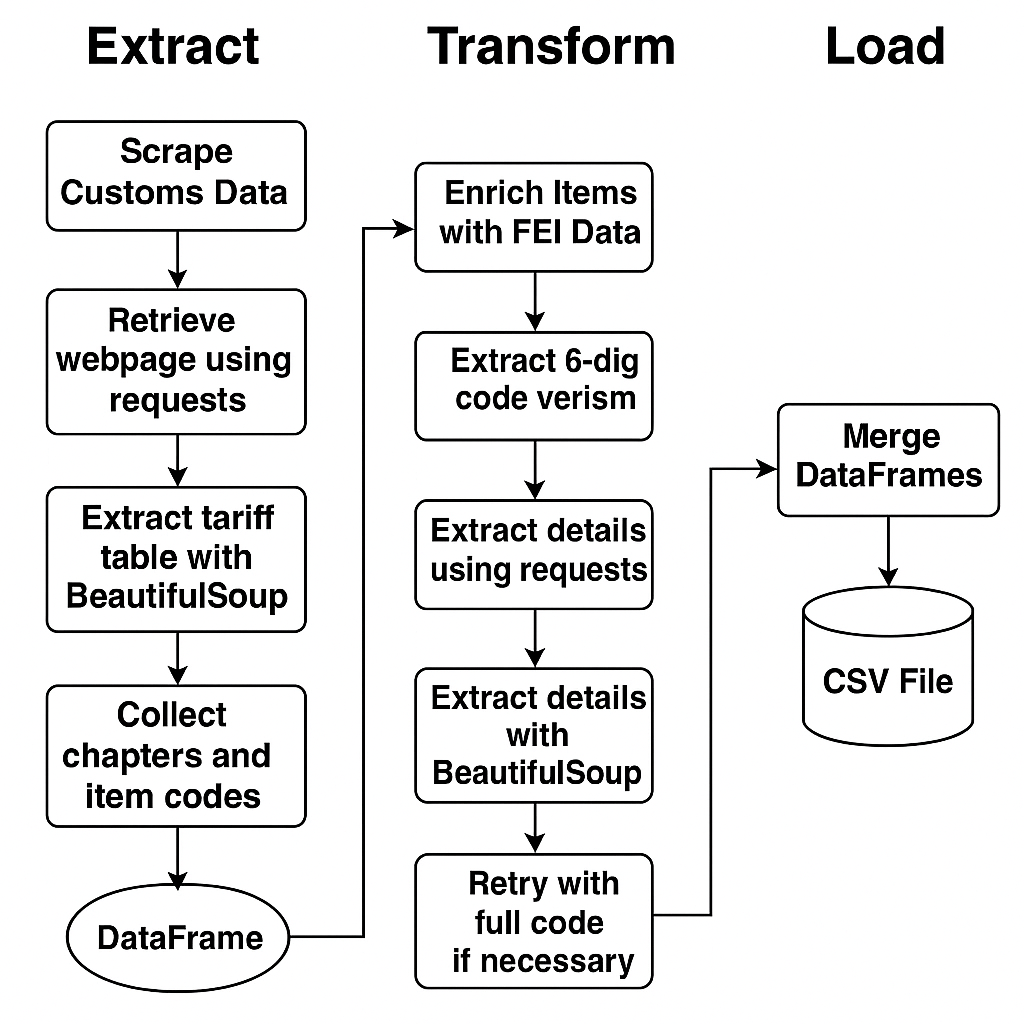
#### **Primary Source (Customs.gov.eg)**

* **Website**: [Egyptian Customs Tariff Portal](https://customs.gov.eg/Services/Tarif)
* **Data Extracted**:
  + HS Codes (Item)
  + Chapter metadata
  + Page reference
* **Structure**:
  + Paginated tables (99 chapters × 100 pages max)
  + HTML tables with class table

#### **Secondary Source (FEI.org.eg)**

* **Website**: [FEI Tariff Lookup](http://www.fei.org.eg/tariff/tariff.php)
* **Data Extracted**:
  + Product descriptions
  + Measurement units
  + Customs fee percentages
  + VAT rates and conditions
  + Trade agreements (e.g., GAFTA, COMESA)
* **Structure**:
  + Unstructured HTML
  + Key fields in <div class="span6 content">
  + Agreements in paired <div class="span6"> elements

### **3. Pipeline Workflow + Process flow diagram**



#### **Phase 1: HS Code Extraction (Customs.gov.eg)**

**Input**: Dynamically generated URLs (chapter + page combinations)  
**Process**:

1. Iterate through chapters (1–99) and pages (1–100).
2. For each URL:
   * Send HTTP request, parse HTML with BeautifulSoup.
   * Detect valid tables via table class.
   * Skip if table is empty or malformed.
   * Extract Item (HS Code) and clean (remove /).
   * Store with Chapter and Page metadata.
3. Terminate chapter scrape if no data found.  
   **Output**: DataFrame with:

* Item (HS Code)
* Chapter
* Page

#### **Phase 2: Tariff Detail Scraping (FEI.org.eg)**

**Input**: HS Codes from Phase 1.  
**Process**:

1. For each HS Code:
   * Fetch URL: http://www.fei.org.eg/tariff/tariff.php?hscode={HS\_CODE}.
   * **Field Extraction**:
     + **Description**: Text after Description: in div.span12.
     + **Unit/Custom Fee/VAT**: Values from labeled div.span6 elements.
     + **Agreements**:
       - Extract pairs of agreement names + rates.
       - Filter for Egypt-specific clauses.
   * **Fallback Logic**:
     + Retry with 6-digit HS code if:
       - All fields = "N/A", **or**
       - VAT missing but other fields exist.
   * Default missing data to "N/A".  
     **Output**: Dictionary per HS Code → Combined into DataFrame.

#### **Phase 3: Data Consolidation & Export**

**Input**: HS Codes (Phase 1) + Tariff Details (Phase 2).  
**Process**:

1. Merge DataFrames on Item (HS Code).
2. **Data Formatting**:
   * Wrap Item in quotes ("0101210000") to prevent Excel truncation.
   * Standardize numeric fields (e.g., 5% → 5).
3. Export to CSV with:
   * quoting=1 (Excel-compatible).
   * UTF-8 encoding.  
     **Output**: egyptian\_tariff\_data.csv with:

* HS Code (Item)
* Chapter/Page metadata
* Description
* Customs Fee (%)
* VAT (%)
* Agreements

### **4. Error Handling**

* **Empty Pages**: Skip after detection.
* **HTTP Errors**: Retry (3×), then log.
* **Missing Fields**: Default to "N/A".
* **VAT Inconsistencies**: Trigger 6-digit fallback.