Assignment 2: Enterprise Data Report

Introduction

Managing urban traffic systems in cities like London demands scalable and intelligent data analysis strategies. The *Department for Transport's (2023) Road Traffic Statistics: London* report reveals complex trends related to road types, vehicle volumes, and congestion. To turn this large-scale data into useful insights, enterprise-level frameworks such as exploratory data analysis (EDA), cloud infrastructure, and business intelligence tools must be employed.

This report critically evaluates the London traffic dataset and proposes a data architecture based on EDA principles. It examines how cloud platforms, data lakes, data warehouses, and BI tools can be integrated to support traffic monitoring and policy development. Each section addresses architecture, methodology, representation, and system limitations in relation to transport data management.

Description and Analysis of Road Traffic and Congestion Issues

The London region continues to face significant traffic congestion, with 2022 seeing 31.9 billion vehicle miles, still below the 2019 pre-pandemic level of 34.4 billion but showing slow recovery. Despite lower overall traffic, congestion has worsened during peak times in areas like Westminster, Camden, and Southwark.

Data from the Department for Transport reveals that A roads carry about 90% of London's traffic, dominating despite being a small part of the network. Van (LGV) traffic has nearly doubled since the early 2000s, driven by urban deliveries, while HGV traffic remains stable but still contributes notably to congestion and emissions.

Weekly and seasonal patterns show heavier traffic on weekdays and during school terms, emphasizing the need for better traffic management and data-driven infrastructure planning to address congestion hotspots.

These patterns highlight the need for improved traffic management systems and integrated, data-driven decision-making.

EDA Framework and Tooling

This paper applies a structured EDA framework based on Tukey's (1977) model, progressing through acquisition, cleaning, univariate/bivariate analysis, visualisation, and pattern detection. Tukey's approach was selected for its emphasis on early-stage data understanding through visual and statistical exploration, which is particularly suited to the observational and unstructured nature of traffic datasets. While this structure ensures systematic handling of the London road traffic dataset, it risks becoming overly linear, failing to account for iterative refinement cycles common in modern data science practice (Wickham, 2014).

The primary analysis environment selected is Python, using libraries such as Pandas for data manipulation. Python is widely used for its open-source flexibility, scalability, and integration with machine learning libraries (Waskom, 2021). Its ability to automate repetitive tasks makes it particularly well-suited to handling the high volume of traffic data recorded across multiple dimensions (e.g., date, road type, vehicle category).However, this flexibility comes with increased complexity.

Additionally, Power BI will be used for visual analytics. While Python is effective for in-depth statistical work, Power BI offers a business-friendly interface for dashboard creation (Microsoft, 2023). However, one limitation is that it requires pre-aggregated datasets to function effectively, which adds an ETL (extract, transform, load) burden.

In sum, the dual-tooling approach enables both technical depth and business accessibility, but its success hinges on strong documentation, consistent metadata standards, and version control to avoid analytical divergence.

Data Architecture and Model Design

To accommodate the complexity of road traffic data, the proposed architecture integrates a cloud-based data lake for raw ingestion with a dimensional data warehouse for structured analytics, following modern enterprise data strategy principles (Inmon, 2005).

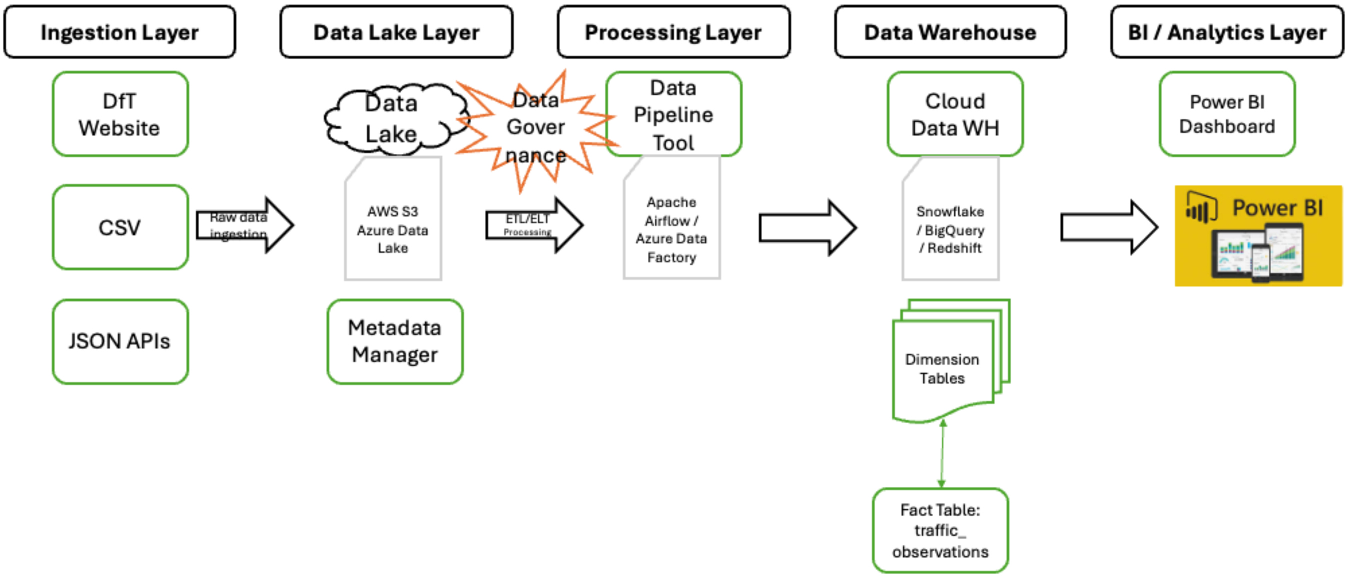

Figure 1: Proposed hybrid data architecture for London road traffic analysis (Hamberger, 2025)

The first layer in this architecture is the data ingestion stage, where traffic data is sourced directly from the Department for Transport (DfT) via downloadable CSV files and JSON APIs. These raw datasets include records such as traffic_volume_by_local_authority.csv, road_type_statistics.json, and detailed temporal vehicle flow data. This data is ingested into a cloud-based data lake (e.g., AWS S3 or Azure Data Lake) using schema-on-read logic, allowing for maximum flexibility in exploratory data analysis, machine learning, and long-term archival (Miloslavskaya and Tolstoy, 2016).

To prevent the data lake from becoming a disorganised “data swamp,” the architecture includes a metadata manager. Data governance mechanisms are also embedded at this stage (Nargesian et al., 2019).

Once ingested and catalogued, the data flows into a cloud-based data warehouse using ETL or ELT pipelines. Tools like Apache Airflow or Azure Data Factory manage transformations and ensure referential integrity. Warehouses such as Snowflake, BigQuery, or Redshift are ideal for high-performance querying, large-scale joins, and BI tool integration (Kimball and Ross, 2013).

The structured data is modelled using a star schema, optimised for reporting and analysis:

Fact Table: *fact_traffic_observations*
Contains measurable data such as traffic counts, average speeds, and congestion levels, linked via foreign keys.

Dimension Tables:

dim_date: calendar data, weekdays, term-time indicators
dim_vehicle: vehicle class and emissions category
dim_road: road type, classification, and route hierarchy
dim_edh: administrative zone, congestion charge status

This model supports scalable and performant querying, particularly for time-series and geographical analysis using Power BI, which enhances the interpretability of trends such as peak-hour congestion (Zeng et al., 2022).

Critical Analysis of Methodology and Design Choices

The proposed architecture adopts a hybrid model that separates raw data storage (via a cloud-based data lake) from structured analytics (through a dimensional data warehouse). This layered approach is grounded in industry best practices (Inmon, 2005; Kimball & Ross, 2013). However, while theoretically robust, its practical deployment in public-sector contexts presents several methodological challenges.

One of the foundational assumptions is that data lakes provide the agility required for exploratory analysis and machine learning applications. Miloslavskaya and Tolstoy (2016) support this view, noting that schema-on-read allows analysts to experiment without rigid schemas. Yet, as Giebler et al. (2021) caution, such flexibility is a double-edged sword. Without rigorous metadata standards and governance, data lakes risk becoming “data swamps,” especially when ingesting heterogeneous sources like CSVs and JSON APIs with inconsistent schemas.

Moreover, the reliance on ETL/ELT pipelines assumes the availability of highly skilled data engineers to build and maintain these processes. While tools like Apache Airflow or Azure Data Factory can automate workflows, they introduce system complexity that may not be sustainable without ongoing investment (Dennyson, 2024). This challenge is echoed by Batty (2018), who warns that public-sector organisations often underestimate the long-term cost of managing data infrastructure.

The decision to use a star schema for the data warehouse is consistent with dimensional modelling best practices (Kimball and Ross, 2013), offering simplicity, query performance, and direct integration with Power BI. However, some scholars argue this approach can limit analytical depth. For instance, Nargesian et al. (2019) critiques traditional warehousing methods as being too rigid for modern, iterative analytics, favouring more flexible NoSQL or graph-based models for representing multidimensional relationships, such as those between road type, weather, and time-dependent congestion levels.

Furthermore, the selected cloud platforms (e.g., Snowflake, BigQuery) bring benefits such as scalability and elasticity, but they also raise concerns over vendor lock-in, cost predictability, and data sovereignty. A more open-source or hybrid on-prem/cloud deployment could mitigate these risks. (Gillwald and Moyo, 2016)

In comparison, the emerging lakehouse paradigm (Armbrust et al., 2021) presents an integrated alternative by unifying storage and analytics in one layer. This simplifies architecture and reduces duplication. However, as noted by Seddon and Currie (2016), lakehouses remain relatively immature, especially in terms of supporting BI tools and structured reporting, making them less for transport policy environments.

Data Representation Strategy and Visualisation Choices

Effective data representation is essential for transforming complex road traffic data into actionable insights. In this report Power BI was selected for its strong integration with dimensional data models and its ability to produce accessible, interactive dashboards. Visual elements such as time-series plots, pythomaps, and bar charts support analysis of temporal and geospatial traffic trends, which are central to transport policy evaluation (Batty, 2018). These representations follow visualisation best practices and offer functionality such as drill-through and slicers, enabling users to interrogate data across boroughs, time periods, and vehicle types.

However, Power BI is poorly suited in handling large, dynamic datasets or real-time streaming data, which restricts its utility for operational monitoring. While Few (2009) advocates for more flexible platforms such as Tableau for deeper analytical insights, these tools carry a steeper learning curve and may not be suitable for non-technical users. The selection of Power BI therefore represents a trade-off, privileging usability and integration over extensibility and real-time analytical depth.

Strengths and Limitations of the Data Analysis Strategy

The data analysis strategy demonstrates several strengths, particularly in its modularity and stakeholder alignment. The hybrid architecture supports both exploratory workflows (via the schema-on-read logic in the data lake) and high-performance querying (through the warehouse's star schema), striking a balance between flexibility and control (Kimball and Ross, 2013). The dual-tooling approach, Python for in-depth EDA and Power BI for communication, further supports both technical analysis and decision-making, aligning with the principle of data democratisation (Waskom, 2021; Few, 2009).

Nonetheless, this separation introduces a governance burden. Metadata consistency, data lineage, and transformation traceability must be managed to avoid inconsistencies across environments. Additionally, the architecture is also not optimised for real-time traffic management; future iterations should consider integrating streaming tools such as Apache Kafka or Azure Stream Analytics to support live operational use cases (Dennyson, 2024).

Conclusion

This report explored challenges and opportunities in managing London's road traffic data through a structured enterprise data analysis strategy. A hybrid data architecture was proposed, combining a cloud-based data lake for raw data and a dimensional data warehouse for analytics and reporting. Using an EDA framework with Python and Power BI enabled thorough exploratory analysis and clear visualisation. The design supports modularity, performance, and stakeholder engagement but has drawbacks such as architectural complexity, cloud reliance, and limited real-time analytics. Despite these, the solution provides a solid foundation for data-driven transport planning. Future work could enhance the strategy with predictive analytics, real-time monitoring, and integration of sensor and GPS data to improve adaptive urban traffic management.

Wordcount: 1612

References:

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

Armbrust, M. *et al.* (2021) *Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics*, *11th Annual Conference on Innovative Data Systems Research (CIDR '21)*. Available at:
http://cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf (Accessed: June 4, 2025).

Batty, M. (2016) 'Big data and the city,' *Built Environment*, 42(3), pp. 263-337. Available at: https://doi.org/10.2148/benv.42.3.321.

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

Dennyson, R. (2024) 'Real-Time Transaction Visualization: Kafka, Azure Event Hub, and Power BI Integration,' Medium, 16 November. Available at: https://medium.com/@robertdennyson/real-time-transactions-visualization-kafka-azure-event-hub-and-power-bi-integration-49d7f8623529 (Accessed: May 28, 2025).

Few, S. (2009) Now You See it: Simple Visualization Techniques for Quantitative Analysis.

Giebler, C. *et al.* (2021) 'The Data Lake Architecture Framework.,' *BTW*, pp. 351-370. Available at: https://doi.org/10.18420/btw2021-19.

Gillwald, A. and Moyo, M. (2016) Modernising the Public Sector through the Cloud, Cloud Policy for the Public Sector. report. Available at:https://researchictafrica.net/wp-content/uploads/2017/10/Cloud-Computing-in-the-public-sector-final-25052017_V03.pdf (Accessed: May 28, 2025).

Inmon, W.H. (2005) *Building the data warehouse*. 4th edn. New York: Wiley.

Kimball, R. and Ross, M. (2013) *The Data Warehouse Toolkit: the definitive guide to dimensional modeling*. Available at: http://cds.cern.ch/record/1623624. (Accessed: May 29, 2025).

Miloslavskaya, N. and Tolstoy, A. (2016) 'Big data, fast data and data lake concepts, *Procedia Computer Science*, 88, pp. 300–305. Available at: https://doi.org/10.1016/j.procs.2016.07.439.

Nargesian, F. *et al.* (2019) 'Data lake management,' *Proceedings of the VLDB Endowment*, 12(12), pp. 1986–1989. Available at:https://doi.org/10.14778/3352063.3352116.

Power BI - Data Visualization | Microsoft Power Platform (2023). Available at: https://powerbi.microsoft.com/en-us/what-is-power-bi/(Accessed: May 26, 2025).

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

Seddon, J.J.J.M. and Currie, W.L. (2016) 'A model for unpacking big data analytics in high-frequency trading,' *Journal of Business Research*, 70, pp. 300–307. Available at: https://doi.org/10.1016/j.jbusres.2016.08.003.

Stonebraker, M. and Çetintemel, U. (2018) ''One size fits all': an idea whose time has come and gone,' in *Association for Computing Machinery eBooks*, pp. 441–462. Available at: https://doi.org/10.1145/3226595.3226636.

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

Waskom, M. (2021) 'seaborn: statistical data visualization,' *The Journal of Open Source Software*, 6(60), p. 3021. Available at:https://doi.org/10.21105/joss.03021.

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 June 2025).

Zeng, R. *et al.* (2022) 'Performance optimization for cloud computing systems in the microservice era: state-of-the-art and research opportunities,' *Frontiers of Computer Science*, 16(6). Available at: https://doi.org/10.1007/s11704-020-0072-3.