Mobility Project
Summary Article

Carlos Torregrosa and Julia García

November, 2025

Master's Degree in Computational Engineering and Industrial Mathematics

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

1 Introduction

This project focuses on the design, implementation, and analysis of a large-scale mobility data platform based on a lakehouse architecture. Using publicly available mobility, geographic, and socioeconomic datasets for Spain, the project aims to transform raw, heterogeneous data into structured analytical outputs capable of supporting complex, decision-oriented questions.

The primary objective of the work is not limited to data ingestion or exploratory querying, but rather to demonstrate how an end-to-end analytical infrastructure can be built to reliably answer realistic business questions. In particular, the project seeks to extract meaningful insights about temporal mobility patterns, spatial connectivity imbalances, and indicators of industrial activity across different geographic resolutions.

To achieve this goal, a layered lakehouse architecture was progressively developed, combining cloud-based storage, analytical processing, metadata management, and workflow orchestration. Raw data is ingested and preserved without modification, then cleaned, integrated, and normalized before being transformed into analysis-oriented datasets. The entire pipeline is fully automated and parameterized, allowing analyses to be reproduced for different spatial units, temporal ranges, and geographic areas.

The remainder of the report is organized as follows. Chapter 2 describes the data infrastructure and lakehouse architecture, detailing the Bronze, Silver, and Gold layers as well as the orchestration mechanisms. Chapter 3 presents the data methodology, including data sources, preprocessing strategies, integration choices, and analytical assumptions. Chapter 4 introduces the analytical framework and formulates the business questions addressed in the project. Chapter 5 discusses how the adopted architecture enables expert-level insights, while Chapter 6 outlines the main limitations encountered and the lessons learned throughout the development process.

2 Infrastructure Overview

This chapter presents the data infrastructure designed and implemented throughout the project. The proposed architecture follows a lakehouse paradigm, combining scalable data storage with analytical processing capabilities. Although the project was initially framed within an academic context, the resulting infrastructure closely resembles a professional data platform, both in terms of design decisions and operational behavior. The architecture evolved progressively as new challenges emerged, leading to a robust, scalable, and reproducible solution.

2.1 General Architecture

The overall architecture is based on a lakehouse approach, where raw data ingestion, structured refinement, and analytical consumption are clearly separated into dedicated layers. While the project originated as an academic assignment and many of the technologies and concepts were learned during its development, the final outcome is oriented toward answering realistic, decision-driven analytical questions rather than merely executing isolated SQL queries.

The architectural design was not fixed from the beginning but evolved iteratively as data volume increased and operational issues surfaced. This evolutionary approach led to the adoption of principles such as scalability, traceability, and modularity. A key objective was to ensure that the same pipeline could be validated on small time windows and later scaled to a full year of data without structural changes.

From a technological perspective, data storage is handled through an Amazon S3 bucket, enabling shared access and avoiding local storage limitations. DuckDB is used as the analytical engine due to its efficiency and seamless integration with Parquet-based lakehouse workflows. Metadata management is supported by a PostgreSQL database hosted on Neon, which is used to store and coordinate structural and catalog information required by the lakehouse setup. Workflow orchestration is managed using Apache Airflow, while spatial visualizations are produced with Kepler.gl. These technologies were selected both because they were introduced during the course and because they provided an effective balance between simplicity and real-world applicability.

2.2 Bronze Layer: Raw Data Ingestion

The Bronze layer is responsible for ingesting raw data exactly as it is obtained from external sources. The primary data source is the Spanish Ministry of Transport (MITMA), from which daily mobility datasets for the year 2023 were downloaded at multiple spatial resolutions, including municipalities, districts, and GAUs. Each dataset is provided as a separate daily file, leading to a large number of ingestion units.

Additional data sources include a calendar dataset for the Community of Madrid, which required preprocessing to ensure correct parsing, as well as geographic data containing polygon geometries. Socioeconomic data, such as population and income at the municipal level, were obtained from the National Statistics Institute (INE), along with a correspondence file linking MITMA and INE identifiers. Finally, provincial boundary geometries were sourced from a public GitHub repository.

Prior to Bronze, a preliminary pre-Bronze stage was introduced to store all original files in their native formats, predominantly CSV. The Bronze layer then ingests these raw files and converts them into Parquet tables stored in S3. No cleaning, deduplication, or semantic transformations are applied at this stage. As a result, Bronze preserves null values, duplicated records, and original schemas, ensuring full data fidelity and reproducibility.

Data in Bronze is organized by domain-specific directories, such as mobility flows by spatial resolution, population, income, and geographic relationships. This organization facilitates inspection and verification of the ingestion process while maintaining a clear separation of data domains. Once data reaches Bronze, it is treated as immutable and serves as a reliable reference point for downstream processing.

2.3 Silver Layer: Data Refinement and Integration

The Silver layer is designed following a strict guiding principle: each piece of information should have a single, well-defined place within the data model. All data ingested into Bronze is restructured and normalized in Silver to eliminate redundancy while preserving completeness.

At this stage, extensive data cleaning operations are performed, including the removal of null values, elimination of duplicated records, and normalization of units to the International System. Column formats are unified, and consistent data types are enforced across all tables. One of the most challenging tasks involved resolving inconsistencies between identifiers originating from different sources, particularly between MITMA and INE datasets. A unified identifier strategy was adopted to ensure consistent joins across all tables.

Silver introduces a structured relational model composed of a limited number of well-defined tables, including mobility trips, calendar information, places, population, and income. Spatial data from multiple sources, such as

provinces and other geographic entities, are consolidated into a single places table. Similarly, mobility data from different spatial resolutions are unified into a single trips table, while preserving metadata that indicates their original spatial granularity.

Although Silver integrates geographic information and mobility data, socioeconomic attributes remain in separate tables to avoid unnecessary duplication. The resulting Silver layer is fully cleaned, consistent, and suitable for direct analytical use. In fact, Silver can be considered a stable analytical interface, allowing meaningful queries without requiring access to Bronze or Gold.

2.4 Gold Layer: Analytical and Decision-Oriented Outputs

The Gold layer represents a shift from data preparation to analysis-driven data products. While Gold tables retain structural similarities with Silver, particularly regarding mobility data, they are explicitly designed to support specific analytical objectives rather than general exploration.

Only selected transformations are materialized in Gold, primarily involving aggregations and derived metrics such as total and average trip counts by time period. These precomputed metrics significantly reduce query execution times and avoid repetitive calculations. Importantly, these transformations are not suitable for Silver, as they would either introduce redundancy or remove the granularity required for other analyses.

Separate Gold tables are maintained for municipalities, and districts. This design choice reflects the fact that the analytical questions are defined at distinct spatial levels and enables more efficient queries by reducing the volume of data scanned. Gold is therefore tightly coupled to the business questions it aims to answer and is not intended to be a generic analytical layer.

Although Gold tables are optimized for automated reporting and downstream analysis, they are not self-contained. Contextual information such as geographic attributes or calendar details still resides in Silver, reinforcing the complementary nature of the two layers. The benefits of Gold are evident in performance improvements, with some queries being reduced from execution times of 15 minutes to 2.

2.5 Orchestration and Metadata Management

Workflow orchestration plays a central role in the implemented infrastructure. As data volume increased, executing monolithic ingestion scripts became impractical due to frequent failures and partial executions that were difficult to diagnose. Apache Airflow was introduced to address these challenges by decomposing the pipeline into fine-grained, manageable tasks.

In the ingestion process, each input file is treated as an independent task. This design prevents cascading failures and eliminates large-scale rollbacks, as errors are isolated to specific files. Airflow also enables long-running ingestion processes to execute unattended, allowing other tasks to proceed in parallel and failed tasks to be selectively re-run.

The DAG structure enforces clear dependencies between stages, ensuring that downstream processes only execute once prerequisite steps have completed successfully. Parameterization is a key feature of the analytical DAGs, allowing users to specify date ranges, spatial resolutions, and geographic polygons at execution time. This flexibility is essential for adapting the analysis to different scenarios and aligns with the project requirements.

Airflow further provides built-in traceability through execution logs, task states, and historical records. These features facilitate debugging, reproducibility, and version tracking of the pipeline. Overall, orchestration is not merely a convenience but a foundational component that enables the lakehouse architecture to scale, remain reliable, and support complex analytical workflows.

3 Data Methodology Overview

This chapter describes the methodological principles guiding data usage, preprocessing, integration, and analysis throughout the project. Beyond the technical infrastructure, particular attention is paid to the rationale behind data-related decisions, analytical assumptions, and inherent limitations derived from the nature of the available datasets.

3.1 Data Sources

Multiple public data sources were combined to support the analysis. The primary source of mobility data is the Spanish Ministry of Transport (MITMA), from which detailed daily trip records for the entire year 2023 were obtained. These datasets provide mobility flows at different spatial resolutions, including municipalities, districts, and GAUs.

Socioeconomic variables, such as population and income at the municipal level, were sourced from the Spanish National Statistics Institute (INE). A correspondence table linking MITMA identifiers with INE identifiers was also obtained from MITMA, enabling cross-source integration. Additionally, a calendar dataset from the Community of Madrid was used to distinguish between working days, weekends, and holidays. Geographic information, including polygon geometries, was primarily sourced from MITMA, while provincial boundaries were obtained from a public GitHub repository.

All sources are considered reliable. MITMA, INE, and the Community of Madrid provide official data, while the GitHub repository was evaluated based on its structure and consistency before being incorporated. Although the INE datasets were generally cleaner and easier to process, the MITMA mobility data required more extensive handling due to inconsistencies and missing values.

The datasets cover the entire year 2023 and span the whole Spanish territory, allowing for spatial analyses at multiple administrative levels.

3.2 Ingestion and Preprocessing Strategy

The adopted strategy follows a clear principle: ingest first, clean later. Raw data files are ingested exactly as obtained and stored without modification, ensuring that original datasets remain available as a reference point. Initial validation was conducted using a reduced subset of data (ten days) to verify the correctness of the pipeline before scaling ingestion to the full year.

Prior to the Silver layer, preprocessing is limited to format conversion. For example, CSV files are transformed into queryable tables, but no cleaning or semantic transformations are applied. This approach preserves data fidelity while enabling reproducibility.

Completeness of ingestion is verified through exploratory queries, such as checking date coverage and comparing row counts against official documentation. Airflow plays a key role in this process, as each input file is handled as an independent task. Missing days in the original MITMA datasets (e.g., specific dates in late October and early November) were detected through orchestration logs rather than manual inspection, highlighting the importance of automated workflow monitoring.

3.3 Data Cleaning and Integration

Data cleaning and integration are primarily performed in the Silver layer. The most common issues addressed include null values, duplicated records, inconsistent formats, and incompatible identifier schemes across sources. Whenever possible, data is transformed into standardized formats; for instance, numerical variables are converted to consistent units and data types.

Rows containing irrecoverable inconsistencies, such as textual content where numerical values are expected, are removed entirely. Similarly, rows with missing critical attributes are discarded to avoid downstream analytical errors. The proportion of removed records is negligible relative to the overall dataset size, and therefore not considered to introduce significant bias.

An exception to this strict filtering policy occurs when aggregating municipalities into provinces. In this case, rows lacking a province assignment are retained, as removing them would result in substantial information loss for certain geographic areas.

Integration across sources is achieved through a unified identifier strategy. Although multiple identifier systems were present, INE identifiers were selected as the reference, following guidance provided during the course. This decision was critical to ensuring consistent joins across mobility, demographic, and geographic datasets.

3.4 Analytical Assumptions

Several analytical assumptions underpin the conducted analyses. First, mobility records provided by MITMA are assumed to be representative of real population movements, as they reflect actual observed displacements rather than simulated or survey-based estimates. Consequently, each recorded trip is treated as a real movement event.

Trips are assumed to have equal analytical weight at the raw level. Differences in population size and urban scale are not addressed directly at this stage but are explicitly accounted for in the formulation of the Business Questions and subsequent analyses.

Aggregating trips by hour or day is assumed to preserve the essential temporal patterns of mobility. While aggregation may obscure fine-grained dynamics, the original disaggregated data remains available in Silver to support alternative analyses if required.

Finally, the year 2023 is treated as a stable and representative period, without exceptional disruptions comparable to those observed during the COVID-19 pandemic. As such, no special temporal corrections are applied.

3.5 Limitations and Potential Biases

Several limitations arise directly from the data. Some dates are missing entirely from the original mobility datasets, and certain cells contain incomplete information. These gaps cannot be recovered and are therefore accepted as inherent constraints.

Potential biases related to data cleaning decisions are considered minimal. The removal of rows with null values affects only a negligible fraction of the data and is unlikely to alter aggregate patterns. Holidays and special days are explicitly identified through the calendar dataset and treated separately in the analyses to avoid misleading comparisons.

Spatial representation varies across zones, with some areas containing fewer recorded trips than others. These differences are acknowledged and addressed through aggregation strategies in the Gold layer. While some zones and specific days exhibit distinct mobility patterns, these variations are central to the analytical objectives rather than sources of unintended bias.

3.6 Methodological Rationale

Direct analysis of raw CSV files was deemed impractical due to structural inconsistencies, missing values, incompatible identifiers, and lack of normalization. Similarly, omitting the separation between Bronze, Silver, and Gold layers would compromise reproducibility, performance, and analytical clarity.

The adopted methodology enables clean, reliable, and well-structured data to be produced and reused efficiently. Silver provides a stable analytical foundation, while Gold significantly reduces computation time by materializing frequently used aggregations. This layered approach allows for rapid reconfiguration of analyses across different date ranges and spatial resolutions.

The methodology is tightly coupled with the underlying architecture. While some Business Questions could technically be answered using only Silver, execution times would exceed acceptable limits. The Gold layer is therefore essential to meeting performance requirements and enabling timely analytical outputs.

4 Analytical Framework and Business Questions

This chapter presents the analytical framework adopted in the project and describes how it is applied to address the defined business questions. The analysis is primarily descriptive in nature, focusing on characterizing observed mobility patterns. However, it also incorporates comparative and exploratory elements, allowing the identification of structural differences across spatial units, temporal patterns, and types of days.

The analyses are conducted by jointly considering temporal and spatial dimensions. Mobility patterns are examined both across different geographic units (municipalities or districts) and across time (hours of the day, workdays versus non-working days). This joint perspective is essential, as mobility behavior cannot be meaningfully interpreted without accounting for both where and when trips occur. The analytical unit therefore varies across business questions, depending on the specific objective of each analysis.

4.1 Analytical Framework

The analytical approach combines descriptive statistics, comparative analysis, and pattern detection techniques. While no predictive modeling is performed, the framework enables the identification of regularities, imbalances, and distinctive mobility signatures across space and time.

Mobility is analyzed through aggregated trip counts, which are examined both temporally (hourly distributions within a day) and spatially (flows between geographic units). The framework assumes that mobility patterns differ substantially depending on the type of day (e.g., workday, weekend, or transitional days) and the functional role of a location (residential, industrial, or service-oriented).

The combination of spatial and temporal aggregation allows the extraction of meaningful patterns while maintaining interpretability. Aggregations are carefully designed to preserve the essential structure of mobility behavior, enabling efficient analysis without sacrificing analytical validity.

4.2 Business Question 1: Identification of Typical Day Patterns

Business Question 1 (BQ1): *Are there distinct types of days characterized by different temporal mobility patterns?*

The objective of this business question is to identify whether different categories of days exhibit systematically different mobility structures. Rather than predefining day types (e.g., weekdays or weekends), the analysis seeks to infer typical day profiles directly from the data.

The analysis uses the total number of trips aggregated by hour of the day, resulting in a 24-dimensional temporal profile for each day. Depending on the selected spatial resolution, the aggregation is performed at either the municipal or district level.

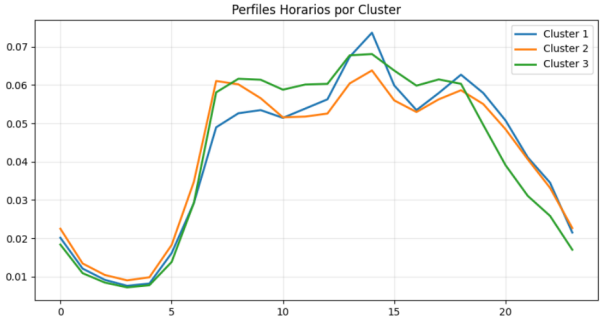
A K-means clustering algorithm is applied with a fixed number of clusters ($K = 3$). This choice is motivated by the hypothesis that mobility patterns can be broadly grouped into three categories: typical workdays, non-working days (such as weekends or holidays), and intermediate or transitional days (e.g., Saturdays or days preceding holidays). The clustering confirms this intuition, producing clearly differentiated temporal profiles.

The output of this analysis is a single PDF report containing a comparative visualization of the average hourly mobility profiles for each cluster. The resulting curves reveal distinct mobility signatures, both in terms of total volume and temporal distribution, supporting the existence of multiple representative day types.

BQ1: Reporte de Clustering de Movilidad

Resumen de Segmentacion: (29 días analizados)

- Cluster 1: 75 municipios | Media viajes/día: 159784
- Cluster 2: 66 municipios | Media viajes/día: 92189
- Cluster 3: 28 municipios | Media viajes/día: 141164



4.3 Business Question 2: Detection of Gravity-Based Connectivity Mismatches

Business Question 2 (BQ2): *Which pairs of municipalities exhibit weaker-than-expected connectivity given their demographic and socioeconomic characteristics?*

This business question aims to identify potential mismatches between observed mobility flows and expected connectivity, based on a gravity-model formulation. The analysis focuses on municipal-level interactions and evaluates whether certain origin–destination pairs are under-connected relative to their expected potential.

The gravitational potential between two municipalities i and j is defined as:

$$G_{ij} = \frac{P_i \cdot R_j}{d_{ij}^2},$$

where P_i is the population of the origin municipality, R_j is the average income of the destination municipality, and d_{ij} represents the average travel distance between them.

Observed mobility flows are normalized through a global scaling factor k, yielding a classical mismatch indicator:

$$M_{ij} = \frac{T_{ij}}{k \cdot G_{ij}},$$

where T_{ij} denotes the observed number of trips.

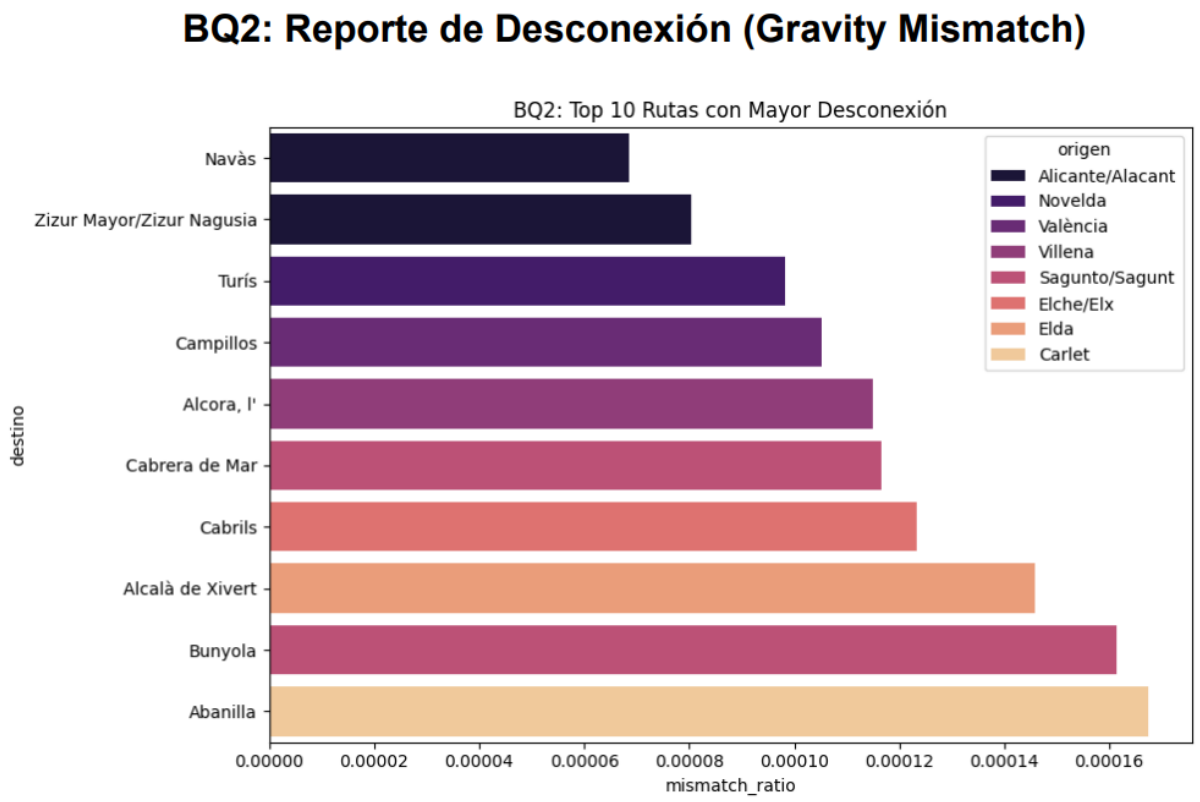
To emphasize routes involving highly populated areas, the mismatch is further weighted by the square root of the product of the populations of the two municipalities:

$$WM_{ij} = M_{ij} \cdot \sqrt{P_i \cdot P_j}.$$

This weighted formulation assigns greater importance to under-connected routes affecting large populations, reflecting their higher potential socioeconomic impact.

Scope limitation: This analysis is intentionally restricted to the municipal level. Although mobility data are available at district resolution, the demographic and income variables required by the gravity model are defined at the municipal scale. Since no reliable disaggregation exists for these variables at the district level, extending the analysis would lack conceptual validity. Consequently, when the analysis is performed at district resolution, Business Question 2 is not evaluated.

The outputs of this analysis include a ranked list of critical routes, a spatial visualization highlighting the most significant mismatches, and a PDF report summarizing the results.



4.4 Business Question 3: Identification of Industrial Activity through Mobility Patterns

Business Question 3 (BQ3): *Can municipalities or districts with strong industrial activity be identified through mobility patterns?*

The goal of this business question is to assess whether mobility data can reveal locations with a strong industrial or activity-based character. The underlying hypothesis is that such areas exhibit significantly higher mobility volumes on workdays compared to weekends.

For each spatial unit, average daily mobility volumes are computed separately for workdays and non-working days. The ratio between these two quantities is then calculated, capturing the relative intensity of workday-related mobility.

Rather than classifying all locations, the analysis focuses on identifying candidates with the strongest industrial signature. A percentile-based approach is adopted, selecting the top 5% of locations with the highest workday-to-weekend ratios. This choice balances interpretability and analytical focus, avoiding both arbitrary thresholds and excessive result sets.

The analysis produces ranked lists, temporal profiles, and spatial visualizations. Results show that industrial and activity-driven areas, including zones with industrial parks or major universities, consistently emerge among the top-ranked locations. The distinction becomes particularly clear when analyzing data at district resolution, where localized activity hubs can be identified.

4.5 Operationalization of the Business Questions through the Gold DAG

The analytical framework is operationalized through a set of automated workflows implemented in Apache Airflow. Each business question is encapsulated within an independent TaskGroup, allowing analyses to be executed in parallel while preserving internal logical dependencies.

Workflows are fully parameterized by date range, spatial polygon, and spatial resolution. This design enables flexible, scenario-specific analyses without modifying the underlying code. Gold-layer tables, specifically designed to support the business questions, significantly reduce execution times and make automated report generation feasible.

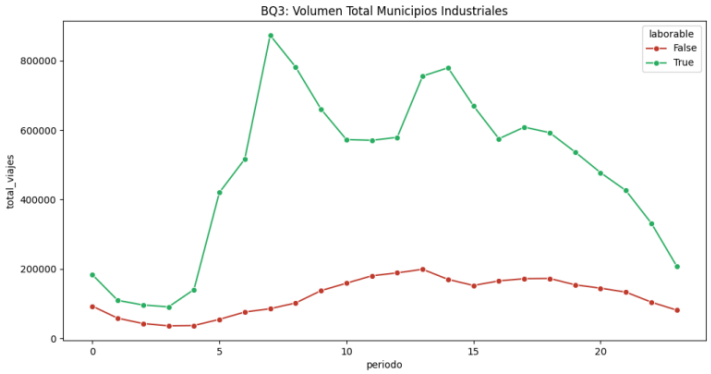
Each TaskGroup follows a sequential structure comprising data extraction, analytical computation, asset generation, and report creation. Task dependencies are explicitly encoded, ensuring that partial or inconsistent outputs cannot be produced. In case of failures, individual tasks can be re-executed without rerunning the entire pipeline.

Overall, the Gold DAG constitutes the final link between the lakehouse architecture and the analytical objectives of the project, translating structured data into reproducible and actionable insights through a robust orchestration layer.

BQ3: Reporte Industrial Consolidado

Top 5% Municipios:

- Almussafes: Ratio 1.97
- Cheste: Ratio 1.61
- Riba-roja de Túria agregacion de municipios: Ratio 1.56
- Sollana: Ratio 1.51
- Jijona/Xixona: Ratio 1.43
- Quart de Poblet: Ratio 1.42
- Betxí: Ratio 1.41
- Silla: Ratio 1.41
- Banyeres de Mariola: Ratio 1.41



5 Discussion: Enabling Expert Insights through a Lakehouse Architecture

This project goes beyond the execution of isolated analyses by demonstrating how a well-designed lakehouse architecture enables expert-level insights that would be difficult, inefficient, or conceptually fragile under a less structured setup. The value of the work lies not only in the results obtained, but in the analytical flexibility and interpretability made possible by the underlying architecture.

A key contribution of the lakehouse design is its ability to support exploratory and iterative analysis. Rather than answering a single predefined question, the system allows analysts to reformulate hypotheses, adjust parameters, and explore alternative perspectives without requiring structural changes to the data pipeline. This flexibility is essential for expert analysis, where insights often emerge through comparison, refinement, and iteration rather than through a single execution.

The separation into Bronze, Silver, and Gold layers plays a fundamental role in enabling this process. The Bronze layer preserves the original data as an immutable reference, ensuring traceability and auditability. The Silver layer provides a clean, normalized, and integrated representation of the data, allowing analysts to explore relationships and validate assumptions with confidence. Finally, the Gold layer translates these structured datasets into purpose-built tables designed to directly support analytical questions. This separation allows each layer to serve a distinct analytical purpose while maintaining coherence across the system.

An important outcome of this design is the ability to recognize when certain analyses are conceptually valid and when they are not. For example, while mobility data are available at both municipal and district levels, the gravity-based mismatch analysis relies on demographic and socioeconomic variables defined at the municipal scale. The architecture makes this distinction explicit, allowing the analysis to be correctly restricted to the appropriate spatial level. Rather than forcing uniformity, the system supports informed analytical boundaries, which is a hallmark of expert reasoning.

The orchestration layer implemented with Apache Airflow further strengthens the analytical value of the system. By encoding dependencies, parameterization, and execution logic directly into DAGs, analyses become reproducible, auditable, and scalable. Insights are no longer tied to manual execution or individual experimentation, but are instead generated through controlled and repeatable workflows. This shift from ad-hoc analysis to operationalized analytics is critical for transforming data exploration into reliable knowledge generation.

Moreover, the architecture enables the discovery of non-obvious patterns that were not explicitly anticipated at the outset of the project. Examples include the identification of transitional day types in mobility patterns, the emergence of localized industrial activity at district level, and the differentiation between spatial resolutions in terms of analytical interpretability. These findings arise not from predefined assumptions, but from the system's capacity to support multi-dimensional exploration.

From an expert perspective, one of the most significant insights is that increased data granularity does not automatically translate into better analysis. The project illustrates how meaningful results depend on the alignment between data resolution, available contextual variables, and the analytical question being addressed. A robust architecture not only enables deeper analysis, but also helps prevent misleading conclusions by enforcing methodological coherence.

In summary, the lakehouse architecture developed in this project serves as an enabler of expert insights by combining scalability, flexibility, and methodological rigor. It allows complex mobility data to be explored from multiple angles, supports informed analytical decision-making, and bridges the gap between raw data ingestion and actionable understanding. The resulting system closely resembles real-world analytical platforms used in professional data engineering and decision-support environments, highlighting the practical relevance of the approach.

6 Main Limitations

Despite the robustness of the final architecture and analytical workflows, the project faced several important limitations throughout its development. These limitations were not merely technical obstacles, but learning points that significantly influenced both the design decisions and the final structure of the system. This section describes the main challenges encountered, the solutions adopted, and the lessons learned for future projects of similar scale.

6.1 Limitation 1: Rollback Issues and Large-Scale Data Ingestion

Problem description. One of the most critical challenges encountered during the project was the management of large-scale data ingestion. Initial ingestion attempts were performed using coarse-grained scripts that processed multiple files in a single execution. As data volume increased, this approach led to frequent rollback issues: partial failures caused entire ingestion processes to restart, while already uploaded data remained in the storage layer. This resulted in duplicated data, uncontrolled growth of storage usage, and a lack of clarity regarding which files had been successfully ingested.

Adopted solution. The issue was addressed by redesigning the ingestion workflow using Apache Airflow. Data ingestion was decomposed into fine-grained, file-level tasks, where each task processed a single file and could be executed independently. This ensured idempotent behavior: failures affected only the specific file involved, without compromising the integrity of the rest of the pipeline.

Lessons learned. The main lesson from this limitation is the importance of designing ingestion pipelines that are incremental, idempotent, and resilient from the outset. When working with large datasets, dividing processes into small, independently executable tasks is essential to avoid cascading failures and to maintain control over data consistency.

6.2 Limitation 2: Incomplete Integration of GAU-Level Data

Problem description. Although GAU-level mobility data were fully ingested into the Bronze layer, their integration into the Silver layer proved problematic. During the Silver transformation phase, GAU files exhibited abnormally long processing times, with ingestion becoming progressively slower as more files were processed. While no explicit errors were raised, performance degradation made the process infeasible beyond a certain point in time.

Adopted solution. Given that GAU-level analysis was not strictly required to answer the core business questions, a pragmatic decision was made to exclude GAU data from the Silver and Gold layers. GAU files remain stored in Bronze for potential future investigation, but they are not part of the analytical pipeline.

Lessons learned. This limitation highlights the importance of understanding both data structure and performance implications early in the pipeline design. When computational constraints arise, prioritizing analytical relevance over completeness can be a necessary and valid decision. Future work could focus on investigating optimized ingestion strategies or alternative representations for GAU data.

6.3 Limitation 3: Exhaustion of Free-Tier Metadata Storage Resources

Problem description. The project relied on a PostgreSQL instance hosted on Neon for metadata management. Extensive testing, re-ingestion, and schema restructuring led to the exhaustion of the free-tier usage limits. This issue emerged at a late stage of the project, during final validation runs, posing a potential risk to timely completion.

Adopted solution. Given the advanced state of the project and the limited remaining workload, the decision was made to temporarily upgrade the Neon plan rather than migrating the metadata to a new project. This ensured continuity and avoided introducing additional risks at a critical stage.

Lessons learned. This experience underscores the need to manage resource-intensive testing carefully when using free-tier services. In future projects, large-scale testing should initially be performed on restricted datasets, and major structural changes should preferably be executed in isolated environments to avoid unnecessary resource consumption.

6.4 Limitation 4: Late Inclusion of Spatial Resolution as an Analytical Parameter

Problem description. Initially, the Silver layer was designed without explicitly preserving the original spatial resolution of mobility records (municipalities, districts, or GAUs). When the requirement to parameterize analyses by spatial resolution emerged, it became evident that this information had been lost during transformation.

Adopted solution. To address this issue, the Silver mobility ingestion process was redesigned to include an explicit attribute indicating the original spatial resolution of each record. This required re-ingesting the complete mobility dataset into Silver.

Lessons learned. This limitation illustrates the importance of preserving all potentially relevant metadata during the Silver transformation phase. Even attributes that may not seem immediately useful can become critical for downstream analytical flexibility. Careful upfront design of Silver schemas is essential to avoid costly reprocessing.

6.5 Limitation 5: Inapplicability of Gravity-Based Analysis at District Level

Problem description. Business Question 2 relies on demographic and socioeconomic variables, specifically population and income, which are only available at the municipal level. While mobility data exist at district resolution, no reliable disaggregation of these variables is available.

Adopted solution. Rather than forcing an ill-defined analysis, the decision was made to restrict Business Question 2 to municipal-level analysis only. When district-level analysis is selected, this business question is intentionally not evaluated.

Lessons learned. This limitation highlights the importance of conceptual validity over technical feasibility. Not all analyses can or should be applied uniformly across spatial scales. Clear methodological boundaries are essential to ensure meaningful and defensible results.

6.6 Limitation 6: Infrastructure Deployment and Hardware Resource Constraints

Problem description. An attempt was made to migrate the analytical pipeline to AWS EC2 instances to leverage cloud scalability. However, this transition encountered two primary categories of errors. First, significant friction arose regarding credential management; there was a lack of clarity concerning the specific environment variable names expected by the `duck_runner` image, resulting in failures such as the `CONTR_POSTGRES` connection error. Second, the project faced recurrent Out-of-Memory (OOM) failures driven by the sheer scale of the mobility datasets. This was exacerbated by the restrictions of the AWS Free Tier, which limited the deployment to instances with minimal resources leading to errors like `KeyError: 'memory'` when the environment failed to provide the necessary allocation parameters.

Adopted solution. To overcome these bottlenecks and ensure the timely completion of the project, analytical efforts were refocused on local execution. The pipeline was deployed on a high-performance local machine equipped with 32GB of DDR5 RAM. This hardware configuration, currently considered a "gold standard" for intensive data processing, provided the stability and memory overhead required to handle the datasets without the constraints and connectivity issues experienced in the cloud environment.

Lessons learned. Managing large-scale data ingestion and transformation remains a significant technical hurdle that requires precise alignment between software requirements and hardware capabilities. While the attempt provided valuable practical knowledge on the use of Elastic Compute in the cloud with EC2 instances, it also demonstrated that high-complexity projects are often not the best environment for initial cloud experimentation. Applying these cloud learnings to a smaller-scale project would have allowed for a more structured mastery of environment variable injection and instance configuration without the persistent frustration of memory-related crashes. This experience highlights the importance of evaluating resource tiers early in the design phase to match the infrastructure to the data scale.

6.7 Summary

Overall, the limitations encountered throughout the project played a crucial role in shaping the final architecture and methodology. Addressing these challenges required iterative refinement, pragmatic decision-making, and a strong emphasis on robustness and reproducibility. The lessons learned provide valuable guidance for future large-scale data engineering and analytical projects.

7 Conclusions

This project has presented the design and implementation of a complete lakehouse-based analytical platform for large-scale mobility data. From raw data ingestion to automated report generation, the work demonstrates how architectural decisions, data methodology, and orchestration strategies can be combined to produce reliable and reproducible analytical outcomes.

From an analytical perspective, the project has delivered several relevant insights. Distinct types of days with clearly differentiated mobility patterns were identified without imposing predefined categories. Connectivity mismatches between municipalities were detected using a gravity-based formulation that incorporates demographic and socioeconomic factors, highlighting routes with potentially high societal impact. Additionally, mobility patterns proved effective in identifying areas with strong industrial or activity-driven characteristics, particularly when analyzed at finer spatial resolutions.

Beyond the specific results, a key contribution of the work lies in the architecture itself. The separation between Bronze, Silver, and Gold layers enabled scalability, traceability, and analytical flexibility. Automated workflows implemented through Apache Airflow transformed complex analyses into reproducible processes, allowing insights to be generated consistently across different scenarios. The project illustrates that well-designed data infrastructure is a prerequisite for expert-level analysis, not merely a technical convenience.

From a learning perspective, this project represented a huge challenge and a significant opportunity for growth. Most (if not all) of the concepts applied throughout the project were new and had to be understood, tested, and refined progressively as the system was being built. The development process required continuous problem-solving across multiple layers of the architecture, from data ingestion and storage to metadata management, orchestration, and analysis. This iterative exposure to real technical constraints, unexpected failures, and performance limitations provided a depth of practical understanding that would not have been achievable through isolated exercises or purely theoretical work.

A key outcome of the project is the acquisition of a holistic view of data systems. Rather than treating data engineering, analytics, and visualization as independent components, the project highlighted the importance of coherence between architectural design, methodological rigor, and analytical objectives. This end-to-end perspective constitutes one of the most valuable learning results of the work.

Several limitations were encountered throughout the development, including rollback issues during ingestion, performance constraints with certain data resolutions, and the inapplicability of some analyses at specific spatial levels. Rather than weakening the project, these challenges shaped its final design and informed key methodological decisions. Addressing these limitations required iterative refinement and reinforced the importance of robustness, modularity, and conceptual validity.

Future work could extend the platform by incorporating additional data sources, exploring alternative aggregation strategies, or enabling new types of analyses. In particular, further investigation into GAU-level data integration and the inclusion of finer-grained socioeconomic variables could expand the scope of the system. Nevertheless, the current platform already provides a solid foundation for scalable mobility analysis and closely resembles real-world analytical infrastructures used in professional data engineering and decision-support contexts.