**Final 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)**

* Collect all unique variable/column names from all reshaped DataFrames.
* Send these names to OpenAI for harmonization.
* Receive the mapping of original names → harmonized names.

1. **Apply Harmonized Names**

* For each DataFrame:
* Rename columns/variables using the harmonized 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 'NaN').

1. **Clean the Merged Master DataFrame**

* Apply value mapping rules (e.g., 10k → 10000, n/a → 'NaN').
* Standardize codes, strip spaces, ensure all empties are 'NaN'.
* Handle metadata columns (extract, fill, and remove metadata rows if needed).
* (Any other cleaning logic you want, now applied to the merged master.)

1. **Remove True Duplicates**

* Identify duplicates by firm\_id and time key(s).
* Keep the first occurrence; move extras to a “Duplicate Records” block.

1. **Audit Missing Data**

* Count 'NaN' in key fields (e.g., revenue, employees, sex).
* Issue warnings if needed.

1. **Export Results**

* Write a single MASTER.csv file:
* Top: clean dataset.
* After: # Duplicate Records section.
* All empties are empty, and source is always the last column.
* Send the file back to the frontend.

