

TFM — Predicción de retrasos en Cercanías Madrid (RENFE)

Estructura de datos del pipeline de captura

Documento de referencia para el equipo · Generado a partir de capturas reales del 14/06/2026

1. Por qué este documento

Este documento describe la estructura real de los datos que ya estamos capturando 24/7 desde el 13/06/2026, tomando como referencia las capturas compactadas del día 14/06/2026 (día completo). El objetivo es que todo el equipo tenga una referencia común de qué columnas existen en cada fuente, qué significan, su nivel de completitud (% de valores nulos), y qué decisiones de modelado o limpieza se derivan de ello.

Las cuatro fuentes que se muestran corresponden a las tres capas del feed GTFS-Realtime de RENFE (trip_updates, vehicle_positions, alerts) y a la observación meteorológica horaria de AEMET. Todas están ya disponibles en formato Parquet en la capa 'processed' del servidor, particionadas por día.

2. trip_updates — Retrasos por tren y parada (VARIABLE OBJETIVO)

Esta es la fuente más importante: contiene el retraso (delay, en segundos) de cada tren respecto a la siguiente parada en su trayecto. Es la base de la variable a predecir.

48.411 filas · 16 columnas · 469 capturas (1 captura/minuto) el 14/06/2026

2.1 Diccionario de columnas

Columna	Tipo	% nulos	Descripción
fetched_at_utc	datetime	0%	Momento exacto en que nuestro colector descargó este snapshot.
feed_timestamp	epoch (str)	0%	Timestamp de cabecera GTFS-RT publicado por RENFE para este snapshot.
entity_id	str	0%	Identificador interno de la entidad dentro del feed (p.ej. TUUPDATE_...).
trip_id	str	0%	Identificador del trayecto/tren. Clave para cruzar con GTFS estático (trips.txt) y obtener la línea.
route_id	—	100%	NO viene poblado por el feed. La línea debe obtenerse cruzando trip_id con GTFS estático.
start_date / start_time	—	100%	Tampoco vienen poblados. Mismo caso que route_id.
trip_schedule_relationship	categorico	0%	Estado del trayecto: SCHEDULED (programado normal), ADDED (servicio añadido, p.ej. autobús de sustitución) o CANCELED.
stop_id	str	0.3%	Identificador de la parada a la que se refiere este registro. Cruzable con stops.txt del GTFS estático.
stop_sequence	—	100%	No poblado por el feed.
arrival_delay_s	Int64 (segundos)	11.9%	RETRASO EN LA LLEGADA A LA PRÓXIMA PARADA. Esta es la variable objetivo. Puede ser negativo (el tren va adelantado).
arrival_time	epoch (str)	10.9%	Hora estimada (real) de llegada a la próxima parada.

Columna	Tipo	% nulos	Descripción
departure_delay_s / departure_time	—	99-100%	Casi nunca poblados; el feed apenas informa de salidas, solo de llegadas.
stop_schedule_relationship	categorico	89.4%	Cuando viene poblado, el valor observado es SKIPPED (parada que el tren se va a saltar).

2.2 Valores observados en columnas categóricas (14/06)

Columna	Valor	Nº filas	% del total
trip_schedule_relationship	SCHEDULED	47.803	98,7%
trip_schedule_relationship	ADDED	457	0,9%
trip_schedule_relationship	CANCELED	151	0,3%
stop_schedule_relationship	SKIPPED	5.137	10,6% (resto: nulo)

2.3 Estadísticos de arrival_delay_s (la variable objetivo) — 14/06

Estadístico	Valor
Filas con dato (no nulo)	42.666 (88,1%)
Mínimo	-39 min (-2.520 s) — adelanto
Mediana (p50)	0 s — la mayoría de observaciones van a tiempo
p90	+6 min (360 s)
p99	+28 min (1.680 s)
Máximo	+4 h 36 min (16.560 s)

Comparado con el día anterior (13/06), no se observa el patrón de delays anómalos en torno a $\pm 24h$ que afectaba al 6,5% de las filas ese día. Pendiente de EDA: confirmar si ese patrón correlaciona con trip_schedule_relationship distinto de SCHEDULED o con estados anómalos en vehicle_positions.

2.4 Muestra de filas (mismo tren, snapshots consecutivos)

fetch_at_utc	trip_id	schedule_rel.	stop_id	arrival_delay_s	arrival_time
2026-06-14T02:14:31Z	6062D26578C3	SCHEDULED	13110	0	1781403590
2026-06-14T02:15:31Z	6062D26578C3	SCHEDULED	13110	60	1781403637
2026-06-14T02:16:31Z	6062D26578C3	SCHEDULED	13110	-60	1781403558
2026-06-14T02:17:32Z	6062D26578C3	SCHEDULED	13110	-60	1781403558

Nota: este tren (6062D26578C3) corresponde a una franja de Bilbao (lat ~43.2, ver vehicle_positions abajo), no Madrid — recuerda que el feed es nacional. El sufijo final del trip_id (C3) coincide con la línea indicada en vehicle_label (C3-26578), lo cual da una pista fuerte para extraer la línea por regex como aproximación rápida, en paralelo al cruce formal con trips.txt.

3. vehicle_positions — Posición y estado de cada tren

Contiene la posición GPS, velocidad y estado operativo de cada tren en movimiento. Es la fuente principal de features de ‘estado de la red’ (propagación de retrasos, congestión).

42.769 filas · 15 columnas · 469 capturas el 14/06/2026

3.1 Diccionario de columnas

Columna	Tipo	% nulos	Descripción
fetched_at_utc / feed_timestamp	datetime / epoch	0%	Igual que en trip_updates: momento de captura y timestamp de cabecera del feed.
entity_id	str	0%	Identificador de la entidad (p.ej. VP_C4-23665).
vehicle_id	str	0%	Número de vehículo/composición (p.ej. 23665).
vehicle_label	str	0%	Etiqueta legible: línea + nº vehículo + andén si aplica (p.ej. ‘C3-26578-PLATF.(2)’). Otra vía para extraer la línea.
trip_id	str	0%	Mismo identificador de trayecto que en trip_updates — clave de cruce entre ambos feeds.
route_id	—	100%	No poblado, igual que en trip_updates.
latitude / longitude	float	0.1%	Posición GPS del tren en el momento de la captura.
bearing / speed	—	100%	No poblados por el feed. No usar como features previstas; el feed real no los publica.
current_status	categorico	0%	Estado del tren respecto a la parada: STOPPED_AT, INCOMING_AT o IN_TRANSIT_TO. Feature potente de estado de red.
current_stop_sequence	—	100%	No poblado.
stop_id	str	0%	Parada de referencia para el current_status. Mismo espacio de IDs que en trip_updates.
vehicle_timestamp	epoch (str)	0%	Timestamp de la posición reportada por el vehículo (puede diferir ligeramente del feed_timestamp).

3.2 Valores de current_status (14/06)

Valor	Nº filas	% del total	Interpretación
STOPPED_AT	20.785	48,6%	Tren detenido en la parada indicada.
INCOMING_AT	11.507	26,9%	Tren aproximándose a la parada.
IN_TRANSIT_TO	10.477	24,5%	Tren en trayecto hacia la parada (entre estaciones).

3.3 Muestra de filas (mismo tren que en la sección 2.4)

fetched_at_utc	vehicle_label	trip_id	lat	lon	current_status	stop_id
2026-06-14T02:14:31Z	C3-26578	6062D26578C3	43.2101	-2.8888	STOPPED_AT	13111
2026-06-14T02:15:31Z	C3-26578	6062D26578C3	43.2056	-2.8897	IN_TRANSIT_TO	13111
2026-06-14T02:16:31Z	C3-26578-PLATF.(2)	6062D26578C3	43.1791	-2.9029	INCOMING_AT	13110
2026-06-14T02:17:32Z	C3-26578-PLATF.(2)	6062D26578C3	43.1791	-2.9029	INCOMING_AT	13110

El cruce trip_id entre trip_updates y vehicle_positions permite construir, para cada snapshot, la imagen completa de un tren: dónde está, hacia dónde va y cuánto retraso lleva en la próxima parada.

4. alerts — Incidencias en texto libre (entrada del NLP)

Contiene las alertas/avisos oficiales de RENFE: obras, servicios sustitutivos, incidencias, etc. Es la entrada principal del componente NLP del proyecto, ya que los campos estructurados (cause, effect) llegan vacíos en la práctica.

223.774 filas (explotadas por ruta afectada) · 55 alertas distintas (entity_id) · 411 capturas el 14/06/2026

4.1 Diccionario de columnas

Columna	Tipo	% nulos	Descripción
fetched_at_utc / feed_timestamp	datetime / epoch	0%	Momento de captura y timestamp de cabecera del feed.
entity_id	str	0%	Identificador único de la alerta (p.ej. AVISO_494107, AUTOBUS_4062D24001C2). 55 valores distintos el 14/06.
cause	—	100%	No poblado por el feed. NO USAR como feature estructurada.
effect	categorico	99,9%	Casi siempre vacío; cuando aparece, el único valor observado es MODIFIED_SERVICE (156 filas).
active_start	epoch (str)	0%	Inicio del periodo de validez de la alerta.
active_end	—	100%	No poblado: las alertas no indican cuándo terminan, hay que inferirlo por su desaparición del feed.
informed_route_id	str (cod. línea+núcleo)	2,9%	Línea(s) afectada(s), en formato propio de RENFE (p.ej. 10T0031C2). Una alerta puede afectar a varias rutas → cada una genera una fila.
informed_stop_id	—	97,2%	Casi siempre vacío; cuando aparece, identifica una estación concreta afectada.
informed_agency_id / header_text	—	100%	No poblados.
description_text	str (texto libre, ES)	0%	TEXTO COMPLETO DE LA ALERTA. Esta es la entrada real del NLP: aquí está toda la información (tipo de incidencia, líneas, estaciones, servicio alternativo).

4.2 Ejemplos de description_text (14/06)

entity_id	informed_route_id	description_text (fragmento)
AVISO_494107	10T0031C2	#MadC2 #MadC8 Por obras de mejora entre las estaciones de Alcalá de Henares y Guadalajara. Los trenes, en ambos sentidos, sufren demoras y detenciones...
AUTOBUS_4062D24001C2	(varias)	Se prestará servicio de autobús para este recorrido...
AVISO_442947	(varias)	SERVICIO POR AUTOBÚS C-3: Plan alternativo de...

Cada entity_id (alerta) suele afectar a varias rutas (informed_route_id), de ahí que 55 alertas distintas generen 223.774 filas explotadas. Para el NLP conviene trabajar a nivel de entity_id + description_text (deduplicando), y usar informed_route_id solo como feature auxiliar de cruce, no como unidad de análisis.

El prefijo del texto (#MadC2 #MadC8, SERVICIO POR AUTOBÚS C-3...) es una pista de bajo coste para extraer líneas afectadas mediante regex como primera aproximación, antes de aplicar NLP más elaborado.

5. observacion_horaria (AEMET) — Features meteorológicas

Observación meteorológica horaria de las estaciones AEMET del corredor de Cercanías. Aporta las features exógenas de clima al modelo.

1.104 filas · 12 estaciones distintas · 16 columnas · 14/06/2026

5.1 Diccionario de columnas

Columna	Tipo	% nulos	Descripción
captura_stamp	str	0%	Marca temporal de la captura del colector.
estacion_id	str	0%	Código de estación AEMET (indicativo climatológico).
estacion_nombre	str	0%	Nombre de la estación (p.ej. MADRID RETIRO, MADRID/BARAJAS).
fecha_hora	datetime	0%	Fecha y hora de la observación.
temp_aire_c / temp_max_c / temp_min_c	float (°C)	0%	Temperatura del aire, máxima y mínima del periodo. Núcleo sólido, sin nulos.
precip_mm	float (mm)	0%	Precipitación acumulada. 'lp' (inapreciable) ya convertido a 0.0 por el compactador.
viento_vel_ms / viento_racha_ms / viento_dir_grados	float	16,3%	Velocidad media, racha máxima y dirección del viento.
humedad_rel_pct	float (%)	0%	Humedad relativa.
presion_hpa	float (hPa)	47,8%	Presión atmosférica. Alta tasa de nulos (~48%) — candidata a descartar o imputar por estación cercana.
latitud / longitud / altitud_m	float	0%	Coordenadas y altitud de la estación. Útiles para elegir la estación más cercana a cada tramo de vía.

5.2 Estaciones disponibles (14/06)

estacion_id	estacion_nombre
3110C	BUITRAGO
3125Y	SAN SEBASTIÁN DE LOS REYES
3126Y	EL GOLOSO (AUTOMÁTICA)
3129	MADRID/BARAJAS
3168D	GUADALAJARA
3170Y	ALCALÁ DE HENARES-ENCÍN
3182Y	ARGANDA DEL REY
3191E	COLMENAR VIEJO/FAMET
3194U	MADRID C. UNIVERSITARIA
3195	MADRID RETIRO
3330Y	ROZAS DE PUERTO REAL
3338	ROBLEDO DE CHAVELA

Se documentaron 15 estaciones del corredor; el 14/06 se observan datos de 12. Pendiente revisar si las 3 restantes (Retiro 3195 aparece sí, pero conviene listar cuáles faltan) no están reportando ese día o si el filtro geográfico del colector las excluye. 1.104 filas / 12 estaciones ≈ 92 capturas/estación/día (~cada 15 min), más frecuente que la hora documentada — revisar si hay registros duplicados por hora.

5.3 Muestra de filas

fecha_hora	estación	temp (°C)	precip (mm)	viento (m/s)	humedad (%)	presión (hPa)
2026-06-13T12:00Z	BITRAGO	26.4	0.0	1.9	24	909.1
2026-06-13T12:00Z	SAN SEBASTIÁN DE LOS REYES	30.0	0.0	(nulo)	21	(nulo)
2026-06-13T12:00Z	EL GOLOSO (AUTOMÁTICA)	28.6	0.0	2.2	23	(nulo)
2026-06-13T12:00Z	MADRID/BARAJAS	30.4	0.0	2.3	19	954.4

6. Implicaciones para el modelado — puntos de discusión

- La variable objetivo (`arrival_delay_s`) está disponible en el 88% de las filas de `trip_updates`, con mediana 0s y cola larga hacia retrasos positivos (p99 ~28 min). Hay que decidir si modelar como regresión del delay, clasificación por umbral (p.ej. ≥ 5 min), o ambas.
- `route_id`, `start_date`, `start_time` y `stop_sequence` NUNCA vienen poblados por el feed. La línea de Cercanías debe obtenerse cruzando `trip_id` (y/o `vehicle_label`) con el GTFS estático (`trips.txt`), con el sufijo del `trip_id` como pista adicional vía regex.
- `vehicle_positions` aporta `current_status` (`STOPPED_AT` / `INCOMING_AT` / `IN_TRANSIT_TO`) como feature de estado en tiempo real, cruzable por `trip_id` con `trip_updates`.
- `bearing` y `speed` de `vehicle_positions` NO existen en la práctica (100% nulos) — eliminar de cualquier lista de features prevista basada en la documentación teórica de GTFS-RT.
- `cause` y `header_text` de `alerts` no existen; `effect` casi nunca. El NLP debe trabajar sobre `description_text` (texto libre en español), a nivel de alerta única (`entity_id`), no de fila explotada.
- AEMET aporta 12 estaciones con temperatura/precipitación/humedad sin nulos; presión tiene ~48% de nulos y viento ~16% — decidir tratamiento (descarte de presión, imputación de viento).
- Pendiente de EDA: investigar si el patrón de delays anómalos (± 24 h) visto en el 13/06 (6,5% de filas) reaparece en otros días y si correlaciona con `trip_schedule_relationship` $\neq$ `SCHEDULED`.

Documento generado a partir de las capturas compactadas del 14/06/2026 (día completo). Próxima actualización tras integrar el GTFS estático y disponer de varios días de captura.