UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CAMPUS D'ALCOI

TECHNICAL REFERENCE & DATA DICTIONARY

Lakehouse Architecture Map

Authors:

María López Hernández
Joan Sánchez Verdú
Fernando Blanco Membrives

BIG DATA ENGINEERING AND TECHNOLOGIES

January 16, 2026

Abstract

This document serves as the primary technical reference for the implementation of the Mobility Lakehouse. It provides a comprehensive specification of the data architecture, adhering to the "Medallion" design pattern (Bronze, Silver, and Gold). The document details the schema definitions for each layer, the data lineage from source systems (MITMA and INE), and the transformation logic required to convert raw ingestion files into analytical data marts used for infrastructure gap analysis and functional zoning.

Contents

1 Data Layering Strategy

The platform adheres to the "Medallion" Lakehouse architecture (Bronze → Silver → Gold). This document serves as the technical reference for the data models, detailing schema definitions, data types (DuckDB dialect), and transformation logic.

1.1 Bronze Layer (Raw Ingestion)

The Bronze layer functions as an immutable staging area. Data is ingested from external sources (MITMA, INE, Madrid Open Data) in its native format. To mitigate pipeline failures arising from schema drift or unexpected data types, strict typing is deferred to the Silver layer. Consequently, attributes in this layer are cast as `VARCHAR`, preserving the original formatting (e.g., European decimal commas, varying date formats).

Common Audit Columns

To ensure auditability and data lineage, every table in this layer includes the following system columns:

- `ingestion_timestamp` (`TIMESTAMP`): The UTC timestamp recording the insertion time.
- `source_url` (`VARCHAR`): The specific file URI or API endpoint origin.

1.1.1 Mobility Data (High Volume)

Table: `bronze_mobility_data`

Description: Stores the raw daily origin-destination matrices. This table is physically partitioned by the `fecha` column for query optimization.

Column	Type	Description
`fecha`	`VARCHAR`	Date of the trip (Format: YYYYMMDD).
`periodo`	`VARCHAR`	Hourly interval (00-23).
`origen`	`VARCHAR`	Source MITMA Zone ID.
`destino`	`VARCHAR`	Destination MITMA Zone ID.
`distancia`	`VARCHAR`	Distance category (e.g., "005-010").
`actividad_origen`	`VARCHAR`	Imputed purpose at origin (e.g., "casa").
`actividad_destino`	`VARCHAR`	Imputed purpose at destination.
`estudio_origen_posible`	`VARCHAR`	Auxiliary study flag.
`estudio_destino_posible`	`VARCHAR`	Auxiliary study flag.
`residencia`	`VARCHAR`	Zone ID of the traveler's residence.
`renta`	`VARCHAR`	Income decile of the traveler.
`edad`	`VARCHAR`	Age group interval.
`sexo`	`VARCHAR`	Gender (1/2).
`viajes`	`VARCHAR`	Expansion factor (String w/ comma decimals).
`viajes_km`	`VARCHAR`	Total kilometers traveled.

1.1.2 Spatial & Reference Data

Table: `bronze_geo_municipalities`

Description: Ingested from ESRI Shapefiles. This is the only table containing a native spatial type upon ingestion.

Column	Type	Description
`ID`	`VARCHAR`	The MITMA Zone Code.
`geom`	`GEOMETRY`	Binary geometry object (Polygon/MultiPolygon).

Table: `bronze_zoning_municipalities`

Description: Provides the mapping between numerical IDs and text names.

Column	Type	Description
`column0`	`VARCHAR`	Index column from raw CSV.
`ID`	`VARCHAR`	MITMA Zone Code.
`name`	`VARCHAR`	Official Municipality Name.
`filename`	`VARCHAR`	Name of the source file.

Table: `bronze_mapping_ine_mitma`

Description: A crosswalk table linking Transport Ministry codes (MITMA) with Statistics Institute codes (INE).

Column	Type	Description
`seccion_ine`	`VARCHAR`	Census Section ID.
`distrito_ine`	`VARCHAR`	Census District ID.
`municipio_ine`	`VARCHAR`	INE Municipality Code.
`distrito_mitma`	`VARCHAR`	Transport District ID.
`municipio_mitma`	`VARCHAR`	Transport Municipality ID.
`gau_mitma`	`VARCHAR`	Great Urban Area (GAU) code.
`filename`	`VARCHAR`	Name of the source file.

1.1.3 Socio-Economic & Temporal Data

Table: `bronze_population_municipalities`

Description: Raw population counts. Headers are absent in the source file.

Column	Type	Description
`column0`	`VARCHAR`	Zone Code.
`column1`	`VARCHAR`	Population count.
`filename`	`VARCHAR`	Name of the source file.

Table: `bronze_ine_rent_municipalities`

Description: Economic indicators sourced from INE.

Column	Type	Description
`Municipios`	`VARCHAR`	Composite string (Code + Name).
`Distritos`	`VARCHAR`	District ID (if applicable).
`Secciones`	`VARCHAR`	Section ID (if applicable).
`Indicadores...`	`VARCHAR`	The name of the metric (e.g., Average Rent).
`Periodo`	`VARCHAR`	Reference year.
`Total`	`VARCHAR`	Numeric value (contains dots for thousands).
`filename`	`VARCHAR`	Name of the source file.

Table: `bronze_work_calendars`

Description: Calendar data for holiday identification. Columns 5-8 represent empty fields often found in loosely structured CSVs.

Column	Type	Description
`Dia`	`VARCHAR`	Date string (DD/MM/YYYY).
`Dia_semana`	`VARCHAR`	Name of the weekday.
`laborable...`	`VARCHAR`	Workday status.
`Tipo_de_Festivo`	`VARCHAR`	Flag for National/Local holidays.
`Festividad`	`VARCHAR`	Name of the holiday.
`column5-8`	`VARCHAR`	Parsing artifacts from raw CSV.
`filename`	`VARCHAR`	Name of the source file.

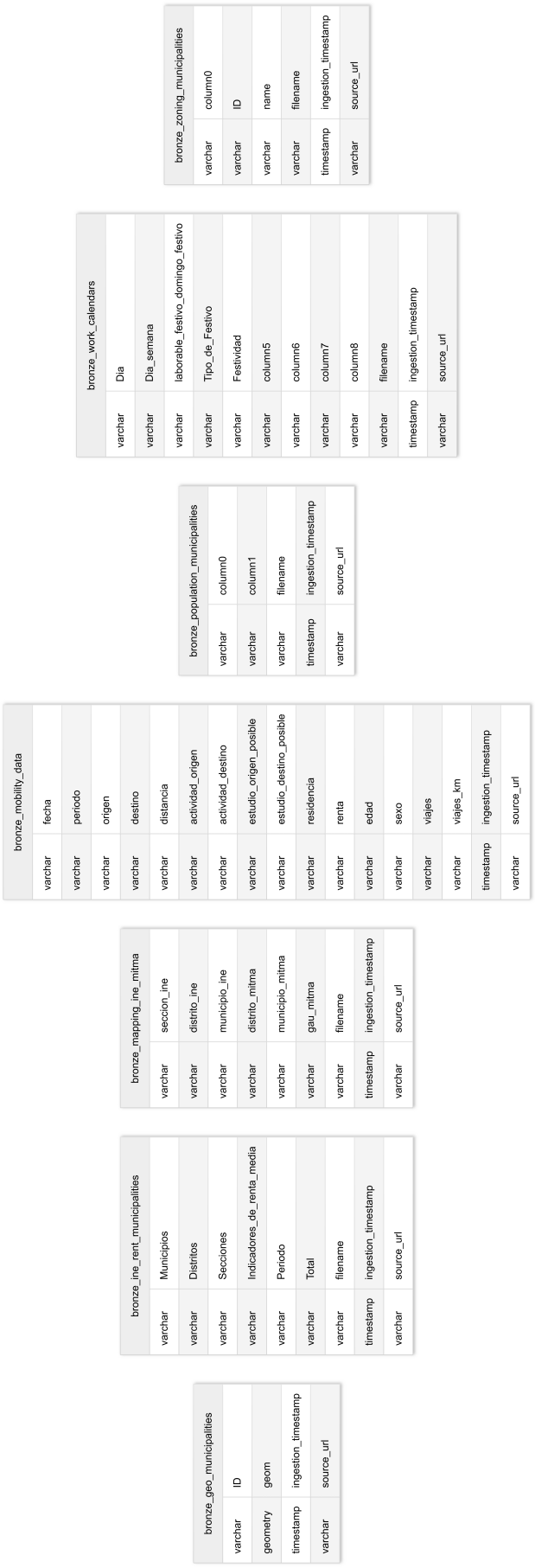

Figure 1: Entity Diagram of the Bronze Layer. One table for each file.

1.2 Silver Layer (Cleaned & Integrated)

The Silver layer implements a dimensional modeling approach (Star Schema). In this stage, data is cast to strict types, cleaned of formatting artifacts (e.g., normalization of numeric separators), and enriched with derived spatial calculations. To ensure lineage and version control, all tables include a `processed_at` (`TIMESTAMP`) column.

1.2.1 Dimension Tables (Context)

Table: `silver.dim_zones`

Purpose: The central spatial authority table for the Lakehouse.
Transformation: Integrates `geo`, `zoning`, and `mapping` sources. A Surrogate Key is generated to decouple the analytical model from source system identifiers.

Column	Type	Description
`zone_id`	`BIGINT`	**PK.** Surrogate Key generated via `ROW_NUMBER()`.
`mitma_code`	`VARCHAR`	Original Transport Ministry ID.
`ine_code`	`VARCHAR`	National Statistics ID (Cleaned).
`zone_name`	`VARCHAR`	Standardized human-readable name.
`polygon`	`GEOMETRY`	WKB Boundary for spatial joins.
`centroid`	`GEOMETRY`	Calculated center (`ST_Centroid`) for point-based distance logic.

Table: `silver.dim_zone_distances`

Purpose: Pre-computed Distance Matrix ($N \times N$) to accelerate gravity model calculations.
Transformation: Generated via a Cartesian product of `dim_zones`.

Column	Type	Description
`origin_zone_id`	`BIGINT`	**FK** reference to `dim_zones`.
`dest_zone_id`	`BIGINT`	**FK** reference to `dim_zones`.
`dist_km`	`DOUBLE`	Euclidean distance in km. **Logic:** `GREATEST(0.5, dist)` is applied to prevent division-by-zero errors in downstream models.

Table: `silver.dim_zone_holidays`

Purpose: Provides temporal context for mobility patterns (e.g., distinguishing workdays from holidays).
Transformation: Subsets `work_calendars` for "National Holidays" and broadcasts them to all zones via a Cross Join.

Column	Type	Description
`zone_id`	`BIGINT`	**FK** reference to `dim_zones`.
`holiday_date`	`DATE`	The specific date of the holiday event.

1.2.2 Metric Tables (Auxiliary Data)

Table: `silver.metric_population`

Grain: One row per Zone per Year.
Cleaning Logic: Applies Regex filter `[a-zA-Z]` to remove invalid header rows found in the raw CSV.

Column	Type	Description
`zone_id`	`BIGINT`	**FK** reference to `dim_zones`.
`population`	`BIGINT`	Official population count.
`year`	`INTEGER`	Reference year (e.g., 2023).

Table: `silver.metric_ine_rent`

Grain: One row per Zone per Year.
Cleaning Logic: Removes thousands separators (dots) and filters dataset to retain only "Net average income per person".

Column	Type	Description
`zone_id`	`BIGINT`	**FK** reference to `dim_zones`.
`income_per_cap`	`DOUBLE`	Normalized net income value.
`year`	`INTEGER`	Reference year.

1.2.3 Fact Table (Core Mobility)

Table: `silver.fact_mobility`

Purpose: The transactional heart of the Lakehouse.
Grain: One row per trip flow (Origin → Destination → Time Interval).
Transformation: Converts string dates/hours into Timezoned Timestamps. Invalid zones are excluded via Inner Joins to `dim_zones`.

Column	Type	Description
`period`	`TIMESTAMP`	**Logic:** `fecha` + `periodo` converted to "Europe/Madrid" time zone.
`partition_date`	`DATE`	Physical Partition Key derived from `fecha`.
`origin_id`	`BIGINT`	**FK** reference to `dim_zones`.
`dest_id`	`BIGINT`	**FK** reference to `dim_zones`.
`trips`	`DOUBLE`	Number of trips (Precision Double).

1.2.4 Data Observability

Table: `silver.data_quality_log`

Purpose: Metadata registry for monitoring pipeline health and data constraints.

Column	Type	Description
`check_timestamp`	`TIMESTAMP`	Audit execution time.
`table_name`	`VARCHAR`	Target entity of the quality check.
`metric_name`	`VARCHAR`	Type of check (e.g., "null_rate", "row_count").
`metric_value`	`DOUBLE`	The quantitative result of the check.
`notes`	`VARCHAR`	Contextual warnings or error messages.

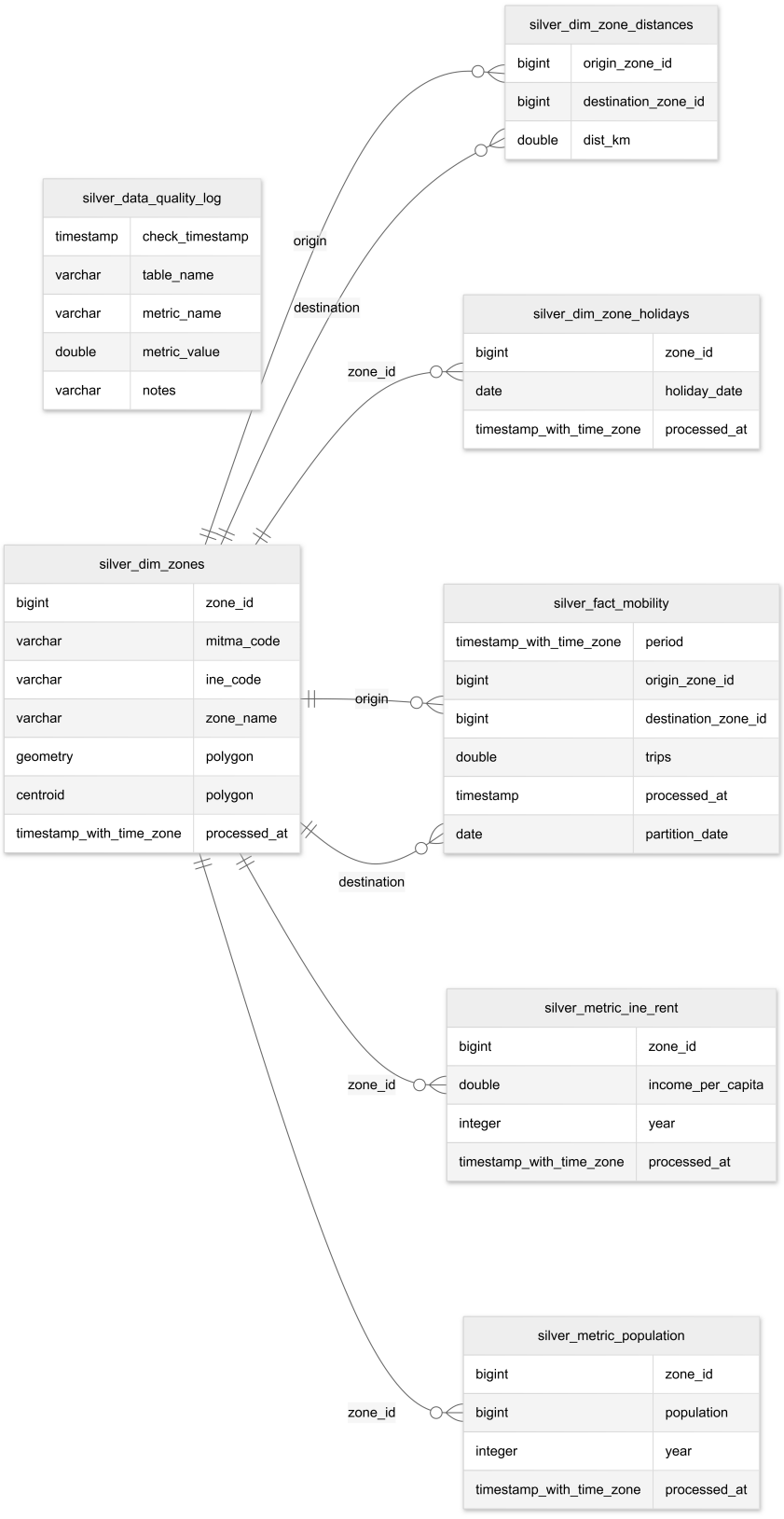

Figure 2: Entity-Relationship Diagram (ERD) of the Silver Layer. The schema follows a Star Schema design, centering on the `fact_mobility` table linked to spatial (`dim_zones`) and temporal dimensions.

1.3 Gold Layer (Data Mart)

The Gold layer serves as the final destination for analytics, specifically engineered to address distinct research questions. In contrast to the normalized Silver layer, these schemas are denormalized and pre-aggregated to optimize query performance for OLAP operations and visualization tools (e.g., Kepler.gl, Plotly).

Below is the technical definition of the three core analytical products materialized in this layer:

1.3.1 Temporal Analysis (Cluster Profiles)

Table: `gold.typical_day_patterns`

Purpose: Persists the centroids of mobility profiles identified via Unsupervised Learning (K-Means). This table aggregates high-volume transactional data into lightweight hourly curves representing standard behavioral archetypes (e.g., "Workday" vs. "Holiday").
Grain: One row per Cluster per Hour.

Column	Type	Description
`cluster_id`	`INTEGER`	The label assigned by the K-Means algorithm (e.g., 0, 1, 2).
`hour`	`INTEGER`	Temporal interval (0-23).
`avg_trips`	`DOUBLE`	The normalized average trip volume for this specific hour/cluster combination. Used for plotting demand curves.
`processed_at`	`TIMESTAMP`	Audit timestamp.

1.3.2 Infrastructure Gap Analysis (Gravity Model)

Table: `gold.infrastructure_gaps`

Purpose: The outcome of the Gravity Model implementation. This table synthesizes mobility flows with socio-economic dimensions to quantify the disparity between *theoretical potential* and *observed usage*.
Grain: One row per Origin-Destination pair.

Column	Type	Description
`org_zone_id`	`BIGINT`	Origin Zone FK.
`dest_zone_id`	`BIGINT`	Destination Zone FK.
`population`	`BIGINT`	Population at origin (Demand Driver).
`rent`	`DOUBLE`	Income per capita at destination (Attractor).
`total_trips`	`DOUBLE`	Observed mobility volume (from Silver).
`dist_km`	`DOUBLE`	Distance impedance between zones.
`mismatch`	`DOUBLE`	**KPI:** Calculated as $ActualTrips / PotentialScore$. A value $\ll 1.0$ suggests an infrastructure deficit.

1.3.3 Functional Classification (Network Roles)

Table: `gold.zone_functional_class`

Purpose: A semantic layer that profiles every municipality based on its role in the metropolitan network (e.g., Net Labor Importer vs. Exporter).
Grain: One row per Zone.

Column	Type	Description
`zone_id`	`BIGINT`	Primary Key.
`zone_name`	`VARCHAR`	Human-readable name for map visualization.
`internal`	`DOUBLE`	Count of intra-zonal trips.
`outflow`	`DOUBLE`	Count of trips leaving the zone.
`inflow`	`DOUBLE`	Count of trips entering the zone.
`net_flow`	`DOUBLE`	Metric: $(In - Out)/Total$. Positive values indicate "Activity Hubs".
`retention`	`DOUBLE`	Metric: $Internal/(Out + Internal)$. Indicates self-sufficiency.
`label`	`VARCHAR`	Categorical classification derived from decision tree logic (e.g., "Bedroom Community").

Figure 3: Gold Layer Schema. These tables represent the final analytical products, denormalized and enriched with derived metrics (e.g., mismatch ratios, functional labels) ready for visualization.