# Enterprise Data Analysis Strategy for Road Traffic Management in London

Tobias Zeier, the 5th of June 2025
University of Essex Online

## 1 Introduction

The London region, known for its dense population and extensive transport network, faces growing challenges in managing congestion, ensuring road safety, and meeting environmental goals. The Department for Transport's 2023 report on road traffic statistics provides key insights into these issues. This report seeks to critically evaluate those statistics and to design an enterprise data analysis (EDA) strategy to improve the management of road traffic. It examines enterprise data architectures, including data lakehouses, cloud technologies, and analytics tools, and proposes a robust EDA model aligned with business intelligence (BI) goals. Emphasis is placed on congestion-related data and the strategy's capacity to support data-driven decision-making.

## 2 Description of Road Traffic Statistics: Congestion and Current Issues

The 2023 Road Traffic Statistics report highlights that London's road network continues to experience significant congestion, particularly during peak commuting hours. Inner London is especially affected due to higher population density and restricted road space. Key issues identified include the rising volume of private vehicle journeys and poor bus reliability in heavily congested areas. Heavy Goods Vehicles (HGVs) and delivery vans, largely driven by the growth of e-commerce, have also added considerable pressure to the road network (Weltevreden, 2024). Notably, average speeds on the Strategic Road Network in London decreased by 0.07% between 2018 and 2019, while average delays increased by 12% over the same period, making London the slowest average speed and highest delays region in the United Kingdom. These earlier figures are intentionally referenced, because more recent data has been skewed by the COVID-19 pandemic's impact on travel behaviour and traffic patterns. Although average traffic in London in 2023 was 0.5% higher than in 2022, it remained 5.4% lower than pre-pandemic levels. Crucially, some measurements are still performed manually, and real-time traffic data remains underutilised, limiting the city's ability to implement dynamic, data-driven responses to evolving congestion patterns. The EDA strategy outlined in this document will address these issues and recommend a solution.

## 3 EDA Framework

While the CRISP-DM framework remains a widely adopted standard for structured data mining projects, it presents several limitations in the context of modern, real-time urban systems (Saltz, 2024). It lacks explicit support for streaming data, agile development, and deployment automation – all of which are essential in a dynamic environment like London's road network. Moreover, CRISP-DM offers little guidance on governance, data ethics, or regulatory compliance, which are critical in public-sector applications. For these reasons, a more contemporary and operationally integrated framework such as the Team Data Science Process (TDSP) will be better suited. TDSP, developed by Microsoft, incorporates agile principles, version control, cloud compatibility, and end-to-end lifecycle management, making it highly effective for deploying scalable and compliant enterprise data solutions (Hotz, 2025). TDSP comprises the following five key steps:

- **Business Understanding:** Define the objectives: reducing congestion, improving journey times, optimising traffic flow and enhancing road safety.
- **Data Acquisition and Understanding:** Ingest data such as Traffic Volume Data, Speed Data, Travel Time Data, Incident Data, Public Transport Data, Environmental Data, Static Data, Weather Data, cleaning it and determine if it can answer the current question
- **Modelling:** Apply analytical techniques, predict rush-hour congestions, group accident-prone zones.
- **Deployment:** Integrate insights into operational systems (e.g. traffic light control, policy decision making systems).
- **Customer Acceptance:** Does the system meet business needs?

Alternative frameworks were considered but ultimately deemed less suitable for this context. The KDD Process (Knowledge Discovery in Databases), while foundational in the field of data mining, places less emphasis on business understanding and practical deployment. SEMMA (Sample, Explore, Modify, Model, Assess), developed by SAS, focuses heavily on modelling and assumes a pre-cleaned dataset, making it less adaptable for the varied and complex nature of real-world traffic data. Meanwhile, the OODA Loop (Observe, Orient, Decide, Act) is valuable for real-time decision-making in dynamic environments but lacks the detailed structure needed for comprehensive data preparation and analysis workflows (Swamynathan, 2019).

# 4 Data Architecture and Models

The proposed enterprise data strategy is based on a cloud-hosted lakehouse model, combining the flexibility of data lakes with the structured querying and governance of data warehouses. The unified architecture is well-suited to handle the vast variety of structured and unstructured data generated by London’s road traffic system, including sensor outputs, GPS data and incident logs (Armbrust et al., 2021). Unlike traditional hybrid models, which separate storage and processing layers, lakehouses reduce technical complexity by enabling direct querying of both raw and curated data (Oreščanin & Hlupić, 2021).

Lakehouses also meet enterprise governance requirements. Platforms such as Databricks and Snowflake now offer fine-grained access control, audit trails, and GDPR-compliant data lineage (Mazumdar et al., 2023). They integrate with legacy systems using standard connectors (e.g., JDBC, ODBC, API), supporting tools like Excel and Power BI (Martin, 2023). Its scalability allows it to accommodate more data like pollution and traffic noise.

Compared to a hybrid model, the lakehouse offers a cleaner, more future-proof solution with fewer moving parts. It supports both BI dashboards and machine learning workloads without fragmenting data infrastructure (Schneider et al., 2024). On the other hand, lakehouses are relatively newer concepts and require significant data engineering expertise. This means that there are potentially few experts on the market who have mastered the technology. While hybrid approaches with data lakes and data warehouses have been around for a longer time, there are more projects that have already successfully implemented the same or similar architecture (Herden, 2020).

It is highly recommended to deploy the solution using a major cloud service provider (CSP) such as Google Cloud Platform (GCP), Amazon Web Services (AWS) or Microsoft Azure. The rationale for this is flexible scalability, cost-effective pay-as-you-go models, managed services reducing operational overhead and access to newest features (Sharma, 2021). Potential risks such as vendor lock-in and overconfidence in data security must be addressed. Despite the fact that CSPs are investing billions of dollars to increase data security, it remains a shared responsibility between service provider and customer. Furthermore, potential compliance issues with storing sensitive traffic data offshore, or data sovereignty concerns in cloud environments, need to be critically evaluated (Upreti et al., 2025).

# 5 Data Analysis Design and Methodology

The lakehouse will follow a medallion architecture, separating raw, cleaned, and curated data layers (Sirbu, Taleanu and Pop, 2024). Bronze holds raw unstructured and structured traffic feeds. Silver standardises and validates the data. Gold hosts aggregated, refined data optimised for business intelligence and machine learning.

The data analysis design centres on creating key data assets and their interconnections to answer critical questions. The Gold Layer tables will form a star schema model, linking fact tables to dimension tables using keys. Key datasets include traffic logs, incidents, time, location, and vehicle metadata. This model allows for efficient temporal and spatial querying, enabling planners to identify trends, hotspots, and risk factors

The methodology follows an iterative approach, aligned with TDSP:

1. **Phase 1:** Descriptive Analytics: Focus on "what happened." Build dashboards for current traffic, speeds, and congestion hotspots.
2. **Phase 2:** Diagnostic Analytics: Focus on "why it happened." Deep-dive into anomalies, correlating incidents/weather with congestion.
3. **Phase 3:** Predictive Analytics: Focus on forecasting "what will happen." Develop models to predict future congestion or incident likelihood.
4. **Phase 4:** Prescriptive Analytics: Focus on recommending "what to do." Use models to suggest optimal traffic signal timings or diversion routes.

This phased approach meets immediate reporting needs while progressively building towards sophisticated, proactive traffic management.

## 6 Data Representation Choices

The selection of data representation strategies (time-series, geospatial, multivariate, tabular) is crucial for extracting meaningful insights and presenting them effectively (Dadashova et al., 2020).

- **Time-Series Representation:** Enables identification of temporal patterns and anomalies, fundamental for operational traffic management and informing urban planning. One limitation is that it can be overwhelming if not properly aggregated, potentially obscuring macroscopic trends.
- **Geospatial Representation:** Provides intuitive visual understanding of congestion and incident locations, invaluable for identifying bottlenecks and directing resource deployment (Hazaymeh et al., 2022). Limitation: Privacy concerns arise with raw GPS data, and maps may become cluttered without proper filtering.
- **Multivariate Analysis:** Explores complex relationships between various traffic-influencing factors (e.g., weather, events). Moves beyond simple statistics to diagnostic understanding. Limitation: Requires statistical understanding; complex models can be difficult for non-technical users to interpret.
- **Aggregated Tabular Data:** Offers concise summaries of KPIs for quick overview and reporting, ideal for stakeholders. Limitation: Loses granular detail needed for specific interventions, potentially masking underlying patterns.

To effectively communicate insights from London's road traffic data, selecting appropriate visualisation tools is critical. The following platforms are widely used in enterprise settings(Ajayi, Abieba and Alozie, 2025).

**Power BI (Microsoft Azure Integration)**

- Best for: Interactive dashboards, real-time monitoring, and seamless integration with Azure services.
- Advantage: Supports geospatial mapping (e.g., congestion heatmaps) when combined with Mapbox.
- Limitation: Requires licensing costs for advanced features.

**Tableau (Advanced Geospatial Analytics)**

- Best for: High-impact visual storytelling and granular spatial analysis (e.g., accident hotspots).
- Advantage: Superior custom map layers compared to Power BI, simple drag-and-drop functionality.
- Limitation: Higher cost and steeper learning curve.

**Grafana (Real-Time IoT Dashboards)**

- Best for: Streaming data from ANPR cameras and traffic sensors.
- Advantage: Optimised for low-latency monitoring.
- Limitation: Less intuitive for non-technical users.

Recommendation: Power BI combined with Mapbox is the optimal choice for the Department of Transport due to its cost-efficiency, Azure compatibility, and real-time capabilities.

# 7 Conclusion

In conclusion, effective management of London's road traffic requires a robust, scalable, and future-proof data strategy. This report has critically evaluated the road traffic statistics and outlined an enterprise data analysis approach grounded in a lakehouse architecture. By adopting a unified platform that supports governance, real-time insights, and analytical agility, this strategy aligns with both current operational needs and long-term digital transformation goals. With the appropriate tools, frameworks, and data models in place, public agencies can make better-informed decisions to reduce congestion, improve journey reliability, and enhance urban mobility across the capital.

**Word Count:** 1623

**References:**

Department for Transport (2023) *Road traffic statistics*. Available at: https://roadtraffic.dft.gov.uk/regions/6 (Accessed 5 June 2025).

Weltevreden, J. (2024) *European E-Commerce Report 2024.* Amsterdam: University of Applied Sciences & Ecommerce Europe. Available at: https://ecommerce-europe.eu/wp-content/uploads/2024/10/CMI2024_Complete_light_v1.pdf (Accessed 3 June 2025).

Saltz, J. (2024) *CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects*. Available at: https://www.datascience-pm.com/crisp-dm-still-most-popular/ (Accessed 5 June 2025).

Hotz, N. (2025) What is TDSP? Available at: https://www.datascience-pm.com/tdsp/ (Accessed 5 June 2025).

Swamynathan, M. (2019) *Mastering machine learning with Python in six steps: a practical implementation guide to predictive data analytics using Python*. 2nd edn. Bangalore: Apress. Available at: https://doi.org/10.1007/978-1-4842-4947-5

Armbrust, M., Ghodsi, A., Xin, R. and Zaharia, M. (2021) 'Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics', *Conference on Innovative Data Systems Research (CIDR)*. Virtual Event, 11-15 January. Available at: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf (Accessed: 4 June 2025).

Oreščanin, D., and Hlupić, T. (2021) 'Data Lakehouse - a Novel Step in Analytics Architecture', *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*. Opatija, Croatia, 27 September – 01 October. IEEE. 1242-1246. Available at: https://doi.org/10.23919/MIPRO52101.2021.9597091

Mazumdar, D., Hughes, J., and Onofre, J. B. (2023). 'The Data Lakehouse: Data Warehousing and More'. *arXiv:2310.08697*. Available at: https://doi.org/10.48550/arXiv.2310.08697

Martin, N. (2023) *Lakehouse architecture for simplifying data science pipelines*. Master thesis. Uppsala University. Available at: https://www.diva-portal.org/smash/get/diva2:1820925/FULLTEXT01.pdf (Accessed 4 June 2025).

Herden, O. (2020) ‘Architectural Patterns for Integrating Data Lakes into Data Warehouse Architectures’, *Big Data Analytics*, pp. 12-27. Available at: https://doi.org/10.1007/978-3-030-66665-1_2

Schneider, J., Christoph Gröger, Lutsch, A., Schwarz, H., and Bernhard Mitschang (2024). ‘The Lakehouse: State of the Art on Concepts and Technologies’, *SN Computer Science*, 5, 449. Available at: https://doi.org/10.1007/s42979-024-02737-0

Sharma, A., Singh, U. K., Upreti, K., Kumar, N. and Singh, S. K. (2021) ‘A Comparative analysis of security issues & vulnerabilities of leading Cloud Service Providers and in-house University Cloud platform for hosting E-Educational applications’, *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*. Hassan, India, 24-25 October. IEEE. 552-560. Available at: https://doi.org/10.1109/MysuruCon52639.2021.9641545

Upreti, K., Jain, R., Kumar, R., Goyal, K., Deepika, and Gupta, K. (2025) ‘The Evolution of Cloud Computing: A Study of Aspirational Technologies and Practical Achievements’, *2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT)*. Bhopal, India, 7-9 March. IEEE. 898-903. Available at: https://doi.org/10.1109/CSNT64827.2025.10968661

Sirbu, D. -I., Taleanu, A. -T. and Pop, F. (2024) ‘Replication as Lineage Mechanism for Materialized Views in Lakehouse Architectures’, *2024 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. Craiova, Romania, 4-6 September. IEEE. 1-7. Available at https://doi.org/10.1109/INISTA62901.2024.10683854

Dadashova, B., Li, X., Turner, S., and Koeneman, P. (2020) ‘Multivariate time series analysis of traffic congestion measures in urban areas as they relate to socioeconomic indicators’, *Socio-Economic Planning’ Sciences*, 75, 100877. Available at: https://doi.org/10.1016/j.seps.2020.100877

Hazaymeh, K., Almagbile, A., and Alomari, A. H. (2022) ‘Spatiotemporal Analysis of Traffic Accidents Hotspots Based on Geospatial Techniques’, *ISPRS International Journal of Geo-Information*, 11(4), 260. Available at: https://doi.org/10.3390/ijgi11040260

Ajayi, O. O., Abieba, O. O. and Alozie, C. E. (2025) ‘The Impact of Data Visualization on Decision-Making in Software Engineering: A Review of Tools and Techniques’, *International Journal of Academic and Applied Research (IJAAR)*, 9(4), pp. 79-87. Available at: http://ijeais.org/wp-content/uploads/2025/4/IJAAR250409.pdf (Accessed 4 June 2025).