

# Informe Técnico — Pipeline de Datos COVID-19 (OWID)

---

**Estudiante:** Isaac Villa

**Fecha:** 31/08/2025

**Git:** <https://github.com/IsaacVilla11/proyect-dagster.git>

## Arquitectura del pipeline

Assets y dependencias:

- 1) leer\_datos → lectura canónica (OWID) y normalización de 'date'.
- 2) datos\_procesados → normaliza tipos, filtra fechas futuras, de-duplica por (country,date), elimina nulos clave y filtra Ecuador + Perú.
- 3) metrica\_incidencia\_7d → media móvil de 7 días de incidencia por 100k.
- 4) metrica\_factor\_crec\_7d → cociente de sumas móviles 7d (actual vs previa).
- 5) reporte\_excel\_covid → exportación a data/reporte\_covid.xlsx.

[leer\_datos] → [datos\_procesados] → [metrica\_incidencia\_7d] ↵  
[metrica\_factor\_crec\_7d] ↗ [reporte\_excel\_covid]

## Decisiones de validación

**Entrada (leer\_datos):**

- sin\_fechas\_futuras → compara max(date) por día con tolerancia configurable.
- claves\_no\_nulas\_y\_unicidad → country/date/population presentes, population>0 y unicidad por (country,date).

**Salida (metrica\_incidencia\_7d):**

- incidencia\_rango\_esperado →  $0 \leq \text{incidencia\_7d} \leq 2000$ .

Hallazgos: fechas futuras por agregaciones; revisiones con casos negativos (ajuste sólo en métricas); entidades no-país con población inválida.

## Consideraciones de arquitectura

- pandas: suficiente para rolling/groupby y volumen del dataset.
- DuckDB: recomendable al escalar datasets/joins o para SQL columnas.
- Soda: alternativa declarativa, pero los Asset Checks cubren los objetivos del curso.

## Resultados

Cobertura de datos (datos\_procesados):

- Ecuador: 581 filas | fechas: 2021-01-20 → 2023-12-29
- Perú: 1035 filas | fechas: 2021-02-09 → 2023-12-10

Incidencia acumulada 7d por 100k — última observación:

- Ecuador: última 2023-12-29 → incidencia\_7d=0.000
- Perú: última 2023-12-10 → incidencia\_7d=0.000

Factor de crecimiento 7d — última observación:

- Ecuador: semana\_fin 2023-01-27 → casos\_semana=0, factor\_crec\_7d=0.000
- Perú: semana\_fin 2023-05-27 → casos\_semana=0, factor\_crec\_7d=0.000

**Tabla — Incidencia 7d (primeros 3 por país)**

date	country	incidencia_7d
2021-01-27 00:00:00	Ecuador	7,556
2021-01-28 00:00:00	Ecuador	7,929
2021-01-29 00:00:00	Ecuador	8,505
2021-02-15 00:00:00	Peru	19,941
2021-02-16 00:00:00	Peru	20,74
2021-02-17 00:00:00	Peru	20,177

**Tabla — Factor de crecimiento 7d (primeros 3 por país)**

























semana_fin	country	casos_semana	factor_crec_7d
2021-02-04			
00:00:00	Ecuador	9557	1,014
2021-02-05			
00:00:00	Ecuador	9117	0,922
2021-02-06			
00:00:00	Ecuador	8725	0,822
2/22/2021 12:00:00			
AM	Peru	48694	1,042
2/24/2021 12:00:00			
AM	Peru	48011	0,988
2/25/2021 12:00:00			
AM	Peru	48256	1,021

## Conclusión

El pipeline automatiza punta-a-punta el análisis para Ecuador y Perú con controles de calidad explícitos y salidas reproducibles. El filtrado de fechas futuras y el ajuste de casos negativos exclusivamente en las métricas aseguran interpretaciones consistentes. La

combinación de incidencia\_7d (magnitud estandarizada) y factor\_crec\_7d (dirección de tendencia) permite monitorear la situación de forma clara y accionable.

## Anexos:

Assets				View lineage	Reload definitions
Search and filter assets				+ Materialize selected	Refreshing data...
Asset name	Code location / Asset group	Status			
 <b>datos_procesados</b> Limpia nulos/duplicados, filtra Ecuador + país comparativo y devuelve las col...	 covid_dagster  default	 Materialized 31 ago, 2:04 p.m.			
 <b>leer_datos</b> Lee el CSV canónico de OWID (sin transformar).	 covid_dagster  default	 Materialized 31 ago, 2:04 p.m.			
 <b>metrica_factor_crec_7d</b> Factor de crecimiento semanal: semana_actual / semana_previa.	 covid_dagster  default	 Materialized 31 ago, 2:04 p.m.			
 <b>metrica_incidencia_7d</b> Incidencia acumulada a 7 días por 100k hab.	 covid_dagster  default	 Materialized 31 ago, 2:04 p.m.			
 <b>reporte_excel_covid</b> Exporta a Excel los resultados finales: datos_procesados, incidencia_7d y fa...	 covid_dagster  default	 Materialized 31 ago, 2:04 p.m.	