

Data Mining – Deber 4**1. Evidencia – Ingesta de años 2015–2025 (Yellow/Green):****Imprimir conteos por año/mes, duración:**

```
=====
INICIANDO INGESTA MASIVA NYC TLC -> POSTGRES
=====
```

```
Inicio: 2025-11-11 13:19:00.976454
Servicios: ['yellow', 'green']
Rango: 2015-2025
Run ID: run_001
=====
```

```
Checkpoint cargado: 258 archivos registrados.
```

```
=====
Servicio: YELLOW
=====
```

```
yellow 2015-01: Ya procesado (SKIP) [12741035 filas, 494.63848s]
yellow 2015-02: Ya procesado (SKIP) [12442394 filas, 339.948858s]
yellow 2015-03: Ya procesado (SKIP) [13342951 filas, 364.81764s]
yellow 2015-04: Ya procesado (SKIP) [13063758 filas, 350.321475s]
yellow 2015-05: Ya procesado (SKIP) [13157677 filas, 359.700977s]
yellow 2015-06: Ya procesado (SKIP) [12324936 filas, 334.870962s]
yellow 2015-07: Ya procesado (SKIP) [11559666 filas, 312.419438s]
yellow 2015-08: Ya procesado (SKIP) [11123123 filas, 308.906263s]
yellow 2015-09: Ya procesado (SKIP) [11218122 filas, 303.17454s]
yellow 2015-10: Ya procesado (SKIP) [12307333 filas, 336.394519s]
yellow 2015-11: Ya procesado (SKIP) [11305240 filas, 307.264311s]
```

```
=====
Servicio: GREEN
=====
```

```
green 2015-01: Ya procesado (SKIP) [1508493 filas, 45.402486s]
green 2015-02: Ya procesado (SKIP) [1574830 filas, 47.060252s]
green 2015-03: Ya procesado (SKIP) [1722574 filas, 51.600149s]
green 2015-04: Ya procesado (SKIP) [1664394 filas, 50.882582s]
green 2015-05: Ya procesado (SKIP) [1786848 filas, 53.026007s]
green 2015-06: Ya procesado (SKIP) [1638868 filas, 54.647252s]
green 2015-07: Ya procesado (SKIP) [1541671 filas, 48.479503s]
green 2015-08: Ya procesado (SKIP) [1532343 filas, 49.18342s]
green 2015-09: Ya procesado (SKIP) [1494927 filas, 46.274572s]
green 2015-10: Ya procesado (SKIP) [1630536 filas, 52.876946s]
green 2015-11: Ya procesado (SKIP) [1529984 filas, 48.926557s]
green 2015-12: Ya procesado (SKIP) [1608297 filas, 54.280331s]
green 2016-01: Ya procesado (SKIP) [1445292 filas, 48.692978s]
```

Summary final:

```
=====
RESUMEN FINAL DE INGESTA
=====
Exitosos:      0
Omitidos:      258
No encontrados: 0
Fallidos:      0
Rango:         2015-2025
Servicios:     yellow, green
Run ID:        run_001
Checkpoint:    /home/jovyan/work/checkpoint_ingesta.json
Duracion total: 0.00s (0.00 min)
Fin: 2025-11-11 13:19:00.984427
=====
```

2. Evidencia - Imprimir conteos por año/mes, duración y summary final (OBT):

Imprimir conteos por año/mes, duración:

```
[2025-11-11 19:14:38] Iniciando script build_obt.py (OPTIMIZADO CON COPY)
[2025-11-11 19:14:38] Argumentos: Namespace(mode='full', year_start=2020, year_end=2022, overwrite=True)
[2025-11-11 19:14:38] Intentando conectar a PostgreSQL...
[2025-11-11 19:14:38] Conexion exitosa a PostgreSQL con optimizaciones
[2025-11-11 19:14:38] =====
[2025-11-11 19:14:38] INICIANDO CONSTRUCCION OBT (MODO OPTIMIZADO)
[2025-11-11 19:14:38] =====
[2025-11-11 19:14:38] Rango: 2020-2022
[2025-11-11 19:14:38] Overwrite: True
[2025-11-11 19:14:38] RUN_ID: run_001
[2025-11-11 19:14:38] =====

[2025-11-11 19:14:38] Creando tabla analytics.obt_trips...
[2025-11-11 19:14:38] Tabla analytics.obt_trips creada/verificada

[2025-11-11 19:14:38] =====
[2025-11-11 19:14:38] Servicio: YELLOW
[2025-11-11 19:14:38] =====

[2025-11-11 19:14:38] --- Anio 2020 ---
[2025-11-11 19:14:38] Procesando: yellow 2020-01
[2025-11-11 19:14:38] - Extrayendo datos con COPY...
[2025-11-11 19:14:39] - Insertando datos con COPY...
[2025-11-11 19:14:43] - COMPLETADO: 6,405,008 filas en 92.6s (1525240 filas/seg)
[2025-11-11 19:14:43] Procesando: yellow 2020-02
[2025-11-11 19:14:43] - Extrayendo datos con COPY...
[2025-11-11 19:14:44] - Insertando datos con COPY...
[2025-11-11 19:14:48] - COMPLETADO: 6,299,367 filas en 88.4s (1536431 filas/seg)
[2025-11-11 19:14:48] Procesando: yellow 2020-03
[2025-11-11 19:14:48] - Extrayendo datos con COPY...
[2025-11-11 19:14:49] - Insertando datos con COPY...
[2025-11-11 19:14:51] - COMPLETADO: 3,007,687 filas en 83.1s (1253203 filas/seg)
```

```

[2025-11-11 19:18:30] =====
[2025-11-11 19:18:30] Servicio: GREEN
[2025-11-11 19:18:30] =====

[2025-11-11 19:18:30] --- Anio 2020 ---
[2025-11-11 19:18:30] Procesando: green 2020-01
[2025-11-11 19:18:30]   - Extrayendo datos con COPY...
[2025-11-11 19:18:30]   - Insertando datos con COPY...
[2025-11-11 19:18:31]   - COMPLETADO: 447,770 filas en 95.9s (497522 filas/seg)
[2025-11-11 19:18:31] Procesando: green 2020-02
[2025-11-11 19:18:31]   - Extrayendo datos con COPY...
[2025-11-11 19:18:31]   - Insertando datos con COPY...
[2025-11-11 19:18:32]   - COMPLETADO: 398,632 filas en 98.4s (498290 filas/seg)
[2025-11-11 19:18:32] Procesando: green 2020-03
[2025-11-11 19:18:32]   - Extrayendo datos con COPY...
[2025-11-11 19:18:32]   - Insertando datos con COPY...
[2025-11-11 19:18:33]   - COMPLETADO: 223,496 filas en 80.7s (446992 filas/seg)

```

Summary final:

```

[2025-11-11 19:22:47] =====
[2025-11-11 19:22:47] RESUMEN FINAL
[2025-11-11 19:22:47] =====
[2025-11-11 19:22:47] Total filas insertadas: 96,478,663
[2025-11-11 19:22:47] Finalizado: 2025-11-11 19:22:47
[2025-11-11 19:22:47] =====

```

3. Evidencia – Docker Compose ejecutándose:

The screenshot shows the Docker Desktop interface. On the left is a sidebar with navigation options: Ask Gordon, Containers (selected), Images, Volumes, Builds, Models, MCP Toolkit, Docker Hub, Docker Scout, and Extensions. The main panel is titled 'Containers' and shows a summary of container usage: 'Container CPU usage: 24.44% / 1200% (12 CPUs available)' and 'Container memory usage: 751.06MB / 7.51GB'. Below this is a table of running containers.

	Name	Container ID	Image	Port(s)	CPU (%)	Memory usage...	Memory (%)	Disk read/w	Actions
<input type="checkbox"/>	obt-builder-run1	b07e3fb263a8	deber4-obt		0%	9.66MB / 7.69GB	0.12%	20MB / 0B	[Stop] [Refresh] [Delete]
<input checked="" type="checkbox"/>	deber4	-	-	-	24.44%	741.4MB / 21.39K	10.01%	842MB / 17	[Stop] [Refresh] [Delete]
<input type="checkbox"/>	ny_taxi_postg	3059debd28df	postgres:11	5432:5432	24.39%	340.9MB / 7.69Gi	4.33%	260MB / 4	[Stop] [Refresh] [Delete]
<input type="checkbox"/>	pgadmin	22d2fa44c215	dpape/pgadmin4	8080:80	0.02%	232.6MB / 7.69Gi	2.95%	120MB / 12	[Stop] [Refresh] [Delete]
<input type="checkbox"/>	pyspark-noteb	adfe6aaedd2	jupyter/pys	4040:4040	0.03%	167.9MB / 6GB	2.73%	462MB / 17	[Stop] [Refresh] [Delete]

4. Evidencia – Postgress:

Esquema RAW:

Query:

```
SELECT table_name, column_name, data_type
FROM information_schema.columns
WHERE table_schema = 'raw'
ORDER BY table_name, ordinal_position;
```

Data Output:

table_name	column_name	data_type
green_taxi_trip	improvement_surchar...	double precision
green_taxi_trip	total_amount	double precision
green_taxi_trip	payment_type	bigint
green_taxi_trip	trip_type	double precision
green_taxi_trip	congestion_surcharge	integer
green_taxi_trip	run_id	text
green_taxi_trip	source_year	integer
green_taxi_trip	source_month	integer
green_taxi_trip	ingested_at_utc	timestamp without time zo...
green_taxi_trip	cbt_congestion_fee	double precision
taxi_zone_lookup	LocationID	bigint
taxi_zone_lookup	borough	text
taxi_zone_lookup	zone	text
taxi_zone_lookup	service_zone	text
taxi_zone_lookup	ingested_at_utc	timestamp without time zo...
taxi_zone_lookup	run_id	text
yellow_taxi_trip	VendorID	bigint
yellow_taxi_trip	trip_pickup_datetime	timestamp without time zo...
yellow_taxi_trip	trip_dropoff_datetime	timestamp without time zo...

Esquema ANALYTICS:

Query:

```
SELECT table_name, column_name, data_type
FROM information_schema.columns
WHERE table_schema = 'analytics'
ORDER BY table_name, ordinal_position;
```

Data Output:

table_name	column_name	data_type
ocb_trips	rate_code_id	double precision
ocb_trips	payment_type	bigint
ocb_trips	trip_type	double precision
ocb_trips	store_and_fuel_flag	character varying
ocb_trips	passenger_count	double precision
ocb_trips	trip_distance	double precision
ocb_trips	fare_amount	double precision
ocb_trips	extra	double precision
ocb_trips	mta_tax	double precision
ocb_trips	tip_amount	double precision
ocb_trips	tolls_amount	double precision
ocb_trips	improvement_surchar...	double precision
ocb_trips	congestion_surcharge	double precision
ocb_trips	airport_fee	double precision
ocb_trips	total_amount	double precision
ocb_trips	trip_duration_min	double precision
ocb_trips	avg_speed_mph	double precision
ocb_trips	tip_pct	double precision
ocb_trips	run_id	character varying

5. Evidencia – Tabla comparativa de todos los modelos (propios y sklearn) con métricas:

Tabla comparativa completa:

Modelo	Tipo	RMSE_Val	MAE_Val	R2_Val	Tiempo (s)
Ridge_Scratch	From-Scratch	2.203395	1.121511	0.975824	0.198644
Ridge_Sklearn	Scikit-Learn	2.203395	1.121511	0.975824	0.230908
Lasso_Sklearn	Scikit-Learn	2.209805	1.119803	0.975684	2.135232
Lasso_Scratch	From-Scratch	2.209948	1.119719	0.975680	53.260162
SGD_Scratch	From-Scratch	2.230374	1.133703	0.975229	499.708629
ElasticNet_Scratch	From-Scratch	2.281987	1.149258	0.974069	35.541106
ElasticNet_Sklearn	Scikit-Learn	2.282110	1.149357	0.974066	2.188510
SGD_Sklearn	Scikit-Learn	8.621060	1.776164	0.629904	78.973437

6. Evidencia – Gráficos de residuales/errores por bucket:

