Implementation and Rationale Document for London Traffic EDA Model

Introduction

This document provides the rationale behind the implementation of the Enterprise Data Analysis (EDA) model for London traffic data. It justifies the strategic choices regarding storage, transformation, analytics, governance, and deployment, incorporating cloud computing, data lakes, and warehouses. The decisions are aligned with the aims of enhancing real-time decision-making, improving congestion management, and supporting data-driven public policy.

Data Sources and Value Consideration

The source data originates from the UK Department for Transport (DfT) and includes datasets such as `dft_aadf`, `dft_rawcount`, `dft_countpoints`, and `dft_vehicle_type`. These datasets offer granularity across traffic volume, location, direction, and vehicle types. According to Fleckenstein, Obaidi & Tryfona (2023), valuing such datasets involves assessing completeness, usage frequency, and business impact. In this context, London traffic data is considered high-value due to its recurring application in transport planning and urban resilience.

Architecture Overview

1. Data Lake for Raw Storage

Cloud-native platforms, such as Azure Data Lake and AWS S3, offer cost-efficient and schema-on-read flexibility. This approach aligns with the need for

storing unstructured and semi-structured CSV/Excel files without rigid schemas (Sawadogo & Darmont, 2021).

2. ETL and Transformation Layer

The transformation process is built on Python and Pandas, orchestrated via Apache Airflow. This choice offers scalability, modular design, and observability. Coleman et al. (2016) highlight that smaller entities benefit from open-source, Python-based ecosystems due to cost and skill availability.

3. Data Warehouse (Azure Synapse)

Structured data is written to a star-schema model. The fact table contains traffic volume metrics, linked with dimension tables for borough, time, road type, and vehicle type. This model supports multi-dimensional slicing and improves dashboard responsiveness. As Nambiar & Mundra (2022) argue, schema-on-write is essential for BI performance.

4. Analytics and Reporting Layer

Power BI and Tableau are used to visualise time trends, heatmaps, and choropleth maps. These tools allow integration with SQL-based warehouses and offer role-based access control.

5. Governance and Compliance

Azure Purview is adopted to manage metadata, lineage tracking, and ensure compliance with UK GDPR. Loewenstein et al. (2015) emphasise the importance of protecting location and mobility data, especially in public IoT environments.

EDA Model Alignment

The implementation adheres to the CRISP-DM framework:

- Business Understanding: London suffers from congestion, particularly in Inner boroughs.

- Data Understanding: Identified coverage and directional gaps.

- Preparation: Null value imputation, joining datasets, and filtering.

- Modeling: Created time series, heatmaps, and correlation matrices.

- Evaluation: Based on spatial fairness, predictive potential, and policy relevance.

- Deployment: Visualisation dashboards hosted on cloud environments.

This structure ensures repeatability, modularity, and scalability.

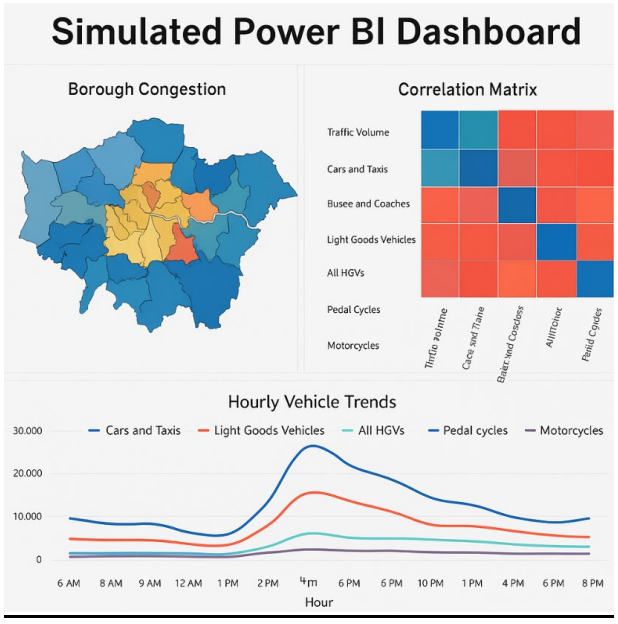

Figure 2: Simulated Power BI dashboard illustrating borough congestion, vehicle trends, and correlation matrices.

Cloud Strategy Justification

The model runs on Microsoft Azure, offering platform-native tools like Synapse, Purview, and Data Factory. Azure integrates well with Power BI and supports automated pipeline scheduling, reducing manual effort.

The IDC (2011) report emphasises that organisations with cloud-based big data capabilities are more agile, better positioned to respond to regulatory requirements, and can scale horizontally. In contrast, on-premise solutions increase latency and operational costs.

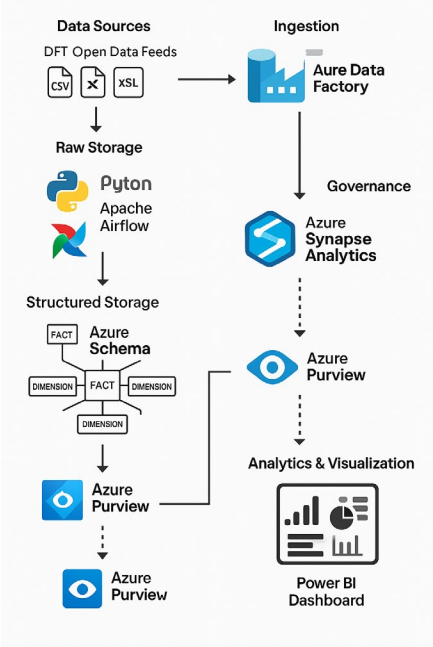

Figure 1: Technical implementation of the London EDA pipeline using Azure-native components, highlighting structured storage, orchestration, and metadata governance.

Ethical and Privacy Considerations

Real-time traffic analysis often invokes surveillance concerns. Loewenstein et al. (2015) argue that perceived value of privacy varies with context, and this makes anonymisation, aggregation, and role-based access mandatory. Using TinyML on edge devices (Dutta & Bharali, 2021) further helps by processing data locally and transmitting only summarised metadata to the cloud.

Limitations and Mitigation

- High Costs: Azure services incur cost; mitigated through tiered storage and scheduled compute.

- Skill Shortages: Addressed via modular Python pipelines and training.

- Data Incompleteness: Imputed using statistical models and cross-validated with other borough-level data.

- Privacy: Role-based access and encryption by default reduce GDPR risk.

Comparative Architecture Critique

While Azure Synapse offers robust performance and seamless integration with Power BI, its elasticity and auto-scaling under concurrent loads are often outperformed by Snowflake. Recent benchmarks (Chattopadhyay et al., 2023) reveal that Snowflake exhibits superior performance in high-concurrency environments due to its automatic compute scaling and decoupled storage architecture. However, Synapse remains preferable in this context due to its

tighter integration with Azure-native tools such as Purview and Data Factory, enabling a unified governance and compliance framework—a critical consideration for public sector deployment.

Conclusion

The EDA model implementation for London traffic integrates cloud-native architecture with a validated data science methodology. Its rationale is based on performance efficiency, ethical considerations, and practical deployment in public sector environments. By leveraging modular open-source tools and enterprise-grade cloud services, the model is designed to be sustainable, scalable, and policy-aligned.

Word Count 771

References

Coleman, S. et al. (2016). How Can SMEs Benefit from Big Data? Challenges and a Path Forward. *Quality and Reliability Engineering International*, 32(6), pp. 2151–2164. (Accessed: 07th July 2025).

Dutta, L. & Bharali, S. (2021). TinyML Meets IoT: A Comprehensive Survey. *Internet of Things*, 16, 100461. https://doi.org/10.1016/j.iot.2021.100461 (Accessed: 18th June 2025).

Fleckenstein, M., Obaidi, A. & Tryfona, N. (2023). A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model. Harvard Data Science Review, 5(1). https://doi.org/10.1162/99608f92.c18db966 (Accessed: 18th June).

IDC (2011). Big Data: What It Is and Why You Should Care. IDC White Paper. Available at: https://www.idc.com

Loewenstein, G. et al. (2015). What Is Privacy Worth? Journal of Legal Studies, 42(2), pp. 249–273. https://doi.org/10.1086/671754 (Accessed: 07th July 2025).

Nambiar, A. & Mundra, D. (2022). An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing,* 6(4), p.132. (Accessed: 4th July 2025).

Sawadogo, P. & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, 56, pp. 1207–1241. (Accessed: 5th July 2025).