CONSTRUCTION OF AN ETL PIPELINE FOR TRAFFIC DATA ANALYSIS USING PUBLIC SOURCES

Jonas de Godoi[1], Marcos Monteiro Junior[1], Gabrielly de Queiroz Pereira[1]

[1]Departamento de Informática – Universidade Estadual de Ponta Grossa (UEPG)

godoijonas16@gmail.com, mmjunior@uepg.br, gqpereira@uepg.br

Abstract. *The growth of vehicle flow in cities has intensified challenges related to urban mobility, requiring data-driven solutions to support traffic planning and management. This work presents the development of an Extraction, Transformation, and Load (ETL) pipeline for the automated processing of public transit data from the city of Curitiba. The solution was implemented with open-source technologies—Python, Apache Airflow, and MySQL—aiming to ensure scalability, automation, and integration between different data sources. The results demonstrated the pipeline's effectiveness in collecting, transforming, and storing information, enabling the generation of analytical visualizations in Power BI that highlight spatial and temporal patterns of traffic occurrences. The study reinforces the importance of Data Engineering as a tool to support decision-making and the optimization of urban mobility.*

Keywords: Data Engineering, ETL Pipeline, Urban Mobility, Traffic Data Analysis.

1. INTRODUCTION

The Brazilian economy is predominantly urban, with cities contributing 90% of the GDP, which reflects a significant centralization of productive activities in these spaces. This urban configuration is reinforced by the demographic data that, in 2000, 81% of the Brazilian population was already urban (IBGE, 2000). In this context of densification and economic dynamism, (Meneses, 2003) highlights the importance of urban traffic management as a key element for the proper functioning of social and economic life in Brazil's main cities.

(Instituto de Pesquisa Econômica Aplicada (Ipea), 2010) points out that population mobility in Brazil has undergone major transformations since the mid20th century. These changes are seen primarily as a reflection of the rapid and intense urbanization process, the uncontrolled growth of urban centers, and the increased adoption of individual motorized transport.

According to (Júnior, 2016), the increased demand for the movement and transport of individuals and goods, carried out by motor vehicles, drives the need to build new roads and modernize existing ones, in addition to implementing progressively more advanced traffic control systems. This growing complexity in urban and transport management, combined with the digitalization of processes, has generated an unprecedented volume of data. According to Silva (2022), the growing volume of data drives the need to develop robust, high-capacity systems. These systems are essential for collecting, processing, and analyzing this vast amount of information, with the final goal of extracting relevant knowledge that can support and optimize decision-making processes. The relevance of data in contemporary society is so significant that its importance is even compared to that of valuable commodities like oil, or even capital

goods Arrieta-Ibarra et al. (2018). According to (Lee, 2020), the extensive use of data can bring considerable advantages to both the general population and the governmental sphere. These benefits extend across various sectors, such as public safety, health, and sustainable development, as the use of data supports the formulation of new policies, resource allocation, the development of smart cities, and the implementation of preventive measures. As a result, greater efficiency in public management and more transparency in resource administration are achieved. However, it is crucial to emphasize that the full realization of this potential depends on the proper treatment of the data, because, according to Bueno (2023, p. 13), "When it is in its raw form, it is not possible to harness its full potential." This need for transformation and refinement of raw data is fundamental for value generation.

As per Bie et al. (2022), one of the prominent challenges today is the lack of qualified professionals to manage large volumes of data and, fundamentally, to design processing solutions that are effective and scalable. Such solutions are crucial for converting this data into applicable knowledge for the benefit of individuals, institutions, corporations, and society as a whole. Additionally, the author highlights that a significant portion of the effort in these activities, estimated at 80. According to Azevedo (2024), the mastery of key Data Engineering methods and techniques by computer science professionals graduating from universities is an essential requirement for them to adequately meet the demands imposed by both society and the market.

Promoting clarity and objectivity in the presentation of essential information is a key benefit provided by data engineering and Business Intelligence. According to Bueno (2023), by enabling companies to collect and expose data transparently, these disciplines help reduce risks such as fraud and better anticipate unforeseen events.

The strong dependence on data-driven insights to guide decisions and achieve business success is a characteristic of organizations in all sectors in the digital age. Nylen & Holmström (2015) highlight that, within data management for modern analytics, the Extraction, Transformation, and Load (ETL) pipeline is recognized as one of its most vital components.

According to Pereira, Silva & Oliveira (2021), the usual configuration of an Extraction, Transformation, and Load (ETL) pipeline includes a source for raw data, a computational processing stage scheduled to occur at regular intervals, and a repository for storing data after it is processed. Thumburu (2020) explains that, through this mechanism, organizations can collect unprocessed data from multiple sources, adapt it to a usable format, and then store it in a unified data repository for analytical purposes. Additionally, ETL pipelines are considered indispensable as they facilitate accurate and timely analyses, making them a central element of any data management plan.

The value of ETL pipelines has expanded due to the dizzying increase in data generation and the requirement for companies to operate with celerity in a competitive market context. The ability to quickly analyze integrated, highquality data is fundamental for organizations to make informed decisions, leading to operational optimization and improved results (Raj et al., 2020). Considering the direct impact of data management on the efficiency of various sectors, and in view of the growing

mobility challenges in urban centers, this Final Project is dedicated to the application of Data Engineering for treating and analyzing transit information.

2. OBJECTIVES

The main objective of this work is to develop an Extraction, Transformation, and Load (ETL) pipeline to collect, process, and store public transit data, enabling the subsequent analysis of traffic patterns and, thus, contributing to the optimization of urban mobility.

To achieve the proposed general objective, were established the following specific objectives:

• Identify and select relevant and accessible public transit data sources for the study;

• Design and implement the ETL pipeline stages using technologies such as Python for processing scripts, MySQL as the database, and Apache Airflow for orchestration and automation;

• Process and transform the collected data, aiming for its structuration in a database optimized for queries and analyses;

• Develop basic visualizations or reports that demonstrate the traffic patterns identified from the processed data, exemplifying their potential to assist in decision-making;

• Preliminarily evaluate the efficiency and scalability of the proposed ETL solution, considering the potential for processing growing volumes of data.

3. DATA ENGINEERING

According to Reis & Housley (2022), ”Data engineering is a set of operations aimed at creating interfaces and mechanisms for the flow and access to information.”

Expanding on this definition, (Chauque, 2023) characterizes Data Engineering as the discipline that encompasses the development, implementation, and maintenance of systems. Such systems are designed to transform raw data into reliable, high-standard information that feeds analysis and machine learning applications. The same author highlights that the Data Engineer plays the crucial role of keeping this data not only available but also ready for consumption by data scientists, analysts, and other users, managing the stages of obtainment, storage, and preparation. Consequently, the data engineering process is established as the foundation for data analysis projects and the development of machine learning models.

4. BIG DATA

Based on the analysis of information about the evolution of data management, it can be affirmed that, although the term Big Data is more recent, the foundations for its development were established between the 1960s and 1970s. During this period, technological advancement drove the emergence of the first data centers and the creation of the relational database model, fundamental milestones that allowed for the storage and organization of ever-larger volumes of information, which were already being generated at that time (Silva; Farina; Florian, 2022).

According to Caldas & Silva (2016), the concept of Big Data is centered on datasets of massive volume that are produced through various technological practices and processes. Among the sources that generate this data are social media, internet access, telephony, operational technologies, and other distributed information sources.

The Big Data phenomenon can be understood as the process of converting a large quantity of raw data, originating from diverse social sources such as the perceptions of customers, suppliers, and competitors, into a set of structured information. This transformation is carried out through specific technologies, techniques, and algorithms that, in addition to enabling the process, are capable of managing the immense volume of data. The final objective is for companies to be able to use this information strategically (Brandao˜ et al., 2019). Therefore, the concept of Big Data is directly related to recent technological advancements for handling massive and diverse types of datasets.

According to (Demchenko et al., 2013), the concept of Big Data can be understood through five essential elements, known as the 5 V's. Volume refers to the immense quantity and size of the data, while Velocity relates to the speed at which it is created and needs to be processed. Variety, in turn, points to the multiple formats and diverse structures the data can present. Additionally, Veracity deals with the reliability and quality of this information, and finally, Value corresponds to the potential to extract significant and useful knowledge from the dataset. As shown in Figure 1.

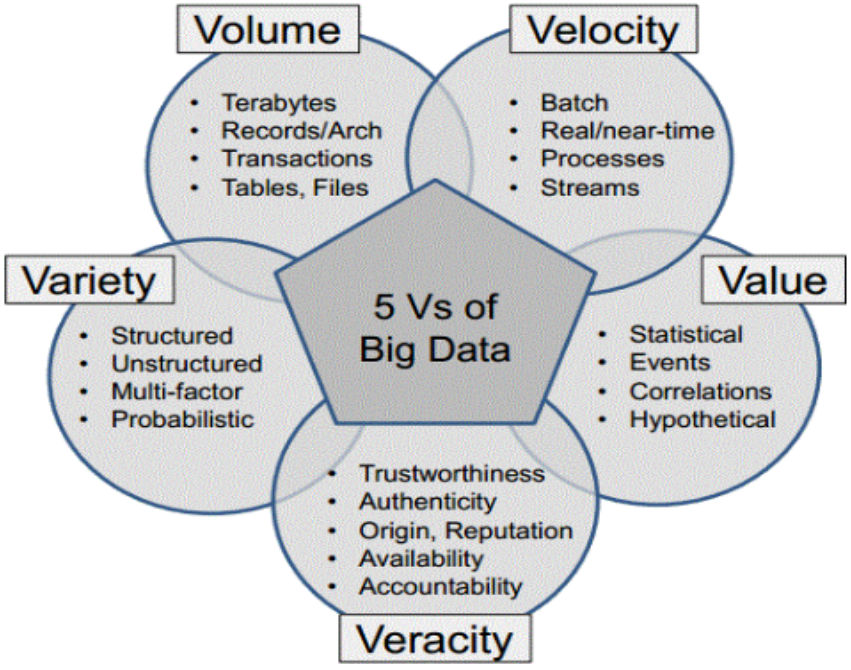

Figure 1: The 5 V's of Big Data. Source:(Demchenko et al., 2013)

5. BIG DATA ANALYTICS

The process known as Big Data Analytics designates the set of activities focused on analyzing large volumes of data, encompassing both structured and unstructured

databases. The fundamental objective of this practice is to extract high value information and insights that serve as a basis to guide companies, enabling more strategic and assertive decision-making (TOTVS, 2023).

According to (Rajaraman, 2016), there are four main types of data analysis:

Descriptive Analysis: Focused on answering the question ”What happened?”, it organizes and presents past data in a visually comprehensible way, such as in charts and dashboards, to facilitate interpretation. A classic example is the presentation of population census data, which classifies inhabitants by criteria such as age, gender, and income.

Predictive Analysis: It uses existing data to extrapolate and predict what is likely to happen in the near future. Tools like time series analysis, neural networks, and machine learning algorithms are employed to, for example, anticipate a customer's needs in e-commerce or manage political campaigns by analyzing electorate perceptions.

Exploratory Analysis (or Discovery): It aims to find unexpected relationships and patterns in large datasets. By analyzing diverse sources such as feedback, tweets, and emails, companies can discover trends in customer behavior, allowing them to anticipate actions (like canceling a service) and create offers to reverse that decision.

Prescriptive Analysis: It goes beyond prediction and seeks to answer the question “What should we do?”. It identifies opportunities and suggests actions to optimize the solution to a problem and achieve a specific goal. A common example is the dynamic pricing of airline tickets, which uses historical travel data to maximize profits.

6. DATA FOR URBAN MOBILITY ANALYSIS

Data collection is an essential component of Intelligent Transportation Systems, providing the foundation for traffic analysis and management. Various technologies are employed to capture information about the transportation system. Traditionally, technologies such as inductive loop detectors and pneumatic tubes have been used to obtain basic data like traffic volume and spot speed, although they present challenges such as high cost and implementation impact (Sumalee; Ho, 2018). With technological advancements, new data sources have become prominent. Video cameras, for example, allow the extraction of information on flow, speed, and vehicle types through image processing. Similarly, automatic license plate recognition (ALPR) and RFID data enable the analysis of routes and travel times (Sumalee; Ho, 2018).

More recently, data from GPS, Bluetooth, WiFi, and mobile telephony have gained prominence due to their capacity for individualized and continuous tracking, allowing for behavioral analyses and traffic conditions in greater detail (Sumalee; Ho, 2018).

As pointed out by (U.S. Department of Transportation, Federal Highway Administration, 2006), the effectiveness of data collection for mobility analysis depends not only on the technology itself but also on its correct application and installation. The

document emphasizes that factors such as environmental conditions, road characteristics, and maintenance costs are determinants in the choice between pavement-installed sensors, such as inductive loops, and non-intrusive technologies, such as cameras and radar, directly impacting the reliability of the data generated for traffic management.

Additionally, transactional systems such as electronic toll collection (ETC) have established themselves as a high-value data source for traffic engineering. Unlike point sensors, ETC data provides precise records of unique vehicle passages through multiple points on a road network, enabling detailed travel time analyses, calculation of origin-destination matrices, and the identification of route patterns with high reliability. As highlighted by (Wang; Luo; Yang, 2025), the vast amount of ETC data—characterized by exact vehicle identification, broad spatio-temporal coverage, and stability under all weather conditions—represents a valuable resource that, when explored, significantly improves the accuracy of traffic condition estimation and prediction.

Beyond the diversity of data sources, extracting useful knowledge for urban mobility analysis depends on a well-defined methodological process to treat and analyze raw data. In this regard, (Oliveira et al., 2023) propose a methodology that ranges from pre-processing to the final data application. The study also conducts a comparison among various Python libraries, such as PyMove, ScikitMobility, and MovingPandas, associating the available tools with each phase of the analysis process. The research concludes that significant gaps exist in current tools, such as the lack of support for trajectory classification and anomaly detection, which points to opportunities for developing more comprehensive methods in the field.

Given the variety of technologies presented, it is concluded that the future of mobility analysis lies not in a single data source, but in their fusion. The combination of data from multiple sources is the essential approach to build a complete and resilient understanding of urban traffic dynamics.

7. DATA PIPELINES

A data pipeline is a fundamental process in modern data engineering, functioning like an assembly line for efficiently collecting, processing, and delivering information (Barbosa, 2020). Essentially, its structure is composed of three key elements: a source, where data is captured; processing steps, where it is transformed; and a destination (or sink), where the treated data is sent (Databricks, 2024). The sources from which data is extracted are diverse, including APIs, files, and both SQL and NoSQL databases. However, this raw data is rarely ready for immediate use, making the processing phase indispensable. During this flow, it is crucial to track data lineage to document its provenance, transformations, and storage location, whether in a data lake, a data warehouse, or local systems (Stryker, 2024). The process occurs in a sequence of steps, where the output of one serve as the input for the next in a continuous manner. To optimize performance, phases that are independent of each other can be executed in parallel, increasing the efficiency of the pipeline as a whole (Barbosa, 2020). Although

the pipeline process is very comprehensive, some examples will be shown in the following sections.

8. BATCH DATA PIPELINES

According to (Richman, 2025), batch pipelines are designed to move large volumes of data at scheduled intervals, being a simpler and more suitable approach for historical reporting. The author explains that this method is computationally intensive, as it usually requires scanning the entire source system with each execution. Historically, this forced companies to run these processes during off-peak hours to avoid overloading their local servers.

9. STREAMING PIPELINES

According to (Vera-baquero; Colomo-palacios; Molloy, 2016), when it is necessary to handle data changes in real-time, the streaming pipeline is the ideal solution, as it processes data continuously as it is collected. In this way, real-time data analysis accelerates the generation of insights and decision-making. There is a wide range of technologies that can be used to operationalize this approach, one of which is Kafka, which, according to (Kreps et al., 2011), was developed as a distributed messaging solution, specifically designed to manage the collection and delivery of high loads of log data while maintaining low latency. According to (Data science academy, 2024), Kafka was designed for high message and data throughput, supporting horizontal scalability through partitioning. This ability to handle thousands of messages per second makes it ideal for real-time streaming applications, such as monitoring and analytics. Resilience is another central feature, guaranteed by replicating data on disk among the nodes of a cluster to ensure fault tolerance. Its fundamental architecture operates with producers, responsible for creating the events, and consumers, who read them. Figure 2 shows its architecture.

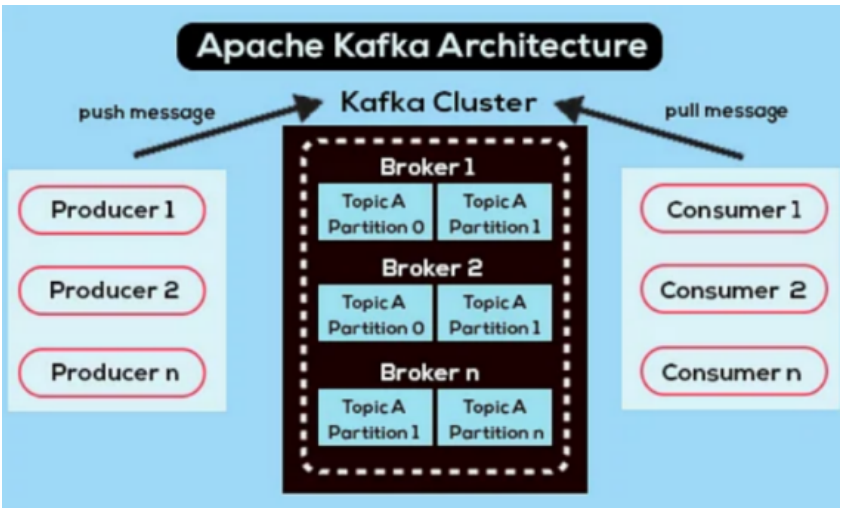

Figure 2: Kafka Architecture. Source:(Data science academy, 2024)

10. EXTRACTION, TRANSFORMATION, AND LOAD (ETL) PIPELINES

One of the main legacies of the fourth industrial revolution is the significant increase in data generation, as its technological innovations enable organizations to collect, create, and analyze data instantly, which has caused exponential growth in data volume over the years. (Bueno, 2023) points out, however, that managing this information faces obstacles related to storage, maintenance, and analytical capabilities. In response to these difficulties, methodologies such as ETL (Extraction, Transformation, and Load) emerged, providing an organized means to handle such information.

According to (Silva, 2022), the Extraction, Transformation, and Load (ETL) pipeline represents a specific category of data pipeline, designed to transfer data from an original source to a destination location. At this location, the data becomes accessible to end-users, who can use it to solve business questions. The ETL process, as its name indicates, is distinctly divided into three stages. The initial phase, Extraction, aims to collect data from various origins, such as systems, applications, and sensors. Subsequently, in the Transformation stage, the collected raw data is processed; it undergoes cleaning, mapping, and conversion, often to fit a predefined schema, which ensures its quality, integrity, and alignment with operational needs. Finally, the Load phase consists of storing the already transformed data in an appropriate destination, such as a database system or other applications.

In summary, the ETL process functions as a technique for consolidating information from various sources into a central point and modifying it into a format suitable for analysis, as explained by (Bueno, 2023). The author also emphasizes that such a process is vital for maintaining data integrity and consistency when used in an environment geared towards data analysis. Figure 3 shows the process.

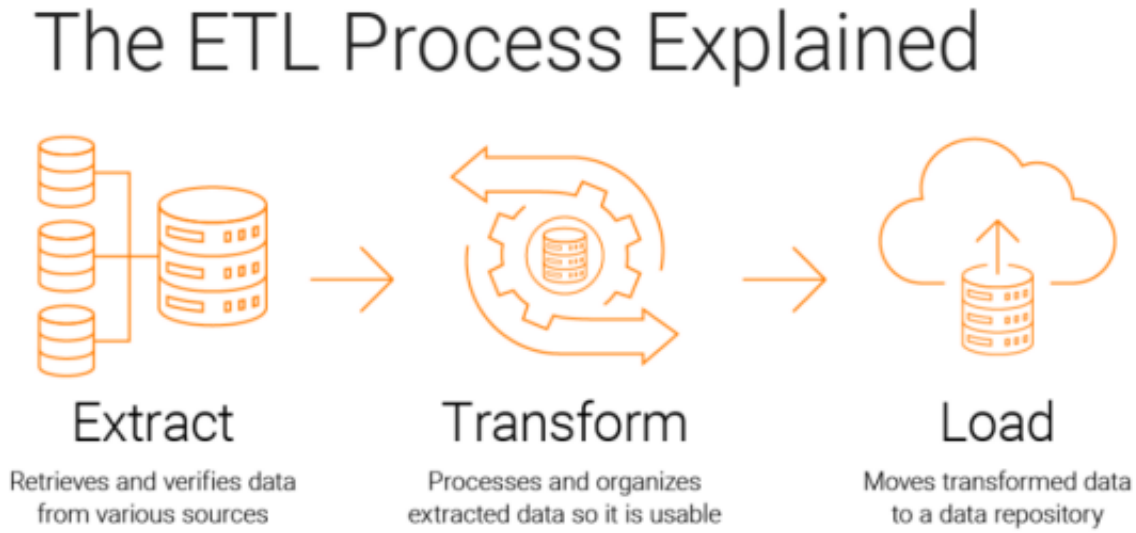

Figure 3: Diagram of an ETL Process Workflow. Source:(Airbyte, 2024)

11. TECHNOLOGIES FOR BUILDING ETL PIPELINES FOR TRAFFIC DATA

Modern urban mobility management, a fundamental pillar for the development of Smart Cities, is intrinsically dependent on an organization's ability to integrate, process, and analyze vast volumes of data in real-time and at scale (International Transport Forum; OECD, 2020). The digital transformation of urban centers aims to use technology to enhance the efficiency of public services and the quality of life for citizens, an objective that can only be achieved with a robust and reactive data infrastructure (Ribeiro; Braghetto, 2022).

According to (John, 2025), to deal with the intrinsic diversity of data formats, organizations frequently employ ETL (Extract, Transform, Load) tools and middleware solutions. Such resources constitute an essential solution for standardizing information from heterogeneous systems. ETL processes operate by extracting data from multiple sources, converting it into a cohesive format, and subsequently loading it into a centralized repository, such as a Data Warehouse. In parallel, middleware platforms act as intermediaries, enabling efficient communication and continuous data exchange between distinct applications. Consolidated market technologies for this purpose include Apache NiFi, Talend, and Informatica.

In the specific domain of transit, this need becomes critical. Data sources include real-time feeds from vehicular GPS (GTFS-RT), data from sensors in traffic lights (JSON/XML), electronic ticketing records with proprietary formats, and information from meteorological APIs. The construction of pipelines with the aforementioned tools is, therefore, indispensable for unifying this variety of sources and generating valuable insights into urban mobility management (Google developers, 2024).

Apache NiFi is an open-source data integration platform designed to automate the flow of data between systems. Its main distinguishing feature is its intuitive, low-code graphical user interface (GUI), which allows for the construction of complex data pipelines, known as dataflows, through a flow-based programming model (Confluent,).

Informatica PowerCenter is an established data integration platform known for its reliability and scalability. Its modular, service-oriented architecture is composed of three main components: the Repository Service, which manages metadata centrally; the Integration Service, a scalable execution engine that orchestrates data flows; and the Domain and Node architecture, which ensures high availability and load balancing (Dataterrain, 2025b).

According to (Dataterrain, 2025a), Talend is another leading data integration platform, recognized for its vast library of connectors and its deployment flexibility (on-premises, cloud, or hybrid).

12. ANALYSIS AND VISUALIZATION OF TRAFFIC DATA

According to (Labor Engenharia, 2023), to understand the relevance of data analysis in traffic management, it is fundamental to understand the dynamics of vehicle and pedestrian flow in cities. Urban mobility is a complex system influenced by multiple variables, such as periods of the day, days of the week, events, meteorological

conditions, road work, and accidents. Data analysis emerges as an essential tool for public agencies, as the collection of real-time information allows for pattern identification, problem anticipation, and more effective decisionmaking for traffic management.

The data visualization step is crucial for translating complex analyses into comprehensible information. According to (Number Analytics), the graphical representation of traffic data uses specific techniques for each objective, such as heat maps to identify congestion points, scatter plots to analyze the relationship between different variables, and bar charts to compare data across distinct categories.

Interactive dashboards are control panels that consolidate multiple data visualizations (such as bar charts, line graphs, maps, and tables) into a single, cohesive interface. Tableau and Power BI: They are considered leaders in the Gartner Magic Quadrant for Analytics and BI platforms (Edmond; Crabtree, 2024). Power BI, developed by Microsoft, is recognized for its user-friendly interface (based on drag-and-drop), strong integration with the Microsoft ecosystem (Excel, Azure), and a cost-effective licensing model, making it a popular choice for many organizations (United Techno, 2025). Tableau, in turn, is acclaimed for its superior flexibility, ability to create advanced and highly customized data visualizations, and its robustness for exploring large and complex datasets. The decision between the two generally depends on the budget, existing IT infrastructure, and the complexity of the organization's analytical needs (DynaTech Systems, 2024).

13. RELATED WORK

In this chapter, a literature review is conducted, focusing on related research that explores the construction of data pipelines and the analysis of traffic data.

A first case study is the Final Project by (Silva, 2022) at the Universidade Federal do Ceara (UFC), which developed a data stream processing pipeline for a mobile urban public transit application. The objective was to notify users in realtime about a vehicle's proximity. The architecture was designed for low latency, using Apache Kafka as a messaging system to receive location data (in JSON format) from mobile devices, Apache Storm as the streaming processing engine to analyze events, and MongoDB as the database to store the results. The clear focus of this work was the immediate reaction to individual events, a typical realtime processing use case.

A second case study, documented by (Tesfaye, 2024), describes the construction of an ELT (Extract, Load, Transform) pipeline and a data warehouse for analyzing traffic data collected by drones. The objective was to create an analytical repository to improve traffic flow. Python was used for processing scripts, Apache Airflow for workflow orchestration and scheduling, and PostgreSQL as the data warehouse. Additionally, the project used the dbt (data build tool) to execute data transformations directly within PostgreSQL, after loading. The focus of this work was the orchestration of batch tasks and the construction of a repository optimized for complex analyses.

Another relevant approach is that of (Jales, 2019) who, in their Final Project at the *Universidade Federal do Rio Grande do Norte* (UFRN), focused on applying data

science techniques for the exploratory analysis of geographic urban transportation data. Using a large public dataset on New York City taxi rides from 2009 to 2014, the study details the complete analysis cycle: from data acquisition and cleaning to visualization and generating insights. The methodology was implemented in the Python language, using libraries like Pandas for manipulation, Matplotlib for creating charts, and, notably, Folium for plotting interactive maps and Shapely for handling polygons and geographic areas. The work's main contribution does not lie in building a system or predictive model, but in demonstrating a methodological process for extracting information and patterns from a large volume of data. Visualizations such as heatmaps were generated to identify points of high trip concentration (like Manhattan and airports) and timeseries analyses that revealed the seasonality of fares and the number of rides by day, month, and period of the day.

When comparing the present work with that of (Silva, 2022), it is observed that, although both address the construction of data pipelines applied to urban mobility, the adopted objectives and architectures differ significantly. Silva's (2022) work focuses on stream processing (streaming) of mobile device data, with a focus on real-time applications, such as notifying passengers about the proximity of public transport vehicles. For this, it employs technologies like Apache Kafka and Apache Storm, allied with a NoSQL database (MongoDB). The solution developed in this Final Project, on the other hand, adopts a batch ETL approach, using Python, MySQL, and Apache Airflow for orchestration, with public transit data provided by municipal agencies as its source. The focus, in this case, is on historical analysis and report generation to support urban planning. This comparison highlights how different demands—immediate reactivity versus strategic analysis—imply distinct architectural choices, even though both are based on Data Engineering principles.

In relation to the work of (Tesfaye, 2024), it is observed that, although both use Apache Airflow for orchestration, the approaches differ regarding the data transformation strategy: while Tesfaye (2024) employs the ELT approach, transferring data first to the data warehouse and transforming it later using dbt, this work follows the traditional ETL model, processing the data before loading it into the database. Furthermore, the data sources also differ: drones in Tesfaye's (2024) case and municipal public data in the present study. This methodological difference highlights how technological choices must be guided by the application context and the analytical objectives of each solution.

Finally, in comparison to the study by (Jales, 2019), it is observed that that work focuses on exploratory data analysis, without implementing a robust and automated pipeline. This Final Project, however, proposes the construction of a complete ETL infrastructure, allowing for the automation of data collection and transformation and enabling recurring and scalable analyses. Thus, while Jales (2019) highlights the potential of data to generate insights through visualizations and statistics, this work differentiates itself by providing a structured solution that can serve as a continuous basis for future applications in urban mobility.

Expanding the comparison beyond final projects, the research by Ribeiro & Braghetto (2022) presents a scalable architecture for data integration in smart cities, focused on real-time processing of heterogeneous sources. In contrast to the solution proposed by them, which aims for reactivity and support for dynamic urban services, the present work adopts a batch ETL pipeline orchestrated by Apache Airflow. This methodological choice proves more suitable for the objective of performing historical analyses of public transit data, generating strategic reports for urban planning. While the architecture of Ribeiro & Braghetto (2022) responds to the need for large-scale continuous monitoring, the approach discussed in this work offers a pragmatic and robust solution for extracting value from already consolidated data, demonstrating a complementary path for applying data engineering to improve urban mobility.

Like this work's proposal, Vasconcelos, Ramos & Coutinho (2023) also developed a data pipeline for public transport traffic analysis. However, the architecture they adopted is based on a classic Big Data ecosystem, using Apache Kafka for streaming data ingestion, HDFS as a distributed storage system, and Apache Spark for parallel processing. In contrast, this project opted for an approach orchestrated with Apache Airflow and storage in a relational database (MySQL). This distinction highlights different data engineering strategies: while the solution by Vasconcelos, Ramos & Coutinho (2023) is designed for scalability and processing large volumes of real-time data, the approach in this work proves robust and pragmatic for historical analysis and generating interactive dashboards for urban planning.

14. MATERIALS AND METHODS

In this section, the materials and methodology used to achieve the objectives proposed in this work are presented. The project's development consisted of constructing an Extraction, Transformation, and Load (ETL) pipeline for public transit data from the city of Curitiba, following the steps of data source identification, technology architecture definition, and implementation of the processing and storage flow.

14.1. Data Sources

For this study, open data sources provided by Curitiba's municipal agencies were used, as per the first specific objective. The sources were selected for their relevance to urban mobility analysis and their diversity of formats, representing a realistic data integration scenario.

14.1.1. Source 1: Geographic Data from IPPUC

Description: Georeferenced data from the road system, such as the Road Hierarchy (Hierarquização Viária) and Neighborhood Division (Divisa de Bairros), which classifies and delimits the city's roads and neighborhoods, were used (Instituto de Pesquisa e Planejamento Urbano de Curitiba (IPPUC), 2025). This data is essential for geographically contextualizing the occurrence analyses. Format: The data was obtained in Shapefile (.shp) format, a standard for geographic information systems (GIS) data.

14.1.2. Source 2: Sigesguarda Occurrence Database

Description: This dataset, from the Municipal Secretariat of Social Defense and Transit, details all occurrences registered by the Municipal Guard (Curitiba (Município). Secretaria Municipal de Defesa Social e Trânsito, 2025). It includes crucial information such as the nature of the occurrence (Traffic, Support, Damage, etc.), the date, the time, and the location (neighborhood and street). Format: The data is available in comma-separated values (.csv) format.

14.2. Project Architecture

The project's architecture was designed to be automated and scalable, using a set of open-source technologies widely adopted in data engineering projects, as shown in Figure 4.

Programming Language (Python): All extraction and transformation logic was developed in Python (version 3.12). Essential libraries were used, such as Pandas for manipulating tabular data, GeoPandas for reading and processing geographic data (Shapefiles), and SQLAlchemy in conjunction with mysqlconnector-python for interacting with the database.

Database (MySQL): MySQL (version 8.0) was chosen as the database management system for centralized and structured data storage.

Workflow Orchestrator (Apache Airflow): For automating, scheduling, and monitoring the pipeline, Apache Airflow was used. The tool allowed the workflow to be defined as a Directed Acyclic Graph (DAG), ensuring that the pipelines for each data source were executed on a schedule. The environment was configured and run using Docker, ensuring portability and service isolation.

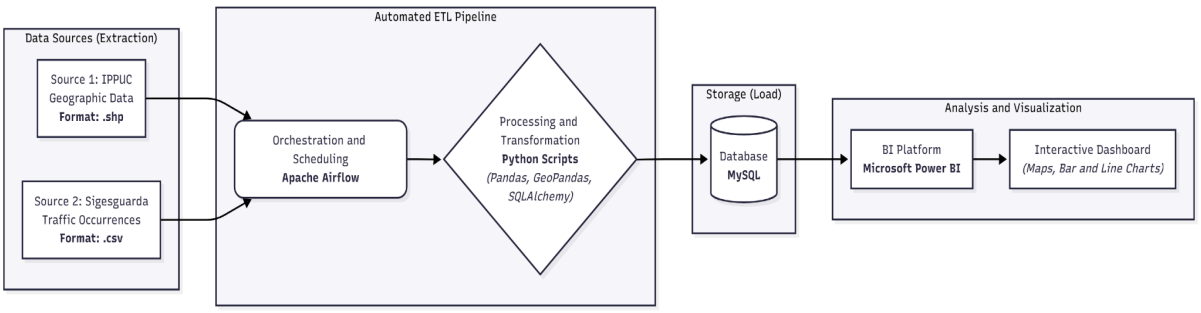

Figure 4: Flowchart of the Proposed ETL Pipeline Architecture. (Source: The author, 2025).

14.3. ETL Pipeline Development

The ETL process was orchestrated by Apache Airflow, divided into independent pipelines for each data source to handle the specificities of each format and ensure project modularity.

14.3.1. Pipeline 1: Processing Geographic Data from IPPUC

The first pipeline was responsible for processing geographic data from IPPUC. The objective was to extract information from Curitiba's road system from its original

Shapefile format, transform it, and load it in a structured way for cross references with occurrence data.

14.3.2. Extraction of Geographic Data

The extraction was performed using the GeoPandas library, which specializes in manipulating geospatial data in Python. A function was developed to read the SIST VIARIO CLASSIFICADO.shp file, which contains the geometries (lines) of each city road, loading the data into a GeoDataFrame.

14.3.3. Data Transformation and Preparation

After extraction, the data underwent a transformation step. Only the columns of interest were selected, and they were renamed to follow a clearer standard. The most important transformation was converting the geometry column to the WKT (Well-Known Text) format, a textual representation that allows storing geographic information in standard databases.

14.3.4. Loading Geographic Data

In the final stage of this pipeline, the processed GeoDataFrame, with its geometry already in WKT format, was loaded into a specific table in the MySQL database.

14.4. Pipeline 2: Processing Sigesguarda Occurrence Data

This second pipeline was designed to extract, treat, and load the transit occurrence data from Sigesguarda. The extraction started from the original CSV file, read into a Pandas DataFrame. The transformation step, in turn, focused on cleaning, standardizing, and enriching the information, chief among them being the treatment of date and time columns to create a standardized timestamp and new temporal variables. Finally, the resulting DataFrame was loaded into its respective table in the central database.

14.5. Final Stage: Load Strategy

In the final stage of both pipelines, the transformed DataFrames were loaded into a MySQL database. This approach was adopted to centralize and persist the data in a structured manner, ensuring it was available consistently for analysis. The connection and data insertion were managed by a specific function using the SQLAlchemy library, ensuring efficient communication with the database and data integrity for later consumption by tools like Power BI.

15. DATA ANALYSIS AND VISUALIZATION

To demonstrate the analytical potential of the processed data and achieve the fourth specific objective, the analysis and visualization stage was conducted using the Microsoft Power BI platform. The choice of this tool is based on its market leadership, as noted in this work's literature review, and its user-friendly interface and ability to create interactive reports.

The process consisted of connecting Power BI directly to the MySQL database, which stores the data already treated and validated by the ETL pipeline. Within the Power BI environment, data modeling was performed, including creating relationships

between the occurrence and road system tables, and building an interactive dashboard. Visualizations were developed, such as a custom Shape Map to identify points of high occurrence concentration by neighborhood, as well as bar and line charts for analyzing temporal and occurrence-type patterns.

16. RESULTS AND DISCUSSION

In this chapter, the results obtained from the analysis of traffic occurrence data in Curitiba, processed through the previously described ETL pipeline, are presented. The visualizations, developed on the Microsoft Power BI platform, aim to identify patterns and trends in the incident records. The results will be presented in thematic subsections, beginning with the geographical distribution, followed by temporal analysis and analysis by occurrence type, in line with the objectives proposed by this work.

16.1. Geographical Analysis of Occurrences

The first analysis seeks to understand the spatial distribution of incidents in the municipality. For this, a choropleth map (shape map) was created to illustrate the concentration of the number of occurrences by neighborhood. This visualization, presented in Figure 5, cross-references the geographic data of neighborhood boundaries, provided by IPPUC, with the occurrence records from Sigesguarda, allowing for the identification of areas with higher and lower incidence during the year 2025.

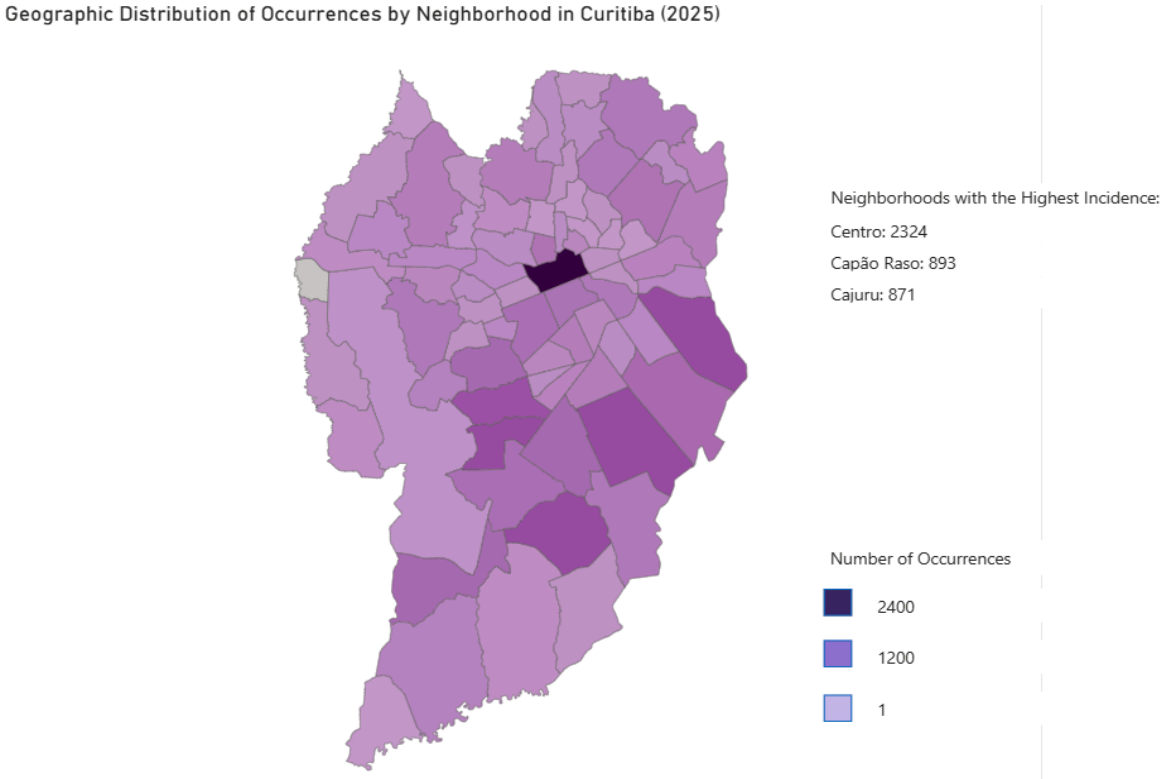

Figure 5: Quantitative distribution of traffic occurrences by neighborhood in Curitiba in 2025. Source: Prepared by the author (2025).

The map analysis reveals a significantly uneven distribution of occurrences across the territory. The "*Centro*" (Downtown) neighborhood emerges as the area of highest criticality, recording a total of 2,324 occurrences in the analyzed period, a volume drastically superior to any other region in the city.

In addition to the central area, the *Capão Raso* and *Cajuru* neighborhoods also stand out as high-incidence points, with 893 and 871 occurrences, respectively. In contrast, neighborhoods located in the more peripheral areas of the municipality, especially in the far south and west, show a considerably lower incidence. This geographical distribution suggests a strong correlation between the volume of occurrences and the density of vehicle and pedestrian flow in the different regions of the city.

16.2. Analysis by Occurrence Type

After analyzing the geographical distribution, the incidents were categorized to identify the most frequent types of occurrences registered by the Municipal Guard. Figure 6 presents the quantitative distribution of records for the top 10 categories, consolidated for the entire year of 2025.

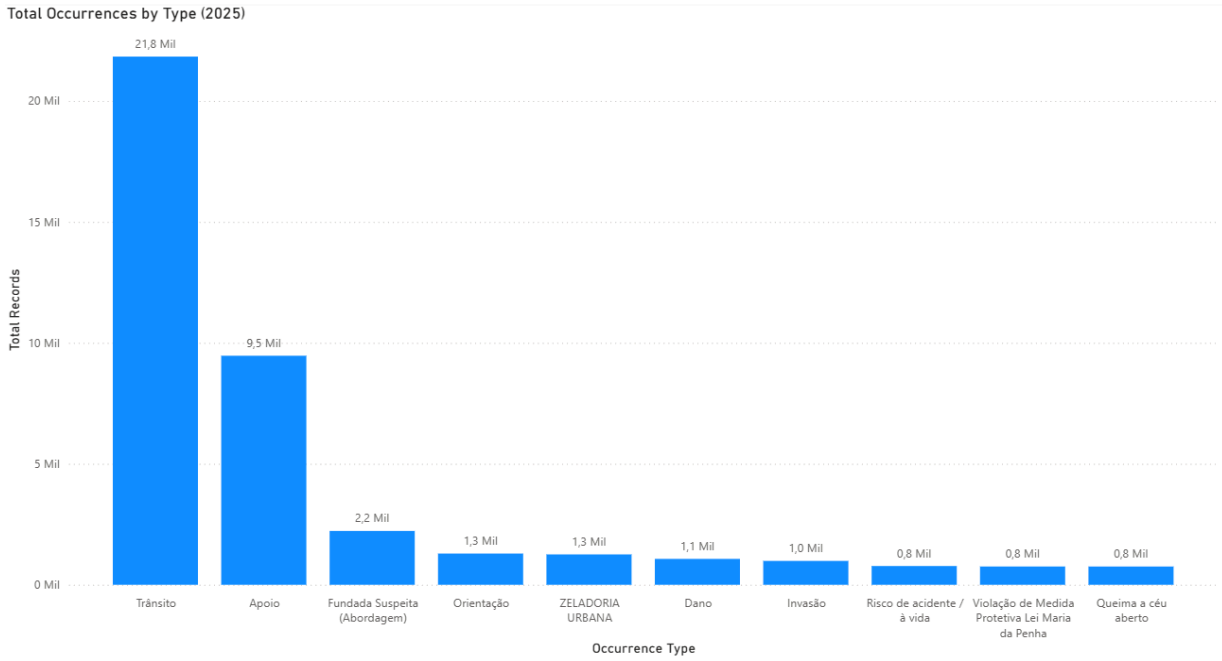

Figure 6: Total records by occurrence type in 2025. Source: Prepared by the author (2025).

The analysis of the distribution reveals a predominant concentration in a small number of categories. The "*Transito*" (Traffic) nature stands out expressively as the most recurrent, totaling almost 22,000 records (21,800). This volume is more than double the second most frequent category, "*Apoio*" (Support), which accounted for 9,500 records.

After the two main categories, a sharp drop in volume is observed, with "*Fundada Suspeita (Abordagem)*" (Reasonable Suspicion - Approach) appearing

with 2,200 records, followed by other categories with even smaller volumes. The prominence of the "Transito" category validates the relevance of this work's ^ theme, confirming that incidents related to urban mobility constitute the largest share of registered occurrences and, therefore, warrant in-depth analysis.

16.3. Temporal Analysis of Occurrences

After the spatial analysis, the study delved into the temporal patterns of the records. This section investigates how occurrences are distributed over time, beginning with intraday analysis, i.e., throughout the 24 hours of the day.

16.4. Distribution of Traffic Occurrences by Hour of the Day

To identify the periods of highest and lowest incidence of traffic incidents throughout a typical day, a line graph was generated consolidating the "*Transito*" (Traffic) category occurrences by hour, as presented in Figure 7.

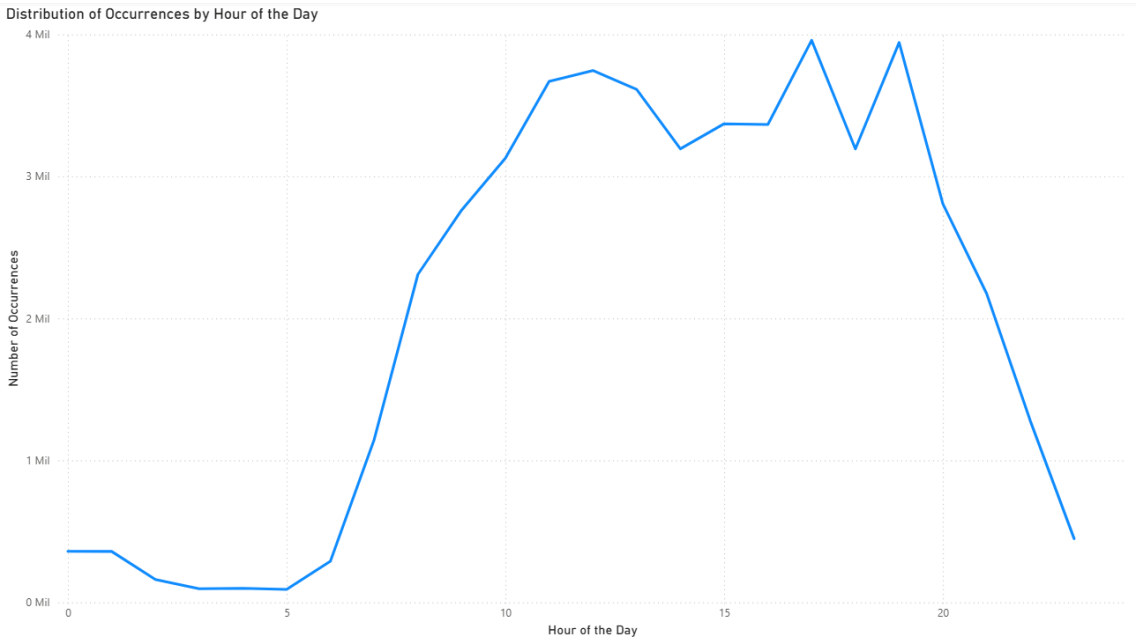

Figure 7: Distribution of the number of Traffic occurrences by hour of the day in Curitiba (2025). Source: Prepared by the author (2025).

The graph demonstrates a daily pattern directly associated with the dynamics of urban mobility. The period of lowest incident risk occurs in the early morning, with the lowest volume recorded between 3h and 5h a.m. Starting at 6h, with the beginning of the morning commute, the number of occurrences grows sharply and continuously, reaching the first peak of the day around 11h, with almost 4,000 records.

A notable drop is observed during lunchtime (12h-13h), followed by an afternoon period of high intensity and volatility. The maximum peak of traffic incidents occurs around 17h, surpassing the 4,000-occurrence mark. Unlike the morning peak, the afternoon presents other high-critical moments, such as a second expressive peak at 19h. After this last one, the quantity of occurrences enters a steep decline, progressively reducing throughout the night. This behavior shows that the periods of greatest risk for

traffic incidents are concentrated during peak travel times, especially in the late afternoon, which encompasses both the end of the workday and the beginning of the evening flow.

16.5. Cross-Analysis by Road System

To investigate whether the road's functionality influences the nature of occurrences, a cross-analysis was performed between the occurrence type and the road system classification, as defined by IPPUC. "*Setorial*" type roads, for example, are designed to connect different neighborhoods, while "*Coletoras*" (Collector roads) distribute traffic locally. Figure 8 compares the proportion of the main occurrence types for selected road categories.

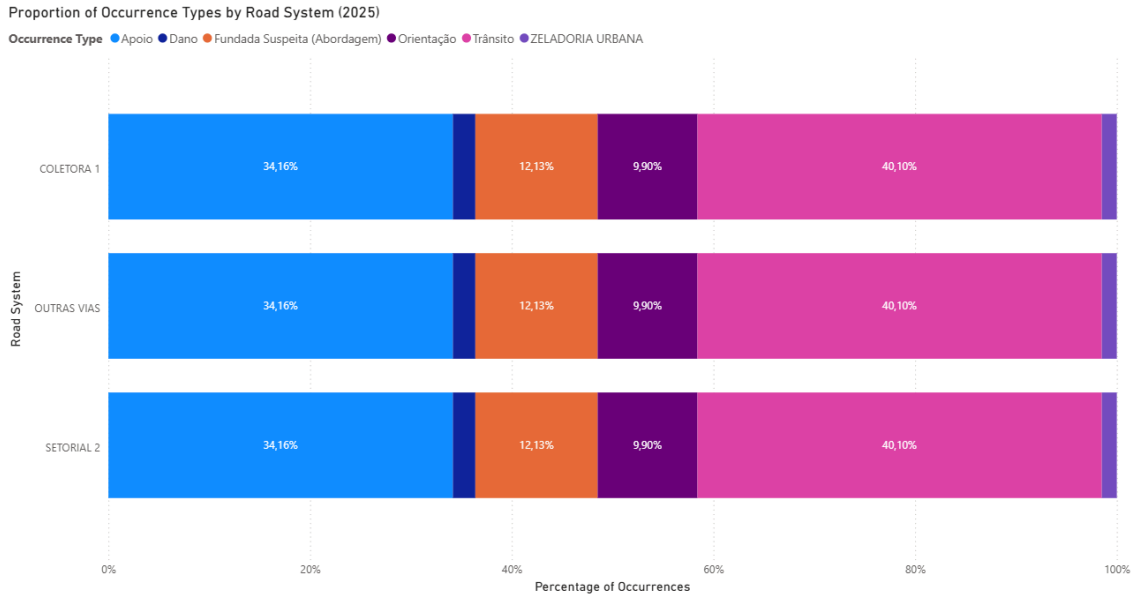

Figure 8: Proportion of occurrence types by selected road system categories (2025). Source: Prepared by the author (2025).

The most notable result revealed by Figure 8 is the surprising consistency in the proportion of occurrence types across the different road classifications analyzed. Contrary to the hypothesis that higher-flow roads, such as "*Setoriais*" (Sectoral), might have a higher proportion of traffic incidents, the distribution remains identical.

Precisely, in all road categories— *Coletora* 1, *Setorial* 2, and *Outras Vias* (Other Roads)—the "*Transito*" (Traffic) occurrence accounts for 40.10% of the to- ˆ tal records, followed by "*Apoio*" (Support) with 34.16%. The categories "*Fundada Suspeita (Abordagem)*" (Reasonable Suspicion - Approach) and "*Orientação*" (Guidance) also show identical proportions in all scenarios, with 12.13% and 9.90%, respectively.

This finding suggests that, although the absolute volume of occurrences may vary by road type, the proportional distribution of the main types of service provided by the Municipal Guard is not impacted by the road hierarchy. This indicates that the service request patterns maintain a stable and homogeneous profile across different

urban contexts in the city. This is valuable information for resource allocation planning, which will be discussed in the conclusions of this work.

17. CONCLUSION

This Final Project aimed to design and implement an Extraction, Transformation, and Load (ETL) pipeline for public transit data from the city of Curitiba, in order to demonstrate the potential of Data Engineering in the analysis and optimization of urban mobility. The solution, developed with open-source technologies such as Python, Apache Airflow, and MySQL, and visualized with Microsoft Power BI, proved effective in collecting, processing, and presenting valuable information from heterogeneous data sources.

The results obtained confirmed relevant patterns regarding the city's traffic dynamics. The geographical analysis showed a significant concentration of occurrences in the 'Centro' neighborhood, as well as on important road axes such as *Capão Raso* and *Cajuru*, highlighting the points of greatest pressure on the road system. The categorization of occurrences showed that 'Traffic' incidents are, by far, the most frequent, validating the relevance of this work's focus. Furthermore, temporal analysis revealed incident peaks coinciding with high-flow periods, with the afternoon period being the most critical. Finally, cross-analysis by road system revealed that the proportion of occurrence types remains relatively stable across different road hierarchies, suggesting a homogeneous service profile by the Municipal Guard in various urban contexts.

Given these results, it can be affirmed that all objectives were successfully achieved. Relevant data sources from IPPUC and *Sigesguarda* were identified and processed (Objective 1). The ETL pipeline was implemented using the proposed technologies, with Python scripts orchestrated by Apache Airflow (Objective 2). The data was transformed, cleaned, and structured, with emphasis on converting geometries to the WKT format and date/time feature engineering (Objective 3). Visualizations were also developed in an interactive dashboard, which presented traffic patterns clearly and objectively (Objective 4). Finally, the Docker-based architecture and automation with Airflow provided the solution with the necessary scalability and efficiency to handle growing data volumes, preliminarily validating its design (Objective 5).

However, some limitations are acknowledged. The analysis was based on a temporal scope of only one year (2025), which prevents the identification of long-term trends. Furthermore, the scope was restricted to two data sources; integrating other databases, such as meteorological information, city event data, or mobility app records, could enrich the analyses and reveal new correlations.

For future work, it is suggested to expand the pipeline to include the aforementioned data sources, as well as applying machine learning models to the already processed data, enabling predictive analyses, such as estimating high-risk accident zones throughout the day. Another natural evolution would be the transition of the pipeline from a batch model to a real-time streaming architecture, using

technologies like Apache Kafka, which would allow for continuous monitoring and more immediate responses to traffic incidents.

In contrast to solutions based on stream data processing, such as that of Silva (2022), which prioritize real-time applications, and with approaches focused on drone usage and subsequent transformation into data warehouses, like that of Tesfaye (2024), this work differentiates itself by adopting a batch ETL pipeline oriented towards the use of public transit data. This choice proved adequate for historical analysis and for generating strategic reports to support urban planning, offering a complementary contribution to other proposals in the literature.

In summary, this work not only fulfilled its technical objectives but also highlighted the role of Data Engineering in building smarter cities. By enabling the transformation of raw data into useful information for urban planning, the developed solution contributes to more informed decisions and the improvement of mobility and quality of life for the population.

18. ACKNOWLEDGEMENTS

I am grateful to all who contributed, directly or indirectly, to my undergraduate degree. My sincere thanks.

REFERENCES

Airbyte. (2024). What is an ETL Pipeline? Definition, Examples, and How to Build One. (https://airbyte.com/data-engineering-resources/etl-pipeline).

Arrieta-Ibarra, I. et al. (2018). Should we treat data as labor? moving beyond "free". AEA Papers and Proceedings, v. 108, p. 38-42.

Azevedo, T. F. (2024). Ensino de Engenharia de Dados nas universidades brasileiras e o mercado: estado atual e propostas de modernização. 340 p. Dissertação (Dissertação (Mestrado em Informática)) - Universidade Federal do Amazonas, Manaus, AM.

Barbosa, T. E. (2020). Mas Afinal, o Que é Pipeline de Dados? (https://blog.dsacademy.com.br/mas-afinal-o-que-e-pipeline-de-dados/).

Bie, T. D. et al. (2022). Automating data science. Communications of the ACM, v. 65, n. 3, p. 76-87. (https://dl.acm.org/doi/10.1145/3495256).

Brandão, R. et al. (2019). Uso do big data no contexto de inteligência competitiva: revisão sistemática da literatura. In: Proceedings of the 16th International Conference on Information Systems and Technology Management - CONTECSI. [S.I.: s.n.].

Bueno, D. M. C. (2023). Automação e engenharia de dados para a análise das disparidades salariais de gênero no mercado de trabalho brasileiro: um estudo de 2012 a 2021. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Controle e Automação) Instituto de Ciência e Tecnologia, Universidade Estadual Paulista, Sorocaba. 53 p.

Caldas, M. S.; Silva, E. C. C. (2016). Fundamentos e aplicação do Big Data: como tratar informações em uma sociedade de yottabytes. Bibl. Univ., Belo Horizonte, v. 3, n. 1, p. 65-85, jan./jun.

Chauque, H. C. (2023). Engenharia de Dados: Proposta de um Pipeline ETL para integração de dados das vendas na CDM: Caso de estudo: Cervejas de Moçambique. Maputo: [s.n.]. Estágio Profissional Licenciatura em Engenharia Informática, Universidade Eduardo Mondlane, Faculdade de Engenharia, Departamento de Engenharia Electrotécnica.

Confluent. What Is Apache NiFi? How It Works, and When to Use It. (https://www.confluent.io/learn/apache-nifi/).

Curitiba (Município). Secretaria Municipal de Defesa Social e Trânsito. (2025). Sigesguarda: Atendimentos 153. (https://dadosabertos.curitiba.pr.gov.br/conjuntodado/detalhe?chave=b16ead9d-835e-41e8-a4d7-dcc4f2b4b627).

Data science academy. (2024). Processamento de Streaming de Eventos com Apache Kafka. (https://blog.dsacademy.com.br/processamento-de-streaming-de-eventos-com-apache-kafka/).

Databricks. Data Pipelines. 2024. ⟨https://www.databricks.com/br/glossary/data-pipelines⟩. Acesso em: 23 de junho de 2025.

Dataterrain. (2025). Informatica PowerCenter Architecture: Components and Benefits. (https://dataterrain.com/informatica-powercenter-architecture-components-benefits).

Demchenko, Y. et al. (2013). Addressing big data issues in scientific data infrastructure. In: Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS 2013). [S.I.: s.n.]. p. 48-55. First International Symposium on BigData and Data Analytics in Collaboration (BDDAC 2013).

DynaTech Systems. (2024). Power BI vs Tableau: Which is better in 2024? (https://dynatechconsultancy.com/blog/power-bi-vs-tableau-which-is-better-in-2024).

Edmond, S.; Crabtree, M. (2024). Power BI vs. Tableau: Which One Should You Choose? (https://www.datacamp.com/blog/power-bi-vs-tableau-which-one-should-you-choose).

Google developers. (2024). Referência do GTFS Realtime. (https://developers.google.com/transit/gtfs-realtime/reference?hl=pt).

IBGE. (2000).Censo demográfico 2000 - Resultado Universo: População Residente, por Situação do Domicílio e Sexo, segundo os Grupos de Idade - Brasil. Rio de Janeiro, RJ: Instituto Brasileiro de Geografia e Estatística (IBGE).

Instituto de Pesquisa e Planejamento Urbano de Curitiba (IPPUC). (2025). Geodownloads. (https://ippuc.org.br/geodownloads/geo.htm).

Instituto de Pesquisa Econômica Aplicada (Ipea). (2010). Acessibilidade no transporte urbano de passageiros: um panorama da política pública federal. In: Brasil em

Desenvolvimento: Estado, Planejamento e Políticas Públicas. Brasília: Ipea. v. 2, p. 407-428. (https://repositorio.ipea.gov.br/handle/11058/3777).

International Transport Forum; OECD. (2020). Leveraging Digital Technology and Data for Human-Centric Smart Cities: The Case of Smart Mobility. [S.I.]. Prepared under the auspices of the Saudi Arabia G20 Presidency.

Jales, D. M. (2019). Abordagens para análise de dados geográficos em transportes urbanos. Natal, RN: [s.n.]. Trabalho de Conclusão de Curso (Engenharia de Computação).

John, B. (2025). Challenges and solutions in data integration for heterogeneous systems. ESP Journal of Engineering & Technology Advancements, v. 2, n. 4, March.

Júnior, S. M. B. (2016). Trabalho de Conclusão de Curso, Sistema de contagem de fluxo de veículos. Pato Branco, PR: [s.n.]. (https://riut.utfpr.edu.br/jspui/bitstream/1/14629/1/PB_COENC_2016_2_09.pdf).

Kreps, J. et al. (2011). Kafka: A distributed messaging system for log processing. In: Proceedings of the NetDB. [S.l.: s.n.]. v. 11, p. 1-7.

Labor Engenharia. (2023). Análise de dados e inteligência de tráfego: como usar a tecnologia para melhorar a mobilidade urbana. (https://laborengenharia.com/analise-de-dados-e-inteligencia-de-trafego-como-usar-a-tecnologia-para-melhorar-a-mobilidade).

Lee, J. W. (2020). Big data strategies for government, society and policy-making. The Journal of Asian Finance, Economics and Business, v. 7, n. 7, p. 475-487. (https://doi.org/10.13106/jafeb.2020.vol7.no7.475).

Meneses, H. B. (2003). Interface lógica em ambiente SIG para bases de dados de sistemas centralizados de controle do tráfego urbano em tempo real. 183 p. Dissertação (Dissertação (Mestrado em Engenharia de Transportes)) Universidade Federal do Ceará, Centro de Tecnologia, Fortaleza.

Number Analytics. The Ultimate Guide to Traffic Data Analysis for Highway Engineering. (https://www.numberanalytics.com/blog/ultimate-guide-traffic-data-analysis-highway-engineering).

Nylén, D.; Holmström, J. (2015). Digital innovation strategy: a framework for diagnosing and improving digital product and service innovation. Business Horizons, Elsevier, v. 58, n. 1, p. 57-67.

Oliveira, E. et al. (2023). Tratamento e análise de dados de mobilidade urbana: Uma metodologia teórica e prática. In: Minicursos da X Escola Regional de Computação do Ceará, Maranhão e Piauí (ERCEMAPI 2023). Porto Alegre: Sociedade Brasileira de Computação (SBC). cap. 3, p. 55-79.

Pereira, C.; Silva, J.; Oliveira, M. (2021). An event-driven serverless etl pipeline on aws. Applied Sciences, MDPI, v. 11, n. 1, p. 191. (https://www.mdpi.com/2076-3417/11/1/191).

Raj, A. et al. (2020). Modelling data pipelines. In: 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). [S.I.: s.n.].

Rajaraman, V. (2016). Big data analytics. Resonance, Springer, v. 21, n. 8, p. 695-716.

Reis, J.; Housley, M. (2022). Fundamentals of Data Engineering. 1. ed. USA: Wilkey.

Ribeiro, M. B.; Braghetto, K. R. (2022). A Scalable Data Integration Architecture for Smart Cities: Implementation and Evaluation. Journal of Information and Data Management, v. 13, n. 2, p. 207-223.

Richman, J. (2025). Data Pipelines Explained: What They Are and How They Work. (https://estuary.dev/blog/what-is-a-data-pipeline/).

Silva, C. M. d. (2022). Trabalho de Conclusão de Curso (Graduação), Pipeline de processamento de dados de dispositivos móveis em transporte coletivo urbano. Russas, CE: [s.n.]. (http://repositorio.ufc.br/handle/riufc/64493).

Silva, M. V.; Farina, R. M.; Florian, F. (2022). BIG DATA: FUNDAMENTOS E APLICAÇÃO NAS EMPRESAS. RECIMA21 Revista Científica Multidisciplinar, v. 3, n. 12, dec. ISSN 2675-6218. (https://doi.org/10.47820/recima21.v3i12.2427).

Stryker, C. (2024). O que é um pipeline de dados? (https://www.ibm.com/br-pt/topics/data-pipeline).

Sumalee, A.; Ho, H. W. (2018). Current trends in intelligent transportation systems (ITSS) and smart cities. IATSS Research, v. 42, n. 2, p. 67-71, jun. (https://doi.org/10.1016/j.iatssr.2018.05.005).

Tesfaye, A. (2024). Traffic Data Pipeline and Warehouse. Owlcation. (https://owlcation.com/stem/Traffic-Data-Pipeline-And-Warehouse).

Thumburu, S. (2020). A comparative analysis of etl tools for large-scale edi data integration. J Innov Technol, v. 3, n. 1.

TOTVS. (2023). Big Data Analytics: o que é, como funciona e suas vantagens. (https://www.totvs.com/blog/inteligencia-dados/big-data-analytics/).

United Techno. (2025). Power BI vs Tableau: A Detailed Comparison Guide. (https://www.unitedtechno.com/power-bi-vs-tableau-detailed-comparison-guide/).

U.S. Department of Transportation, Federal Highway Administration. (2006). Traffic Detector Handbook: Third Edition-Volume I. McLean, VA.

Vasconcelos, F. F.; Ramos, V. T.; Coutinho, F. J. (2O23). Os desafios e soluções para a implementação de big data analytics em cidades inteligentes. In: Companion Proceedings of the 38th Brazilian Symposium on Databases. [S.I.: s.n.]. p. 50-56.

Vera-baquero, A.; Colomo-palacios, R.; Molloy, O. (2016). Real-time business activity monitoring and analysis of process performance on big-data domains. Telematics and Informatics, Elsevier, v. 33, n. 3, p. 793-807.

Wang, Y.; Luo, R.; Yang, X. (2025). Urban traffic state sensing and analysis based on etc data: A survey. Applied Sciences, v. 15, n. 12, p. 6863.