BUDT703 Project-0506 Team - 13 Report

Team 13 Members:

Anisha Bahl
Malay Patel
Shikha Rani
Nikhil Nagesh

Project Name:

TransitPulse KPI & Incident Monitoring System

Overview:

This document will go over how our data was processed, structured, and some insights and recommendations we came up with based on our SQL code and analysis.

Tables Used:

The TransitPulse System integrates WMATA operational performance, ridership details, and crime information into a unified schema-relational model. Our model includes 9 core tables and 2 bridge tables:

1. **TimelineReference:** This table represents each month in the system. It gives us a shared time reference so all monthly metrics such as ridership, incidents, on-time performance, and crime can be tied to the same reporting period.
2. **MetroStation:** This table lists all metro stations. Stations act as the main location link across several datasets like ridership counts, crime reports, and service incidents.
3. **TransitLine:** This table contains the different metro lines. Lines help us group and study on-time performance results and service incidents because these metrics are usually reported at the line level.
4. **CrimeCategory:** This table is a lookup for crime types. It allows us to keep the crime data organized into consistent categories when we analyze trends at stations.
5. **StationSnapshot:** This table gives monthly activity for each station. It connects a station to a specific month and stores measurements such as station-level ridership for that period.
6. **SystemRidership:** This table stores total monthly ridership for the whole system, for example rail and bus totals. It uses the same monthly time reference so we can compare system-wide patterns with station-level activity.

7. **OnTimePerformance:** This table stores the on-time performance for each month. Since on-time information is reported by line and month, the table links each record to both a line and a month.
8. **ServiceIncident:** This table stores counts of service incidents and delays for each month. Each incident record is linked to a line and may also reference a station depending on the level of detail in the data.
9. **ReportedCrime:** This table stores the monthly number of crimes reported at each station. Each record connects a station, a crime category, and a month.
10. **StationLine:** This is a bridge table that connects stations to the lines they serve. Some stations belong to more than one line, so this table captures those many to many relationships.
11. **StationCrime:** This also is a bridge table, which is being used to organize crime related metadata. It links stations and crime categories where needed, especially when we want to describe which types of crimes are relevant for which stations.

Data Loading Process:

To prepare our database for the project, we followed a straightforward set of steps. The goal was to make sure the tables were created correctly, the data fit the structure we designed, and everything connected the way it should.

1. Creating the Schema: We started by creating the Team13 schema in SQL. Before loading anything, we removed older versions of the tables. We dropped them in the right order so that foreign key dependencies did not cause errors. This gave us a clean place to load all new data.

2. Building All Tables: Next, we created each table based on our ERD. This included the main dimension tables like TransitLine, MetroStation, CrimeCategory, and TimelineReference. After that, we created the fact tables such as StationSnapshot, SystemRidership, OnTimePerformance, ServiceIncident, and ReportedCrime. We also added the bridge tables StationLine and StationCrime to handle many-to-many relationships. Primary keys and foreign keys were added during this step to make sure relationships worked as expected.

3. Loading the Data: Once the structure was ready, we inserted data into the dimension tables first because the rest of the tables depend on them. After loading the dimensions, we added the timeline data with monthly entries. Then we inserted data into the fact tables. This step connected ridership, crime, on-time performance, and incident data to the correct dates, lines, and stations. By loading in this order, all foreign key checks passed without issues.

4. Creating Views for Extra Calculations: Instead of storing calculated columns inside the tables, we created several SQL views. These views helped us compute things like percent change in ridership, on-time performance, monthly crime breakdowns, and crime rates per million riders. Using views made the database cleaner and easier to maintain, and it also helped Tableau pull calculated values directly without extra processing.

Schema and Design Decisions References:

Design Principles that are used:

1. All monthly data is tied to one shared time table. This keeps every fact table aligned on the same monthly reporting period.
2. Stations and lines serve as the main dimensions for most of the analysis. They connect ridership, incidents, crime, and performance data in a consistent way.
3. Bridge tables are used to handle many-to-many relationships so we do not duplicate station or line records.
4. We avoid storing derived values inside tables. Instead, the plan is to create SQL views for calculated attributes when needed.
5. Referential integrity rules and business rules help manage how updates or deletions move through the schema.
6. The overall structure follows a star-schema pattern, which supports faster queries for reporting and analytics.

Enforced Business Rules: These rules match the referential integrity matrix in the design specification.

1. A TimelineReference record cannot be removed if any fact table is still using that month.
2. If a station is deleted, the stationId in ServiceIncident and ReportedCrime becomes NULL rather than removing the fact records.
3. If a station is deleted, related rows in StationLine and StationCrime are removed, since they only serve as link tables.
4. When dimension attributes are updated, the updates flow into the fact tables that depend on them.

Findings and Analysis Insights: Using the sample data from 2025 along with earlier data from 2015 to 2021, we reviewed several key performance patterns. Some of the findings and insights gathered are:

Ridership Findings:

1. In the 2015 portion of the data, stations such as Metro Center, Union Station, and Dupont Circle show the highest ridership levels. This is expected since these locations sit in major activity areas and handle large commuter volumes. Stations in more residential zones, such as Glenmont and Takoma, show moderate or lower ridership. These observations match what we see in the StationSnapshot sample.
2. The year-over-year results from the vw_StationPctChangeYoY view show noticeable declines around 2021 for many stations. This lines up with the broader pandemic effect on public transit. After that year, several downtown stations begin to recover more quickly, which reflects a gradual return of riders to central job and event centers.

System Ridership Trends:

1. The vw_RidershipChange view shows that Metrorail ridership typically rises from January through March across different years.
2. Metrobus follows a similar upward pattern during this period, though the changes are less pronounced.
3. MetroAccess remains the most consistent mode, with minimal month-to-month fluctuation.

On-time performance insights:

The Red Line (L01) has slightly lower on-time performance compared to the other lines. It tends to stay just under the 90 percent level in the sample data. The Silver, Green, and Orange lines show somewhat better reliability. A large part of the drop in overall on-time results is linked to cancelled trips rather than minor delays.

Incident Findings:

The Red Line (L01) also reports the highest number of incidents in the sample, with 215 recorded in one of the periods. The issues that show up most often are track maintenance, mechanical problems, and signal failures. These observations connect closely with the concerns described in mission question 3, which focuses on reliability challenges.

Crime Findings:

In the sample crime data, Gallery Place (ST16) has the highest number of property crimes, mostly larceny. Other stations such as Metro Center, Union Station, and Dupont Circle show moderate levels. When looking at the vw_CrimeRatePerMillion view, stations with lower ridership often have higher crime rates after normalizing for rider volume, while busier stations end up with lower crime rates per million riders even if they show more total incidents.

The most common crime types across stations are larceny, assault, and robbery. More severe crimes such as homicide and arson appear only rarely in the dataset.

<u>Combined Insights across ridership, incidents and crime:</u>

Multi-factor hotspot stations:

1. **Metro Center ST15** stands out because it combines high ridership with moderate crime levels and a noticeable number of operational incidents.
2. **Gallery Place ST16** shows up across all metrics, with the highest crime counts, strong ridership, and its role as a busy transfer hub.

3. **Union Station ST18** also ranks as a priority station due to its large rider volumes and recurring property and robbery-related incidents.

Correlation insight:

1. The data suggests that ridership and crime/incidents don't strongly correlate. Once crime is normalized, busy stations don't automatically show higher crime levels.
2. Operational incidents seem to be driven more by infrastructure and line-specific issues rather than the number of riders passing through a station.

Recommendations:

For safety and crime reduction:

1. Increase police visibility at major transfer points, especially at stations such as Gallery Place and Metro Center. These stations attract large crowds and have higher property crime counts in the sample data.
2. Adjust safety measures based on the type of crime. For example, simple reminders and signage can help reduce larceny, while more patrols during late hours can help address incidents involving assault.
3. Build predictive crime tools that use the crime-per-million-rider metric to identify early signs of rising risk at specific stations.

For operational performance enhancements:

1. Focus on resolving Red Line incident patterns by digging deeper into track maintenance issues and reinforcing preventive work on the Blue, Orange, and Silver lines to reduce signal-related problems.
2. Boost overall on-time performance by targeting cancellations, since they're one of the biggest drivers of delays.

For ridership and demand planning:

1. Shift staffing toward stations that are experiencing noticeable year on year ridership growth.
2. Modify service frequency in areas with heavier demand, particularly during peak periods.
3. Leverage Tableau's forecasting tools to better anticipate future ridership trends and plan accordingly.

Data and analytics enhancements:

1. Bring in more variables such as weather, event schedules, and maintenance logs. These can explain patterns that do not appear in the core dataset.
2. Add real-time data inputs so dashboards reflect current conditions instead of relying only on past records.
3. Set up automated alerts that notify WMATA staff when crime counts, incident levels, or on-time performance change sharply from normal patterns.

Conclusion:

The TransitPulse KPI and Incident Monitoring System brings several important datasets together in one place. Ridership, on-time performance, service incidents, and crime information are all connected through a shared structure, which gives WMATA a clearer view of how the system is performing. With the help of SQL views, a solid schema, and dashboard tools on top of it, the system makes it easier to spot stations that need more attention, plan safety improvements, and monitor operational reliability.

This report explained how the data was prepared, how the schema was designed, and what insights came out of the analysis. These pieces form the starting point for future improvements to the TransitPulse system and give WMATA a stronger foundation for data-driven decision making.