Enterprise Data Analysis Strategy for Road Traffic Management in the London Region

Student Name: David Abiodun
Student ID: 12696259

Module Title: Into Data Science
Assignment Title: Summative Assignment: Enterprise Data Report - Case Analysis
Word Count (excluding references): 1,608
Submission Date: 9 June 2025

Introduction

Traffic congestion in London affects travel time, fuel usage, air quality, and productivity. Effective management requires timely data to identify patterns and trends. The Department for Transport (DfT) provides vehicle counts and road type data. This report evaluates London's traffic statistics and proposes an Exploratory Data Analysis (EDA) strategy.

The strategy utilises data lakes, warehouses, cloud services, and analytics tools to facilitate informed decision-making. It applies an EDA model to government datasets, highlighting congestion patterns, data challenges, and how to transform raw data into actionable insights. (Nambiar and Mundra, 2022)

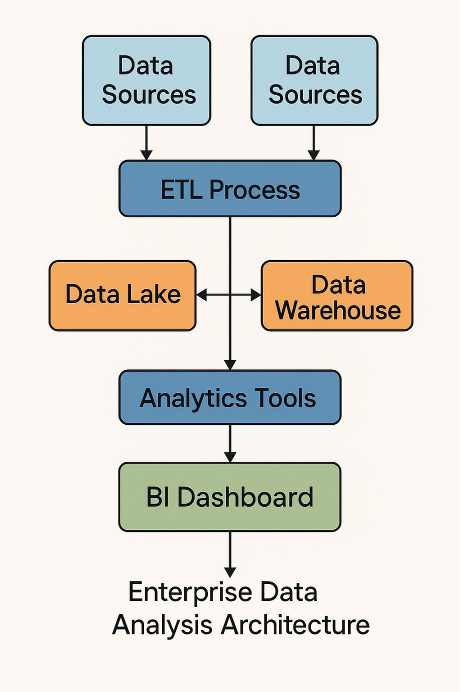


Figure 1: Enterprise Data Architecture Flowchart

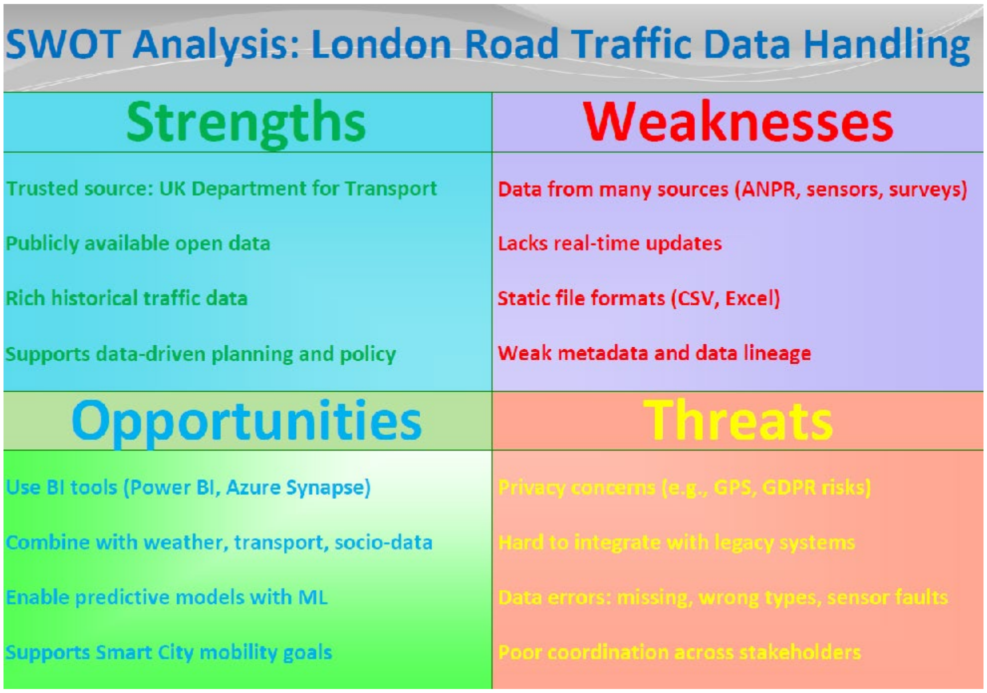
Road Traffic Statistics: Congestion and Current Issues

The London dataset features annual average daily flow (AADF) and hourly vehicle counts by road segment. It shows high traffic volumes on major A roads and motorways in Inner London. There is also an increase in Heavy Goods Vehicles (HGVs) and Light Goods Vehicles (LGVs), with LGVs reflecting the growth of e-commerce.

The directional data shows asymmetrical traffic flow, with more vehicles entering Central London in the morning and leaving in the evening. Boroughs like Westminster, Camden, and Lambeth experience regular congestion, particularly on major routes like the A406 and A13.

Urbanisation has outpaced infrastructure development, resulting in increased traffic pressure on existing roads. Several factors have contributed to increased average travel times during peak hours. Research by Wang et al. (2009) has linked congestion to a higher risk of accidents, underscoring the need for improved monitoring and control strategies.

These congestion insights expose key limitations in the current infrastructure, guiding the design of a scalable data architecture proposed in the next section.



SWOT Analysis: London Road Traffic Data Handling

Strengths	Weaknesses
Trusted source: UK Department for Transport	Data from many sources (ANPR, sensors, surveys)
Publicly available open data	Lacks real-time updates
Rich historical traffic data	Static file formats (CSV, Excel)
Supports data-driven planning and policy	Weak metadata and data lineage

Opportunities	Threats
Use BI tools (Power BI, Azure Synapse)	Privacy concerns (e.g., GPS, GDPR risks)
Combine with weather, transport, socio-data	Hard to integrate with legacy systems
Enable predictive models with ML	Data errors: missing, wrong types, sensor faults
Supports Smart City mobility goals	Poor coordination across stakeholders

Figure 2: SWOT Analysis

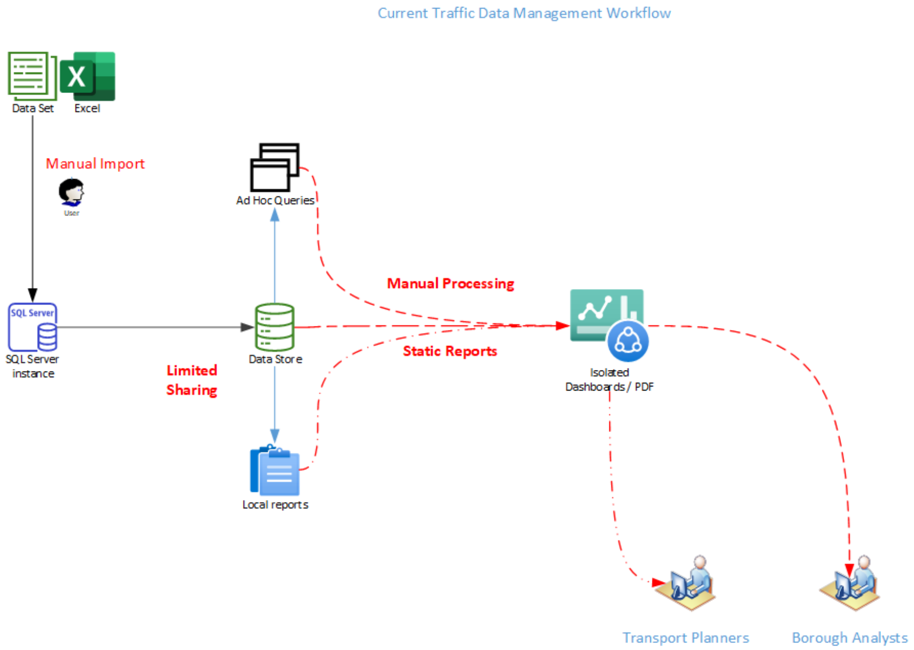


Figure 3: Current Data Management Workflow

Figure 3 shows manual data imports, siloed reports, limited integration, and a lack of automation, highlighting key bottlenecks in the existing system.

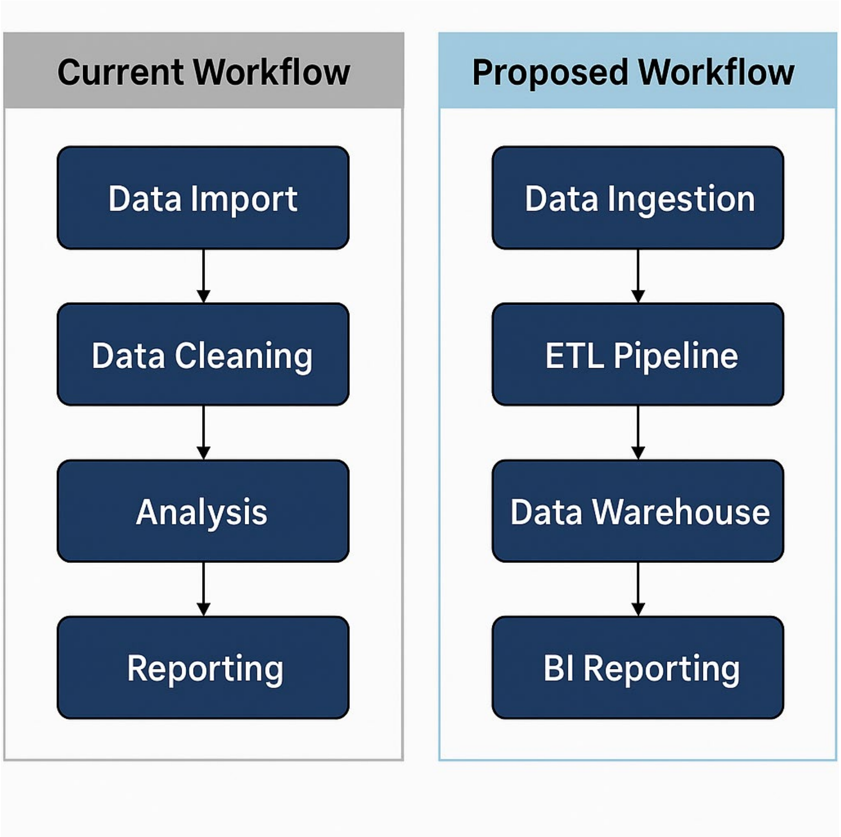
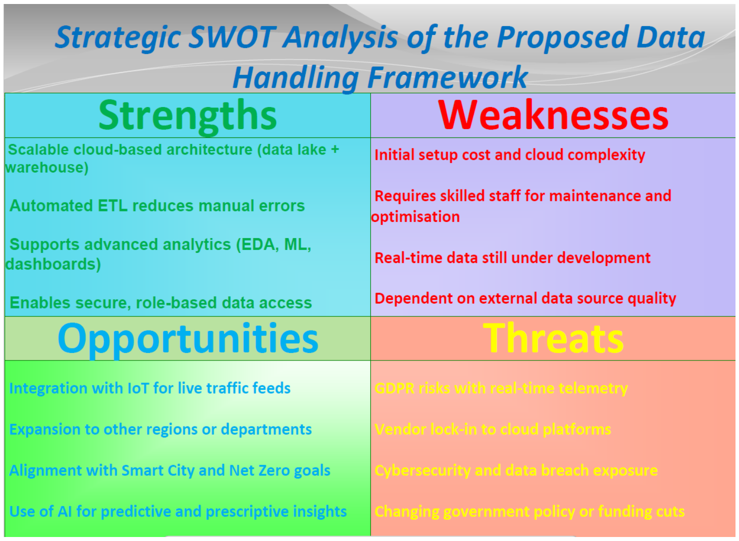


Figure 4: Current vs. Proposed Workflow for Traffic Data Management



Strategic SWOT Analysis of the Proposed Data Handling Framework

Strengths	Weaknesses
Scalable cloud-based architecture (data lake + warehouse) Automated ETL reduces manual errors Supports advanced analytics (EDA, ML, dashboards) Enables secure, role-based data access	Initial setup cost and cloud complexity Requires skilled staff for maintenance and optimisation Real-time data still under development Dependent on external data source quality
Opportunities	**Threats**
Integration with IoT for live traffic feeds Expansion to other regions or departments Alignment with Smart City and Net Zero goals Use of AI for predictive and prescriptive insights	GDPR risks with real-time telemetry Vendor lock-in to cloud platforms Cybersecurity and data breach exposure Changing government policy or funding cuts

Figure 5: Proposed SWOT Analysis

Data Architecture and Models

The strategy employs a hybrid data architecture with cloud-based data lakes and warehouses, organised for ingestion, transformation, storage, analytics, and governance policies that ensure quality and privacy.

Data Lake (Raw Storage): The first layer stores raw files (CSV, Excel) from DfT in a scalable cloud bucket (e.g., AWS S3 or Azure Data Lake Storage) for historical data preservation and audit needs.

Data Warehouse (Structured Queries): A schema-on-write warehouse, such as Amazon Redshift or Azure Synapse, supports Structured Query Language (SQL) queries. It contains:

Fact Table: Aggregated traffic volume per road segment, date, and direction.

Dimension Tables: Road type, vehicle type, time, and count location metadata from `dft_countpoints_region_id_6`.

Data from `aadfbydirection`, `traffic_road_type`, and `vehicle_type` are cleaned and joined using primary keys, such as `count_point_id` and `year`. This supports advanced queries on congestion per borough, time, and vehicle class.

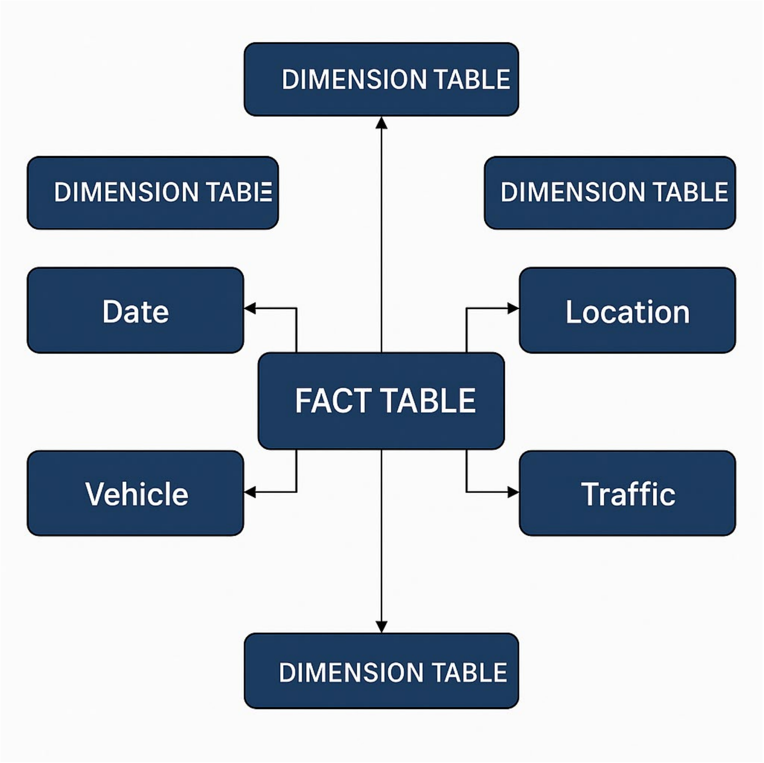


Figure 6: ER Diagram of Fact and Dimension Tables

Cloud and Analytics Integration: Google BigQuery, Azure ML, or Databricks can be integrated for elastic computing and machine learning (ML) workflows. Data ingestion is automated through Python scripts or tools like Apache NiFi, and data quality is enforced using schema validation, null handling, and automated logging within the ETL pipeline.

Tooling Comparison – Azure ML, Databricks, and Open-Source Alternatives

Selecting the right analytics platform depends on factors such as cost, scalability, and governance. Azure ML offers tight integration with Microsoft's cloud ecosystem, making it ideal for organisations already using Azure services. Databricks provides a collaborative, scalable environment with robust ML and Spark support, suitable for large-scale data workflows.

However, both options may create vendor lock-in. Open-source alternatives, such as JupyterLab with scikit-learn or Apache Superset, offer cost-effective flexibility but require more manual configuration and setup for governance. Hybrid strategies that utilise open formats (e.g., Parquet), orchestration tools (e.g., Apache Airflow), and platform-agnostic pipelines can mitigate lock-in while maintaining optimal performance. To clearly illustrate the limitations of the current approach and the advantages of the proposed enterprise architecture.

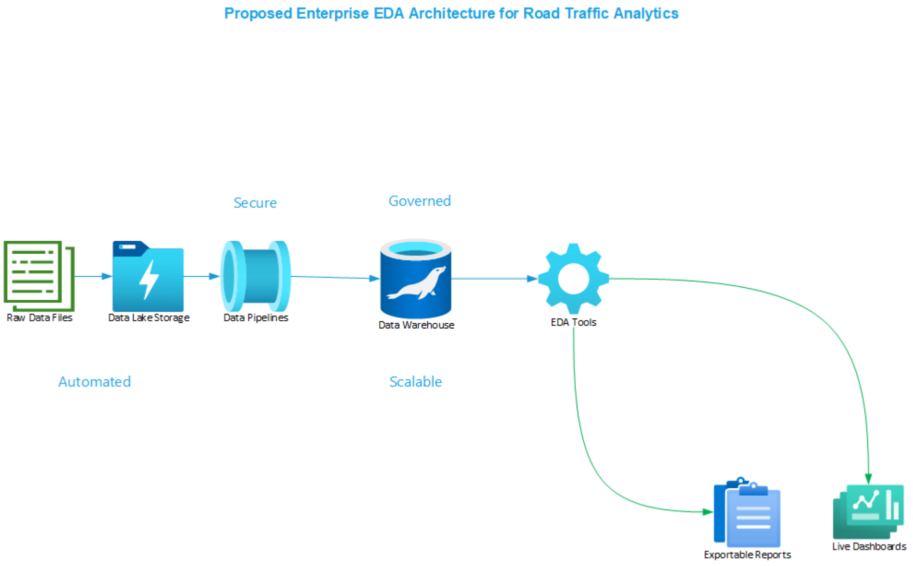


Figure 7: Proposed enterprise data architecture for London Road traffic analytics.

This visual outline is a scalable, automated architecture using cloud storage, ETL pipelines, a centralised warehouse, and EDA tooling to support real-time analytics and policymaking.

Table 1

Aspect	Current Model (As-Is)	Proposed Model (To-Be)
Data Storage	Local files (Excel/CSV on desktops)	Cloud Data Lake + Data Warehouse
Integration Approach	Manual: individual analysts merge datasets	ETL pipeline automates ingestion and joins
Data Cleaning	Manual in Excel or Access	Automated via scripts and validation logic
Tooling	Excel, Access, and some SQL scripts	Python, SQL, Power BI, Tableau
Real-time Capability	None; datasets are historical	Future integration with IoT and API's is possible
Scalability	Limited to team-level reports	Elastic cloud infrastructure
User Access	Siloed; not easily shareable	Controlled, role-based access to dashboards
Governance	None or informal	Data catalogue and versioning in place
Analytical Insight	Basic: mainly summary stats	Advanced: EDA, time-series, geospatial
Support for Forecasting	Not supported	Supports ML models and trend analysis

Table 1 above presents a side-by-side comparison, highlighting how the transformation supports scalability, real-time capabilities, and advanced analytics.

Data Analysis, Design and Methodology

Exploratory Data Analysis (EDA) extracts insights before predictive modelling by summarising metrics, detecting anomalies, and establishing correlations.

Framework Applied: A modified CRISP-DM (Cross-Industry Standard Process for Data Mining) guides the methodology.

Business Understanding: Identify congestion hotspots.

While this strategy adopts the CRISP-DM framework due to its structured and iterative nature, an alternative such as the OSEMN model (Obtain, Scrub, Explore, Model, and

Interpret) could also have been considered. However, CRISP-DM was chosen as it is better aligned with enterprise-scale data mining and offers a more precise separation of business understanding and data preparation phases, which are crucial for policymaking and infrastructure design in government contexts.

Data Understanding: Explore volume by time, direction, and vehicle type.

Preparation: Clean and structure the datasets using Python (Pandas).

Analysis: Use visualisations (Matplotlib, Seaborn) and SQL queries.

Evaluation: Assess the usability of insights for planning and policy.

Example Analysis: From the `dft_rawcount` dataset, we aggregate hourly volume data across a weekday to identify peak traffic hours. The directional data from `aadfbydirection` allows for splitting the volume into inflow/outflow patterns, supporting congestion charging or one-way system planning.

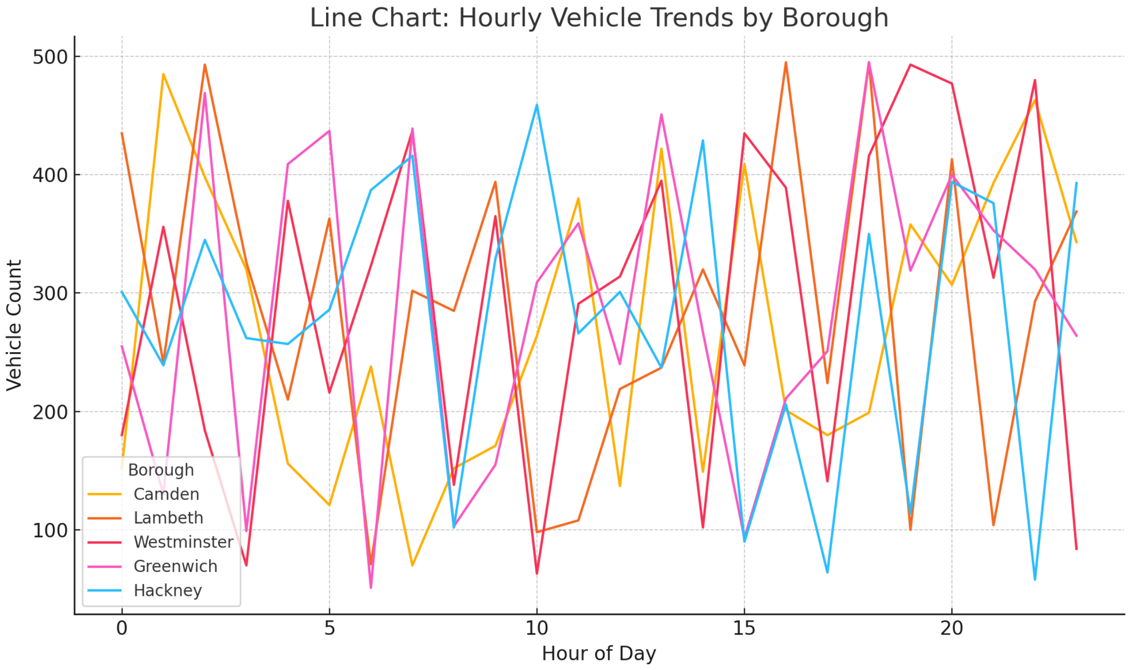


Figure 8: Line Chart of Vehicle Count Trends

We use correlation matrices to examine the relationship between LGV volume and congestion over time and space, while seasonal decomposition helps identify trends and patterns within these relationships. (Gogtay and Thatte, 2017)

Data Representation Choices

Data is presented using interactive visualisations:

- **Heatmaps**: Show hourly congestion by borough.
- **Choropleth Maps**: Show spatial vehicle count patterns.
- **Line Graphs**: Show traffic trends over time.

These representations, deployed on tools like Tableau or Power BI and connected to cloud warehouses, enable both technical and non-technical users to explore traffic anomalies quickly. (Hayes and Capretz, 2015)

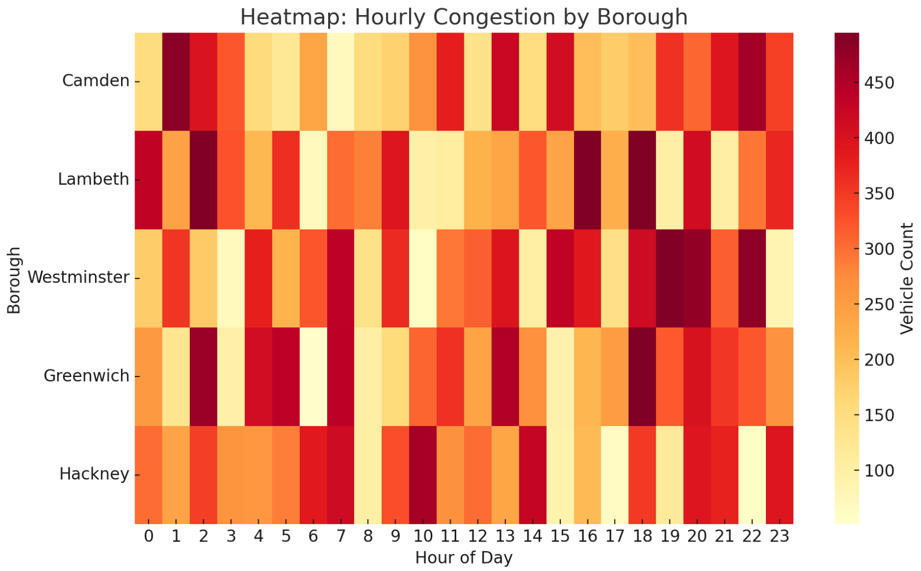


Figure 9: Heatmap of Hourly Congestion by Borough

The choice of time-series plots and spatial maps is due to their effectiveness in displaying multi-dimensional data, particularly for directional patterns and road classifications.

Ethical and Socioeconomic Impacts of Traffic Data Use

Traffic data analytics can enhance policies but also raises ethical and socioeconomic concerns. Real-time data integration (e.g., ANPR, GPS) poses privacy risks if not anonymised or encrypted. The GDPR mandates obtaining consent and minimising data, especially when inferring personal movement patterns.

Socioeconomic disparity can arise if data-driven decisions favour affluent boroughs, leaving underserved areas with fewer traffic interventions. Additionally, there's a risk of algorithmic bias in ML models trained on incomplete or skewed datasets.

Ethical analytics should prioritise transparency, fairness, and public trust. Data governance must encompass consent mechanisms, community consultation, and bias auditing to ensure inclusive urban mobility planning.

Strengths and Limitations of the Strategy

Strengths:

Scalability: Cloud infrastructure enables horizontal scaling to manage large historical datasets.

Reusability: EDA scripts and queries can be reused and version-controlled, allowing for efficient and consistent development.

Flexibility: Data lakes store data in all formats, supporting both structured and unstructured data (e.g., future sensor data).

EDA reveals hidden relationships, such as asymmetries in traffic flow, and supports proactive planning and decision-making.

Limitations:

Granularity: Annual AADF values mask hourly or incident-related spikes.

Data Quality: Some records contain null values or misalignments in `road_category_id`.

Lack of Real-Time Integration: Static datasets do not capture live traffic events or incidents.

Operational Complexity: The proposed cloud system requires initial setup effort and funding.

Skill Requirements: Ongoing maintenance demands experienced staff to manage ETL and cloud workflows.

Sliwa et al. (2019) argue that system-of-systems modelling is crucial for managing hybrid vehicular networks, suggesting that IoT sensors and simulation models can improve traffic operations in London. Future improvements may include real-time feeds (e.g., GPS, ANPR cameras) and predictive modelling using machine learning.

Conclusion

The analytics strategy enhances traffic management and policy-making by revealing congestion zones and patterns, helping allocate budgets for upgrades and enforcement. DfT traffic data shows congestion in Central London and major A roads. A modern enterprise data strategy enables scalable analysis through the use of data lakes and cloud services.

EDA extracts insights for congestion control and planning. Despite limitations such as a lack of real-time data, future machine learning (ML) integration is expected to address these issues. Authorities can use insights to adjust signals, fund upgrades, and expand ULEZ.

Dynamic weekend and freight patterns highlight the need for flexible modelling. Using tools like Azure Purview enhances data governance. (Li et al., 2015).

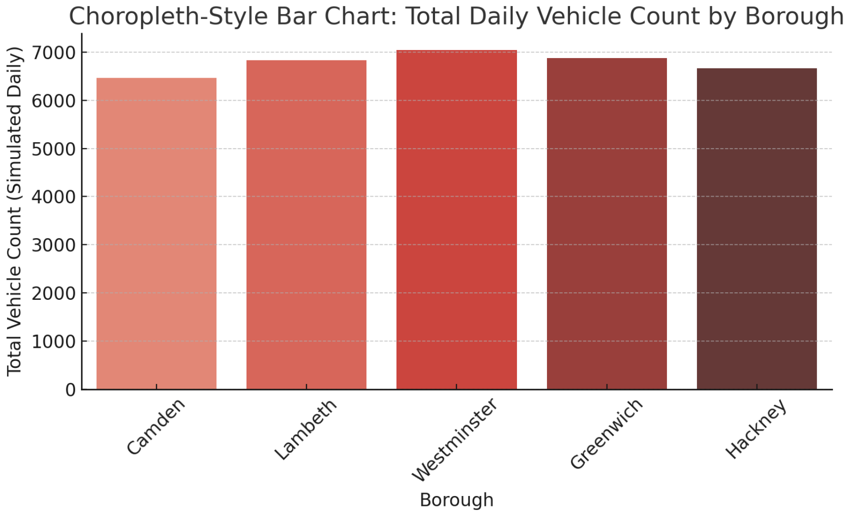


Figure 10: Bar Chart of Total Daily Vehicle Counts by Borough

Lifecycle management involves automated archival policies to move cold data to cost-effective storage and alerts for data quality issues. Using a data catalogue like AWS

Glue or Azure Purview enhances visibility and allows for efficient, secure data queries. (Li et al., 2015).

Looking ahead, the integration of IoT, ML, and cross-modal data will be pivotal in developing adaptive, citizen-centred traffic systems in London.

Word Count 1608 (excluding references)

References

Nambiar, A. and Mundra, D. (2022) ‘An overview of data warehouse and data lake in modern enterprise data management’, Big Data and Cognitive Computing, 6(4), 132. Available at: https://doi.org/10.3390/bdcc6040132 (Accessed: 8 May 2025).

Bhuyan, P., McCoy, E.J., Li, H. and Graham, D.J., 2021. Analysing the causal effect of London cycle superhighways on traffic congestion. *The Annals of Applied Statistics*, 15(4), pp.1999–2022. https://doi.org/10.1214/21-AOAS1450

Wang, C., Quddus, M.A. and Ison, S.G., 2009. Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention*, 41(4), pp.798–808.

Gogtay, N.J. and Thatte, U.M. (2017) ‘Principles of correlation analysis’, *Journal of the Association of Physicians of India*, 65, pp. 78–81. Available at: https://www.kem.edu/wp-content/uploads/2012/06/9-Principles_of_correlation-1.pdf. (Accessed: 8 May 2025).

Hayes, M.A. and Capretz, M.A.M. (2015) ‘Contextual anomaly detection framework for big sensor data’, *Journal of Big Data*, 2(1), 2. Available at: https://doi.org/10.1186/s40537-014-0011-y (Accessed: 8 May 2025).

B. Sliwa, T. Liebig, T. Vranken, M. Schreckenberg and C. Wietfeld, "System-of-Systems Modelling, Analysis and Optimisation of Hybrid Vehicular Traffic," *2019 IEEE International Systems Conference (SysCon)*, Orlando, FL, USA, 2019, pp. 1-8, doi: 10.1109/SYSCON.2019.8836786.

Li, H., Graham, D.J. and Majumdar, A. (2012) ‘The effects of congestion charging on road traffic casualties: a causal analysis using difference-in-difference estimation’, *Accident Analysis & Prevention*, 49, pp. 366–377. Available at: https://doi.org/10.1016/j.aap.2012.02.013.

Li, J., Tao, F., Cheng, Y. and Zhao, L. (2015) ‘Big data in product lifecycle management’, *International Journal of Advanced Manufacturing Technology*, 81(1), pp. 667–684. Available at: https://doi.org/10.1007/s00170-015-7151-x (Accessed: 8 May 2025).

Sawadogo, P. and Darmont, J. (2021) ‘On data lake architectures and metadata management’, *Journal of Intelligent Information Systems*, 56, pp. 1207–1241. Available at: https://doi.org/10.1007/s10844-020-00608-7 (Accessed: 8 May 2025).

Sivanathan, A., Sherratt, D., Gharakheili, H.H., Radford, A., Wijenayake, C., Vishwanath, A. and Sivaraman, V. (2017) ‘Characterising and classifying IoT traffic in smart cities and campuses’, *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 559–564. Available at: https://doi.org/10.1109/INFCOMW.2017.8116438 (Accessed: 8 May 2025).

Dutta, L. and Bharali, S. (2021) ‘TinyML meets IoT: A comprehensive survey’, Internet of Things, 16, 100461. Available at: https://doi.org/10.1016/j.iot.2021.100461 (Accessed: 8 May 2025).