**Data Harmonization Flow**

1. **File Upload & Ingestion**

* User uploads files (.csv, .xlsx, .zip).
* For each file:
* Read the file.
* Keep track of the filename (but do NOT add it as a column yet).

1. **Shape Detection (AI)**

* For each file:
* Extract a sample (headers + a few rows).
* Send this sample to OpenAI for format detection.
* Receive the format type (e.g., "classic\_wide", "key\_value", etc.).

1. **Local Reshaping**

* For each file:
* Use the detected type to reshape the data into tidy long format.

1. **Column Harmonization (AI + Fallback)**

* Collect all unique variable/column names from all reshaped DataFrames.
* **4.1. Context-Aware Harmonization:**
* Send columns (with sample values and context) to OpenAI for harmonization.
* **4.2. Receive AI mappings** (with confidence scores and “unknown” flags).
* **4.3. Fuzzy Matching & Synonym Dictionary Fallback:**
* For any columns with low confidence or marked as “unknown” by the AI, use local fuzzy matching and your synonym dictionary to attempt to map these columns.
* **4.4. Merge AI and fallback mappings** to create the final mapping.

1. **Apply Harmonized Names**

* For each DataFrame:
* Rename columns/variables using the final mapping.

1. **Add Source Column to End (Just Before Merge)**

* For each DataFrame:
* Add a source column (with the filename) as the last column.

1. **Merge All DataFrames**

* Stack all DataFrames vertically (union columns, fill missing with 'NULL).

1. **Clean the Merged Master DataFrame**

* Apply value mapping rules (e.g., 10k → 10000, n/a → 'NULL', etc.).
* Standardize codes (e.g., always "Male"/"Female" for gender), strip spaces, ensure all empties are 'NULL'.
* Handle metadata columns (extract, fill, and remove metadata rows if needed).
* Infer and standardize data types (numeric, date, categorical, etc.).
* (Any other cleaning logic you want, now applied to the merged master.)

1. **Remove True Duplicates**

* Identify duplicates by unique identifier variables eg name ids country etc and time key(s).
* Keep the first occurrence; move extras to a “Duplicate Records” block.

1. **Comprehensive Auditing and Reporting**

* Output a summary report with:
* All harmonization decisions (original → mapped name, confidence score, fallback used).
* All cleaning actions (e.g., value replacements, type conversions).
* Any flagged issues (e.g., revenue < employees, missing key fields).
* Optionally, add a column to flag rows with suspicious or outlier values (e.g., suspicious column for revenue < employees).

1. **Export Results**

* Write a single MASTER.csv file:
* Top: clean dataset.
* After: # Duplicate Records section.
* Include summary/audit sections as comments or separate files if desired.
* All empties are empty, and source is always the last column.
* Send the file back to the frontend.

