**Egyptian Customs Tariff Data Pipeline Design**

### **1. Objective**

Develop an automated data pipeline to scrape, process, and consolidate Egyptian tariff, VAT, and trade agreement information from official government sources into a structured dataset for analysis.

### **2. Data Sources**

* **Primary Source (Customs Data)**:
  + Website: https://customs.gov.eg/Services/Tarif
  + Data: HS Codes (Item), Chapter, and Page metadata
  + Structure: Paginated tables (chapters 1–99, pages 1–100 per chapter)
* **Secondary Source (FEI Tariff Details)**:
  + Website: http://www.fei.org.eg/tariff/tariff.php
  + Data:
    - Product **Description**
    - **Unit** of measurement
    - **Custom Fee** (%)
    - **VAT** details (rate + conditions)
    - **Agreements** (trade deals and exemptions)
  + Structure: Unstructured HTML with key fields in <div class="span6 content">

### **3. Pipeline Workflow**

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

* **Input**: None (seed URLs generated dynamically)
* **Process**:
  1. Iterate through chapters (1–99) and pages (1–100).
  2. For each URL, scrape the HS Code table:
     + Skip if table is empty.
     + Extract Item (HS Code) and remove / characters.
     + Store with Chapter and Page metadata.
* **Output**: DataFrame of HS Codes (Item) + metadata.

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

* **Input**: HS Codes from Phase 1.
* **Process**:  
  For each HS Code:
  1. Fetch the tariff details page.
  2. Extract fields using **defensive scraping**:
     + **Description**: From div.span12 after "Description" label.
     + **Unit/Custom Fee/VAT**: From div.span6 with matching labels.
     + **Agreements**:
       - Pair "Agreement" and "Rate" values from adjacent div.span6.
       - Filter to keep only text from "Egypt" onward.
  3. Handle missing data: Default 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. **Preserve HS Code formatting**:
     + Wrap Item in quotes ("0101210000") to prevent Excel zero-truncation.
  3. Export to CSV with quoting=1 for Excel compatibility.
* **Output**: egyptian\_tariff\_data\_.csv.