

Analyzing U.S. Border Crossing Traffic and Freight Trends Using the Medallion Architecture

1. Business Case and Pipeline Design

Company: NexPort Logistics (fictional)

A cross-border freight and logistics company operating in North America.

Business Scenario

NexPort Logistics specializes in managing shipment schedules, border clearances, and trucking routes for import/export businesses. Border delays caused by traffic surges, under-staffing, or poor planning lead to increased delivery costs and lower customer satisfaction.

Business Goal

The goal is to build a data lakehouse that leverages border crossing data and freight pricing trends to:

- Analyze port-level traffic trends.
- Identify peak congestion periods.
- Correlate border activity with freight pricing changes.
- Predict future traffic and cost patterns based on historical data.
- Inform scheduling, staffing, and cost-saving logistics decisions.

Users of the Lakehouse Outputs

Operations Managers: Adjust driver schedules, lane usage, and border timing.

Executives: Monitor trends and optimize cross-border operations.

BI & Data Analysts: Generate reports and insights from the Gold layer tables.

Datasets

1.Border Crossing Entry Data – U.S. Department of Transportation

URL: <https://data.transportation.gov/Research-and-Statistics/Border-Crossing-Entry-Data/keg4-3bc2>

This dataset provides monthly traffic volumes across U.S. land ports by transportation mode (truck, rail, passenger vehicles, etc.). It is central to analyzing congestion trends, identifying seasonal traffic peaks, and understanding operational stress points across different ports of entry. It forms the foundation of the lakehouse for detecting movement patterns and resource planning opportunities.

2. Freight Transportation Services Index (TSIFRGHT) – FRED

URL: <https://fred.stlouisfed.org/series/TSIFRGHT>

This index tracks price trends in U.S. freight transportation services over time. When combined with crossing data, it allows NexPort Logistics to study how pricing changes correlate with border activity levels. This supports cost forecasting, profitability analysis, and decisions on shipment scheduling during high cost vs. low-cost periods.

Medallion Architecture Design

Bronze Layer – Raw Data Ingestion

Sources:

- Border Crossing Entry Data (CSV from data.transportation.gov)
- Freight Pricing Index (CSV from fred.stlouisfed.org)

Metadata to Add:

- Ingestion timestamp
- Source file name/version

Silver Layer – Cleaned & Enriched Data

Data Cleaning Steps:

1. Normalize Port Names ("Buffalo-Niagara Falls" to "Buffalo")
2. Handle Missing or Invalid Values (nulls in Value or Measure)
3. Parse Dates and Extract Year, Month, Quarter
4. Align and Join Freight Index data on Year + Month for analysis

Gold Layer – Analytical Tables

Monthly_Traffic_By_Port: Monthly totals by port

Vehicle_Type_Summary: Compare traffic types (trucks, buses, etc.)

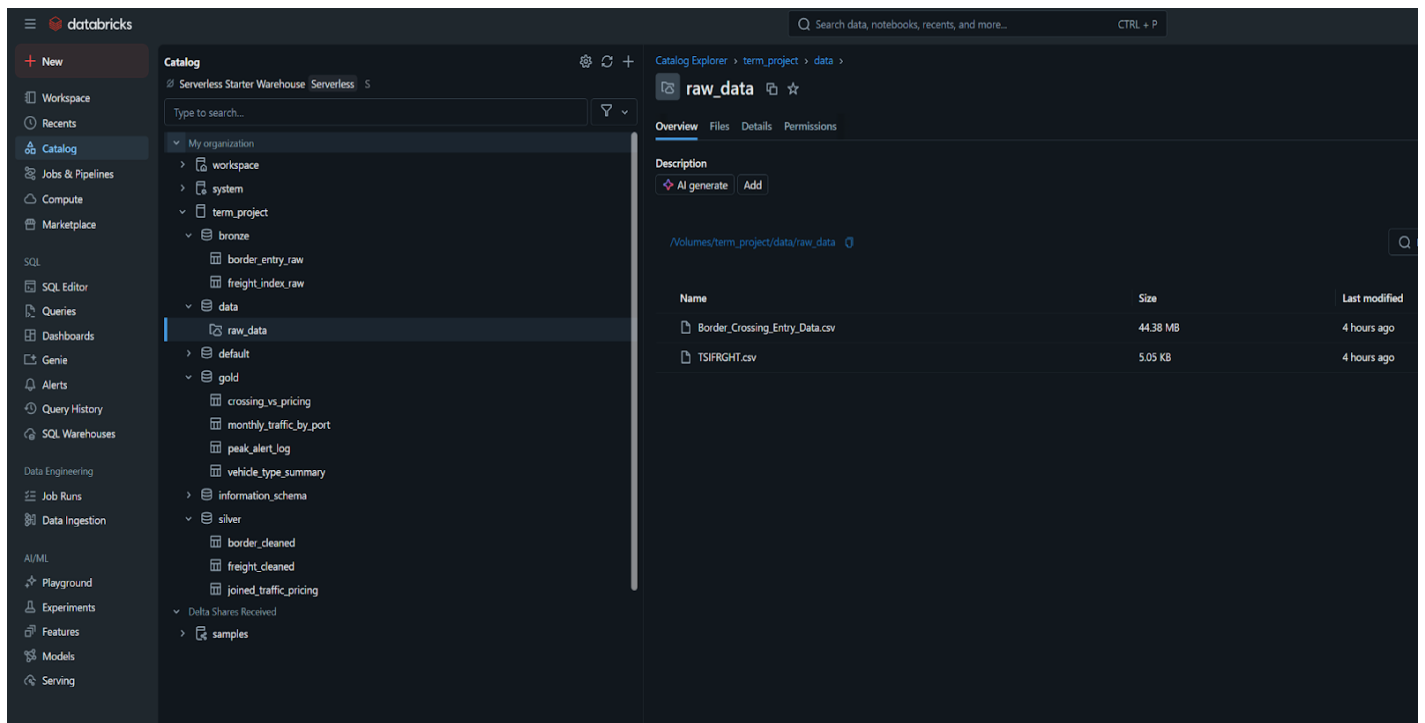
Peak_Alert_Log: Highlight 90th percentile+ traffic days

Crossing_vs_Pricing: Merge freight index with traffic totals

2. Pipeline Implementation

The pipeline was implemented in Databricks using notebooks for each layer:

- Bronze: Loaded raw CSVs and added ingestion timestamp.
- Silver: Cleaned data with proper types, filtered valid dates, and standardized column names.
- Gold: Derived three analytics tables by joining border and freight data, calculating aggregates, and extracting high-volume alerts.



The screenshot displays the Databricks interface. On the left, a sidebar contains navigation links for Workspace, Recents, Catalog, Jobs & Pipelines, Compute, Marketplace, SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, AI/ML, Playground, Experiments, Features, Models, and Serving. The main area is divided into two panels. The left panel, titled 'Catalog', shows a tree view of the data catalog with the following structure:

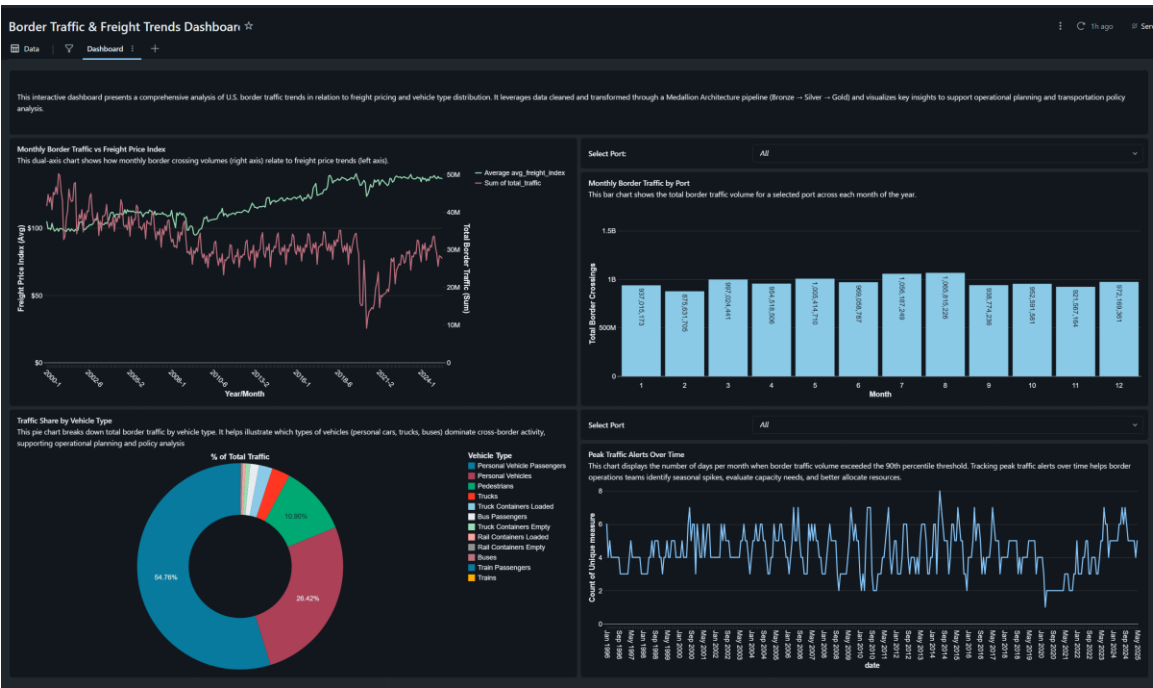
- My organization
 - workspace
 - system
 - term_project
 - bronze
 - border_entry_raw
 - freight_index_raw
 - data
 - raw_data
 - default
 - gold
 - crossing_vs_pricing
 - monthly_traffic_by_port
 - peak_alert_log
 - vehicle_type_summary
 - information_schema
 - silver
 - border_cleaned
 - freight_cleaned
 - joined_traffic_pricing
 - Delta Shares Received
 - samples

The right panel, titled 'raw_data', shows the 'Overview' tab. It includes a search bar, a description section with an 'AI generate' button, and a table listing the files in the 'raw_data' table.

Name	Size	Last modified
Border_Crossing_Entry_Data.csv	44.38 MB	4 hours ago
TSIFRIGHT.csv	5.05 KB	4 hours ago

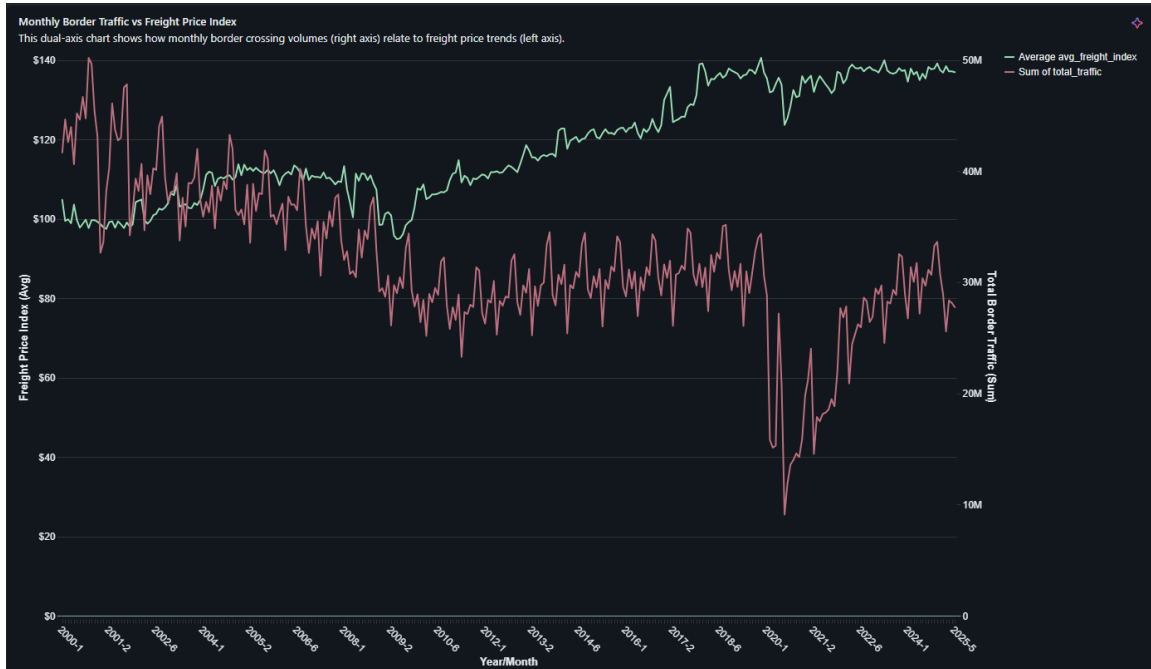
3. Dashboard Implementation

The final dashboard includes 4 visualizations derived from the Gold layer tables:



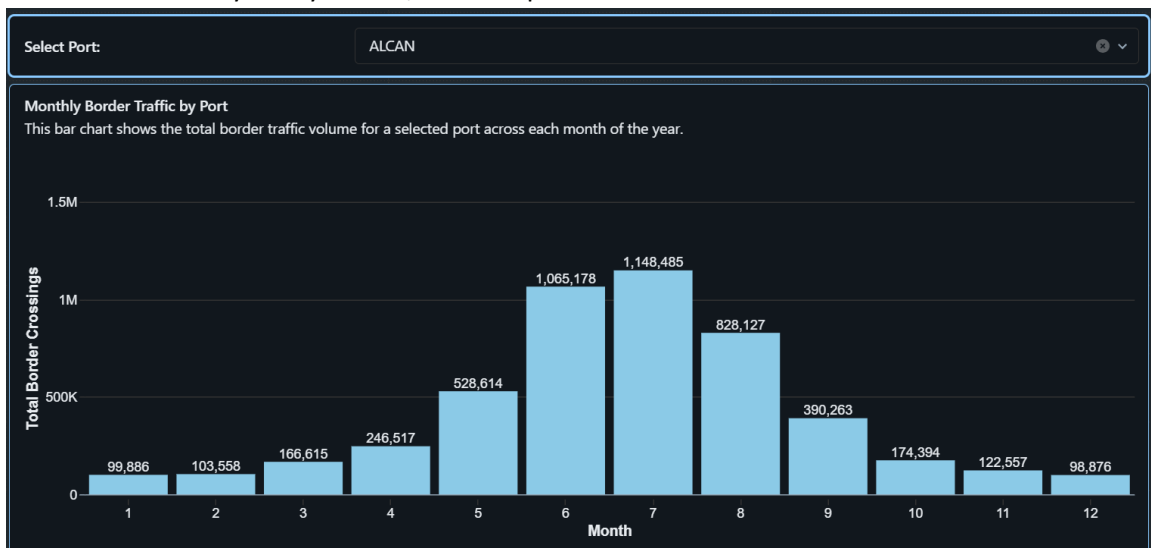
1. Monthly Border Traffic vs Freight Price Index (Dual-Axis Line Chart)

- Audience: Business Analysts, Economists
- Purpose: Analyze relationship between freight costs and border activity over time.
- Reason: Dual-axis allows comparison across different units.



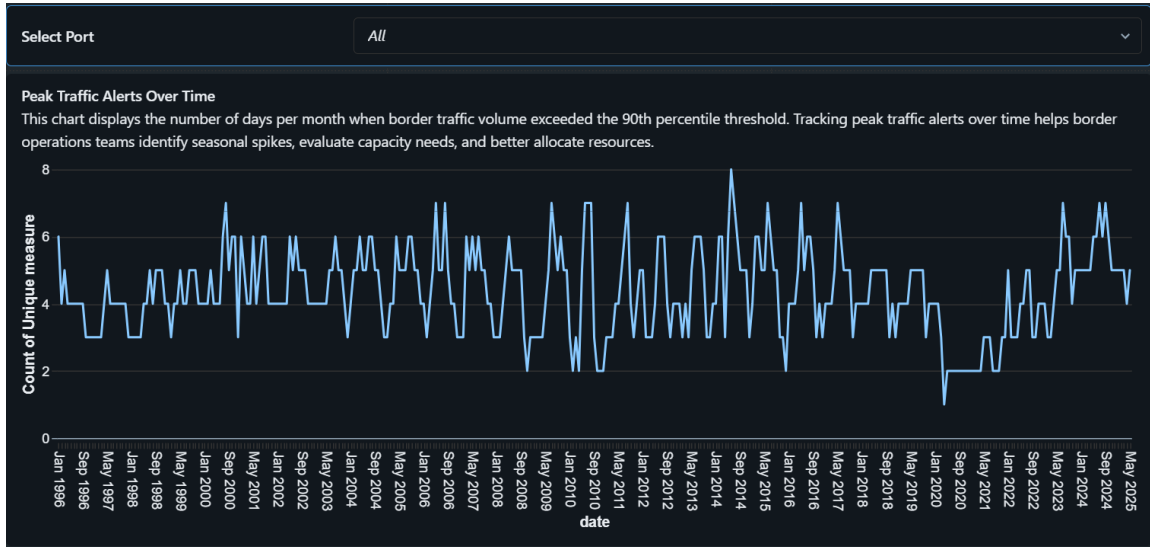
2. Monthly Border Traffic by Port (Bar Chart with Port Filter)

- Audience: Operational Teams
- Purpose: Observe traffic seasonality for specific ports.
- Reason: Clear layout by month, filter simplifies interface.



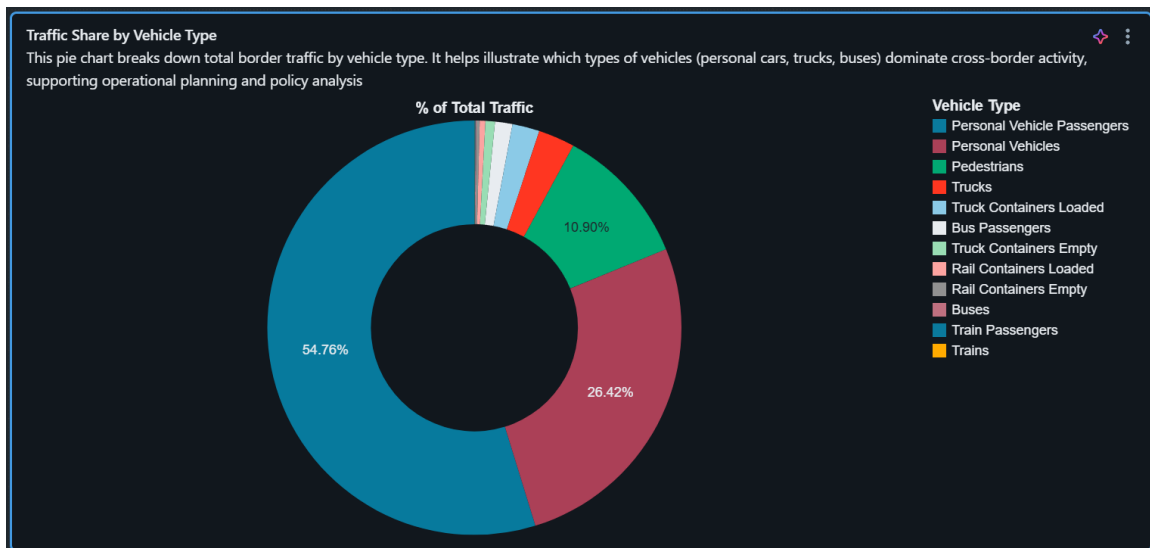
3. Peak Traffic Alerts Over Time (Line Chart)

- Audience: Operations & Planning Teams
- Purpose: Spot high-traffic months (above 90th percentile).
- Reason: Line chart best for long-term anomaly trend.



4. Traffic Share by Vehicle Type (Pie Chart)

- Audience: Policy Analysts
- Purpose: Visualize dominance of vehicle types crossing borders.
- Reason: Pie chart best for categorical distribution.



4. Pipeline Analysis

1. Is the data considered Big Data?

Yes. Based on the 3 Vs:

- Volume: Large historic dataset with millions of entries across years.
- Variety: Different types of measures, vehicle classes, and economic indicators.
- Velocity: Can be updated monthly and expanded with streaming customs reports.

2. Is a data lakehouse appropriate?

Yes. A lakehouse gives the flexibility to bring in different formats and datasets as needed, without having to lock down rigid schemas upfront. Plus, Delta tables make querying fast and reliable, it feels like the best of both worlds between a lake and a warehouse.

3. Ingest Frequency and Logic:

For this project, monthly ingestion makes the most sense. The Border Crossing Entry Data is updated regularly on the DOT portal, and the freight pricing index from FRED also follows a monthly release schedule. So, setting up a monthly pipeline keeps things consistent and manageable.

4. How would you handle batch vs streaming?

For batch, I'd use scheduled jobs to load the data once a month as it's straightforward and reliable. If I ever added streaming data (like real-time traffic feeds or customs alerts), I'd switch to using Databricks Auto Loader or Structured Streaming with schema inference and checkpointing to make sure nothing gets dropped.