Subnational Foreign Trade Data in Latin America: Diagnostic, Challenges, and Technical Pathways

Felipe Ramon de Britto Redondo[1], Orlando da Silva Junior[2]

[1]International Business Center (CIN) Department – Federation of Industries of the State of Minas Gerais (FIEMG) – Minas Gerais, MG - Brazil[1]

[2]Ibmec São Paulo – São Paulo, SP – Brazil

{feliperamondbr, orlandosilvajr}@gmail.com

Abstract. *National foreign trade data (a country's exports/imports) is widely available, but demand for the subnational level (states, municipalities) is growing. This brings volume and variety challenges, due to granularity and differing data handling across countries. This work analyzes the conditions of Latin American subnational foreign trade data to propose technical pathways for integration and standardization aimed at analytical uses. A layered minimal Lakehouse architecture, centered on Data Exploration and Understanding is employed, and case studies (Mexico and Uruguay) are presented. Findings reveal heterogeneity in data collection, compilation and dissemination, highlighting the need for harmonization to enable comparative analysis.*

1. Introduction

Foreign trade is understood as a country's exports and imports of goods and services, along with the regulations that govern them. Following the liberalization of global markets in the 1980s, merchandise trade grew from 28% to 46% of world economy between 1978 and 2023 [Cavusgil et al. 2014, World Bank 2025]. Managing all types of cross-border flows is vital for countries, and trade remains a fundamental one.

Since then, key standards to make trade data more detailed, comparable, and available were introduced, such as the Harmonized System (HS) by the World Customs Organization [WCO 1983] and the *International Merchandise Trade Statistics* (IMTS) manual by the United Nations [UN 2011]. The HS standardized product classification globally, while the IMTS provided methodological guidance for collecting, compiling and disseminating trade flows. Nonetheless, it stops at the national level.

Subnational exports and imports (by states, municipalities, customs zones) are largely missing from global platforms. UN Comtrade Database[2], the most comprehensive international repository, contains no records below the national level. Although it obtains data from national authorities, its extensive metadata survey[3], with hundreds of technical questions to reporting countries, makes no reference to the treatment of subnational data. This omission exposes a critical institutional and statistical gap.

Meanwhile, demand for comparable subnational data has grown across multiple fields, such as international market selection, regional competitiveness, economic

[1] This article reflects the authors' own views and is the result of their professional research. It does not represent the position or opinions of FIEMG, nor any official position of the institution.
[2] https://comtradeplus.un.org.
[3] https://comtrade.un.org/survey/Reports/byQuestion

simulation, and environmental modeling [Jiang et al. 2020; Wittwer 2022; Huber et al. 2023; Francioni and Martín 2024]. Such data enables more precise logistical and environmental analyses and supports regional market assessments within countries.

Despite growing interest, usage remains limited, and literature identifies several barriers such as cost and availability issues [Papadopoulos and Denis 2011]; *methodological nationalism* treating countries as homogeneous units [Kardes 2016]; and that using such data in models is highly data-intensive, which demands further technical and theorical advancements [Jiang et al. 2020].

Given this context, the following research questions (RQs) are central:

RQ1: What is the current state of availability, structure and quality of subnational foreign trade data?

RQ2: How can subnational foreign trade data be standardized and prepared for flexible and scalable analysis across different countries?

To address these, the study analyzes the availability, structure, and quality of Subnational Foreign Trade Data (SFTD), using Latin America as a regional case and proposing technical pathways for its standardization and integration for analytical use.

To support the investigative nature of this work, a layered minimal Lakehouse architecture was employed around exploration, testing and governance axes. Introduced by [Armbrust et al. 2021], Lakehouses combine the flexibility and low cost of Data Lakes with the management and performance features of Data Warehouses. [Schneider et al. 2024] redesign this concept to focus on a technology-agnostic, simple, and accessible architecture. For them, a minimal architecture comprises one component for storage, one for processing and a framework for metadata management and transactional consistency. Regarding storage, the implementation was run locally, DuckDB was used for processing, and Delta Lake served as the framework. The pipeline followed a layered Data Design Pattern to ensure flexibility, traceability and quality across staged transformations.

The three main contributions of this paper are: (i) a technical diagnostic of SFTD across Latin American countries; (ii) a proposal of technical pathways for SFTD standardization and integration; and (iii) a reproducible and accessible data integration framework designed to handle fragmented, complex and heterogeneous public sources.

The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 details the methodology and architecture; Section 4 presents and discusses key findings from the data exploration and handling; Section 5 explores two case studies (Mexico and Uruguay); and Section 6 summarizes the concluding remarks.

2. Related Work

Being a domain-oriented work, the main related literature groups into two main strands: (i) efforts to improve the quality and comparability of international trade statistics; and (ii) studies that explore the analytical potential of SFTD in various empirical contexts.

The first group of studies focuses on diagnosing structural issues in international trade data, emphasizing consistency, and comparability, particularly in UN Comtrade.

[Shaar 2019] identifies discrepancies in bilateral trade comparisons and proposes a reconciliation of UN Comtrade data guided by an index based on most reliable sources.

[Hu et al. 2022] discuss inconsistencies from internal reporting errors and note that comparing national original statistics with Comtrade's processed ones was hindered by data limitations. [Chen et al. 2022] highlight anomalies arising from methodological differences between countries and recommend in-depth, case-by-case examination. Despite these contributions, efforts are limited to national-level data, not reaching SFTD.

The second group of research illustrates the analytical potential of SFTD.

From early to recent reviews [Papadopoulos and Denis 1988, Francioni and Martín 2024], studies note that subnational level remains an untouched area in models to help firms select potential foreign markets. [Huber et al. 2023] build a subnational trade competitiveness index, however, instead of using actual SFTD, they estimate it by combining Comtrade data with subnational employment shares across industry groups.

[Wittwer 2022] disaggregates national trade data into regions by linking port-level flows to regional classifications. Still, this approach was limited to Europe and its regional data is available from a single source, Eurostat[4], a condition rarely found elsewhere. [Jiang et al. 2020] show that environmental models using actual subnational records outperform those relying on proxies but introduces technical and theorical challenges due to volume. More recently, [Gimenez-Perales 2024] and [Ito et al. 2025] use trade microdata to examine firm-level patterns within single countries. While the present study also incorporates firm-level data, it focuses on regional comparisons, therefore, it relies on aggregated SFTD when it is the only available.

3. Methodology

The methodology combines three components: (i) a conceptual basis rooted in international standards that establishes how trade data should be structured and interpreted across countries; (ii) the design of a layered minimal Lakehouse model; and (iii) a structured workflow for data exploration and integration.

3.1. Conceptual Basis: Trade Data Standards and Structures

The IMTS manual published by the United Nations [UN 2011] sets methodological principles for the collection, compilation and dissemination of trade statistics. As countries may adopt different approaches, comparability becomes a challenge [Chen et al. 2022], making IMTS the main reference for data harmonization and interpretation.

Key aspects are, for instance, *Partner Attribution*: for imports, some countries report the *country of origin*, where goods are produced, others the *country of purchase,* where the seller is located. For exports, some report the *country of consumption*, where the goods are expected to be used, others the *country of sale*, where the buyer is located. Alternative choices are *country of last known destination*, *country of shipment* or *country of consignment*. These choices carry distinct implications: companies may record transactions in countries different from where goods are actually produced, thus leading to significant inconsistencies in bilateral trade statistics [Chen et al. 2022].

Another concept is *Valuation*: For exports, the IMTS recommends *Free On Board* (FOB), the value of the goods at the port of departure. For imports, it recommends *Cost,*

[4] https://ec.europa.eu/eurostat

Insurance, and Freight (CIF), which is FOB plus international logistics costs up to the destination port. However, countries may deviate from this guidance [Chen et al. 2022].

A further notable standard is the Harmonized System (HS) [WCO 1983], which provides a hierarchical nomenclature of goods, with increasing levels of specificity. The first two digits (HS2) represent broad categories of products, totaling 97 sectors; the four-digit level (HS4) comprises over 1,250 more specific product groups; and the six-digit level (HS6) more than 5,000 detailed products. Figure 1 illustrates this structure, along with regional and national extensions that can be adopted by countries.

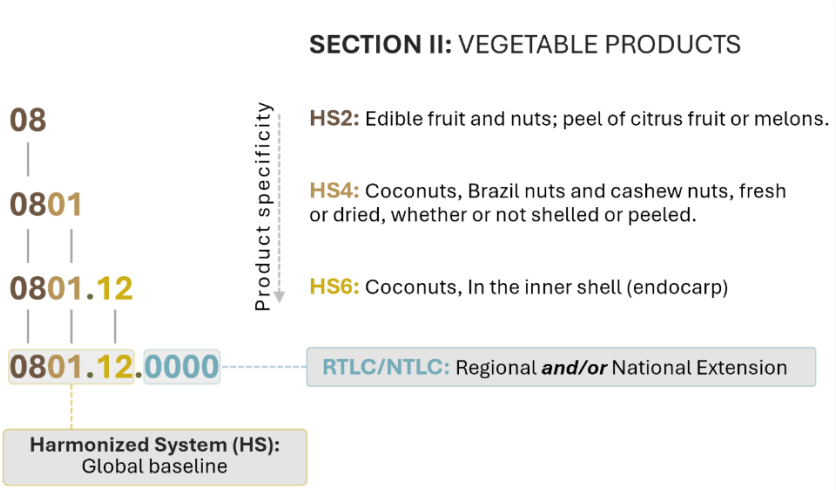

Figure 1. Structure of the Harmonized System and regional and/or national extensions.

Countries may adopt extra digits for regional or national extensions: a Regional Tariff Line Code (RTLC) or a National Tariff Line Code (NTLC). Brazil, Argentina, Paraguay and Uruguay use the Mercosur Common Nomenclature (NCM), extending HS6 to 8 digits. Colombia follows the Andean Community nomenclature (NANDINA), a regional eight-digit code also used by Bolivia, Colombia, Ecuador and Peru, but it adds two more digits to build its own 10-digit NTLC.

Since these extensions use the HS as a baseline, they can be decomposed into any HS level and even mapped to other international nomenclatures[5]. This work adds the International Standard Industrial Classification of All Economic Activities (ISIC), which groups goods by producing sectors, by using an HS6-ISIC correspondence built from an NCM-ISIC mapping developed by the Brazilian Secretariat of Foreign Trade[6].

For geographic units, ISO 3166-2 is applied, an extension of ISO 3166-1 that assigns internationally recognized codes to administrative subdivisions to the state level. This ensures consistent regional classification and enables multi-country comparisons.

3.2. Data Architecture and Design Choices

A local, layered minimal Lakehouse architecture was adopted. Given the domain-oriented and exploratory purpose, this choice prioritized flexibility, simplicity, and accessibility. Also, the project requires handling large data volumes and extensive testing, which brings the need for performance, quality control, and traceability throughout the transformations.

[5] https://unstats.un.org/unsd/classifications/Econ
[6] https://balanca.economia.gov.br/balanca/metodologia/Nota_ISIC-CUCI.pdf

A key advantage of the Lakehouse model is its adaptability to diverse analytical uses [Armbrust et al. 2021; Schneider et al. 2024]. Given the varied applications of SFTD, such as business intelligence, statistical models, and simulation, the architecture was built to accommodate multiple analytical demands from the start.

Inspired by the *minimal Lakehouse architecture* proposed by [Schneider et al. 2024], this work stores data locally and uses DuckDB, an embeddable analytical database, as the main processing engine. DuckDB was chosen for its OLAP optimization, native SQL and Delta support, low infrastructure requirements, and suitability for local, single-node environments, especially for interactive data analysis [Raasveldt and Mühleisen, 2019]. Python complemented transformations and handled extraction and orchestration.

Delta Lake framework guaranteed metadata and transaction management. It wraps Parquet files with a transaction log (Delta Log), forming a Delta table, that provides versioning, time travel (accessing previous states), and ACID guarantees while preserving performance. Since Parquet files are immutable (modifications require rewriting the entire file), Delta Lake avoids physical edits by writing new files with changes and updating the transaction log that points to them. It also supports concurrency, allowing multiple processes to read and write in parallel while ensuring only the latest committed version is visible, thus preventing conflicts and file corruption [Armbrust et al. 2020].

Finally, the pipeline adopts the medallion architecture as a Data Design Pattern to ensure quality and traceability across transformations. This approach organizes data into staged layers, where each stage incrementally refines structure and semantics. The specific processes in each layer are detailed in the following section.

3.3. Data Exploration and Integration Workflow

In line with [Volk et al. 2020], who emphasize the central role of Data Understanding and Testing in complex data workflows, the process model and integration workflow was structured around four intersecting axes as shown below in Figure 2.

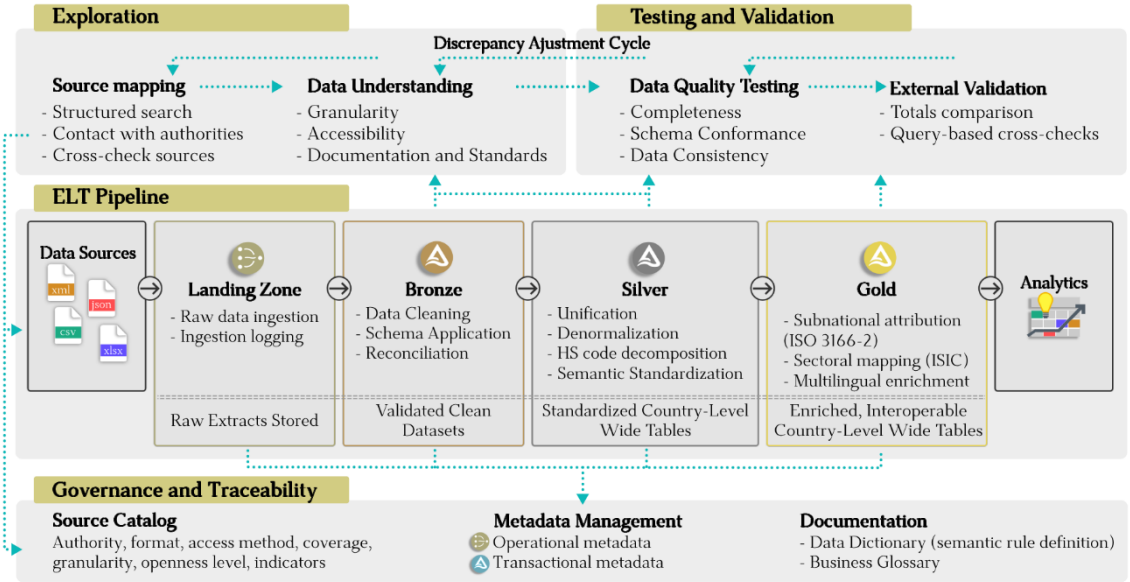

Figure 2. SFTD Process Model and Integration Workflow.

Exploration. This phase involved mapping SFTD sources across Latin America. Smaller islands (the Caribbean) and non-reporting countries to UN Comtrade (Venezuela, Cuba) were excluded. Searches covered open data portals, customs authorities, ministries, national statistics institutes, and other agencies, using boolean operators (*AND*, *OR*, -), wildcards (*exportacion**), filetype filters (*filetype:csv*), region-specific filters, and keywords (*Comercio Exterior*, *estadísticas*). Each source was assessed for disaggregation (geographic, product, partners), access method (APIs, scraped bulk/multiple downloads), documentation quality, and external validation resources.

ELT Pipeline. To handle several formats, a flexible ELT strategy was adopted: Loading precedes Transformation, contrasting with traditional ETL. Country-specific scripts were used, but modular components (processors, logging, downloaders, auxiliary parsers) supported reuse. Each script scans the current layer for the latest stored period and applies an overwrite window (e.g., 12 months) to capture retroactive updates. This is because customs declarations can be amended or canceled months later, often altering data beyond recent periods. Scanning is done in Python for the Landing Zone and with DuckDB's *delta_scan()* from Bronze onward, where data resides in Delta tables.

Processing is incremental and organized into logical chunks (monthly) to handle larger-than-RAM datasets efficiently. For example, *delta_scan()* combined with *predicate pushdown* (SQL *WHERE* filters) minimize memory usage by loading and transforming only the relevant partitions. Intermediate results are materialized as Arrow Tables, a general-purpose in-memory format for interoperability, used here because DuckDB lacked, to the moment, direct Delta write support.

Landing Zone. Country-specific Python scripts extract official data from sources collected during Exploration. They perform trigger-based incremental batch loads and generate extraction logs, storing raw data as retrieved (*schema-on-read*).

Bronze. Here, raw data from the Landing Zone is parsed and cleaned to ensure schema consistency before explicit type casting (*schema-on-write*) and conversion into Delta tables (Parquet + Delta Log). Most parsing is handled by DuckDB and its directory globbing capability that treats matching files as a single logical table, simplifying data management. Less supported formats such as XML and DBF are parsed in Python (*lxml*, *dbfread*, *Polars*). General cleaning and reconciliation of records were applied where raw microdata lacked prior preprocessing.

Silver. This layer produces country-specific, denormalized Wide Tables. Although it introduces some redundancy, this design minimizes the need for complex *JOINs*, enabling faster processing for BI tools, statistical models, and cross-validation with external sources. Transformations were executed via DuckDB's SQL and include:

(i) **Denormalization, decomposition and consolidation.** When source data is normalized, auxiliary dimension tables (e.g., for countries, products) are registered in DuckDB and *JOINed* via SQL to enrich the dataset. Product codes are decomposed into multiple levels (HS2, HS4, HS6, even RTLC, when applicable) to support multi-scale analysis. Imports and exports, if initially stored separately, are unified into consolidated datasets, with a new column (*operation*) distinguishing between the two flows;

(ii) **Semantic standardization.** Common field names were created based on IMTS [UN 2011] and insights from the exploratory phase. For example,

mapping *country of origin* (imports) and *country of consumption* or *last known destination* (exports) to *country_of_origin_or_destination*;

(iii) **Additional naming conventions**. Suffixes (*_id* for national codes, *_iso* for ISO 3166-1 country codes, *_desc* for descriptions) were applied, alongside prefix conventions for subnational divisions (*adm2_* for states, *adm3_* for municipalities). Geographic references vary by country methodology, as will be further explored in the results and case studies. To deal with it, standards were introduced, such as *adm2_comp* to indicate the state/department from where the company operated and *adm2_prod* for the state/department where a product was produced or destined.

Gold. To enable multi-level geographic and sectoral analysis, as well as to improve the datasets accessibility, this layer included the following enrichments:

(i) **Subnational and macro-region mapping.** ISO 3166-2 codes were added (*adm2_* prefix and *_iso* suffix for states/departments) and were grouped into macro-regions (e.g., Northwest, South) using the *adm1_* prefix;

(ii) **Customs offices to state/region mapping.** Customs offices refer to (*customs_location_desc*) where goods are inspected and cleared for import and export. When available, they were linked to their respective states/departments (*adm2_customs_desc*) and macro-regions (*adm1_customs_desc*), thus enabling three geographic references: *adm2_comp*, *adm2_prod* and *adm2_customs*;

(iii) **Sectoral classification.** ISIC codes derived from HS6 were added;

(iv) **Multilingual descriptive fields.** Official descriptions for countries, the HS and ISIC were added in Portuguese, English and Spanish, using suffixes *_desc_en*, *_desc_es*, and *_desc_pt*.

Testing and Validation. This axis involved structured quality checks for completeness, schema conformance, and data consistency, developed continuously throughout the data layers. These validations were executed in a separate testing repository (*data_quality_lab*) using DuckDB and native commands such as *SUMMARIZE*, which quickly produce descriptive statistics for all columns. External validation included replicating trade aggregates via queries and comparing results with official dashboards and UN Comtrade totals.

Governance and Traceability. This axis ensured transparency, reproducibility, and structured documentation. A Source Catalog described each dataset using attributes such as country, authority, format, access method, coverage, product and geographic granularity, openness level, and included indicators. Examples of geographic granularity classifications were *Level 2 (states/departments)*, *Level 3 (municipalities)* or *Others (ports/customs)*. For product granularity: *HS6* for Mexico, *NCM (HS6 + 2)* for Mercosur members, *NTLC (NANDINA + 2)* for Colombia. A Data Dictionary and Business Glossary documented naming conventions and domain concepts. Metadata included extraction logs from Python scripts and transactional records from Delta Lake.

Workflow and Discrepancy Adjustment Cycle. Finally, the arrows in Figure 2 illustrate the interactive and iterative nature of this process, inspired by [Volk et al. 2020]. Issues detected at any stage triggered a *Discrepancy Adjustment Cycle*, prompting revisions to exploration, transformation, or documentation layers as needed.

4. Results and General Diagnostic

This section presents a general diagnostic of SFTD in Latin America, highlighting institutional practices, data characteristics, and key technical challenges.

SFTD can be presented in two general forms: customs or statistical data. Trade data originates from mandatory import/export declarations collected by customs authorities. In this study, *customs data* stays closer to this microdata, offering high granularity but generally requiring preprocessing. *Statistical data* is aggregated and user-friendly but less detailed and often anonymized. For example, Chile's *customs data*[7] includes 178 columns covering monetary, logistical, and negotiation variables, while Mexico's *statistical data*[8] offers only FOB value for both exports and imports.

Dissemination practices vary widely. *Customs data* may or may not be published, and when they are, it's often on customs websites or via Open Data Portals, with limited metadata. *Statistical data*, on the other hand, is more clearly disseminated through official platforms, but generally published by several official sources that may apply different aggregation methodologies.

Access methods differ considerably. Mexico provides a public REST API. For Paraguay[9] and Argentina[10], downloads are unusually slow, possibly due to infrastructure-imposed limits. Uruguay uses FTP[11]. Chile's imposes automation barriers via captchas for its *customs data*. Others may require *web scraping* of multiple fragmented files.

A great variety of file formats, sizes and time range were found. Uruguay delivers *customs data* in XML, Argentina in LST, and Peru[12] in DBF. *Statistical data* formats vary from XLSX (Bolivia[13], Dominican Republic[14]) to CSV (Colombia[15]) and TXT (Guatemala[16]). Mexico offers both JSON and CSV. File sizes and time range also vary: Argentina's *customs data* (2017–2024) totals ~350GB; Peru (2020–2024), ~130GB; Paraguay (1997–2024), ~55GB in CSV; Uruguay (2016–2024), ~66GB. Brazil's *statistical data*[17] (1997–2024) totals ~4.7GB at the state level and ~2.6GB at the municipal level, while Mexico's municipal-level (2006–2024) reaches ~7.5GB.

Substantial processing may be needed. Some *statistical data* are normalized, requiring *JOINs* with dimension tables (Brazil); others are denormalized (Dominican Republic). *Customs data* may be released fully processed (Paraguay) or require substantial preprocessing (Argentina, Uruguay, Peru). Because customs declarations can be amended or cancelled months or even years after their initial submission, trade statistics are subject to retroactive changes. In many countries, customs authorities handle these updates internally and publish datasets using an overwrite strategy, replacing previous months' data with corrected versions. Uruguay and Peru split files between

[7] https://datos.gob.cl/dataset/registro-de-importacion-2024
[8] https://www.economia.gob.mx/datamexico/en/vizbuilder
[9] https://www.dnit.gov.py/web/portal-institucional/datos-abiertos
[10] https://www.afip.gob.ar/operadoresComercioExterior/informacionAgregada/informacion-agregada.asp
[11] ftp://ftp.aduanas.gub.uy./Datos%20Basicos/
[12] http://www.aduanet.gob.pe/aduanas/informae/presentacion_bases_web.htm
[13] https://www.ine.gob.bo/index.php/estadisticas-economicas/comercio-exterior/
[14] https://www.one.gob.do/datos-y-estadisticas/
[15] https://microdatos.dane.gov.co/index.php/catalog/CI-Microdatos
[16] https://portal.sat.gob.gt/portal/operador-economico-autorizado/informacion-comercio-internacional/
[17] https://www.gov.br/mdic/pt-br/assuntos/comercio-exterior/estatisticas/base-de-dados-bruta

executed and modified records, requiring retroactive incremental loading and primary key construction to ensure consistency. Uruguay further includes cancellation records, increasing complexity.

Documentation adds another layer of heterogeneity. Brazil and Mexico[18] provide detailed manuals and data dictionaries. Peru's documentation lacks clarity on how to construct the needed composite primary keys and doesn't indicate that one of the files contains modified records, which was identified through testing. In general, countries that disseminate normalized datasets tend to offer additional challenges, since they are prone to omit dimension references or metadata. For instance, Argentina doesn't document where to find the necessary dimension tables. Additionally, Argentina uses its own country code (INDEC), differing from ISO 3166-1. Two conflicting official code lists were found (AFIP[19] and INDEC[20]). After tests, the INDEC list proved more complete and was adopted. Brazil, despite its centralized documentation, had missing HS6 product codes in dimension tables, which were corrected after reporting to the authorities.

Anonymization directly impacts SFTD completeness. According to IMTS [UN 2011], confidentiality can be "passive" (by request), "active" (by decision of the authority), or both. From more than a hundred respondent countries to UN Comtrade, 43% report using passive, 30% active, and 21% both [UN 2025].

In Latin America, anonymization strategies vary widely. Brazil applies different techniques to different forms of data. For what is the closest to *customs data*[21], only imports data are published, it doesn't include any subnational level, and it applies data masking and statistical aggregation (e.g., mean, median, quartiles). Argentina adopts a similar strategy for *customs data*, but for exports, while imports data reaches even the firm-level. In Brazilian *statistical data*, datasets are split into two. A state-level based on product's origin/destination (*adm2_prod)* and NCM level (*rtlc_ncm*), including logistics indicators; and a municipal-level based on company's location (*adm3_comp*), but limited to HS4 and indicators limited to value and volume. Mexico implements a dynamic anonymization to *statistical data* and Argentina masks even national trade data[22].

Geographic referencing adds further complexity. Many datasets reference only the customs office location. In Uruguay, exporters are not required to declare where the goods were produced, and national agencies must cooperate and exchange data to infer origins of exported [Uruguay XXI 2024]. In Mexico, the same official effort is mentioned, since data refers to exporting/importing firms' location, not the actual origin/destination of goods [INEGI 2024]. Peru includes a "Ubigeo" field to track subnational origin, but inconsistencies have been requiring manual revision before official dissemination[23].

The layered Lakehouse architecture proved effective for harmonizing several sources. By isolating processing into structured stages, it allowed granular data to be progressively standardized and enriched. Semantic standardization, conducted in the Silver layer, enabled field naming harmonization across countries, which significantly reduced the complexity of subsequent enrichment and validation steps. Additionally, the

[18] https://datamexico.org/en/methodology
[19] https://www.afip.gob.ar/genericos/documentos/codigos-maria.pdf
[20] https://www.indec.gob.ar/indec/web/Institucional-Indec-Clasificadores
[21] https://dados.gov.br/dados/conjuntos-dados/estatisticas-de-declaracoes-de-importacao
[22] https://comex.indec.gob.ar/
[23] https://www.gob.pe/institucion/mincetur/colecciones/549-reporte-mensual-de-comercio-regional-rmcr

strategy adopted at this stage to build country-specific wide denormalized tables facilitated consistency checks and analytical processing, as it avoided multiple *JOINS* and enabled direct querying of enriched tables. DuckDB's SQL support proved suitable for testing and exploring data, such as by using the *SUMMARIZE* to quickly grasp on core descriptive statistics on large datasets. Delta Lake framework provided time travel, allowing access to previous table versions even during large-scale transformations, with concurrency ensuring smooth parallel reads and writes.

5. Cases Studies

To demonstrate the implementation, this section presents Mexico and Uruguay as case studies. Each begins with a general diagnostic, followed by a description of the implementation across the ELT pipeline layers. These countries were selected because they contrast strongly yet instructively in terms of data accessibility and the complexity of integration involved.

5.1. Mexico

Mexico presents a robust infrastructure for SFTD, with comprehensive documentation. DataMéxico[24] integrates multiple institutional sources and supports both general users (via dashboards) and technical users (via tools like *VizBuildear* and API endpoints). Interactive charts that exposed API endpoints proved valuable for cross-checking during Testing and Validation.

BCMM, INEGI and Banxico were the sources identified, and they provide *statistical data* at varying aggregation levels. BCMM was selected since it offers the most disaggregated subnational data. Data is structured into three non-aggregable cubes (national, state, and municipal) as part of a dynamic anonymization strategy. The more granular the cube, the higher the data suppression to protect confidentiality. Thus, trade value from the municipal perspective, doesn't add to the state or national perspective. Only FOB values are provided for both imports and exports, with product codes detailed up to HS6 level and monthly frequency with partner country disaggregation.

Experiments showed that anonymization increases with query granularity. Querying monthly imports at the municipal level by HS6 and partner countries returned only 63% of the actual total trade. At the state level, only 79% of the national value was retrieved. Export data showed an even stronger suppression: monthly municipal-level data returned only 42% of the actual total, reflecting higher anonymization due to firm concentration and traceability risks. These tests were performed directly in the Python via the API, prior to ingestion, thanks to Mexico's infrastructure.

Landing Zone. Since the cubes are non-aggregable, six main datasets were constructed: consolidated imports/exports at the municipal and state levels, each by HS2, HS4, and HS6, with monthly frequency and partner disaggregation. National-level cube wasn't considered. The most detailed dataset, municipal, HS6, monthly, by partner, contained 2.4 million import records and 337 thousand export records for 2023 alone.

A Python script extracted data from BCMM's API, retrieving imports and exports as a unified dataset. The API was used to iteratively extract records by year, month, and trade flow, storing CSVs in the Landing Zone (*schema-on-read*).

[24] https://www.economia.gob.mx/datamexico/en

Bronze. DuckDB ingested CSV files from the Landing Zone using directory globbing to combine monthly partitions into a single logical table. Since BCMM's dataset was already denormalized, only explicit type casting was applied via SQL, month by month, to manage memory. Results were materialized as Arrow Tables and (over)written as six Delta tables, separated by HS level and geography due to non-additivity.

Silver. With Mexico's dataset already denormalized and decomposed into HS6, HS4, and HS2 levels, processing was straightforward. Field names applied via SQL followed a standard convention: *adm2_comp_id* and *_desc* for firm location at the state level; *adm3_comp* and *_desc* at the municipal level. Partner countries were standardized as *country_of_origin_or_destination* and transaction value as *fob_value_usd* for both imports and exports.

Gold. In the Gold layer, auxiliary tables with official multilingual product and country descriptions were registered in DuckDB and *JOINed* via SQL. ISIC codes were only mapped to HS6-level Delta tables (municipal and state), as mappings at HS4 were ambiguous due to multiple possible ISIC associations. Since Mexico's dataset already included ISO 3166-2 codes for states, no additional geographic mapping was needed.

5.2. Uruguay

Uruguay provides detailed *customs data*[25], though access is restricted to Uruguay's residents. Upon official request, the authorities granted an FTP public access. Files are structured into three categories: original customs declarations (*ingresados*), amendments (*modificados*), and cancellations (*anulados*), available as daily updates (DD prefix) and monthly consolidated files (DM prefix), each in XML format. Each record includes over 40 normalized fields, covering customs office, countries of origin/consignment, 10-digit NTLC codes (NCM + 2), FOB/CIF values, weights, transport modes, company names, and a detailed breakdown of taxes, surcharges, and value-added TAX.

Landing Zone. A Python script was developed to automate FTP navigation and download all three file types, focusing on the monthly consolidated data, as well as its dimension tables. Data was stored raw (*schema-on-read*). For Uruguay, exports and imports are also unified and spans 2016 to 2024, totaling around 66GB.

Bronze. Since DuckDB cannot parse XML directly, data was processed in Python (*lxml*, *Polars*). Incremental reconciliation was performed using composite primary keys (*NUME_CORRE*, *ANO_PRESE*, *CODI_ADUAN*, *NUME_SERIE*). Modifications and cancellations from recent files often altered records from earlier years, for instance, 2024 updates referencing 2016 transactions, so the script processed data in memory-efficient chunks, applying changes in order (modifications first, then deletions), and updating only affected periods.

Silver. The Silver layer involved denormalization and semantic standardization. Uruguay's 10-digit NTLC was decomposed into NCM (8-digit Mercosur standard), which was then mapped to HS6, HS4, and HS2 levels. Standardized fields included *ntlc_id*, *ntlc_desc*, *rltc_ncm_desc, fob_value_usd*, *net_weight_kg*, among others.

Gold. Since the data lacks subnational fields beyond the customs office, they were mapped to their corresponding states/departments and macro-regions. For example, the

[25] https://www.aduanas.gub.uy/innovaportal/v/18714/1/innova.front/consultas-dua.html

customs office *Punta del Este* (*customs_location_desc*) was mapped to *Maldonado* state (*adm2_customs_desc)* and the region *East* (*adm1_customs_desc_en*). This process covered 17 customs units across 11 states and 5 macro-regions. As a result, from customs locations, it was possible to derive ISO 3166-2 codes (*adm2_customs_iso*) and ensure interoperability. Finally, multilingual descriptions for products, countries and ISIC sectors were added and results were incrementally written to Gold's Delta Lake.

6. Conclusions

The assessment revealed deep heterogeneity across all stages of the SFTD lifecycle: access methods, documentation quality, metadata standards, anonymization practices, and processing needs. Multiple agencies can publish trade data and with distinct methodologies and dissemination practices. *Customs data*, while highly granular, often demands extensive engineering due to raw formats. *Statistical data*, in turn, is generally pre-aggregated, but can be anonymized, limiting detail and analytical scope.

A key limitation lies in geographic attribution, whether to the product's origin or the firm's location. This distinction is rarely made explicit, and in many cases, the underlying information is not collected at all. Some datasets (e.g., Mexico) assign flows only to the operators' location, typically economic hubs where firms are headquartered, rather than actual production sites or destinations, distorting regional analyses. In others (e.g., Uruguay), the only subnational information is the customs office where goods were cleared. While this may occasionally align with product origins for logistical reasons, it leaves central regions underrepresented and creates ambiguity. Additionally, Mexico provides only FOB values, omitting CIF (the IMTS-recommended measure for imports) and quantities, which limits analyses requiring volume-based allocations.

These issues particularly affect studies such as environmental emission tracking or market research aiming to pinpoint true origins and consumption areas, rather than administrative or dispatch locations. Brazil represents a notable exception, as its *statistical data* includes the firm's location, the actual production or destination sites, and the customs zones, offering a potential model for improving data collection in other countries. To help standardize these differences, the study used specific naming conventions: *adm2_comp* for the firm's location, *adm2_prod* for where the goods were produced (or where imports are going), and *adm2_customs* for the location of the customs office.

These institutional and technical challenges highlight the need for clearer international standards, including subnational dimensions in frameworks like IMTS and UN Comtrade. Yet, despite this complexity, the study demonstrated that consolidation and harmonization of heterogeneous and public sources, such as SFTD, are achievable. The layered Lakehouse architecture provided traceability, quality control, and adaptability for exploratory research, statistical modeling, and market intelligence, offering a practical foundation for future regional efforts toward data interoperability.

Future work includes automating scheduled extractions, implementing robust validation rules and discrepancy handling, and extending the approach to additional countries worldwide.

References

Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., van Hovell, H., Ionescu, A., Łuszczak, A., Świtakowski, M., Szafrański, M., Li, X., Ueshin, T., Mokhtar, M., Boncz, P., Ghodsi, A., Paranjpye, S., Senster, P., Xin, R., Zaharia, M. (2020). Delta lake: high-performance ACID table storage over cloud object stores. In Proceedings of the VLDB Endowment, 13(12): 3411-3424.

Armbrust, M., Ghodsi, A., Xin, R., Zaharia, M. (2021). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR, 8(1): 28.

Cavusgil, S. T., Knight, G., Riesenberger, J. R., Rammal, H. G., Rose, E. L. (2014). International business. Pearson Australia.

Chen, C., Jiang, Z., Li, N., Wang, H., Wang, P., Zhang, Z., Zhang, C., Ma, F., Huang, Y., Lu, X., Wei, J., Qi, J., Chen, W. Q. (2022). Advancing UN Comtrade for physical trade flow analysis: review of data quality issues and solutions. Resources, Conservation and Recycling, 186: 106526.

Francioni, B. and Martín, O. (2024). International market, network, and opportunity selection: A systematic review of empirical research, integrative framework, and comprehensive research agenda. Journal of International Management, 30(5): 101174.

Gimenez-Perales, F. (2024). The dynamics of importer–exporter connections. European Economic Review, 161: 104638.

Hu, L., Song, C., Ye, S., Gao, P. (2022). Spatiotemporal statistical imbalance: a long-term neglected defect in UN Comtrade dataset. Sustainability, 14(3): 1431.

Huber, R. A., Stiller, Y., Dür, A. (2023). Measuring subnational trade competitiveness. Scientific data, 10(1): 331.

Instituto Nacional de Estadística y Geografía [INEGI] (2024). Encuesta Trimestral de Comercio Exterior Estatal (ETEF). Available: https://www.inegi.org.mx/programas/etef/. [Accessed: 26 Apr. 2025].

Ito, K., Endoh, M., Jinji, N., Matsuura, T., Okubo, T., and Sasahara, A. (2025). Margins, concentration, and the performance of firms in international trade: Evidence from Japanese customs data. Journal of The Japanese and International Economies, 75: 101340.

Jiang, M., Liu, L., Behrens, P., Wang, T., Tang, Z., Chen, D., Yu, Y., Ren, Z., Zhu, S., Tukker, A., Zhu, B. (2020). Improving subnational input–output analyses using Regional Trade Data: a case-study and comparison. Environmental Science & Technology, 54(19): 12732-12741.

Kardes, I. (2016). Reaching middle class consumers in emerging markets: unlocking market potential through urban-based analysis. International Business Review, 25(3): 703-710.

México (2022). Metodología de la Balanza Comercial de Mercancías de México [BCMM]. Available: https://datamexico.org/en/methodology. [Accessed: 26 Apr. 2025].

Papadopoulos, N. and Denis, J. (1988). Inventory, taxonomy and assessment of methods for international market selection. International Marketing Review, 5(3): 38–51.

Papadopoulos, N. and Martín, O. (2011). International market selection and segmentation: Perspectives and challenges. International Marketing Review, 28(2): 132–149.

Raasveldt, M. and Mühleisen, H. (2019). Duckdb: an embeddable analytical database. In Proceedings of the 2019 international conference on management of data: 1981-1984).

Schneider, J., Gröger, C., Lutsch, A., Schwarz, H., Mitschang, B. (2024). The Lakehouse: State of the art on concepts and technologies. SN Computer Science, 5(5): 449.

Shaar, K. (2019). Reconciling international trade data. ZBW – Leibniz Information Centre for Economics working paper. Hamburg.

United Nations [UN] (2011). International Merchandise Trade Statistics: Concepts and Definitions. Revision 3, United Nations Publication, New York.

Uruguay XXI (2024). Exportaciones de bienes por departamento. Available: https://www.uruguayxxi.gub.uy/uploads/informacion/2dc88869805f86b3ff671e1e7220ceca71f572ea.pdf. [Accessed: 26 Apr. 2025].

Volk, M., Staegemann, D., Bosse, S., Häusler, R., Turowski, K. (2020). Approaching the (Big) Data Science Engineering Process. In IoTBDS: 428-435.

Wittwer, G. (2022). Preparing a multi-country, sub-national CGE model: EuroTERM including Ukraine. Centre of Policy Studies, Working Paper G-334.

World Bank (2025). World development indicators: Trade (% of GDP). Available: https://data.worldbank.org/indicator/NE.TRD.GNFS.ZS. [Accessed: 26 Apr. 2025].

World Customs Organization [WCO] (1983). International convention on the Harmonized Commodity Description and Coding System (as amended through 2018). Brussels: WCO.