Enterprise Data Executive Summary and Implementation Report

Author: *Mark Collins*
Date: *2025-07-13*
Word Count: *1806*

1.0 Executive Summary

1.1 Background and London Traffic

The data from the Department for Transport (Department for Transport, 2023) shows that there were approximately 19.4billion vehicle-miles travelled across 9,200 miles of road in the London region, collected using 3,630 count points.

This data can be useful in forming policy decisions, from economic (Jin and Rafferty, 2017), to environmental (Huang and Loo, 2023), and to safety (Talbot, Filtness and Morris, 2024). In order to understand, implement and measure any improvement, and impact (Bhuyan *et al.*, 2021), an effective Data Management Strategy and corresponding Architecture is required.

1.2 Architecture Model

The recommended architecture is a multi-layered approach consisting of:

1. Data Lake for data ingestion
2. Data Warehouse with curated queries
3. Business Intelligence / analytics for reporting

Details of each Layer

Data Lake

The Data Lake acts as a landing zone for data ingestion. Data of all formats, from all relevant sources is available, from manual submission, to IOT sensors, and API's to other data sources (e.g. accident statistics (Transport for London, no date)).

Depending on the type of data ingested, the Data Lake is to be separated into 'Zones' (Ramchand and Mahmood, 2022). The major benefit of this is that as a public body, the Department for Transport is open to public scrutiny. A zoned Data lake allows all raw data to be maintained in its original format, allowing any decisions made from the data to be traced back to its original provenance.

A Data lake can also use less expensive, expandable storage, allowing the solution to be scalable as the volume of data grows. New data sources can be added at a later date without affecting existing queries.

Data Warehouse

Following some Extract Transform and Load (ETL) functions and queries, data can be loaded into a curated data warehouse. Common queries can be curated and stored within the warehouse, reducing the computational load on the Data lake when data is required.

This layer contains the traditional data structure, defined by star schema and dimension tables, where data can be aggregated ready for consumption.

BI / Analytics

From the Data warehouse data can be loaded into an analytics platform for reporting and visualisation, allowing decisions to be made from the data. Curated in a way to allow questions to be asked and scenarios to be considered and played out, without impacting the Warehouse layer.

Additional data can be fed directly into this layer, such as Map data (Ordnance Survey Limited, 2025) for overlaying the geographic data relating to count points.

1.3 Step-by-step Data Analysis Methods

Data analysis follows the following steps:

Step	Process	Layer
1	Ingest raw CSV/JSON data via cloud E(T)L	Data Lake - Landing Zone
2	Clean and Unify. Standardise time stamps, geocode location, filter etc.	Data Lake - Staged Zone
3	Aggregate tables ready for consumption	Data Warehouse
4	Combine with open source mapping API	BI / Analytics
5	Produce and update BI Dashboard and Reports	BI / Analytics

1.4 Data Flow and Analysis Example

Producing a very simple proof of concept within the Business Intelligence platform KNIME (KNIME AG, 2024), Figure 1:1 shows an example of data flow to provide the Average Annual Daily Flow (AADF) of Vehicles based on the Count Point location.

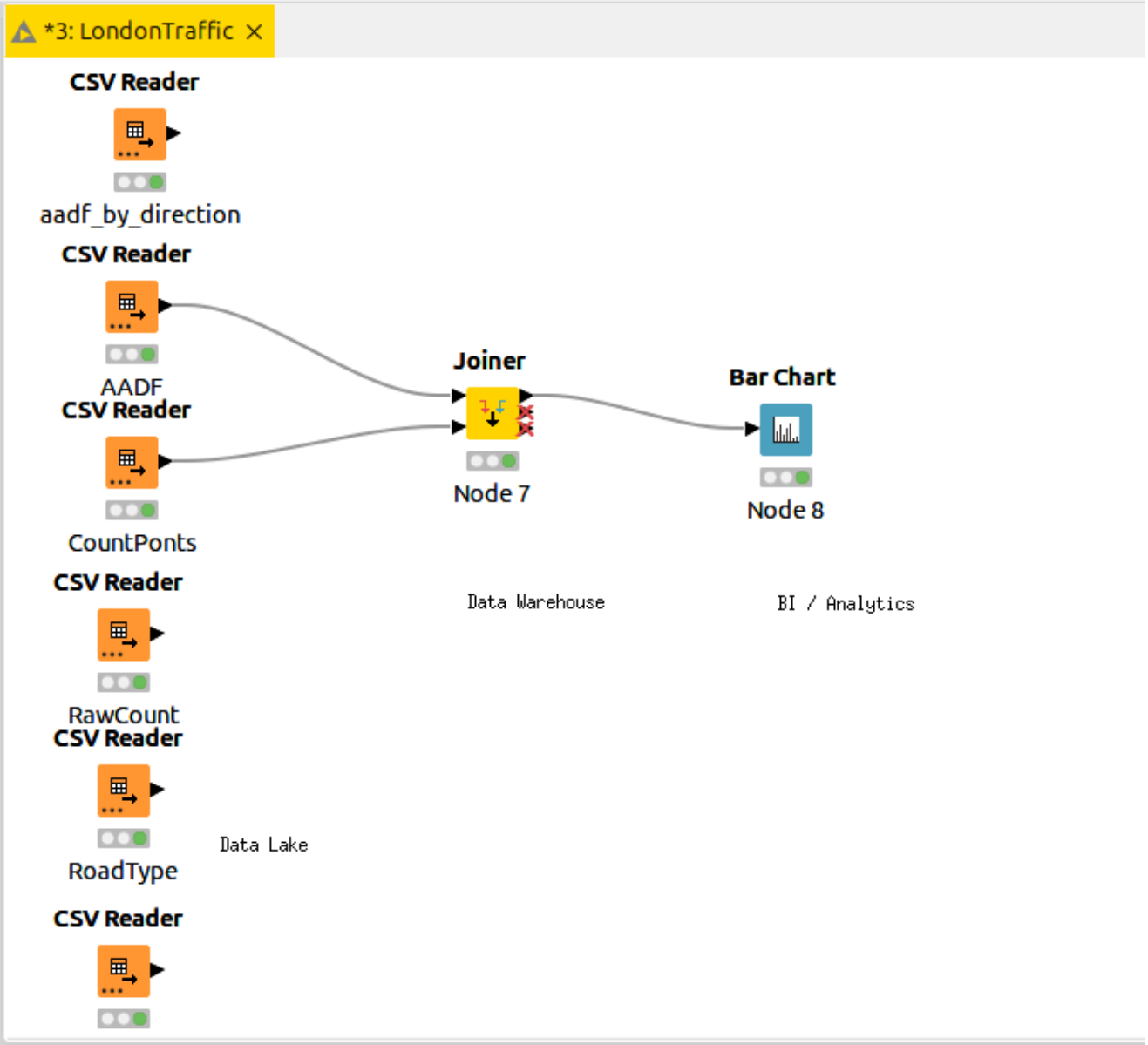

Figure 1:1: *Example Flow of data to provide Average Annual Daily Flow based on location*

Whilst this is a very simple example data flow held on the same hardware, it demonstrates how the raw data can be ingested into the Data lake (CSV tables). The Average Annual Daily Flow table is combined with the geographic data relating to each count point from the CountPoints table. This is then used to produce a visualisation of the data, shown in Figure 1:2.

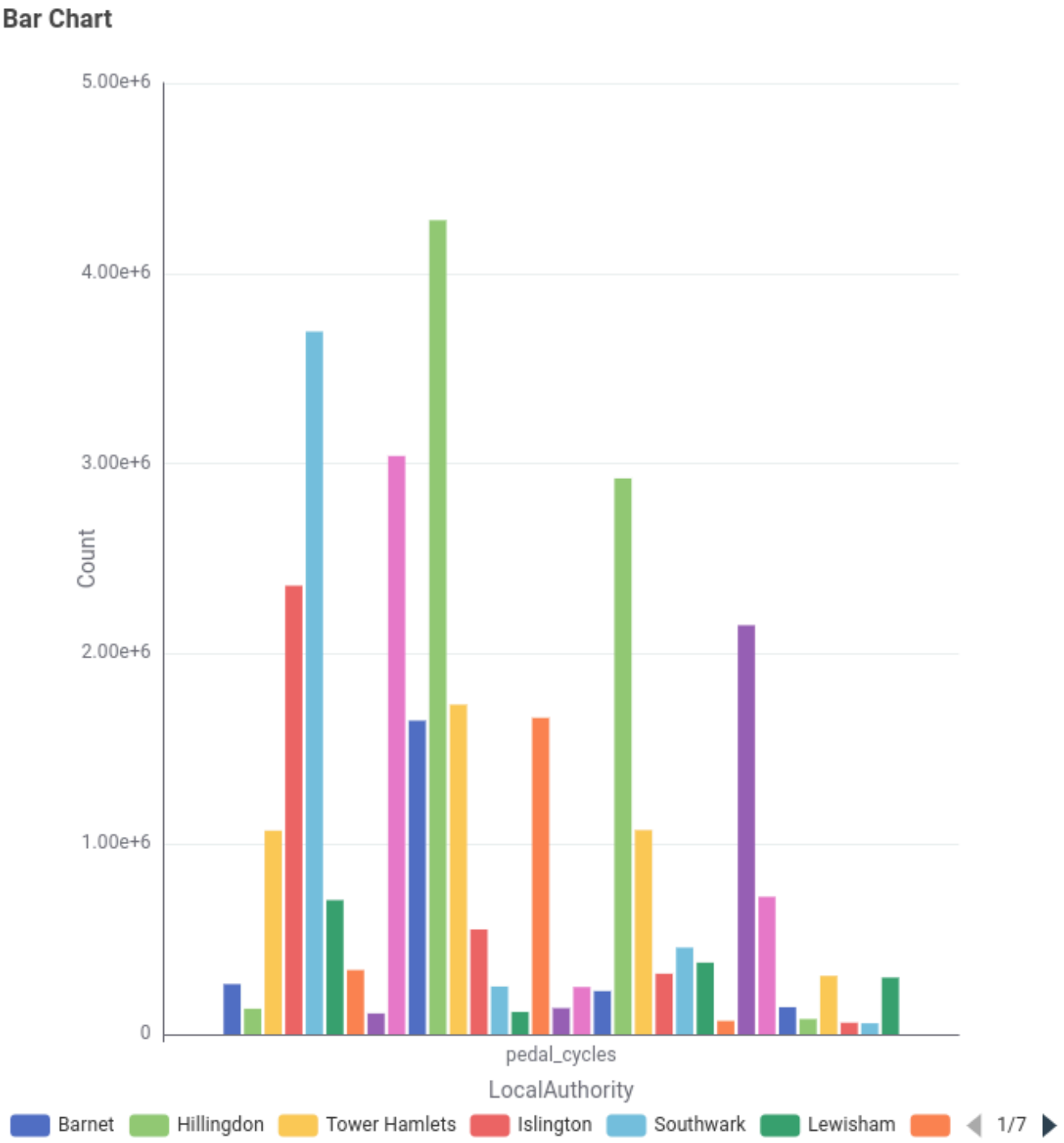

Figure 1:2: *Bar Chart showing number of cycle counts per Local Authority*

Exploring this data, Westminster and Southwark have the highest number of cycles, and Havering with the lowest. A number of questions can be asked based on this Exploratory Data Analysis (EDA), such as *Why* central London has higher cycle numbers than the suburbs, and if policy can be implemented to increase cycling in outer London.

Additionally, the data can be combined with accident data (Transport for London, no date), to explore if policy changes can reduce accidents. Munk (2024) demonstrates that Westminster has the sixth most dangerous junction for Cyclists, and Southwark has the seventh. While it is expected that a high number of cycle accidents would occur in areas where there are more cycle journeys, the Local Authority with the highest number of accidents is Wandsworth, which is close to the middle in terms of number of cycle journeys, indicating this junction merits greater investigation.

1.4 Limitations and Risks

Data Quality: Not all count points and lines of data have a verifiable count method. Some are `Estimated` which may skew results somewhat. It is therefore recommended to filter these lines out during ETL process.

Analytics performance: Having superfluous columns in Data Warehouse tables can reduce performance as greater Volatile Memory is required to load and process. Therefore an additional ETL process should involve removing non-required columns.

Solution acceptance: Many traffic calming measures will have an impact on other road users. For example, many measures to improve cycle safety will have an impact on other road users, and this cost benefit must be fully evaluated to ensure stakeholder buy-in (Elvik, 2000; von Stülpnagel and Rintelen, 2024).

1.5 Conclusion

A multi layered data architecture has been proposed and implemented, providing flexibility, scalability and traceable governance.

The Data pipeline follows a modular approach of `raw` -> `clean` -> `agregate` -> `model + serve` allowing the process to be open to public scrutiny.

Connecting to mapping API's allows for hot-spots to be easily identified and targeted interventions proposed, implemented and impact assessed.

2.0 Implementation Report

2.1 Introduction

The Data Architecture described in Collins (2025) is that of a multi-level architecture, consisting of a Data Lake for data ingestion and cleaning, a Data Warehouse hosting a number of key queries, feeding into an analytics layer for reporting and consumption of the data.

Using the KNIME platform (KNIME AG, 2024) a simplified version of this Architecture has been developed.

2.2 Architecture Overview

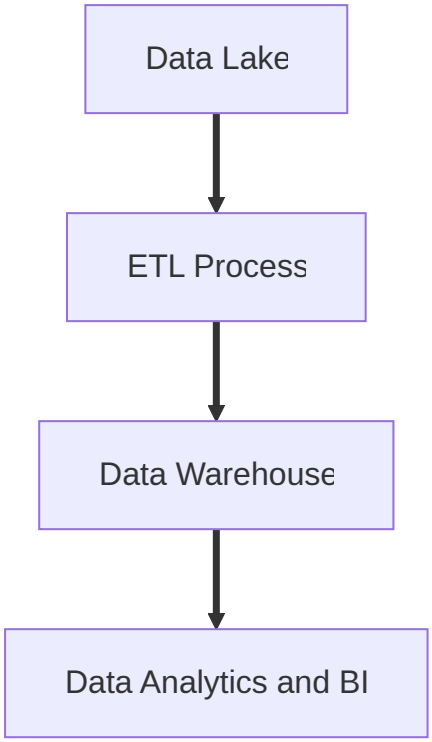

Figure 2:1: *Data Architecture overview*

The sequential flow of data through the Data Architecture involves raw data landing in the Data Lake. A number of Extract Transform Load (ETL) processes are performed landing in the Data Warehouse, ready for consumption for Data Analytics and Business Intelligence (BI) reporting.

2.3 Data Lake and Data Ingestion

The Data Lake acts as a landing zone for the data. Allowing the data to remain in its original raw format within the landing zone of the Data Lake allows for good data governance. All ETL processes can be reproduced, and the the original data is unaffected allowing any decisions made to be traced to the original raw data.

Within the Data Lake, the data can go through an initial cleansing into a staged zone (Ramchand and Mahmood, 2022). This is useful if the data being ingested is coming in multiple formats. The data can be converted into a format that will allow more efficient processing later.

2.3 ETL processes

Within the Extract Transform and Load processes, certain calculations and further filtering can be be performed. For example Volume-Capacity Ratios are used an an indicator of congestion (Noland and Quddus, 2005). This can be calculated by

1. Process the Average Annual Daily Flow (AADF) data within the Data Lake to filter out superfluous columns, leaving:

```
$count_point_id$
$road_catagory$
$road_type$
$estimation_method$ <> "Estimate"
$all_motor_vehichles$
$year$
```

2. Convert daily counts to hourly: `$all_motor_vehicles$ / 24`
3. Estimate the number of lanes for each road type:

```
$road_category$ = "MB" => 1  # Minor B Road
$road_category$ = "MCU" => 1 # Minor Urban
$road_category$ = "PA" => 2  # Primary A Road
$road_category$ = "TA" => 2  # Trunk A Road
$road_category$ = "TM" => 3  # Motorway
```

4. Estimate the Capacity per lane (Makki *et al.*, 2020):

```
$road_category$ = "MB" => 1000
$road_category$ = "MCU" => 1000
$road_category$ = "PA" => 1000
$road_category$ = "TA" => 1000
$road_category$ = "TM" => 2000
```

5. Calculate total road capacity: `$CapacityPerLane$ * $NumLanes$`
6. Calculate Volume-Capacity Ratio: `$VolumePerHour$ / $TotalCapacity$`
7. Join With Count Point data table so V/C ratio can be assigned to location information.
8. Load the data into the Data Warehouse ready for consumption.

2.4 Data Warehouse

Final queries from the ETL process land in the Data Warehouse. Here the data is stored in a cleaned and curated format. In an enterprise environment, this would be a fast access storage medium, with sufficient processing and memory requirements to access the data without delay, and without impacting the ETL process earlier in the data journey.

2.5 Analytics and Business Intelligence

The contents of the Data Warehouse is readily available for further analysis and reporting. For example the Average V/C Ratio can be plotted by Local Authority to show the Authority with the highest congestion rates.
Figure 2:2 shows that the local Authority with the highest average V/C Ratio is Hammersmith and Fulham, with a V/C of 0.842.

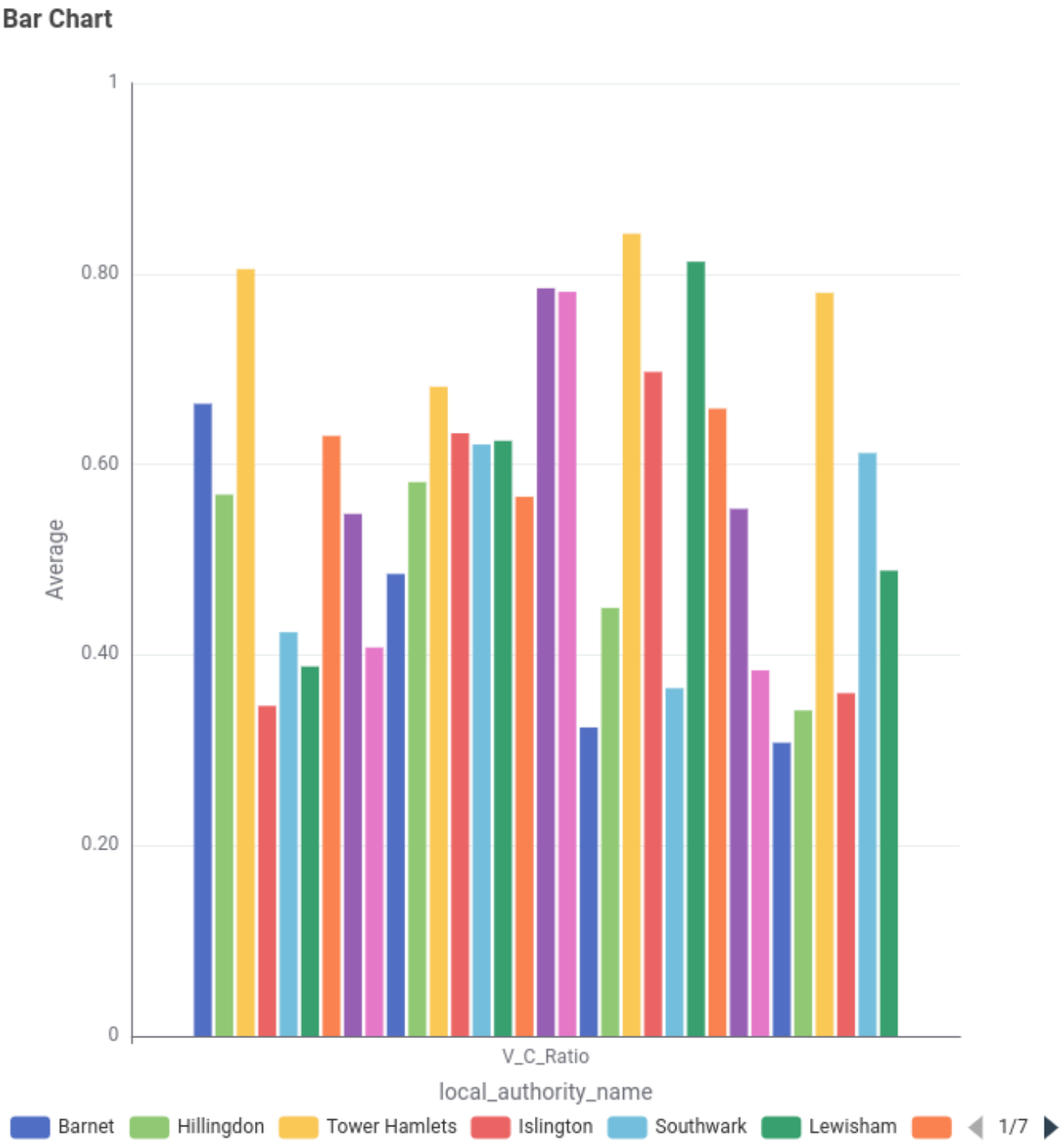

Figure 2:2: *Bar chart showing average V/C per Local Authority*

With the data held within the Data Warehouse, Local Authority can be filtered to only look at this Local Authority, and display by both an average over the Local Authority by year, to look at over-all changes, or by street to find particular roads and junctions driving up the congestion.

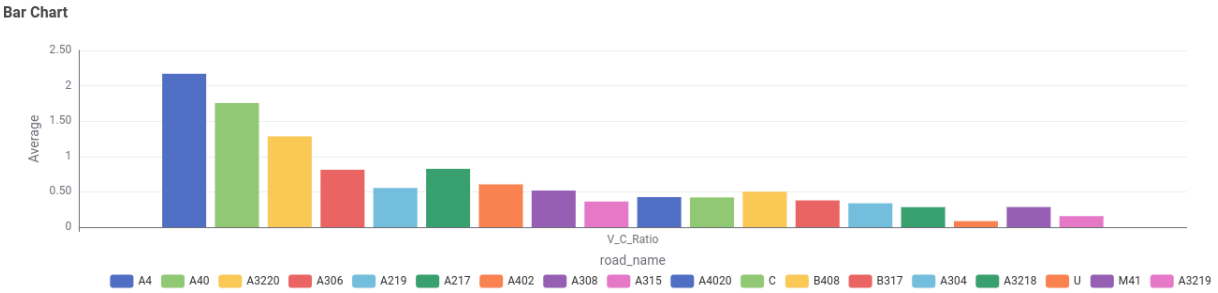

Figure 2:3: *A4 is the most congested road in the Hammersmith and Fulham Authority*

Figure 2:4: *Average V/C ratio acorss Hammersmith and Fulham by year*

2.6 Conclusion

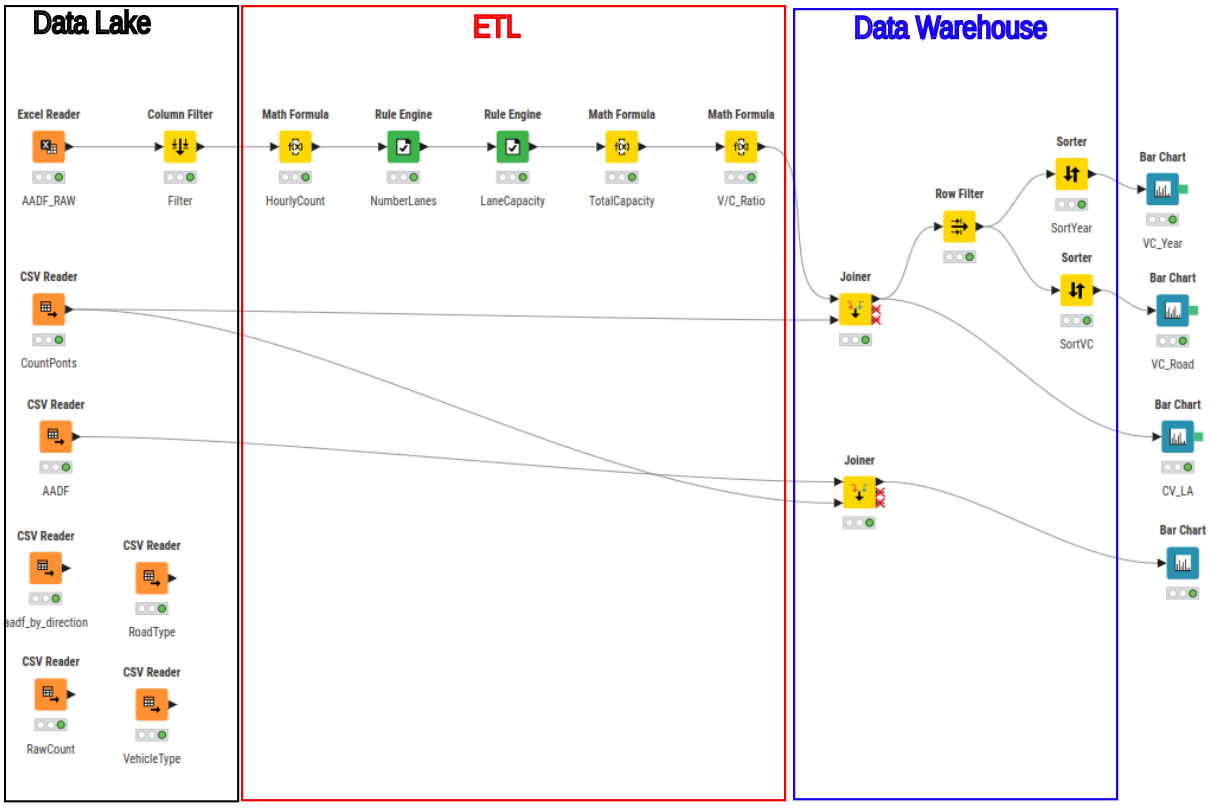

Figure 2:5: *Visual representation of Data Architecture and data flow using KNIME*

A Data Architecture has been developed and implemented using KNIME (KNIME AG, 2024) to manage and analyse the data provided by the Department for Transport covering Road Traffic Statistics in the London Region (Department for Transport, 2023).

The data lake acts as a landing zone for all data, regardless of source, whether this be manual CSV or Excel tables, JSON data, IOT sensors, or API's to other data sources.

The data is then partially cleaned within the Data Lake before undergoing Extraction, Transformation and then Loading into the Data Warehouse.

The Data Warehouse contains the curated data, ready for ingestion by the Analytics / BI layer where decision makers can easily view and interoperate the data.

2.7 Further work and Recommendations

The data Architecture has been created using only the DfT data for the London region. Given the flexibility of the Data Lake, weather or event data can be fed in to measure the impact of these on traffic flow.

Further, accident data (Transport for London, no date) can be loaded into the data lake to observe where traffic calming measures and Cycle infrastructure can be implemented.

Linking in with a mapping API would also allow visualisation of exactly where traffic hot-spots exist.

References

Bhuyan, P. *et al.* (2021) ‘Analysing the causal effect of London cycle superhighways on traffic congestion’, *The Annals of Applied Statistics*, 15(4), pp. 1999–2022. Available at: https://doi.org/10.1214/21-AOAS1450.

Collins, M. (2025) *Enterprise Data Report - Case Analysis*. University of Essex.

Department for Transport (2023) *Road traffic statistics - London region*. Available at: https://roadtraffic.dft.gov.uk/regions/6 (Accessed: 4 June 2025).

Elvik, R. (2000) ‘Which are the relevant costs and benefits of road safety measures designed for pedestrians and cyclists?’, *Accident Analysis & Prevention*, 32(1), pp. 37–45. Available at: https://doi.org/10.1016/S0001-4575(99)00046-9.

Huang, Z. and Loo, B.P.Y. (2023) ‘Urban traffic congestion in twelve large metropolitan cities: A thematic analysis of local news contents, 2009–2018’, *International Journal of Sustainable Transportation*, 17(6), pp. 592–614. Available at: https://doi.org/10.1080/15568318.2022.2076633.

Jin, J. and Rafferty, P. (2017) ‘Does congestion negatively affect income growth and employment growth? Empirical evidence from US metropolitan regions’, *Transport Policy*, 55, pp. 1–8. Available at: https://doi.org/10.1016/j.tranpol.2016.12.003.

KNIME AG (2024) *KNIME Analytics Platform*. Available at: https://www.knime.com/.

Makki, A.A. *et al.* (2020) ‘Estimating Road Traffic Capacity’, *IEEE Access*, 8, pp. 228525–228547. Available at: https://doi.org/10.1109/access.2020.3040276.

Munk, S. (2024) ‘Updated Dangerous Junctions Map’, *London Cycling Campaign*, 19 November. Available at: https://lcc.org.uk/news/new-lcc-junctions-map-shows-cost-of-year-of-inaction/ (Accessed: 10 July 2025).

Noland, R.B. and Quddus, M.A. (2005) ‘Congestion and safety: A spatial analysis of London’, *Transportation Research Part A: Policy and Practice*, 39(7), pp. 737–754. Available at: https://doi.org/10.1016/j.tra.2005.02.022.

Ordnance Survey Limited (2025) *OS Maps API | Data Products | OS*, *Ordnance Survey*. Available at: https://www.ordnancesurvey.co.uk/products/os-maps-api (Accessed: 10 July 2025).

Ramchand, S. and Mahmood, T. (2022) ‘Big data architectures for data lakes: A systematic literature review’, *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1141–1146. Available at: https://doi.org/10.1109/COMPSAC54236.2022.00179.

von Stülpnagel, R. and Rintelen, H. (2024) ‘A matter of space and perspective – Cyclists’, car drivers’, and pedestrians’ assumptions about subjective safety in shared traffic situations’, *Transportation Research Part A: Policy and Practice*, 179, p. 103941. Available at: https://doi.org/10.1016/j.tra.2023.103941.

Talbot, R., Filtness, A. and Morris, A. (2024) ‘Proposing a framework for evidence-based road safety policy-making: Connecting crash causation, countermeasures and policy’, *Accident Analysis & Prevention*, 195, p. 107409. Available at: https://doi.org/10.1016/j.aap.2023.107409.

Transport for London (no date) *Road safety data*, *Transport for London | Every Journey Matters*. Available at: https://www.tfl.gov.uk/corporate/publications-and-reports/road-safety (Accessed: 8 July 2025).