Exploratory Data Analysis of London Traffic: Implementation and Interpretation Report

Introduction and Context: London Road Traffic:

Urban traffic systems are among the most data-rich and operationally complex infrastructures in modern cities. In London, effective traffic management increasingly depends on intelligent data strategies rather than traditional engineering alone. This report builds on previous analysis by applying an Exploratory Data Analysis (EDA) model to road traffic data, supported by cloud-native architecture and enterprise-level data tooling.

This type of hybrid architecture is also familiar in my professional environment at SteelcoBelimed, a manufacturer of sterilisation systems for healthcare. There, we manage similarly complex datasets from telemetry, compliance logs, and machine maintenance records using cloud storage, data lakes, and dimensional warehouses. This parallel underscores the broad relevance of enterprise analytics frameworks in both public and industrial contexts.

The assignment explores traffic patterns such as peak-hour congestion, vehicle-type composition, and spatial distribution in the Greater London area. Comparative insights from Bielefeld, a mid-sized German city, are used to highlight alternative traffic dynamics. The ultimate goal is to derive actionable insights and policy recommendations, such as delivery scheduling or congestion pricing.

The London traffic network, with over 31.9 billion vehicle miles in 2022 (DfT, 2023), faces persistent peak-time bottlenecks, especially in central boroughs like Westminster and Camden. While HGV traffic has remained stable, LGV usage has doubled since the early 2000s, amplifying urban congestion. These trends demand scalable, interpretable data solutions to support infrastructure planning and urban mobility policies.

Detailed description of a typical EDA Model placing emphasis on Storage:

The value of such an architecture becomes especially clear when considering complex industrial environments. In my professional role at SteelcoBelimed, our data infrastructure must manage interconnected data streams from machine performance, maintenance logs, environmental sensors, and production scheduling. Like the London traffic ecosystem, this involves both structured and unstructured data collected across dispersed locations. We rely on a similar hybrid model: cloud storage for telemetry and traceability, data lakes for condition-based monitoring and compliance archives, and dimensional warehouses to generate reports for hospital clients and regulators. The architecture enables end-to-end visibility across our machines' lifecycle, much like how a transport authority needs insight from road use to policy.

The implemented EDA model adheres to Tukey’s (1977) framework, moving from data acquisition and cleaning through analysis and visualization. Storage plays a central role in enabling this process. The model integrates a hybrid cloud architecture:

- **Cloud Storage**: Used for raw data ingestion (e.g., CSV files from DfT and JSON APIs).
- **Data Lake**: Operates with a schema-on-read logic, holding raw and semi-structured traffic data. It is ideal for exploration and historical trend analysis.
- **Data Warehouse**: Structured data is transferred to a dimensional warehouse (e.g., Snowflake, BigQuery) using ETL pipelines. The star schema includes a fact table for traffic observations and dimension tables for roads, vehicle types, dates, and geographic zones.

This layered model supports flexibility (through the lake), performance (via the warehouse), and governance (with cloud-based metadata management). When mapped against the ITIL framework, particularly the Service Design and Operation stages, it becomes evident that the current setup could be improved by integrating continual service improvement (CSI) features. ITIL promotes the inclusion of monitoring, feedback loops, and real-time auditing to ensure alignment between data systems and organisational goals. Embedding these capabilities, such as through automated warehouse monitoring or live performance dashboards, would enhance the architecture's alignment with best practices in enterprise system governance (Axelos, 2025). Visual analytics tools such as Power BI are integrated for data representation.

Although Power BI was selected to provide interactive dashboards, technical limitations prevented full implementation. Instead, Python was used to create static visual models and support data interpretation. This ensured analytical rigour while acknowledging platform constraints.

The central elements of the architecturem data lake, ETL processes, and dimensional modelling, remain unchanged from the earlier implementation in Assignment 1. However, the left (data sources) and right (application outputs) sides of the model were updated to reflect this report's emphasis on urban traffic flows and regional comparisons.

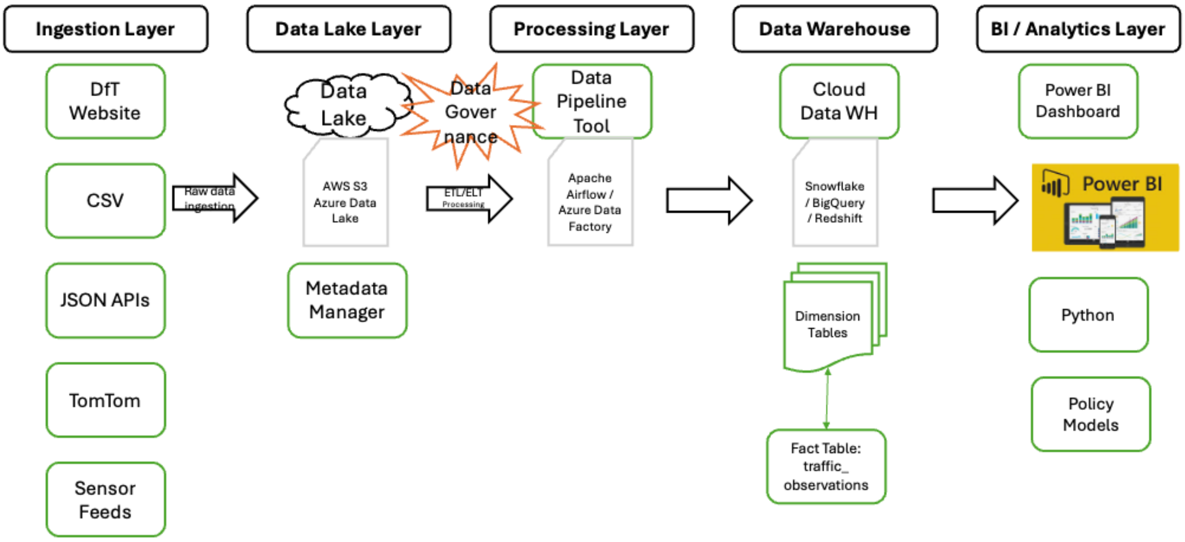


Figure 1: Updated EDA Architecture for London Traffic Analysis - Adapted from Assignment 1 (Hamberger, 2025)

Data Analysis Methods and Deviations

The EDA methodology used in this report follows Tukey's (1977) principles and integrates modern tooling such as Python and Power BI for data exploration and visual storytelling. The analysis is based entirely on publicly available CSV files from the UK Department for Transport (DfT), which offer extensive annual traffic statistics by region, road type, and vehicle classification.

Data Collection: Datasets were retrieved from the DfT London region portal (DfT, 2023), covering road traffic volumes by vehicle type and road category from 1993 to 2023. These files were imported into Power BI to support interactive data exploration and visualisation.

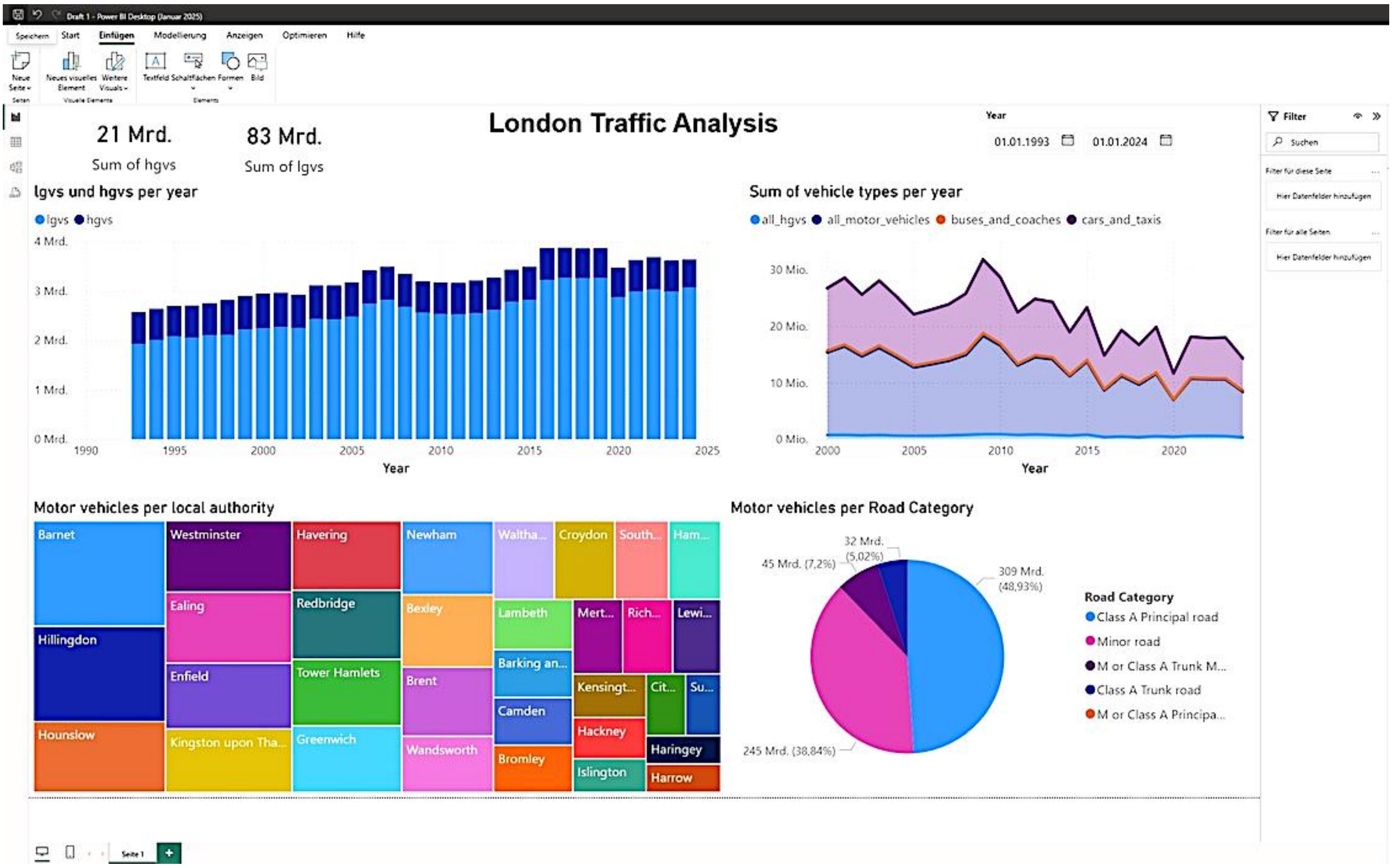


Figure 2: Power BI Dashboard (Hamberger, 2025)

Data Cleaning: Within Power BI, data fields were cleaned by removing estimated or incomplete values, harmonising date and vehicle classification fields, and filtering out records lacking borough identifiers. Data from different files were related using shared dimensions such as year, road type, and vehicle class.

Feature Engineering: Several calculated columns and measures were created in Power BI to enhance analytical insight:

1. lgv_to_total_ratio: Share of LGVs relative to all motor vehicles
2. yearly_change_rate: Annual growth rate in vehicle flow
3. borough_flow_density: Vehicles per borough relative to assumed area
4. vehicle_mix: Proportional breakdown of vehicle types

Segmentation: Visuals and filters were developed to segment the data across:

1. Boroughs (e.g., Westminster, Barnet)
2. Road types (Motorway, A Road, Minor Road)
3. Vehicle types (Car, LGV, HGV, Bus)
4. Temporal trends (yearly change, peak vs. off-peak proxies)

Visualisation: A Power BI dashboard draft was built, displaying a stacked bar chart, area chart, treemap, and a pie chart. These allowed visual exploration of LGV and HGV trends over time, traffic share by borough, and modal breakdowns.

Rationale for Deviation: The original plan involved integrating real-time sensor feeds or JSON APIs. Due to data access limitations and inconsistency in live updates, the method shifted to focus exclusively on cleaned historical data from DfT CSVs. This adjustment preserved data integrity and allowed for the development of a robust and interactive dashboard using Power BI.

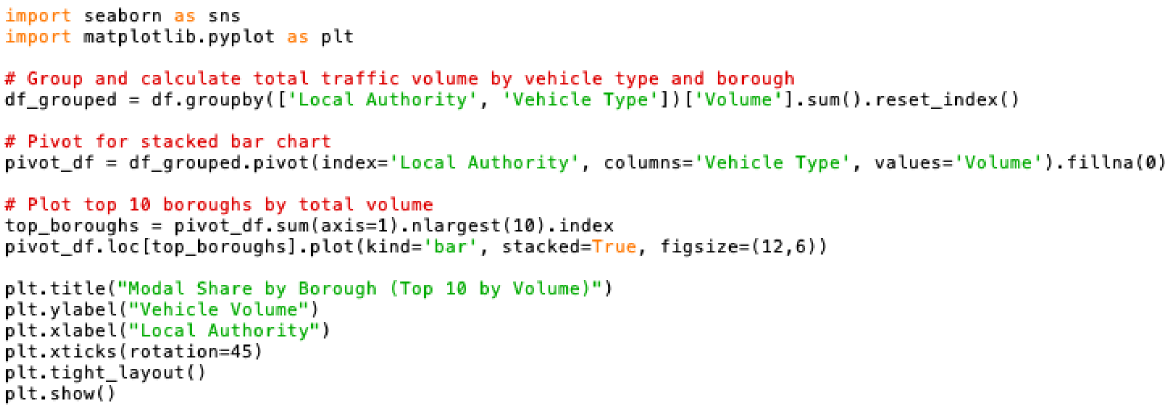
Sequential Presentation of the Data Analysis Results

This section presents the analytical outcomes of the EDA through five visual and quantitative models, reflecting trends in vehicle volume, road usage, and traffic behavior across London’s boroughs. All models are derived from the cleaned and segmented DfT dataset (1993–2023), processed in Power BI and supplemented by using Python (Seaborn and Pandas), which were developed to supplement the Power BI dashboard with additional exploratory plots such as modal share comparisons, density distribution curves, and borough-specific breakdowns. These visuals added statistical depth and flexibility to the descriptive analysis, especially in cases where Power BI's visuals were more summarised or aggregated. Figure 2 illustrates several of the key outputs.

Model	**Description**	**Key Insight**	**Practical Implication**
1	Vehicle Volume by Road Type	49% of traffic on A roads	Target congestion pricing or lane optimisation on key roads
2	Annual Trends in LGVs and HGVs Power BI (Figure 2, top left)	LGV volume doubled since 2000 Aligns with national trends in e-commerce and the growing pressure from delivery services on urban road networks.	Implement smart delivery scheduling
3	Borough-Level Traffic (Treemap visual)	Inner boroughs disproportionately loaded	Promote modal shift via incentives and infrastructure
4	Modal Share Breakdown (pie chart) supporting the findings in Model 1	High private vehicle use, low public share	Use for planning event-based traffic interventions

Table 1: Summary of Data Models, Key Insights, and Practical Implications (Hamberger, 2025)

Additional Python plots confirm the high share of private vehicles and the modest contribution of public transport (buses/coaches), especially outside central zones.



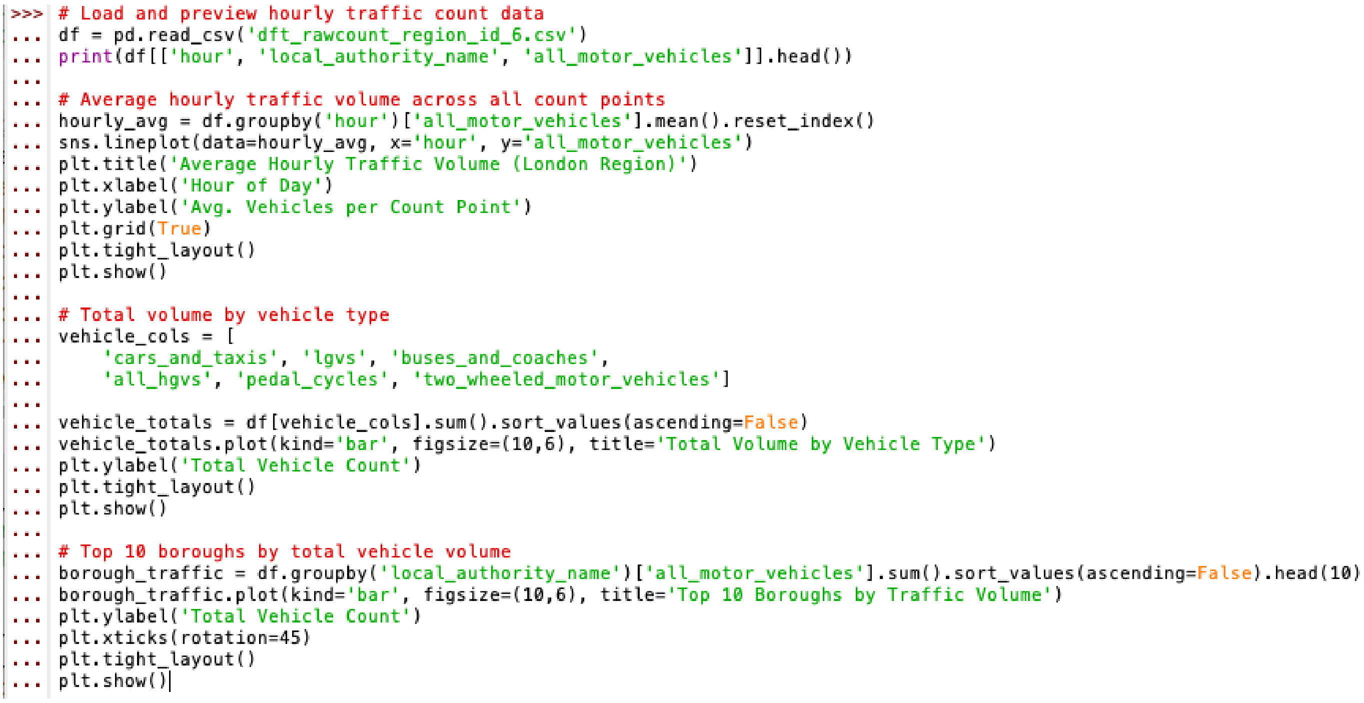
```
import seaborn as sns
import matplotlib.pyplot as plt

# Group and calculate total traffic volume by vehicle type and borough
df_grouped = df.groupby(['Local Authority', 'Vehicle Type'])['Volume'].sum().reset_index()

# Pivot for stacked bar chart
pivot_df = df_grouped.pivot(index='Local Authority', columns='Vehicle Type', values='Volume').fillna(0)

# Plot top 10 boroughs by total volume
top_boroughs = pivot_df.sum(axis=1).nlargest(10).index
pivot_df.loc[top_boroughs].plot(kind='bar', stacked=True, figsize=(12,6))

plt.title("Modal Share by Borough (Top 10 by Volume)")
plt.ylabel("Vehicle Volume")
plt.xlabel("Local Authority")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Figure 3: Python plot of modal share by borough (Hamberger, 2025)

An area chart (Figure 2, top right) shows that car and taxi traffic remain dominant over the last two decades, though subject to seasonal fluctuations. Traffic dips during COVID lockdowns and summer holidays are clearly visible. These long-term patterns reflect both behavioral and regulatory shifts.

Together, these models offer a multidimensional view of London’s traffic ecosystem and lay the groundwork for targeted interpretation in the following section.

To provide further transparency and reproducibility, an alternative data analysis pipeline was implemented entirely in Python, using Pandas and Seaborn. This method was applied to the hourly raw count dataset from the DfT and offers detailed flexibility for statistical exploration without relying on proprietary visualisation platforms. The following examples summarise key components of the analysis:



```
>>> # Load and preview hourly traffic count data
... df = pd.read_csv('dft_rawcount_region_id_6.csv')
... print(df[['hour', 'local_authority_name', 'all_motor_vehicles']].head())
...
... # Average hourly traffic volume across all count points
... hourly_avg = df.groupby('hour')['all_motor_vehicles'].mean().reset_index()
... sns.lineplot(data=hourly_avg, x='hour', y='all_motor_vehicles')
... plt.title('Average Hourly Traffic Volume (London Region)')
... plt.xlabel('Hour of Day')
... plt.ylabel('Avg. Vehicles per Count Point')
... plt.grid(True)
... plt.tight_layout()
... plt.show()
...
... # Total volume by vehicle type
... vehicle_cols = [
...     'cars_and_taxis', 'lgvs', 'buses_and_coaches',
...     'all_hgvs', 'pedal_cycles', 'two_wheeled_motor_vehicles']
...
... vehicle_totals = df[vehicle_cols].sum().sort_values(ascending=False)
... vehicle_totals.plot(kind='bar', figsize=(10,6), title='Total Volume by Vehicle Type')
... plt.ylabel('Total Vehicle Count')
... plt.tight_layout()
... plt.show()
...
... # Top 10 boroughs by total vehicle volume
... borough_traffic = df.groupby('local_authority_name')['all_motor_vehicles'].sum().sort_values(ascending=False).head(10)
... borough_traffic.plot(kind='bar', figsize=(10,6), title='Top 10 Boroughs by Traffic Volume')
... plt.ylabel('Total Vehicle Count')
... plt.xticks(rotation=45)
... plt.tight_layout()
... plt.show()
```

Figure 4: Alternative Python data analysis (Hamberger, 2025)

This Python-based approach provided an open-source alternative to Power BI and enabled a deeper, customisable analysis of the raw hourly data. A comparative evaluation of both tools follows in the next section.

Interpretation of Results and Issues Identified

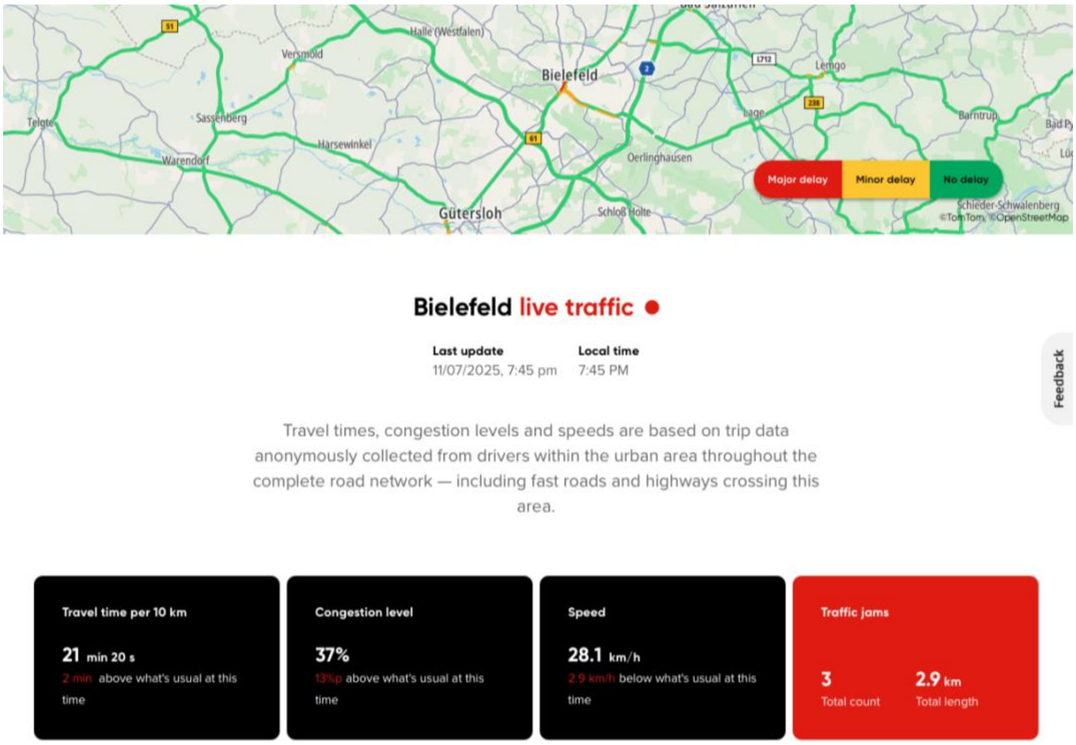
The concentration of nearly 90% of traffic on a limited subset of A-roads and motorways reveals a data centralisation issue: sensor networks and traffic counters are often unevenly distributed, leading to blind spots in data collection. This suggests the need for more decentralised edge data acquisition, supported by cloud-native ingestion pipelines that scale automatically with demand (Ghosh, Basu & O'Mahony, 2009). Without rebalancing the data architecture, analytics risk reinforcing existing biases in infrastructure visibility.

The surge in LGV traffic, driven by e-commerce, challenges current batch-processing architectures. Reliance on static data, such as annual CSVs, limits responsiveness. This demonstrates a clear need for streaming data pipelines(e.g., Apache Kafka, AWS Kinesis) and predictive analytics layers to manage commercial traffic in near real time (Lehe, 2019). From an IT operations view, this aligns with ITIL's Continual Service Improvement (CSI) cycle, requiring infrastructure that adapts as traffic dynamics evolve (Axelos, 2025).

Traffic imbalances across boroughs expose inconsistencies in spatial data quality and metadata governance**.** These discrepancies suggest the need for a shared, geospatial data warehouse that adheres to a unified schema and supports both analysis and operational planning. A data lakehouse approach would be ideal—combining the flexibility of a data lake with the structure and governance of a warehouse (Wickham, 2014).

The dominance of private vehicles over shared modes points to a lack of real-time integration across modal datasets**.** Current architectures reflect a siloed design, where public transport, micromobility, and traffic sensor data are not interoperable. To support modal shift strategies, API-first, microservice-oriented ERP 2.0 architectures are needed, enabling modular, real-time data fusion from diverse sources (Chan, Abu Khadra & Alramahi, 2011; Monk & Wagner, 2001).

To support interpretation, selected screenshots from the TomTom Traffic Index were included to visualise congestion comparisons between London and Bielefeld:



Bielefeld live traffic ●

Last update 11/07/2025, 7:45 pm

Local time 7:45 PM

Travel times, congestion levels and speeds are based on trip data anonymously collected from drivers within the urban area throughout the complete road network — including fast roads and highways crossing this area.

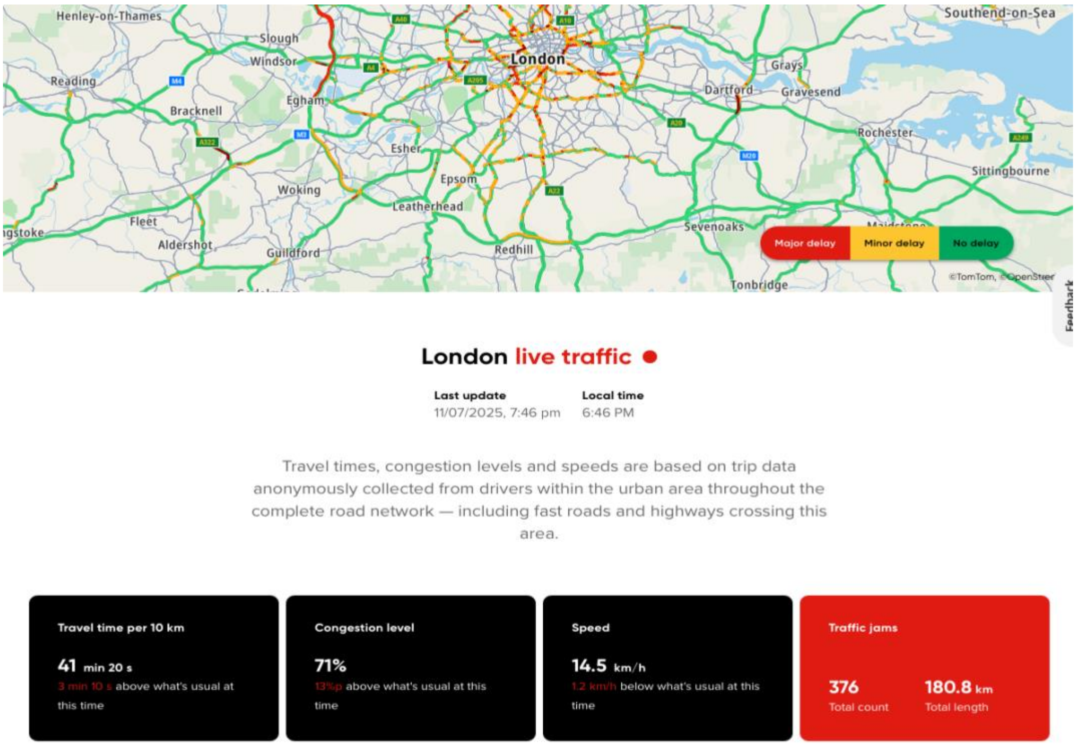
Travel time per 10 km
21 min 20 s
2 min above what's usual at this time

Congestion level
37%
13%p above what's usual at this time

Speed
28.1 km/h
2.9 km/h below what's usual at this time

Traffic jams
3 Total count
2.9 km Total length

Figure 4: TomTom Delay Comparison – London vs. Bielefeld: Bielefeld (TomTom, 2025)



London live traffic ●

Last update 11/07/2025, 7:46 pm

Local time 6:46 PM

Travel times, congestion levels and speeds are based on trip data anonymously collected from drivers within the urban area throughout the complete road network — including fast roads and highways crossing this area.

Travel time per 10 km
41 min 20 s
3 min 10 s above what's usual at this time

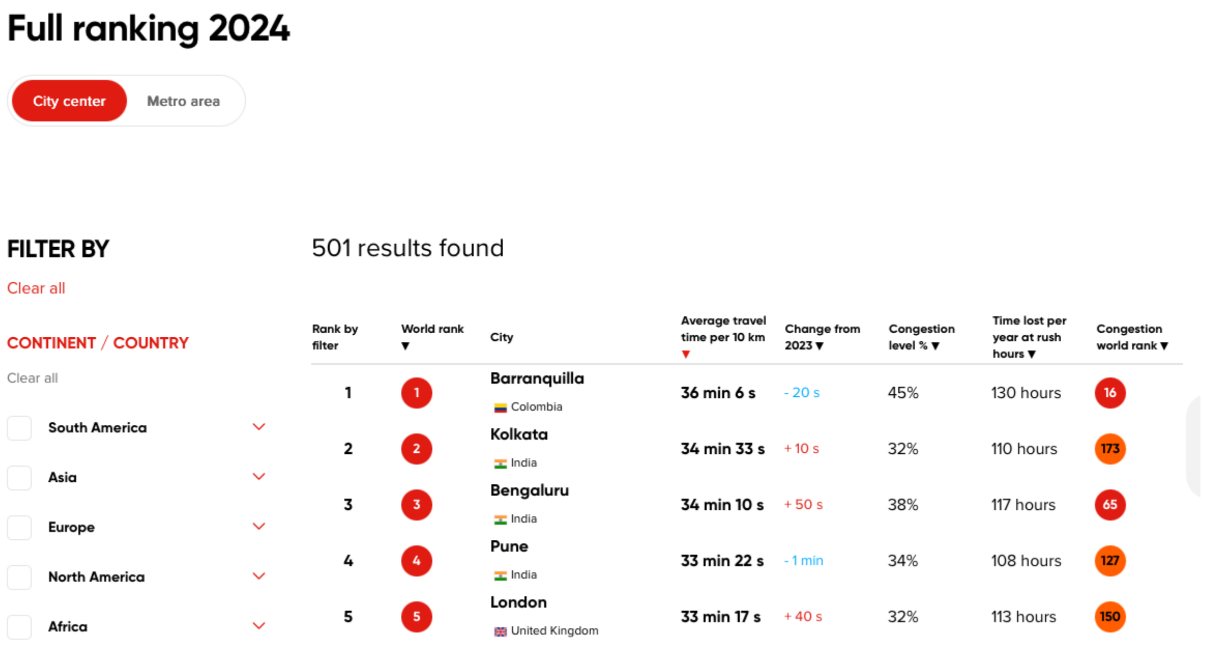
Congestion level
71%
13%p above what's usual at this time

Speed
14.5 km/h
1.2 km/h below what's usual at this time

Traffic jams
376 Total count
180.8 km Total length

Figure 5: TomTom Delay Comparison – London vs. Bielefeld: London (TomTom, 2025)

These screenshots shows that London experienced an average delay of 41 minutes per 10 km on Friday 11 July in the evening, while Bielefeld reported just 21 minutes (TomTom, 2024).



Full ranking 2024

FILTER BY

Clear all

CONTINENT / COUNTRY

Clear all

□ South America
□ Asia
□ Europe
□ North America
□ Africa

501 results found

Rank by filter	World rank ▼	City	Average travel time per 10 km ▼	Change from 2023 ▼	Congestion level % ▼	Time lost per year at rush hours ▼	Congestion world rank ▼
1	1	Barranquilla Colombia	36 min 6 s	- 20 s	45%	130 hours	16
2	2	Kolkata India	34 min 33 s	+ 10 s	32%	110 hours	173
3	3	Bengaluru India	34 min 10 s	+ 50 s	38%	117 hours	65
4	4	Pune India	33 min 22 s	- 1 min	34%	108 hours	127
5	5	London United Kingdom	33 min 17 s	+ 40 s	32%	113 hours	150

Figure 6: TomTom Congestion Ranking Chart (TomTom, 2024)

Figure 6 visualises London’s global top-10 congestion rank compared to Bielefeld's position outside the top 100 (TomTom, 2024).

Across all models, a systemic limitation is the reliance on historic, non-live data. While sufficient for strategic EDA, this hinders tactical, real-time interventions. A cloud-native stack, combining serverless processing, streaming ingestion, and automated reporting, would support scalable, flexible traffic intelligence (Laudon & Laudon, 2006). Integration with machine learning models (e.g., Prophet or ARIMA) could enable predictive traffic load balancing and smarter urban planning.

Finally, when compared to cities like Bielefeld, it’s clear that London’s complexity demands a stronger IT foundation. A smart-city-grade traffic platform, built on open-source tools and deployed in the cloud, would enable borough-level autonomy, shared insights, and continuous optimisation (Batty, 2024).

Solutions and Associated Risks

Solution	Technology & Architecture	Goal	Risks / Limitations	Reference
Dynamic Congestion Pricing	Real-time data feeds, cloud-native deployment, edge sensors	Reduce peak-hour congestion in urban cores	Equity concerns; high system integration cost	Lehe (2019)
Delivery Slot Allocation for LGVs	API-based coordination, hybrid cloud with ITIL governance	Decentralise and smooth delivery traffic flow	Requires cooperation from logistics operators	(Bachofner et al., 2022)
Borough-Level Modal Shift	Data lakes, AI-based simulations, digital twins	Encourage shift to micromobility and public transport	User resistance; policy fragmentation across boroughs	Banister (2008)
Traffic Intelligence Dashboard	Streaming data, predictive analytics (ARIMA, Prophet), integrated BI	Enable real-time oversight and proactive planning	Data privacy; latency; platform funding	Ghosh et al. (2009)

Table 2: Data-Driven Solutions to London Traffic Issues (Hamberger, 2025)

Rationale Summary

These solutions align technical feasibility with policy relevance and data science maturity. Each proposal addresses specific insights from the EDA models and reflects IT service delivery principles such as CSI and service design architecture.

This report implemented an EDA-based strategy to evaluate London’s traffic data, placing it in practical comparison with Bielefeld. A hybrid architecture of cloud storage, data lakes, and warehouses supported scalable analytics and business intelligence. The results demonstrated key traffic inefficiencies and provided evidence-based solutions, although trade-offs must be acknowledged. This framework serves as a foundation for future integration of real-time systems and predictive analytics.

Wordcount: 2142

References:

Hamberger, G. (2025) Enterprise Data Executive Summary and Implementation Report. Data Sciene 2025. Essay submitted to the University of Essex Online.

Batty, M. (2024) *Inventing future cities*. MIT Press.

Bachofner, M. et al. (2022) 'City logistics: Challenges and opportunities for technology providers,' *Journal of Urban Mobility*, 2, p. 100020. Available at: https://doi.org/10.1016/j.urbmob.2022.100020.

Banister, D. (2008) 'The sustainable mobility paradigm,' *Transport Policy*, 15(2), pp. 73–80. Available at: https://doi.org/10.1016/j.tranpol.2007.10.005.

Bielefeld traffic report | TomTom Traffic Index (2025). Available at: https://www.tomtom.com/traffic-index/bielefeld-traffic/ (Accessed: July 12, 2025).

Chan, J.O., Abu-Khadra, H. and Alramahi, N. (2014) 'ERP II readiness in Jordanian industrial companies,' *Communications of the IIMA,* 11(2). Available at: https://doi.org/10.58729/1941-6687.1163.

DFT (2023) *Road traffic statistics – London region*. Available at: https://roadtraffic.dft.gov.uk/regions/6 (Accessed: June 6, 2025).

Ghosh, B., Basu, B. and O'Mahony, M. (2009) 'Multivariate Short-Term traffic flow Forecasting using Time-Series analysis,*' IEEE Transactions on Intelligent Transportation Systems*, 10(2), pp. 246–254. Available at: https://doi.org/10.1109/tits.2009.2021448.

Axelos (2025). ITIL 4 Foundation Certification. Available at: https://www.axelos.com/certifications/itil-service-management/itil-4-foundation (Accessed: July 12, 2025).

Laudon, J. and Laudon, K. (2006) *Management Information Systems: Managing the Digital Firm (10th Edition), Prentice-Hall, Inc eBooks*. Available at: https://dl.acm.org/citation.cfm?id=1197076.

Lehe, L. (2019) 'Downtown congestion pricing in practice,' *Transportation Research Part C Emerging Technologies*, 100, pp. 200–223. Available at: https://doi.org/10.1016/j.trc.2019.01.020.

London traffic report | TomTom Traffic Index (2025). Available at: https://www.tomtom.com/traffic-index/london-traffic/ (Accessed: July 12, 2025).

Monk, E. and Wagner, B. (2001) *Concepts in enterprise resource planning.* Available at: http://ci.nii.ac.jp/ncid/BA84760973.

Traffic Index ranking | TomTom Traffic Index (2024). Available at: https://www.tomtom.com/traffic-index/ranking/ (Accessed: July 12, 2025).

Tukey, J.W. (1977) *Exploratory data analysis*. Addison-Wesley Publishing Company.

Wickham, H. (2014) *Tidy Data*. Journal of Statistical Software, 59(10), pp.1–23. Available at: https://www.jstatsoft.org/article/view/v059i10 (Accessed: 4 July 2025).