AI4Trade — Data Cleaning & Integration Guide (2023–2025)

Repository Companion Document — Indian Westerlies

# 1) Overview

This guide documents the exact steps we used to clean, harmonize, and integrate monthly bilateral trade data for China and the United States, preparing six modeling-ready datasets for the next stage of feature generation. It is the first of a multi-document series intended for our GitHub repository so others can reproduce our setup.

Scope of this document:  
• Ingest OEC CSVs for 2023–2024 (very large files) with chunked reading and save as Parquet.  
• Ingest 2025 monthly Excel exports from TradeMap and normalize to the same schema.  
• Harmonize columns, datatypes, HS codes, flows, and months; aggregate provincial/state rows to national.  
• Integrate 2023–2025 into a unified master and materialize six datasets for modeling.  
• Join feature tables and produce explicit train/test splits (feature details come in a separate document).

# 2) Data Sources & Targets

Primary sources:  
• OEC-provided CSVs for 2023 & 2024 (state- and province-level) for USA and China.  
• 2025 monthly data downloaded as Excel from TradeMap (www.trademap.org).

Target schema (per row): (origin, destination, hs6, hs4, trade\_flow, month, value). All values are in USD at nominal prices. HS-4 is derived from HS-6; national totals are computed by summing state/province rows.

# 3) 2023–2024 Ingestion from Large CSVs

Files for 2023 and 2024 are ~2.8GB each per country/flow. We use pandas.read\_csv with chunksize to stream the data, select columns, enforce dtypes, and aggregate to national level before writing Parquet. Key practices:

• Use chunksize (e.g., 1–5 million rows depending on RAM).  
• Explicit dtype map for codes/IDs as strings to avoid scientific notation and preserve leading zeros.  
• Parse month as datetime using the first day of month convention.  
• Aggregate state/province to national by (origin, destination, hs6/hs4, trade\_flow, month).  
• Derive hs4 = hs6.str[:4].  
• Normalize trade\_flow to {'Export','Import'} and clip negatives to zero.  
• Save monthly-level national Parquet files, partitioned by origin/trade\_flow for faster later I/O.

# 4) 2025 Intake from TradeMap Excel

TradeMap exports (per origin × flow × month) are read from XLSX and conformed to the master schema. We standardize partner codes to ISO-3, clean headers (skip title rows/footers), remove subtotals, drop notes/footnotes, and coerce value columns to numeric. Month is aligned to the first day convention.

Common fixes:  
• Strip thousand separators and currency symbols.  
• Unmerge header cells and rename to canonical names.  
• Map partner names to ISO-3 (maintain a checked lookup table).  
• Ensure HS codes are left-padded to 6; derive hs4.  
• Drop aggregated 'World' totals — retain bilateral rows only.

# 5) Harmonization & National Aggregation

After independent ingestion, all datasets are conformed to the canonical schema and concatenated. Where the raw sources provide sub‑national rows, we sum to national level. We ensure one row per (origin, destination, hs6, trade\_flow, month). We also enforce monotonic monthly periods without gaps in the month index for each (origin, destination, hs6, trade\_flow) series (missing months imply zero).

# 6) Master Integration (2023–2025)

We stack 2023, 2024, and 2025 monthly tables to build a unified master parquet with the canonical columns. De-duplication keys are (origin, destination, hs6, trade\_flow, month); ties are resolved by keeping the maximum value after input source precedence (2025 TradeMap overrides if overlap occurs).

# 7) Deliverables: Six Modeling-Ready Datasets

From the unified master we materialize six artifacts used by downstream feature engineering:  
1) CHN\_export\_master.parquet — China→Partners (Exports, HS6, monthly 2023–2025)  
2) CHN\_import\_master.parquet — Partners→China (Imports, HS6, monthly 2023–2025)  
3) USA\_export\_master.parquet — USA→Partners (Exports, HS6, monthly 2023–2025)  
4) USA\_import\_master.parquet — Partners→USA (Imports, HS6, monthly 2023–2025)  
5) partner\_universe\_lookup.parquet — Valid partner ISO‑3 sets per origin×flow (filtered to OEC rules)  
6) hs\_mappings.parquet — HS6↔HS4 mapping table used for aggregation at submission time

# 8) Feature Join & Train/Test Splits (Preview)

Feature generation is documented separately. Here we record the integration and splitting policy so paths and filenames are clear:  
• Compute all features using data ≤ t−1; no lookahead.  
• Create horizon‑specific targets: h=2 for CHN (y\_target = y(t+2)), h=3 for USA (y\_target = y(t+3)).  
• Merge feature tables back to the four master segment datasets using keys (origin, destination, hs6, trade\_flow, month).  
• Produce explicit train/test split Parquets per segment & horizon: train rows have non‑null targets; test rows are the latest context month (Aug‑2025 for CHN, Jul‑2025 for USA) with y\_target intentionally null.

Naming convention (all with “\_final” suffix):  
• /data/features/features\_{segment}\_train\_h{2|3}\_final.parquet  
• /data/features/features\_{segment}\_test\_h{2|3}\_final.parquet  
Segments: CHN\_export, CHN\_import, USA\_export, USA\_import.

# 9) File Hygiene & Reproducibility

• Fix HS code formatting (string, zero‑padded to 6). Derive hs4 as first 4 chars.  
• Normalize trade\_flow labels to exact {'Export','Import'} casing.  
• Parse month as UTC‑naive first‑day timestamps; validate monotonicity per series.  
• Clip negatives to zero; coerce tiny non‑positive noise to 0 before sMAPE usage downstream.  
• Persist all intermediate outputs as Parquet with snappy compression.  
• Maintain logs for row counts, unique key checks, and month coverage per segment.

# 10) Validation Notes (Sanity Checks)

• Monthly national totals by origin match within tolerance across sources.  
• Top partners per month/year align with known trade patterns.  
• No duplicate keys after aggregation (origin, destination, hs6, trade\_flow, month) are present.  
• Spot‑check HS6→HS4 aggregation reversibility on random samples.

# Appendix A — Expected Directory Layout (Excerpt)

ai4trade/  
 data/  
 raw/ # harmonized national-level parquets (by origin/flow)  
 features/ # feature Parquets + explicit train/test splits  
 logs/ # ingestion logs & row counts  
 docs/ # this document + feature schema, workflow guide

All filenames for the final run include the suffix “\_final”.  
Feature definitions and modeling details are covered in separate documents.