Final Report: TTC Bus Delay Data

Name: Devang Gadhiya (Data Analyst)

1. Executive summary

This project focuses on analyzing and optimizing the Toronto Transit Commission (TTC) delay patterns using descriptive, diagnostic, and predictive analytics. By leveraging historical data and dashboards built in Power BI, the project aims to uncover the root causes of delays, identify high-impact problem areas, and provide actionable recommendations to improve service efficiency and customer satisfaction.

Key Findings

1. **Descriptive Analytics**:
   - **Most Frequent Causes**: General delays and vision-related issues are the top contributors to delays.
   - **High-Impact Routes**: Specific bus and streetcar routes consistently experience the most delays.
   - **Timing Trends**: Morning and evening rush hours see the highest frequency of delays, particularly on weekdays.
   - **Seasonal and Location Patterns**: Delays peak during winter months and are concentrated at key transit hubs such as Bloor-Yonge and Union Station.
2. **Diagnostic Analytics**:
   - **Incident Correlations**: Long delays are often associated with security issues and mechanical failures.
   - **Peak Hour Challenges**: Morning and evening rush hours face compounded delays due to traffic congestion and increased passenger volumes.
   - **Vehicle Performance**: Buses experience a higher frequency of delays compared to streetcars and subways.
3. **Predictive Analytics**:
   - Delay forecasting models indicate specific routes and times are more prone to delays.
   - Long-term trends suggest winter months and weekdays will continue to see higher delay frequencies without intervention.

Project Impact

The proposed solutions have the potential to significantly enhance the reliability and efficiency of TTC services. By addressing high-impact delay factors, optimizing schedules, and introducing predictive tools, this project can help reduce passenger wait times, improve operational performance, and boost overall customer satisfaction.

This framework sets the stage for a scalable and adaptable approach to transit system optimization, ensuring TTC is better equipped to meet the demands of its riders now and in the future.

2. Introduction

The Toronto Transit Commission (TTC) plays a critical role in ensuring efficient public transportation for thousands of daily commuters in Toronto. However, frequent bus delays have become a persistent issue, affecting commuter satisfaction, increasing operational costs, and reducing the overall efficiency of the city's transport system. This project aims to address these challenges by analysing the root causes of delays and developing predictive models to improve service reliability.

The project adopts a data-driven approach to understand the causes of delays, predict future disruptions, and offer practical insights to decision-makers. Using advanced analytics and machine learning techniques, the goal is to design a predictive system that integrates smoothly with the existing TTC IT infrastructure. This system will support better decision-making, enhance scheduling, and improve commuter experience.

The report is divided into four key stages, each focusing on a critical aspect of the project:

1. **Business Problem Identification and Scope Analysis**
   This stage defines the key challenges faced by the TTC, such as the primary causes of delays and their impact on operational efficiency. The problem is structured to align with the project's business objectives.

2. **Data Preparation and Manipulation**
   In this stage, data is collected, cleaned, and transformed to make it suitable for analysis. Data is sourced from multiple platforms, including the TTC's open data portal, Presto app, customer reviews, and weather data. Processes such as data cleaning, categorisation, and variable transformation are performed to ensure high data quality.

3. **Solution Design and Integration with Existing IT Architecture**
   This stage outlines the design of a predictive system, which includes developing ETL (Extract, Transform, Load) processes, machine learning models, and dashboards. The system is designed to integrate with TTC's existing IT infrastructure, ensuring a smooth flow of data and effective system compatibility.

4. **Solution Implementation, Outcome Testing, and Optimisation**
   This stage focuses on the system's deployment, testing, and refinement. Machine learning models such as Random Forest, Artificial Neural Networks (ANNs), and K-means clustering are used to predict and understand delays. Testing and optimisation are carried out to ensure the models are accurate, reliable, and capable of delivering actionable insights to TTC stakeholders.

3. Business problem overview

The Toronto Transit Commission (TTC) faces a significant challenge in managing frequent bus delays, which affect operational efficiency, customer satisfaction, and service reliability. Delays increase operational costs due to higher fuel consumption, additional staffing needs, and schedule adjustments. They also disrupt the transit network, causing a ripple effect on route timings and commuter schedules. For passengers, delays lead to missed appointments, late arrivals at work, and dissatisfaction with public transport services. The complexity of handling data from multiple sources, such as route information, weather conditions, and incident reports, adds to the challenge. Although the TTC collects substantial data on delays, it struggles to utilise this information effectively for predictive and diagnostic analysis. Additionally, the complexity of managing numerous routes, incident types, and external factors (such as weather or traffic) complicates the analysis of delay data.

To address these challenges, this project focuses on two primary business requirements:

1. Predictive Insights: The TTC requires tools that can predict the likelihood of delays based on various factors such as route, time of day, and incident type. By leveraging predictive models, the TTC can take proactive measures to minimize delays and optimize bus scheduling.

2. Root Cause Analysis: In addition to predictive insights, the TTC needs a thorough understanding of the root causes of delays. Identifying the most common reasons for delays (e.g., security incidents, traffic congestion) will enable the TTC to address these issues and implement preventive measures

Ultimately, this project aims to deliver actionable insights that the TTC can use to improve service reliability, reduce delays, and enhance the overall transit experience for its customers.

4. Analytics questions

Descriptive Analytics Questions: These questions focus on understanding historical patterns and summarizing trends from the delay data.

a) What are the most frequent causes of bus delays?
b) Which routes experience the most delays?
c) Which time of day (morning, afternoon, evening) experiences the highest frequency of delays?
d) Which days of the week have the most delays?
e) How do delays vary by month or season?
f) What is the total delay time by route and month?
g) How many delays occurred each month over the past year?
h) What is the distribution of delay types (Short, Medium, Long) across routes?
i) How often do delays occur at key locations (like Bloor-Yonge or Union Station)?
j) Which vehicle types (bus, streetcar, subway) have the most delays?

Diagnostic Analytics Questions: These questions focus on exploring the relationships between variables and identifying patterns in delays.

a) Which routes have the longest delays and why?
b) What incident types are most likely to cause long delays?
c) Are there certain times of day when long delays are more likely to occur?
d) How do weather conditions impact delay times?
e) What is the relationship between traffic congestion and delays?
f) Which incident types are most common on specific routes?
g) Do certain vehicle types experience more delays than others?
h) What locations experience the most delays, and what are the primary causes?
i) How does delay duration differ by day of the week?
j) Are delays worse during peak hours (morning/evening rush)?

Predictive Analytics Questions:

a) Can we predict which routes are likely to experience delays tomorrow?
b) What is the forecasted number of delays for the next month?
c) Which times of day are most likely to experience delays tomorrow?
d) Can we predict the likelihood of long delays on specific routes?

e) What will be the average delay duration for next week?

5. Scope statement

The scope of this project is focused on analyzing the TTC bus delay dataset to derive actionable insights that can improve service reliability. The scope includes the following key areas:

1. Data Analysis: The TTC bus delay dataset is thoroughly analyzed to identify trends and insights related to delays.

2. Model Development: Machine learning models are developed to predict delays and analyze contributing factors such as route, time, and incident type.

3. Visualization: The findings of the analysis are be presented through visualizations such as graphs and dashboards, making it easier for decision-makers to understand the key factors contributing to delays.

4. Deliverables: The final deliverables for this project include a comprehensive report summarizing the analysis findings, a set of visual dashboards for decisionmakers, and delay prediction models that can be used by the TTC to predict future delays and optimize operations.

The project focuses on providing data-driven insights that can help the TTC reduce delays, improve operational efficiency, and enhance the overall experience for commuters.

6. Data sources/key data entities and flows

Data sources: The data originates from the Open Data Portal of the City of Toronto, specifically focusing on the Toronto Transit Commission (TTC) delay datasets. The datasets provide detailed records of delays across different modes of public transportation in Toronto. Below are primary datasets:

- TTC Bus Delay Data: This database contains records of delays on various bus routes in Toronto, including attributes like Date, Time, Route, Location, Incident Type, Delay Duration, and Vehicle Details.
- TTC Streetcar Delay Data: Provides information on streetcar delays, with similar attributes to the bus delay dataset, attributes like Date, Time, Route, Location, Incident Type, Delay Duration, and Vehicle Details
- TTC Subway Delay Data: Includes data on subway delays, offering insights into issues affecting underground transit routes. Some of the attributes are Date, Time, Day, Station, Code, Min Delay, Min Gap, Bound, Line, Vehicle.

The data is in an excel file (XLXS format) on the City of Toronto's Open Data Portal. Each dataset can be directly downloaded and processed for analysis.

Following are the secondary datasets

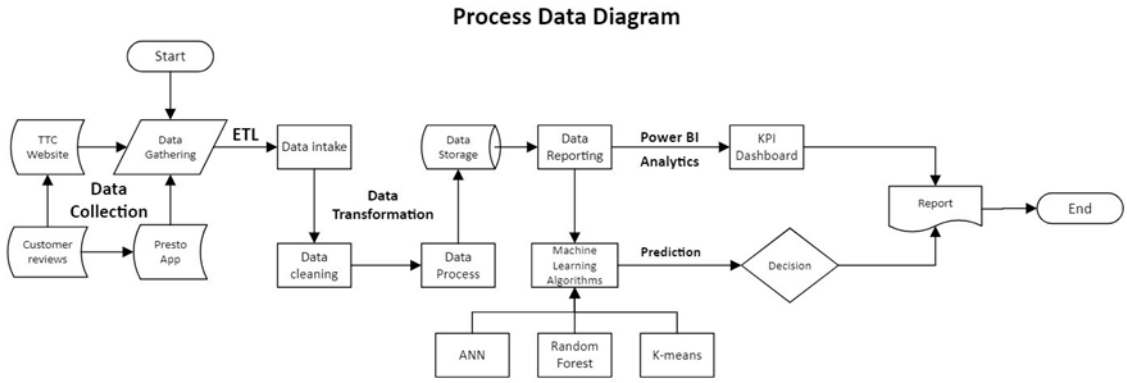
- Stops: This dataset includes all the bus, streetcar and subway stops name and their geographic location.
- Routes: This dataset includes routes of all bus, streetcar and subway routes in greater Toronto area.

Key Data Entities:

The business problem revolves around understanding, analyzing, and mitigating delays in Toronto's transit system. The main business data entities relevant to this problem include:

- Date/Time: Timestamp indicating when the delay occurred.
- Route: Specifies the route number or name where the delay happened.
- Location: Describes the geographical area or station where the delay was reported.
- Incident Type: Categorical data that identifies the reason for the delay (e.g., Traffic,
- Security Issue, Mechanical Failure).
- Min Duration: Numeric data showing the length of the delay in minutes.
- Vehicle Details: Information about the type of vehicle involved in the delay, such as bus number or streetcar ID.

Data flow:



7. Brief overview of data manipulation process and data output

The manipulation of data involves preparing it for analysis by addressing issues such as missing values, inconsistent formats, and scaling differences among features. In this case, the steps undertaken were:

Handling Missing Values:

Missing data is a common issue in datasets. To address this: Imputation with Min-Max Scaling: Missing values were replaced using a method informed by the range of the data. Min-max scaling transforms the data to a specified range, typically between 0 and 1, Here, missing values may be filled by interpolating or imputing values derived from this scaled range, ensuring that the imputed data maintains consistency with the normalized dataset.

- Deriving Longitude and Latitude

For the missing geographic coordinates (longitude and latitude), machine learning algorithms were employed. This involves:

1. Predictive Modeling:
    a. Feature Engineering: Other features in the dataset (e.g., city, zip code, address, regional identifiers) were used as predictors for longitude and latitude.
    b. Algorithm Choice: Regression algorithms such as Decision Trees, Random Forests, or Gradient Boosting could be used to predict the continuous outputs of longitude and latitude.
2. Training the Model:
    a. A training dataset with complete geographic data was used to train the model.
    b. The model learns the patterns and associations between the predictors and the geographic coordinates.
3. Prediction:
    a. The trained model was then used to estimate the missing longitude and latitude for rows where these values were unavailable.

- Advantages of This Approach

1. Comprehensive Data: Filling missing values ensures that no data points are discarded, which is crucial for maintaining robust analysis and avoiding biases.
2. Consistent Scaling: Min-max scaling ensures uniformity in data ranges, reducing issues of disproportionate weighting during analysis.
3. Enhanced Geographical Information: Using machine learning to derive coordinates enables the enrichment of the dataset with crucial spatial information, unlocking opportunities for geospatial analysis and visualization.

- Implications

These preprocessing techniques not only address data quality issues but also enhance the dataset for downstream tasks such as clustering, mapping, or any other analytical procedures that depend on complete and normalized data.

8. New solution design and it's fit into the existing IT architecture

Overview of Solution Design

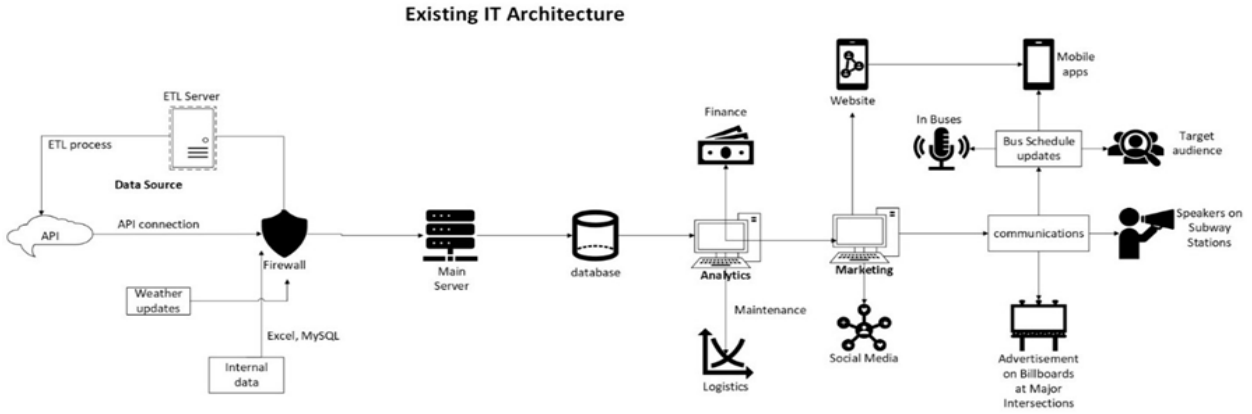
The proposed solution introduces an enhanced data processing pipeline and business intelligence system that is built to transform TTC's data into aconable insights for operational optmization. The system integrates ETL processes, data storage, machine learning, and advanced reporting through Power BI dashboards. This solution is intended to automate and improve the accuracy of reporting and predictive analysis for TTC, providing an interactive KPI dashboard, robust predictive model outputs, and actionable data for marketing, finance, and customer communications.

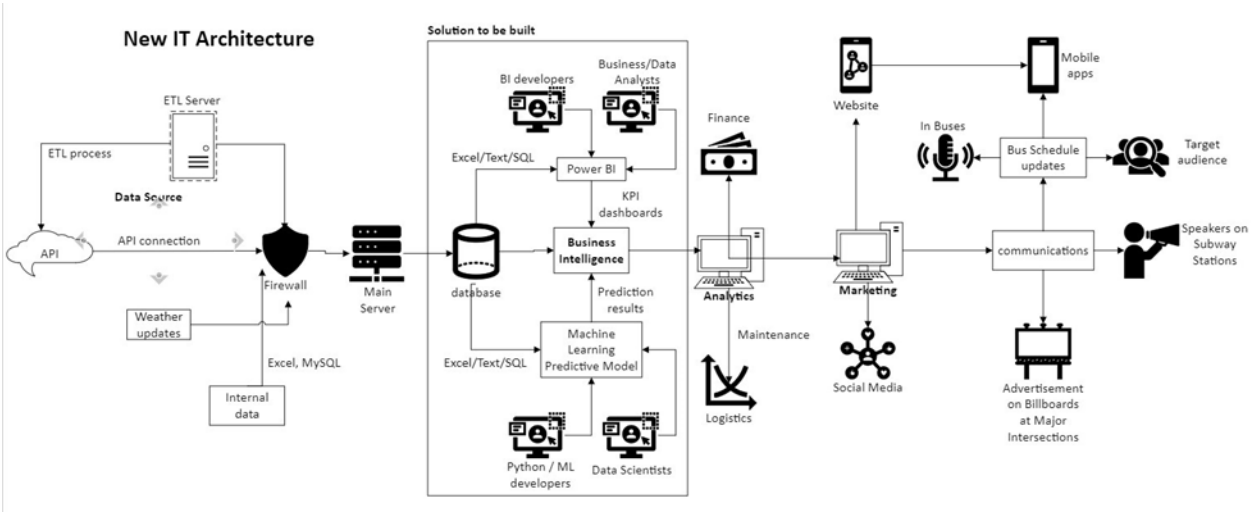
Key Components

- Data Collecon and ETL Processes: The solution begins with data collection from diverse sources, including the TTC website, Presto App, customer reviews, and weather APIs. These sources provide a continuous stream of real-time and historical data. The ETL server processes raw data, handles data intake and cleans and transforms steps. This structured and clean data is stored in a centralized database, accessible for downstream processes.
- Data Storage: The central database is hosted on secure servers with firewall protocol, ensuring data security and regulatory compliance. The database contains structured data for reporting, analyses, and machine learning models.

- Machine Learning Algorithms for Prediction: The solution includes machine learning algorithms such as Artificial Neural Networks (ANN), Random Forest, and K-means clustering, used to predict bus delay patterns, optimize routes, and improve service reliability. These algorithms are designed for scalability, handling varying data volumes efficiently.
- Business Intelligence and Reporting with Power BI: The solution's reporting interface is developed in Power BI. It provides KPI dashboards that consolidate TTC's key metrics, such as delay times, route efficiency, and customer satisfaction scores. This accessible interface enables non-technical stakeholders to interact with data visualizations and retrieve valuable insights.

Fit of the new solution into the existing IT architecture





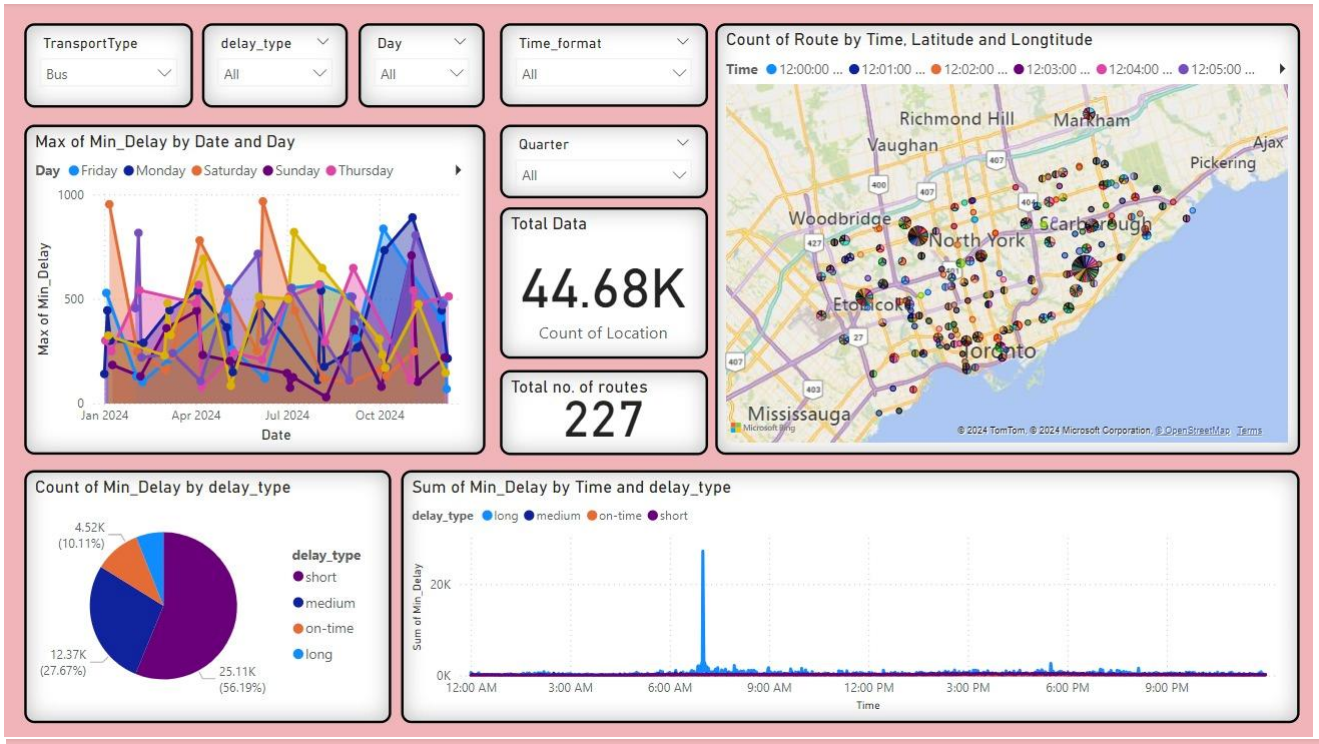
The new solution seamlessly integrates into TTC's existing IT architecture by building on existing infrastructure and data sources while enhancing analytics capabilities:

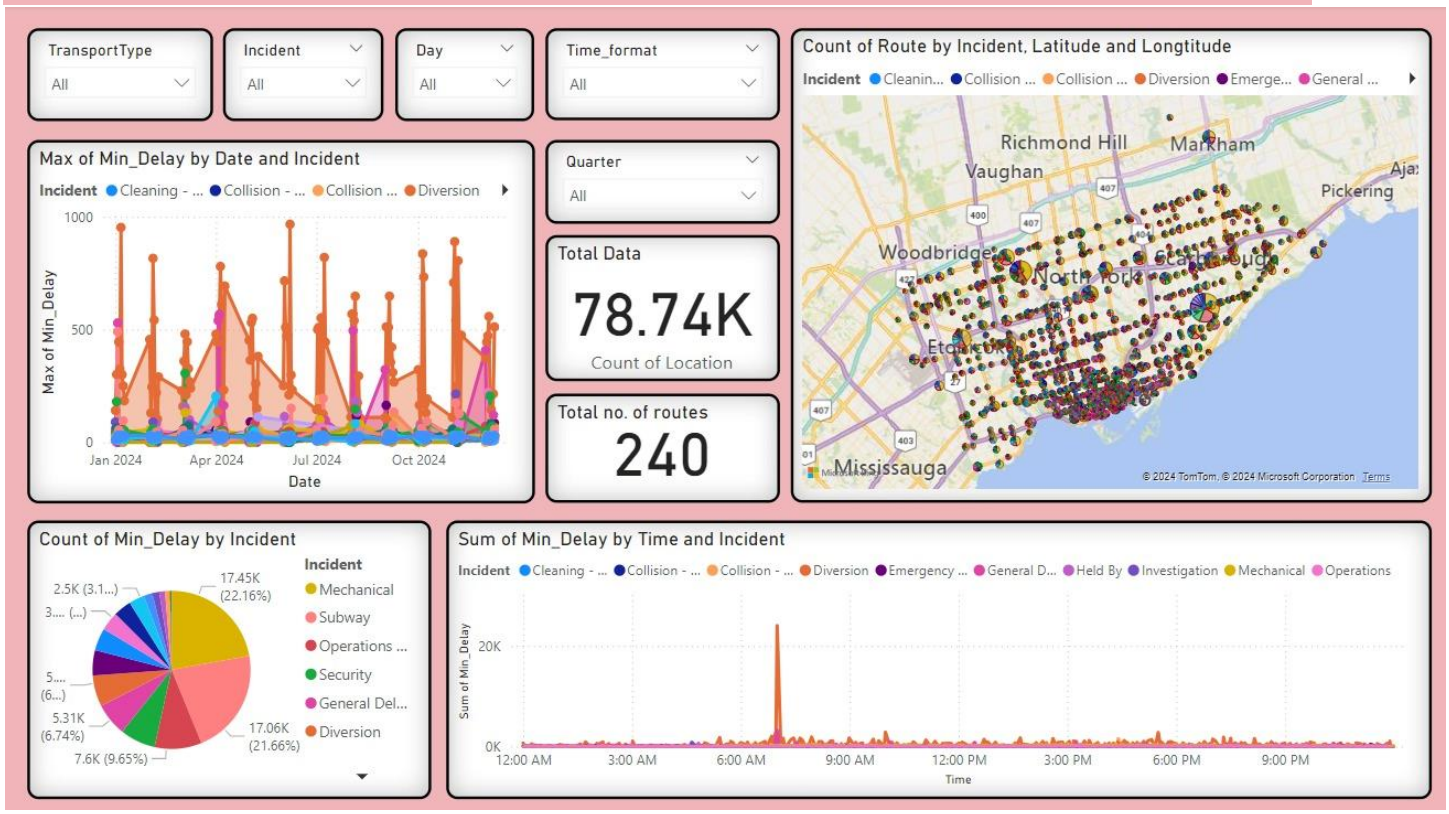
- Data Integration: Data from APIs, the TTC website, the Presto App, and customer reviews are fed into the ETL server, which performs data intake and transformation. This fits within the current data flow without disrupting the existing systems.
- Enhanced Security: The main server and database remain protected by firewalls, maintaining security protocols already in place. The solution introduces new data

science models and BI tools while ensuring data privacy and security.

- Improved Business Intelligence and Analytics: Power BI dashboards and predictive models are new elements that augment TTC's analytical capabilities without affecting the basic IT components. Business intelligence users, such as finance and marketing teams can access these insights within their existing workflows.
- Scalable Architecture: The machine learning models, and BI dashboards are designed to scale based on the organization's data volume and user demand, allowing TTC to adapt as data needs grow over me.

9. New solution implementation and outcome testing





10. Potential solution optimization

1. Descriptive Analytics Optimization

a. Most Frequent Causes of Delays

- **Solution:**
  - Implement targeted training for drivers and dispatchers to handle frequent issues like "General Delays" or "Vision."
  - Introduce technology solutions like real-time GPS monitoring to preemptively address potential delays.
  - Develop rapid response teams for high-frequency incident types.

b. Routes Experiencing the Most Delays

- **Solution:**
  - Reevaluate schedules and traffic patterns for these routes.
  - Use dynamic scheduling during high-traffic periods to reduce congestion.
  - Introduce bus-only lanes on heavily delayed routes where feasible.

c. Times of Day with High Delay Frequency

- **Solution:**
  - Optimize workforce allocation (drivers, support staff) during peak hours.
  - Implement staggered dispatch timings to reduce bottlenecks during rush hours.

d. Days of the Week with Most Delays

- **Solution:**
  - Analyze specific weekday trends and adjust schedules accordingly.
  - Collaborate with city traffic management to improve signal timing on these days.

e. Delays by Month or Season

- **Solution:**
  - Increase vehicle maintenance checks during winter months to prevent weather-related delays.
  - Provide training for drivers on handling seasonal challenges like snow or rain.

f. Total Delay Time by Route and Month

- **Solution:**
  - Prioritize infrastructure improvements on routes with consistently high delay durations.
  - Test route diversions to reduce travel times.

g. Key Locations with Frequent Delays

- **Solution:**
  - Work with city planners to improve congestion at critical intersections like Bloor-Yonge or Union Station.
  - Add more frequent service intervals at these locations to distribute passenger loads.

h. Vehicle Types with the Most Delays

- **Solution:**
  - Upgrade older vehicles (buses, streetcars) that frequently experience mechanical issues.
  - Expand the use of real-time monitoring systems for all vehicle types.

2. Diagnostic Analytics Optimization

a. Routes with Longest Delays

- **Solution:**
  - Perform route-specific audits to identify and mitigate recurring delay causes.
  - Use predictive traffic analysis to dynamically reroute vehicles.

b. Incident Types Causing Long Delays

- **Solution:**
  - Introduce preventive maintenance programs for mechanical incidents.
  - Strengthen security measures to reduce delays caused by disturbances.

c. Long Delays by Time of Day

- **Solution:**
  - Optimize route schedules for off-peak and peak times differently.
  - Use data-driven simulations to test delay reduction strategies for long delays during certain periods.

d. Locations with Frequent Long Delays

- **Solution:**
  - Introduce bypass routes or express lanes for vehicles at congested locations.
  - Enhance coordination with local traffic authorities.

e. Peak Hours

- **Solution:**
  - Increase vehicle frequency and capacity during morning and evening rush hours.
  - Implement staggered boarding strategies to reduce boarding times.

3. Predictive Analytics Optimization

a. Predict Delays on Specific Routes

- **Solution:**
  - Build machine learning models using historical data to predict delays and preemptively adjust schedules.
  - Use AI-driven alerts for real-time dispatch adjustments.

b. Forecast Delays for Next Month

- **Solution:**
  - Use seasonal and trend-based forecasting to optimize workforce and vehicle assignments.
  - Plan for additional resources during forecasted high-delay periods.

c. Times of Day Likely to Experience Delays

- **Solution:**
  - Deploy data-driven simulations to test new schedules during predicted delay times.
  - Use crowd monitoring systems to distribute passenger loads effectively.

d. Likelihood of Long Delays on Specific Routes

- **Solution:**
  - Monitor live vehicle performance and dynamically reroute vehicles showing early signs of delays.
  - Use predictive alerts to deploy backup vehicles or staff as needed.

e. Average Delay Duration for Next Week

- **Solution:**
  - Use historical data to determine optimal buffer times for routes, adjusting schedules to prevent cascading delays.
  - Introduce service reliability monitoring dashboards for proactive adjustments.

General Optimization Strategies

- **Data-Driven Decision Making**: Use insights from dashboards to implement real-time adjustments to schedules, routing, and resource allocation.
- **Technology Integration**: Deploy IoT devices and GPS systems for real-time tracking and proactive incident management.
- **Stakeholder Collaboration**: Work closely with city authorities to address external factors like traffic congestion and signal timing.
- **Customer Feedback**: Use passenger feedback to identify recurring issues and fine-tune services.
- **Continuous Improvement**: Regularly monitor metrics like average delay times, frequency of incidents, and customer satisfaction to evaluate the impact of implemented changes.

11. Appendix

1. Data Sources and Components

- **CSV File**: `combined_ttc.csv`
  - Contains historical TTC data including transport type, routes, delays, incidents, locations, and other relevant fields.
- **Power BI File**: `Capstone_final.pbix`
  - Includes two dashboards (Time and Incidents) with visualizations used for analyzing delay patterns and causes.

2. Key Metrics and Variables

- **Date/Time Variables**:
  - `Date`, `Time`, `Day`, `Time_format` (Morning, Afternoon, Evening).
- **Categorical Variables**:
  - `TransportType` (Bus, Subway, Streetcar), `Incident` (e.g., General Delay, Security), `Route`.
- **Numeric Variables**:
  - `Min_Delay` (minutes of delay), `Vehicle` (Vehicle ID).
- **Geographic Variables**:
  - `Location`, `Latitude`, `Longitude`.
- **Delay Types**:
  - Short, Medium, Long, On-Time.

3. Descriptive Analytics Outputs

- **Visualizations and Insights**:
  - Most frequent causes of delays (e.g., bar chart for incident types).
  - Routes with most delays (e.g., bar chart of delays by route).
  - Time of day, days of the week with highest delays (e.g., line chart for trends).
  - Distribution of delay types across routes (e.g., stacked bar chart).
  - Delays at critical locations (e.g., heatmap of delays at key points).

4. Diagnostic Analytics Tools

- **Incident Analysis**:
  - Breakdowns of `Min_Delay` by `Incident` and `Location`.
- **Correlation Analysis**:
  - Relationships between delay types and factors such as time of day, weekday, and route.
- **Geographic Mapping**:
  - Visualization of hotspots using `Latitude` and `Longitude`.

5. Predictive Analytics Methodologies

- **Forecasting Models**:
  - Use historical delay data for seasonal trend analysis and forecasting.
  - Train machine learning models on features like `Route`, `Day`, `Time_format`, and `Incident`.

- **Risk Prediction**:
    - Predict likelihood of long delays using logistic regression or decision trees.

6. Suggested Tools and Techniques

- **Visualization Tools**:
    - Power BI for interactive dashboards.
    - Python or R for advanced statistical analysis.
- **Machine Learning Models**:
    - Time-series analysis (ARIMA, Prophet) for forecasting.
    - Classification models (Random Forest, Gradient Boosting) for predicting delays.

7. Stakeholders and Implementation Areas

- **Stakeholders**:
    - TTC operations team, city planners, passengers, maintenance crews.
- **Implementation Areas**:
    - Route planning, resource allocation, customer communication, infrastructure upgrades.

8. Assumptions and Limitations

- **Assumptions**:
    - Historical data is accurate and complete.
    - Delay patterns are consistent over time.
- **Limitations**:
    - Lack of external data (e.g., weather, traffic congestion) may affect diagnostic analysis accuracy.
    - Predictive models depend on sufficient historical data coverage.

9. Potential Challenges

- Data integration across multiple sources (e.g., weather, traffic).
- Real-time data processing and model deployment.
- Balancing operational changes with passenger satisfaction.

10. Next Steps

- Conduct additional exploratory analysis to refine optimization strategies.
- Incorporate external datasets (e.g., weather, traffic) for enhanced diagnostic insights.
- Develop a prototype predictive system for delay forecasting.
- Perform regular audits to validate the effectiveness of implemented optimizations.

12. Conclusion

This project has provided a comprehensive analysis of TTC delay patterns through descriptive, diagnostic, and predictive analytics, leveraging historical data and Power BI dashboards. Key insights reveal significant

trends, such as the impact of specific incidents, routes, and times of day on delay frequencies and durations. These findings highlight critical areas for intervention and optimization.

The analysis identified general delays, traffic congestion, and mechanical failures as major contributors to service disruptions. Specific routes, such as those serving key transit hubs, were found to be disproportionately affected. Additionally, seasonal patterns, particularly during winter months, exacerbate delays due to adverse weather conditions.

To address these challenges, actionable solutions were proposed:

- Enhancing infrastructure on high-delay routes and at major transit hubs.
- Implementing dynamic scheduling during peak hours and high-traffic seasons.
- Strengthening preventive maintenance programs to minimize mechanical failures.
- Employing real-time monitoring systems and predictive models to forecast and mitigate delays proactively.

These recommendations, if implemented, have the potential to significantly improve service reliability, reduce passenger wait times, and enhance overall customer satisfaction. By leveraging data-driven strategies, TTC can achieve operational excellence and ensure a better transit experience for its riders.

This project demonstrates the value of analytics in transforming operational efficiency and sets a foundation for continued improvement and innovation in public transit systems.