Ripple Effects

Tariffs & Trade Disruptions in Global Commerce

Course Name & Section: FA 25 6513-C Big Data

Semester Term: Fall 2025

By: Shruti Karmarkar (spk9869), Aviraj Dongare(aud211),

Hrishik Desai(hbd9580)

Abstract

Global trade is increasingly shaped by shifting tariffs, trade wars, and supply chain disruptions that ripple through economies worldwide. This project harnesses Big Data analytics to **quantify and visualize how tariff changes reshape international trade networks** from 2014–2024 across 50+ countries. By merging multi-year datasets from **UN Comtrade** (trade flows by country and product) and **WTO/WITS** (tariff schedules), we aim to uncover patterns, breakpoints, and real-world impacts of tariff policies on global commerce. Using a Spark-based distributed ETL pipeline, the project processes multi-year trade and tariff data efficiently, performing large-scale aggregations and visualizations that reveal how global commerce evolves under changing tariff regimes. Using distributed data processing and interactive geospatial visualizations, the study will reveal which industries, regions, and trading relationships have strengthened or weakened under changing trade conditions turning complex economic data into clear, data-driven insights about globalization's evolving landscape.

Problem Statement

Global trade is constantly affected by changes in tariffs, trade policies, and international agreements. These shifts can disrupt long-standing trade routes, change demand patterns, and create uncertainty in key industries. Yet, most existing studies look at trade data only once a year and use limited economic models, which means they miss the short-term changes and quick reactions that happen after major policy events.

To truly understand these disruptions, we need a **scalable, data-driven system** that combines tariff information with detailed, product-level trade data. Using **Big Data technologies**, this project aims to analyze large volumes of trade and tariff data to detect how trade patterns evolve over time. Our goal is to build a framework that captures these changes more dynamically, showing how international trade relationships shift in response to new policies through data analysis, distributed processing, and visual storytelling.

Objectives

1. To **quantify and visualize** how tariff changes or trade disputes affect import/export volumes across countries and industries.
2. To detect **patterns, anomalies, and shifts** in global trade caused by tariff policies.

3. (Optional advanced goal): Build a **predictive model** that estimates trade volume change given a new tariff rate.

Datasets (Primary & Supplementary)

Dataset	Description	Purpose	Link	Approx. Data File Size
UN Comtrade Database	Contains detailed import/export data by country, year, product (HS code), and trade value.	Main dataset to measure trade flow patterns between 2014- 2024	https://comtradeplus.un.org/	~1.0 million records (~800 MB – 1 GB CSV)
WTO Tariff Database (IDB/CTS)	Contains applied and bound tariff rates by product and country	Provides tariff rate changes (the cause)	https://tao.wto.org/	~200 k – 300 k records (~150 MB – 200 MB CSV)
World Bank WITS (World Integrated Trade Solution)	Combines WTO, UNCTAD, and World Bank trade data	Easier integrated access if we want one pipeline	https://wits.worldbank.org/	~1.2 – 1.5 million combined records (~1 GB – 1.2 GB CSV)

How We'll Use the Data

1. **Data Ingestion:**
    - Download multiple years of trade and tariff data for several countries.
    - Use Apache Spark for ETL (extraction, cleaning, and transformation) with distributed joins and aggregations on multi-million-row trade and tariff datasets.
2. **Integration:**
    - Merge tariff rates (from WTO/WITS) with trade volumes (from UN Comtrade) using common HS codes + country-year pairs.
3. **Analysis:**
    - Identify where tariff hikes led to significant changes in trade flow.
    - Compare pre- and post-tariff trends by country and product category.

4. **Visualization:**
    - Build dashboards or geospatial heatmaps showing trade flow shifts, using Tableau, Power BI, or Plotly Dash.
5. **(Optional Modeling):**
    - Use regression or ML models to predict trade volume changes from tariff changes.

Implementation Using Spark-based ETL and Distributed Aggregation

To ensure scalability and alignment with Big Data frameworks, the ETL pipeline will be implemented using **Apache Spark**. Spark's distributed processing capabilities will allow efficient ingestion, transformation, and aggregation of millions of trade and tariff records across multiple years and countries.

1. **Data Ingestion (Extract)**

- Spark's read.csv() and read.parquet() methods will be used to load large datasets (UN Comtrade, WTO/WITS) directly from **HDFS** or object storage.
- The ingestion will be parallelized by **partitioning on Year and Reporter Country**, ensuring balanced workloads across cluster nodes.
- Metadata (schema, record counts) will be logged using Spark's DataFrame API for reproducibility.

2. **Data Cleaning & Transformation (Transform)**

- Cleaning operations such as null removal, column renaming, HS-code normalization, and type casting will be performed using **PySpark DataFrame transformations** instead of Pandas.
- Country codes and product codes will be standardized using Spark's built-in join and withColumn transformations.
- The combined dataset will be stored in **Parquet format** (columnar, compressed) to improve read performance and reduce storage.

3. **Distributed Aggregation**

- Trade value summaries and tariff impact metrics will be computed using Spark's groupBy().agg() operations at different granularities:
    - Yearly (to identify long-term trends)
    - Partner-country level (to detect rerouting)
    - Product-category level (to measure industry impact)
- Aggregated data will be written back to Parquet for downstream visualization.

4. **Scalability Advantage**

- This Spark-based ETL design allows processing of **1–1.5 million records** efficiently across nodes, avoiding memory bottlenecks.
- The same logic can scale horizontally by adding executors or processing additional years/partners.
- Compared to Pandas, Spark provides a **fault-tolerant, distributed execution plan** suitable for multi-gigabyte trade data.

Tools & Tech Stack

- **Processing:** Apache Spark (PySpark for ETL and distributed aggregation), Hadoop for storage, Pandas for small-scale testing
- **Storage:** HDFS, SQL / NoSQL
- **Visualization:** Tableau, Power BI, or Plotly Dash
- **Modeling:** Python (scikit-learn, statsmodels)
- **Versioning & Workflow:** GitHub, possibly Prefect/Airflow for data pipeline orchestration

Deliverables

1. **Clean merged dataset** (trade + tariff).
2. **Exploratory data analysis report** descriptive statistics, trend plots.
3. Spark-based ETL pipeline scripts demonstrating distributed ingestion, transformation, and aggregation.
4. **Interactive visualization/dashboard** showing trade flow shifts.
5. **Final project report** linking findings to economic insights.
6. **Predictive model forecasting** trade impact of new tariff.