BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Multi-Source Traffic Data Integration for Scalable Urban Analysis

Štěpán Bakaj*

Abstract

Modern traffic management systems require integrating data from numerous heterogeneous sources, which often encounter issues with inconsistent formats and limited scalability. This paper describes the extension of an existing analytical application with a robust back-end capable of asynchronous data processing from platforms such as Waze, NDIC, and the Czech Police. The problem is addressed by designing an architecture based on ELT (Extract, Load, Transform) principles and by having microservices communicate via a message broker. The results include an extended data model integrating a new traffic restrictions entity and a system of universal connectors supporting both push and pull data collection models. The proposed solution provides a scalable foundation for complex spatiotemporal analysis and real-time traffic visualization.

*xbakaj00@vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Effective urban traffic management relies on integrating heterogeneous data from sensors and crowdsourcing platforms to provide a comprehensive situational overview [1]. The predecessor to this work [2] was limited by a rigid architecture and fixed schemas, hindering the integration of diverse formats such as DATEX II or JSON. Traditional ETL processes often lack the flexibility required to handle the high volume and variety of dynamic traffic phenomena [3].

We address these limitations by proposing a microservice oriented ELT architecture centered around a message broker for asynchronous data buffering. This approach enables a generalized canonical data model that utilizes JSONB for residual data storage and ensures traceability through external identifiers. Our primary contributions include implementing scalable connectors for Waze and NDIC, introducing a traffic restrictions entity, and a decoupled pipeline that enables independent horizontal scaling of the extraction and transformation layers.

2. System Architecture and Data Flows

The proposed architecture, illustrated in **Fig. 1**, strictly separates the extraction and transformation phases. This approach addresses the issues of varying availability and update frequencies of external sources.

2.1 Extraction Layer and Message Broker

A dedicated microservice (*Extractor*) was implemented for each source. For instance, the `NDIC Extractor` uses a push model, exposing an HTTP endpoint to receive DATEX II formatted data. Raw data are published to the message broker (Redpanda) without immediate modification. Redpanda was chosen for its high throughput and full Kafka protocol compatibility, ensuring system resilience through message retention even during transformer downtime.

2.2 Transformation and Persistence Layer

Transformer microservices consume messages from relevant topics, performing validation, deduplication, and normalization into a canonical schema. The data are subsequently stored in a PostgreSQL database. For efficient handling of spatiotemporal data, **PostGIS** (for spatial queries and traffic-jam geometries) and **TimescaleDB** (for time-series optimization and historical analysis) are used.

3. Generalized Data Model

To accommodate multiple sources, the model transitioned to a canonical schema. Two key mechanisms ensure its robustness:

- **Traceability:** Every record stores an external_ids. This allows the system to track the original identity of the record in the source system, facilitating updates and cross-referencing.
- **Residual Data Storage:** Each entity contains a `raw` column (JSONB). Instead of storing the full payload, it preserves only those elements that were not mapped to the canonical attributes. This ensures that potentially useful auxiliary information is not discarded during transformation.

4. Record Matching and Deduplication

The system implements a deterministic, rule-based algorithm to match point-based reports (e.g., user alerts) with line-geometry events (e.g., computed traffic jams).

4.1 Adaptive Spatiotemporal Filtering

Candidate selection is performed via spatial SQL queries using PostGIS. The filtering parameters are not static; they are dynamically adjusted based on the event context:

- **Event-Type Specific Time Windows:** The temporal buffer is tailored to the nature of the incident. While transient events like traffic jams use narrow windows (e.g., ± 15 min), persistent infrastructure issues like road hazards or closures utilize extended buffers.
- **Road-Class Adaptive Radius:** The spatial search radius for matching a point to a polyline is scaled according to the road's functional class (e.g., 1000 m for motorways vs. 200 m for local roads).

4.2 Priority-Based Decision Logic

If multiple candidates are found, the system applies a priority hierarchy: (1) explicit identifier links, (2) shared topological segment IDs, and (3) minimal temporal distance. If no match is found, the record is either deduplicated against other standalone reports or stored as a new entity.

5. Evaluation Methodology

In the absence of a pre-labeled ground-truth dataset, a manual audit framework was implemented to evaluate the matching accuracy.

5.1 Sample Construction

An export script creates a sample that is divided into two categories:

- **Merged Group:** Cases where the system performed an active merge.
- **Split Candidates:** "Near-miss" cases where events were close in space and time but remained separate. These are tiered (high/medium/low) based on their proximity.

5.2 Metrics

To evaluate the fusion logic, a web-based interface was implemented for manual annotation. Domain-aware users evaluate the system's output, labeling cases as True Positive (TP), False Positive (FP), or Unclear. Unlike purely automated heuristics, this human-in-the-loop audit incorporates road network topology and spatial context to establish a reliable ground truth. The evaluation focuses on two primary metrics: **Merge Accuracy** (the precision of the fusion logic) and the **Missed Merge Ratio** (the rate at which the system fails to aggregate related events). This process ensures a realistic assessment of system performance by accounting for complex spatial relationships that automated algorithms might overlook. The findings of this manual audit are summarized in **Tab. 1**.

6. Conclusions

The implemented system provides a scalable and flexible foundation for traffic data integration. By using a canonical model with residual data storage and an adaptive matching algorithm, the application can effectively consolidate information from diverse providers, significantly improving the quality of traffic analysis in the Brno region.

Acknowledgements

I would like to thank my supervisor, Ing. Magdaléna Ondrušková, for her guidance and support.

References

[1] Martin Treiber and Arne Kesting. *Traffic flow dynamics: Data, models and simulation*. Springer-Verlag Berlin Heidelberg, 2013. ISBN: 978-3-642-32659-2.

[2] Magdaléna Ondrušková. Analýza a vizualizácia dopravných dát v brne. Diplomová práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2024.

[3] Winnie Daamen, Christine Buisson, and Serge P. Hoogendoorn. *Traffic simulation and data: Validation methods and applications*. CRC Press, 2014. ISBN: 978-1-482-21926-5.