Final Presentation

TTC delays

Devang Gadhiya

Business Problem and Requirements Definition

The primary business problem is frequent bus delays across Toronto, leading to inefficiencies in service and dissatisfaction among riders.
The business requirements include:

Predictive Insights

Root Cause Analysis

Actionable Recommendations

Analytical Questions

- **Descriptive Analysis:**
  - Which time of day (morning, afternoon, evening) experiences the highest frequency of delays?
  - What are the most frequent causes of bus delays?
  - What is the distribution of delay types (Short, Medium, Long) across routes?

- **Diagnostic Analysis:**
  - What incident types are most likely to cause long delays?
  - How does delay duration differ by day of the week?
  - Are delays worse during peak hours (morning/evening rush)?

- **Predictive Analysis:**
  - How can we accurately predict the type of delay (short, medium, or long) based on operational, temporal, and location features such as Latitude, time of day, day of the week, and direction?

Scope Statement

- **Data Analysis:** Analyze the TTC bus delay dataset to uncover trends and insights.
- **Model Development:** Build machine learning models to predict delays and identify key factors (e.g., route, time, incident type).
- **Visualization:** Present insights through graphs and dashboards for decision-makers.
- **Deliverables:** Comprehensive report, visual dashboards, and delay prediction models to support TTC operations.

Data Overview

Primary Dataset : 3 Secondary Dataset : 2

- The dataset includes **date features**, such as the date (01-01-2024), time (2:00), and day of the week (Monday).
- **Geographic features** include routes (YU, BD, 96, 501), station names, and coordinates (latitude and longitude).
- **Incident details** capture the type of event (Mechincal), delay durations (Min_Delay), and delay categories (On-time, Short, Medium, Long).
- **Vehicle information** covers the travel direction (N, E, S) and vehicle identifiers like 6051.
- **Accessibility features** indicate wheelchair accessibility, with 1 denoting accessible stations.

Overview of data manipulation process and data output

City Connect

1. Data Loading

2. Data Cleaning

- String Manipulation
- Handling Missing Data

3. Data Transformation

- Renaming Columns
- Dropping Irrelevant columns from secondary data

4. Date and Time Processing:

By using to_datetime and re are used to convert date strings into datetime objects for easier analysis.

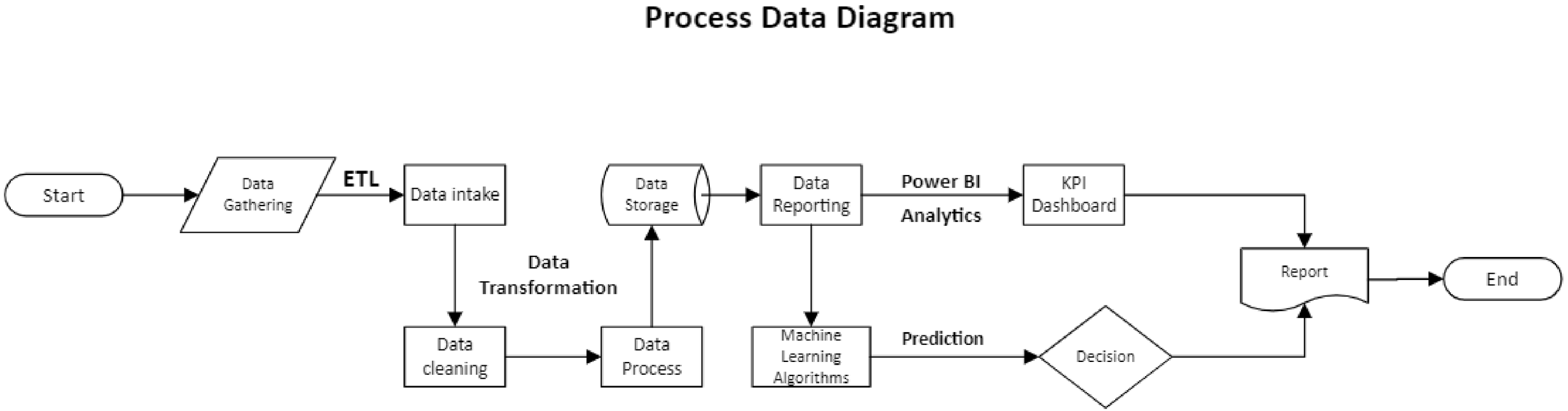
5. Key Operations in the Merging Process :

- Joining Tables
- Handling duplicate records
- Validation

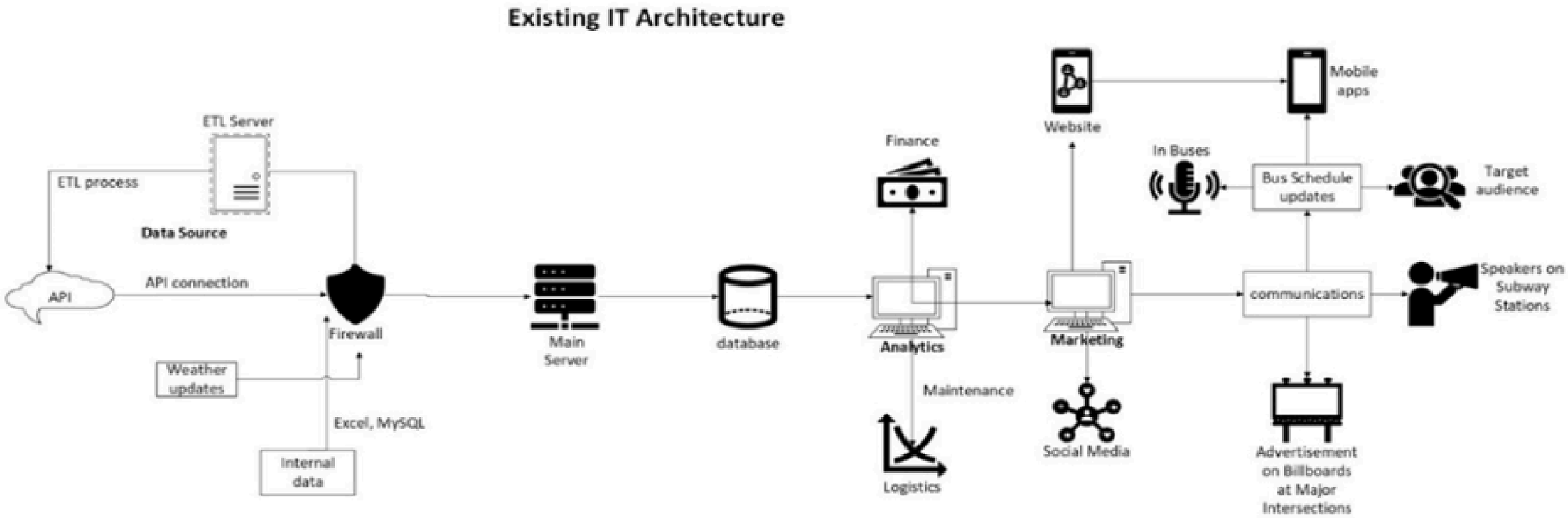
Data Output Summary

- Standardized formats for dates, locations, and Routes.
- Missing data handled or imputed.
- Columns renamed and organized for consistency across datasets.
- Geographical attributes (latitude, longitude) for each stop or station.
- No manual cleaning, created model to do data processing.

Process Data Diagram (Updated)



Existing IT architecture



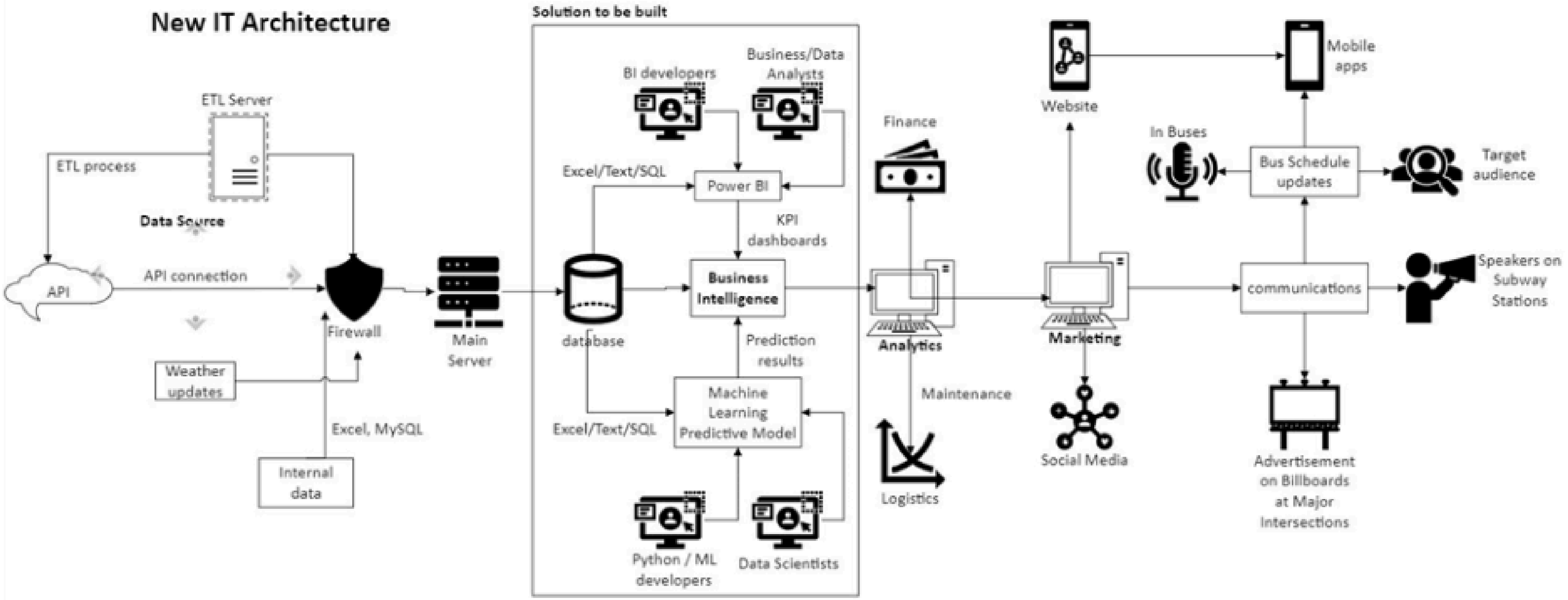
New solution design and it's fit into the existing IT architecture

- **Data Processing & Storage:** Integrates real-time and historical data from diverse sources (e.g., TTC, Presto, weather APIs). Uses ETL processes to clean and transform data, stored securely in a centralized database for analysis and reporting.

- **Machine Learning for Prediction:** Employs scalable algorithms like ANN, Random Forest, and K-means to predict delays, optimize routes, and enhance reliability.

- **Business Intelligence with Power BI:** Provides interactive dashboards for KPIs (e.g., delays, route efficiency, satisfaction scores), enabling actionable insights for marketing, finance, and operations.

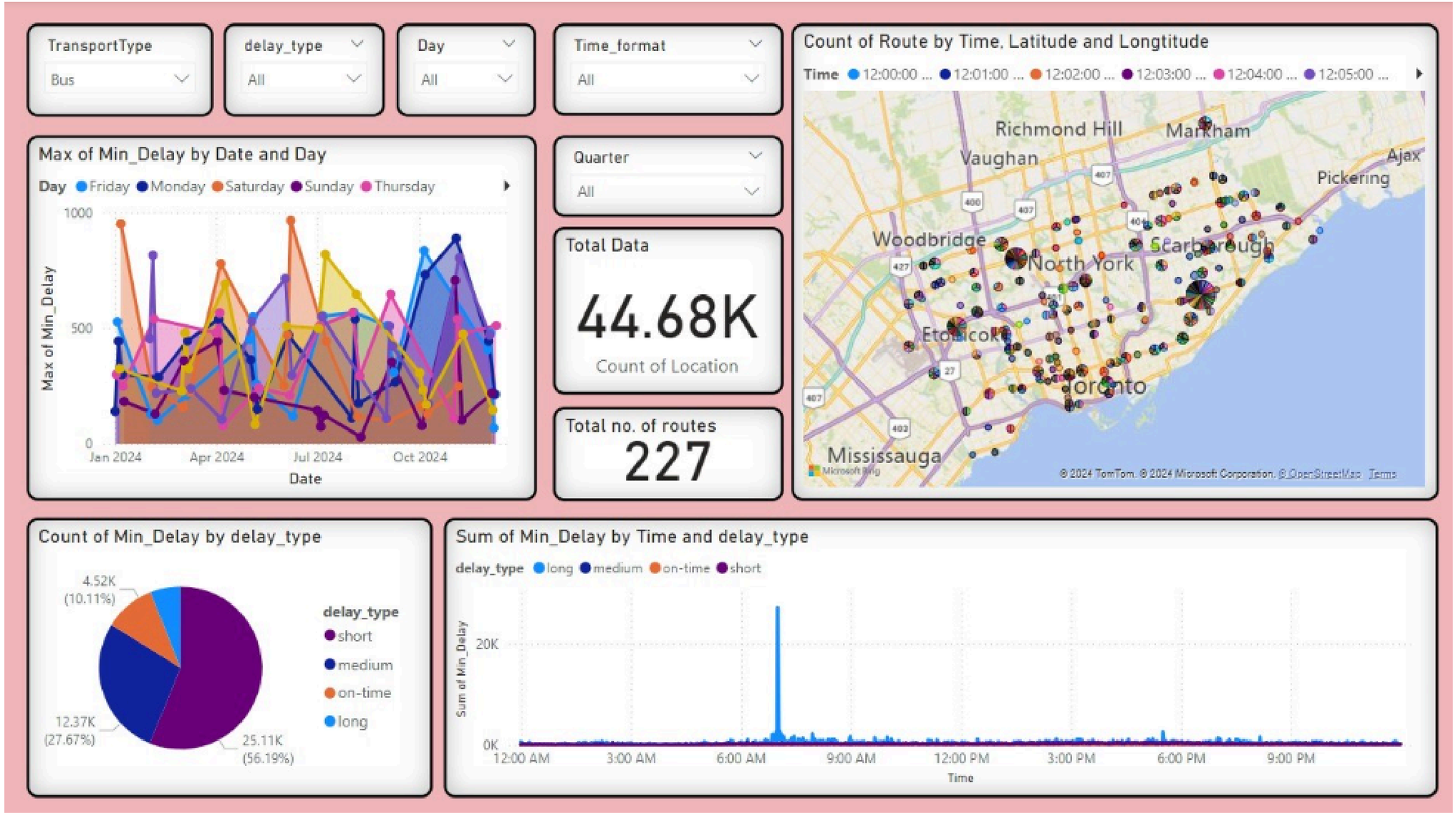
New IT architecture



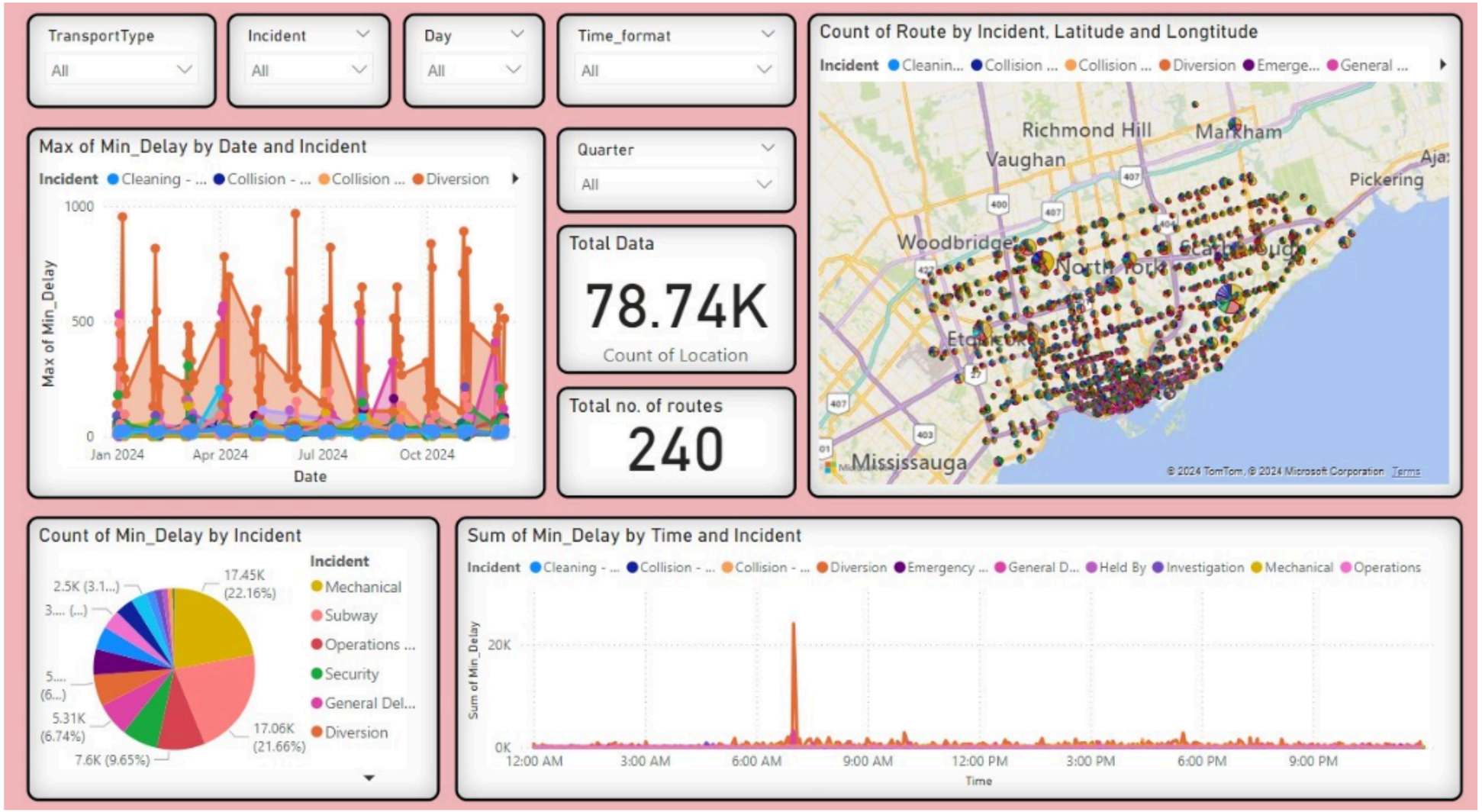
New solution implementation

- **Data Integration:** Seamlessly integrates data from APIs and TTC website into the existing ETL process without disrupting current systems.
- **Enhanced Security:** Maintains existing firewall and security protocols while introducing new data science models and BI tools to ensure privacy and protection.
- **Scalable Business Intelligence**: Power BI dashboards and predictive models enhance analytics without altering core IT components, offering scalable solutions for growing data and user demand.

Outcome testing



Outcome testing



Outcome testing (ML Model)

In [43]:
```
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [44]:
```
# Scale the features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

In [38]:
```
# Build and train the ANN model using MLPClassifier
mlp = MLPClassifier(hidden_layer_sizes=(64), activation='relu', solver='adam', max_iter=1000, random_state=42)
mlp.fit(X_train, y_train)
```

Out[38]:
```
MLPClassifier
MLPClassifier(hidden_layer_sizes=64, max_iter=1000, random_state=42)
```

In [39]:
```
# Make predictions
y_pred = mlp.predict(X_test)
```

In [40]:
```
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f"Test Accuracy: {accuracy:.2f}")
```

```
Test Accuracy: 0.63
```

Potential solution optimization

- **Enhanced Feature Engineering:** Plan to include more detailed temporal and spatial features (e.g., weather conditions, event data) to refine analysis.
- **Model Improvement:** Explore advanced machine learning models (e.g., Gradient Boosting, Hyperparameter Tuning) to enhance predictive accuracy.
- **Real-Time Dashboards:** Develop live Power BI dashboards with real-time data integration for more actionable insights.
- **Geospatial Insights:** Integrate clustering algorithms to better visualize and address delay hotspots.
- **Outcome Focus:** Streamline decision-making with deeper, data-driven insights and predictive analytics for proactive solutions.

Conclusion

This project analyzed TTC delay patterns, identifying key causes such as traffic congestion and mechanical failures, and highlighting high-impact routes and times. Seasonal trends, especially in winter, also contribute significantly to delays. Recommended solutions include enhancing infrastructure, dynamic scheduling, preventive maintenance, and real-time monitoring. Implementing these strategies will improve service reliability, reduce wait times, and enhance customer satisfaction.