Lakehouse Architecture Map

Technical Reference & Data Dictionary

María López Hernández **Joan Sánchez Verdú** **Fernando Blanco Membrives**

Big Data Engineering and Technologies
January 9, 2026

Abstract

This document serves as the primary technical reference for the Mobility Lakehouse. It details the Medallion architecture, schema definitions for Bronze, Silver, and Gold layers, and the transformation logic applied to source data from MITMA and INE.

1 Data Layering Strategy

The platform follows a "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 acts as the immutable staging area. Data is ingested from external sources (MITMA, INE, Madrid Open Data) in its native format. To prevent pipeline failures caused by unexpected data types in the source files, strict typing is deferred to the Silver layer; consequently, almost all columns in this layer are typed as `VARCHAR`, preserving the original formatting (e.g., decimal commas, date formats).

In addition, to ensure auditability and lineage, every table in this layer includes two system columns:

- `ingestion_timestamp` (`TIMESTAMP`): The UTC timestamp of when the record was inserted.
- `source_url` (`VARCHAR`): The specific file URI or API endpoint origin.

Below is the complete dictionary of tables persisted in the Bronze Schema:

1. Mobility Data (High Volume)

`bronze_mobility_data`

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

- `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`): The expansion factor/number of trips (String with comma decimals).
- `viajes_km` (`VARCHAR`): Total kilometers traveled.

2. Spatial & Reference Data

`bronze_geo_municipalities`

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

- `ID` (`VARCHAR`): The MITMA Zone Code.
- `geom` (`GEOMETRY`): The binary geometry object (Polygon/MultiPolygon).

`bronze_zoning_municipalities`

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

- `column0` (`VARCHAR`): Index column from raw CSV.
- `ID` (`VARCHAR`): MITMA Zone Code.
- `name` (`VARCHAR`): Official Municipality Name.
- `filename` (`VARCHAR`): Name of the source file.

`bronze_mapping_ine_mitma`

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

- `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.

3. Socio-Economic & Temporal Data

`bronze_population_municipalities`

Description: Raw population counts. Note the lack of headers in the source file.

- `column0` (`VARCHAR`): Zone Code.
- `column1` (`VARCHAR`): Population count.
- `filename` (`VARCHAR`): Name of the source file.

`bronze_ine_rent_municipalities`

Description: Economic indicators from INE.

- `Municipios` (`VARCHAR`): Composite string (Code + Name).
- `Distritos` (`VARCHAR`): District ID (if applicable).
- `Secciones` (`VARCHAR`): Section ID (if applicable).
- `Indicadores_de_renta_media` (`VARCHAR`): The name of the metric.
- `Periodo` (`VARCHAR`): Reference year.
- `Total` (`VARCHAR`): The numeric value (contains dots for thousands).
- `filename` (`VARCHAR`): Name of the source file.

`bronze_work_calendars`

Description: Calendar data for holiday identification. Extra columns (column5-8) represent empty fields often found in loosely structured CSVs.

- `Dia` (`VARCHAR`): Date string (DD/MM/YYYY).
- `Dia_semana` (`VARCHAR`): Name of the weekday.
- `laborable_festivo_domingo_festivo` (`VARCHAR`): Workday status.
- `Tipo_de_Festivo` (`VARCHAR`): Flag for National/Local holidays.
- `Festividad` (`VARCHAR`): Name of the holiday.
- `column5`, `column6`, `column7`, `column8` (`VARCHAR`): Parsing artifacts from raw CSV.
- `filename` (`VARCHAR`): Name of the source file.

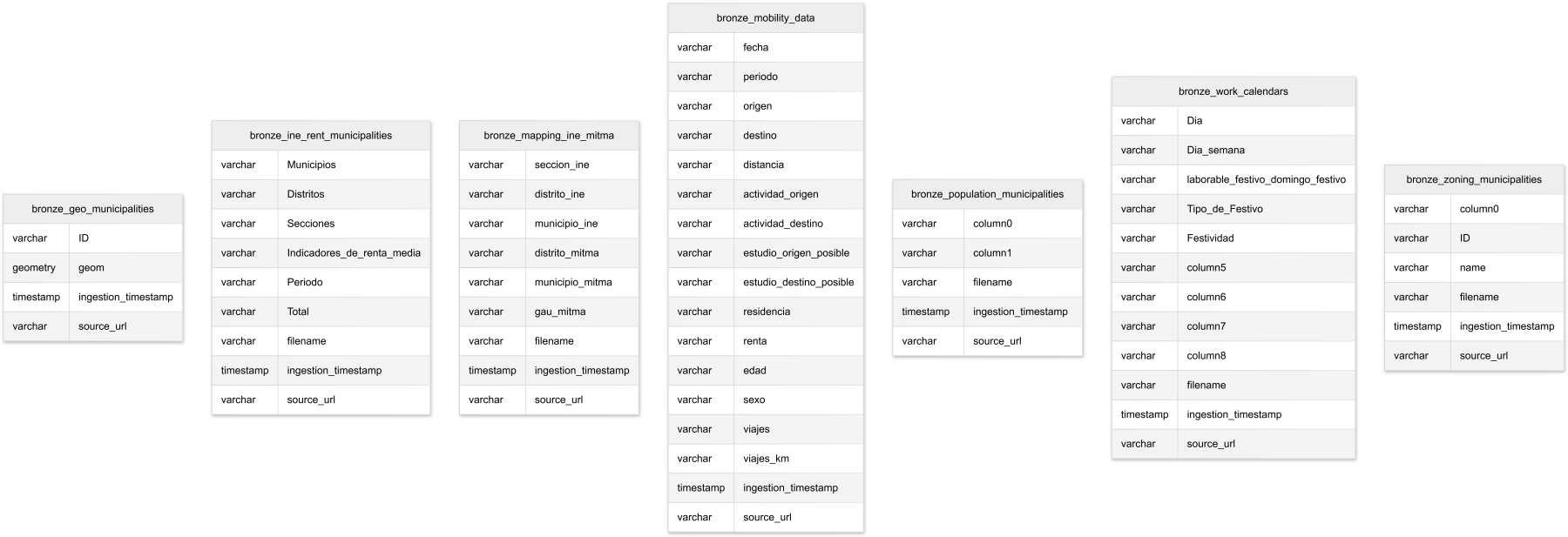

Figure 1: Entity Diagram of the Bronze Layer. One table for each file.

1.2 Silver Layer (Cleaned & Integrated)

The Silver layer implements a Star Schema design. In this stage, data is cast to strict types, cleaned of formatting artifacts (e.g., removing thousands separators), and enriched with spatial calculations. All tables include a `processed_at` (`TIMESTAMP`) column for versioning.

1. Dimension Tables (Context)

`silver.dim_zones`

Purpose: The central spatial authority table.
Transformation: Joins `geo`, `zoning`, and `mapping` tables. Generates a Surrogate Key.

- `zone_id` (`BIGINT`): Sequential integer generated via `ROW_NUMBER()`.
- `mitma_code` (`VARCHAR`): Original Transport ID.
- `ine_code` (`VARCHAR`): National Statistics ID (Cleaned).
- `zone_name` (`VARCHAR`): Human-readable name.
- `polygon` (`GEOMETRY`): Boundary for spatial joins.
- `centroid` (`GEOMETRY`): Calculated center (`ST_Centroid`) for distance logic.

`silver.dim_zone_distances`

Purpose: Pre-computed Distance Matrix ($N \times N$).
Transformation: Cartesian product of `dim_zones`.

- `origin_zone_id` (`BIGINT`): FK ref `dim_zones`.
- `destination_zone_id` (`BIGINT`): FK ref `dim_zones`.
- `dist_km` (`DOUBLE`): Euclidean distance in km. **Logic:** `GREATEST(0.5, dist)` ensures no zero-distances to prevent division errors in Gravity Models.

`silver.dim_zone_holidays`

Purpose: Temporal context for mobility patterns.
Transformation: Filters `work_calendars` for "National Holidays" in 2023 and explodes them to all zones via Cross Join.

- `zone_id` (`BIGINT`): FK ref `dim_zones`.
- `holiday_date` (`DATE`): The specific date of the holiday.

2. Metric Tables (Auxiliary Data)

`silver.metric_population`

Grain: One row per Zone per Year.

- `zone_id` (`BIGINT`): FK ref `dim_zones`.
- `population` (`BIGINT`): Cast from `column1`. **Cleaning:** Regex filter `[a-zA-Z]` removes header rows from raw CSV.
- `year` (`INTEGER`): Reference year (2023).

`silver.metric_ine_rent`

Grain: One row per Zone per Year.

- `zone_id` (`BIGINT`): FK ref `dim_zones`.
- `income_per_capita` (`DOUBLE`): Cast from `Total`. **Cleaning:** Removes dots (".") and filters only "Renta neta media por persona".
- `year` (`INTEGER`): Reference year.

3. Fact Table (Core Mobility)

`silver.fact_mobility`

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

- `period` (`TIMESTAMP WITH TIME ZONE`): **Logic:** `fecha` + `periodo` converted to "Europe/Madrid".
- `partition_date` (`DATE`): Partition Key derived from `fecha`.
- `origin_zone_id` (`BIGINT`): FK ref `dim_zones`.
- `destination_zone_id` (`BIGINT`): FK ref `dim_zones`.
- `trips` (`DOUBLE`): Number of trips cast to double precision.

4. Data Observability

`silver.data_quality_log`

Purpose: Metadata registry for pipeline health.

- `check_timestamp` (`TIMESTAMP`): Audit time.
- `table_name` (`VARCHAR`): Target of the check.
- `metric_name` (`VARCHAR`): e.g., "null_rate", "row_count".
- `metric_value` (`DOUBLE`): The result of the check.
- `notes` (`VARCHAR`): Context or warnings.

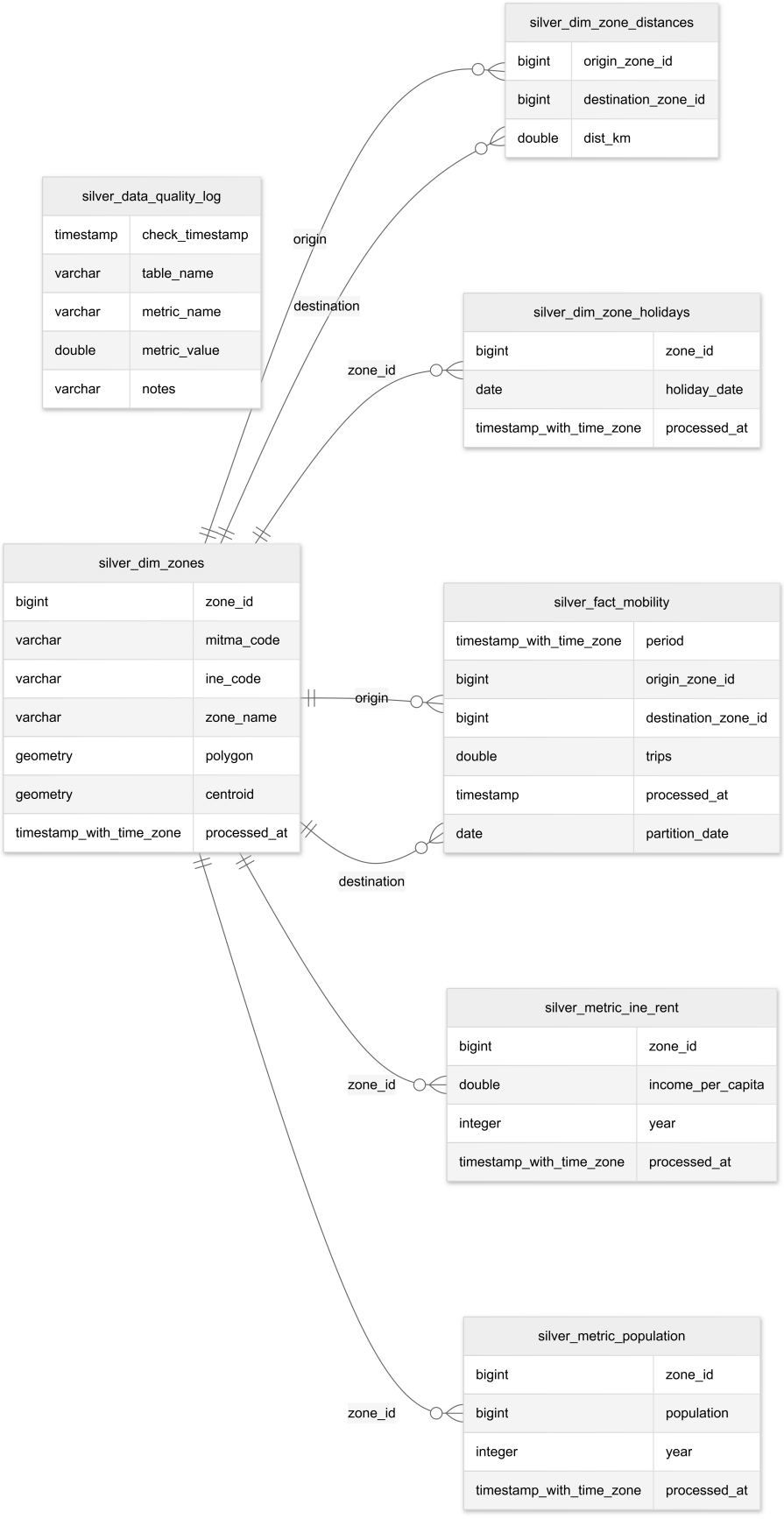

Figure 2: Entity-Relationship Diagram (ERD) of the Silver Layer. The schema follows a Star Schema design, centering on the `fact_mobility` table linked to spatial (`dim_zones`) and temporal dimensions.

1.3 Gold Layer (Data Mart)

The Gold layer is the final destination for analytics, designed to answer specific business questions. Unlike the normalized Silver layer, these tables are denormalized and pre-aggregated to optimize query performance for visualization tools (Kepler.gl, Plotly).

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

1. Temporal Analysis (Use Case 1)

`gold.typical_day_patterns`

Purpose: Stores the centroids of the mobility profiles identified via Unsupervised Learning (K-Means). It reduces billions of raw records into lightweight hourly curves representing standard behaviors (e.g., "Workday" vs "Holiday"). *Grain:* One row per Cluster per Hour.

- `cluster_id` (`INTEGER`): The label assigned by the K-Means algorithm (0, 1, or 2).
- `hour` (`INTEGER`): Hour of the day (0-23).
- `avg_trips` (`DOUBLE`): The average normalized volume of trips for this specific hour/cluster combination. Used for plotting demand curves.
- `processed_at` (`TIMESTAMP`): Audit timestamp.

2. Infrastructure Gap Analysis (Use Case 2)

`gold.infrastructure_gaps`

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

- `org_zone_id` (`BIGINT`): Origin Zone FK.
- `dest_zone_id` (`BIGINT`): Destination Zone FK.
- `total_population` (`BIGINT`): Population at origin (Driver of demand).
- `rent` (`DOUBLE`): Income per capita at destination (Attractor).
- `total_trips` (`DOUBLE`): The actual observed mobility volume from Silver.
- `dist_km` (`DOUBLE`): Distance impedance between zones.
- `mismatch_ratio` (`DOUBLE`): The key analytical KPI calculated as $ActualTrips/PotentialScore$. A value $\ll 1.0$ indicates an infrastructure deficit.

3. Functional Classification (Use Case 3)

`gold.zone_functional_classification`

Purpose: A semantic layer that profiles every municipality based on its role in the metropolitan network (Importer vs. Exporter of commuters). *Grain:* One row per Zone.

- `zone_id` (`BIGINT`): Id of the zone.
- `zone_name` (`VARCHAR`): Human-readable name for map tooltips.
- `internal_trips` (`DOUBLE`): Count of trips starting and ending in the same zone.
- `outflow` (`DOUBLE`): Count of trips leaving the zone.
- `inflow` (`DOUBLE`): Count of trips entering the zone.
- `net_flow_ratio` (`DOUBLE`): Calculated metric: $(In - Out)/Total$. Positive values indicate "Activity Hubs"; negative values indicate "Residential/Bedroom Communities".

- `retention_rate` (`DOUBLE`): Calculated metric: $Internal/(Out + Internal)$. Indicates self-sufficiency.
- `functional_label` (`VARCHAR`): The final categorical classification derived from the decision tree logic (e.g., "Self-Sustaining Cell", "Activity Hub").

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.