# ***Egyptian Customs Tariff Web Scraper ETL Documentation***

*This Python script scrapes tariff data from two Egyptian government websites and combines the results into a structured dataset. The pipeline consists of three core components: HS code extraction, tariff detail enrichment, and data consolidation.*

## ***1. Core Functions***

### *scrape\_customs\_data()*

***Purpose:  
Extracts HS codes from the Egyptian Customs Authority portal (customs.gov.eg).***

***Functionality:***

1. ***URL Generation:***
   * ***Constructs URLs for chapters (1–99) and pages (1–100) using***
   * ***"https://customs.gov.eg/Services/Tarif?page={}&type=1&chapterId={}"***
2. ***Table Scraping:***
   * ***Identifies tables with class="table".***
   * ***Skips empty/malformed tables.***
3. ***Data Extraction:***
   * ***Cleans HS codes (removes /).***
   * ***Captures Chapter and Page metadata.***

***Parameters: None***

***Returns:  
A pandas.DataFrame with:***

| ***Column*** | ***Type*** | ***Example*** |
| --- | --- | --- |
| ***Chapter*** | ***int*** | ***1*** |
| ***Page*** | ***int*** | ***5*** |
| ***Item*** | ***string*** | ***"0101210000"*** |

### *scrape\_fei\_data(hscode)*

***Purpose:  
Fetches detailed tariff data from the Federation of Egyptian Industries (fei.org.eg).***

***Functionality:***

1. ***URL Construction:***
   * ***Formats URLs as http://www.fei.org.eg/tariff/tariff.php?hscode={HS\_CODE}.***
2. ***Field Extraction:***
   * ***Description: Text after Description: in div.span12.***
   * ***Unit/Fee/VAT: Values from labeled div.span6 elements.***
   * ***Agreements: Parses adjacent div.span6 elements for Egypt-specific terms.***
3. ***Fallback Logic:***
   * ***Retries with 6-digit HS code if:***
     + ***All fields are "N/A", or***
     + ***VAT is missing but other fields exist.***

***Parameters:***

* ***hscode (str): HS code to query (e.g., "0101210000").***

***Returns:  
A dictionary with:***

| ***Key*** | ***Type*** | ***Example*** |
| --- | --- | --- |
| ***Description*** | ***string*** | ***"Live horses"*** |
| ***Custom Fee*** | ***string*** | ***"5%"*** |
| ***VAT*** | ***string*** | ***"14% (Standard)"*** |
| ***Agreement details*** | ***string*** | ***"GAFTA: 0%"*** |

## ***2. Execution Pipeline***

### ***Processing Steps***

1. ***HS Code Extraction:***
   * ***Calls scrape\_customs\_data() to retrieve base HS codes.***
2. ***Tariff Enrichment:***
   * ***For each HS code, runs scrape\_fei\_data().***
   * ***Applies fallback logic for incomplete data.***
3. ***Data Merging:***
   * ***Combines results into a final DataFrame.***
4. ***Export:***
   * ***Wraps Item in quotes ("0101210000") to preserve leading zeros in Excel.***
   * ***Saves to egyptian\_tariff\_data\_quoted.csv with quoting=1.***

***Output Insights:***

* ***Coverage: 7,351/8,795 HS codes have published tariff data (as of 2025).***
* ***Missing Data: "N/A" indicates unavailable fields.***

## ***3. Technical Features***

| ***Feature*** | ***Implementation*** |
| --- | --- |
| ***Dual-Source Integration*** | ***customs.gov.eg (HS codes) + fei.org.eg (details)*** |
| ***Pagination Handling*** | ***Auto-detects empty pages/chapters*** |
| ***Error Resilience*** | ***Retries failed requests; defaults to "N/A"*** |
| ***Excel Optimization*** | ***Quotes HS codes; uses CSV quoting=1*** |

## ***4. Error Handling***

| ***Error Type*** | ***Response*** |
| --- | --- |
| ***HTTP Errors (4xx/5xx)*** | ***Logs error; skips after 3 retries*** |
| ***Missing Fields*** | ***Assigns "N/A"*** |
| ***Malformed HTML*** | ***Uses defensive parsing*** |

## ***5. Performance***

* ***Progress Tracking: Logs every 10 HS codes.***
* ***Efficiency: Skips empty pages early.***
* ***Dependencies:***

***requests # HTTP requests***

***BeautifulSoup # HTML parsing***

***# Data manipulation***

## ***6. Usage***

1. ***Run Script:***
2. ***Output:***
   * ***CSV: egyptian\_tariff\_data\_.csv***
   * ***Console: Progress updates + sample data preview.***

## ***7. Maintenance Notes***

* ***Monitor Changes: Verify HTML selectors (table.table, div.span6).***

***Note: Website structure changes may require selector updates.***