

Informe Técnico — Pipeline de Datos COVID-19

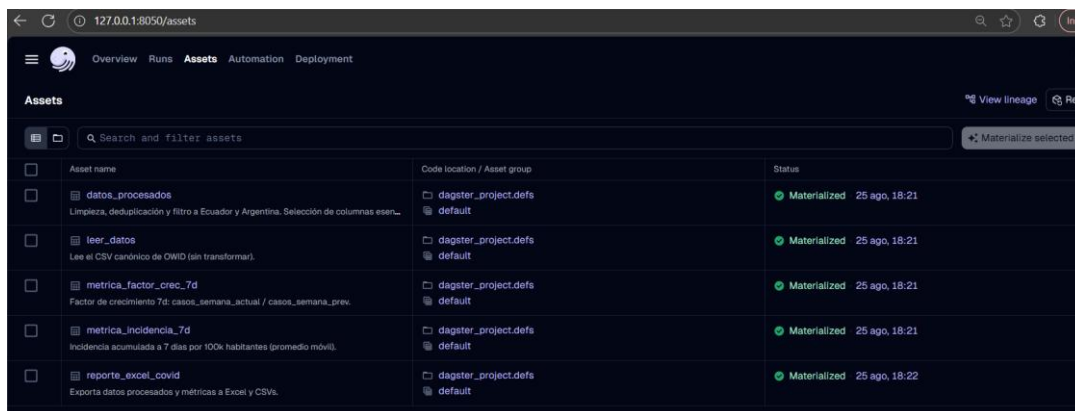
Santiago Pardo

1. Arquitectura del Pipeline

Descripción general:

El pipeline de datos COVID-19 está diseñado para procesar información diaria de casos y vacunación de OWID, centrado en Ecuador y un país comparativo (Argentina). Se implementa en Dagster, aprovechando la filosofía de Software-Defined Assets (SDA), donde cada asset representa un paso claro del pipeline y puede ser monitorizado de forma independiente.

```
2025-08-25 20:24:11 -0500 - dagster-webserver - INFO - Serving dagster-webserver on http://127.0.0.1:8050 in process 8052
2025-08-25 20:24:25 -0500 - dagster.daemon.QueuedRunCoordinatorDaemon - INFO - Priority sorting and checking tag concurrency limits for queued runs.
2025-08-25 20:24:28 -0500 - dagster.daemon.QueuedRunCoordinatorDaemon - INFO - Launched 1 runs.
```



Asset name	Code location / Asset group	Status
datos_procesados Limpieza, deduplicación y filtro a Ecuador y Argentina. Selección de columnas esenciales.	dagster_project.defs default	Materialized 25 ago, 18:21
leer_datos Lee el CSV canónico de OWID (sin transformar).	dagster_project.defs default	Materialized 25 ago, 18:21
metrica_factor_crec_7d Factor de crecimiento 7d: casos_semana_actual / casos_semana_prev.	dagster_project.defs default	Materialized 25 ago, 18:21
metrica_incidencia_7d Incidencia acumulada a 7 días por 100k habitantes (promedio móvil).	dagster_project.defs default	Materialized 25 ago, 18:21
reporte_excel_covid Exporta datos procesados y métricas a Excel y CSVs.	dagster_project.defs default	Materialized 25 ago, 18:22

Diagrama de Assets



2. Justificación de Decisiones de Diseño

- **Elección de Dagster:** Permite estructurar cada paso como un asset independiente y facilita testing, monitoreo y escalabilidad. Los Asset Checks permiten validar la calidad de los datos directamente en la UI.
- **Uso de Pandas:** Brinda herramientas rápidas de manipulación y agregación de datos, suficiente para los datasets diarios de OWID.
- **Validaciones de Entrada y Salida:** Previenen errores e inconsistencias. Los chequeos de entrada aseguran columnas esenciales y valores válidos; los de salida verifican que las métricas estén dentro de rangos razonables.
- **Filtrado a países específicos:** Focaliza el análisis en Ecuador y Argentina, asegurando relevancia para la interpretación de resultados.

- **Limpieza de datos:** Se eliminan duplicados y valores críticos nulos, garantizando métricas confiables y consistentes.

3. Procesamiento de Datos

El asset datos_procesados realiza los siguientes pasos:

1. **Copia del DataFrame original** para evitar modificaciones no deseadas en la fuente de datos.
2. **Renombrado de columnas inconsistentes:** si existe la columna country, se renombra a location para mantener consistencia con las validaciones y métricas.
3. **Eliminación de filas con datos críticos nulos:**
 - new_cases: necesario para cálculo de incidencia y factor de crecimiento.
 - people_vaccinated: permite análisis de cobertura de vacunación.
4. **Eliminación de duplicados:** se conserva la fila más reciente por (location, date) para reflejar revisiones oficiales.
5. **Filtrado a países de interés** (Ecuador y Argentina) para análisis comparativo.
6. **Selección de columnas esenciales:**
 - location, date, new_cases, people_vaccinated, population
 - Esto asegura que el DataFrame resultante sea compacto y eficiente para cálculo de métricas.

4. Validaciones y Control de Calidad

Estas reglas permiten asegurar integridad estructural y lógica del dataset antes de generar métricas, evitando cálculos inválidos o inconsistentes.

4.1 Chequeos de Entrada

Reglas aplicadas:

Regla	Resultado	Severidad
$\max(\text{date}) \leq \text{hoy}$	OK	WARN si falla
Columnas clave location, date, population no nulas	OK	ERROR si falla
Unicidad (location, date)	OK	WARN si hay duplicados
$\text{population} > 0$	OK	ERROR si hay valores ≤ 0
$\text{new_cases} \geq 0$ (permitidos negativos documentados)	OK	WARN si hay negativos

4.2 Chequeos de Salida

- **Incidencia 7d:**
 - Valores deben estar en $[0, 2000]$.
 - Detecta picos anómalos o errores de cálculo.
- **Factor de crecimiento 7d:**
 - Valores deben ser ≥ 0 y finitos.
 - Evita divisiones por cero o valores no definidos.

5. Cálculo de Métricas

A. Incidencia acumulada a 7 días por 100k habitantes

Fórmula:

- $\text{incidencia_diaria} = (\text{new_cases} / \text{population}) * 100000$
- $\text{incidencia_7d} = \text{promedio móvil de 7 días de incidencia_diaria}$

Interpretación:

Esta métrica normaliza los casos según la población y refleja la tendencia reciente de la epidemia.

Ecuador y Argentina, 2021-07-01:

fecha	país	incidencia_7d
2021-07-01	Ecuador	7.53
2021-07-01	Argentina	42.04

B. Factor de crecimiento semanal (7 días)

Fórmula:

- $\text{casos_semana_actual} = \text{suma}(\text{new_cases de los últimos 7 días})$
- $\text{casos_semana_prev} = \text{suma}(\text{new_cases de los 7 días previos})$
- $\text{factor_crec_7d} = \text{casos_semana_actual} / \text{casos_semana_prev}$

Interpretación:

- $\text{factor_crec_7d} > 1 \rightarrow$ la epidemia está creciendo.
- $\text{factor_crec_7d} < 1 \rightarrow$ la epidemia está decreciendo.

Ecuador y Argentina, semana 2021-07-07:

semana_fin	país	casos_semana	factor_crec_7d
2021-07-07	Ecuador	6462	0.63
2021-07-07	Argentina	119726	0.88

6. Exportación de Resultados

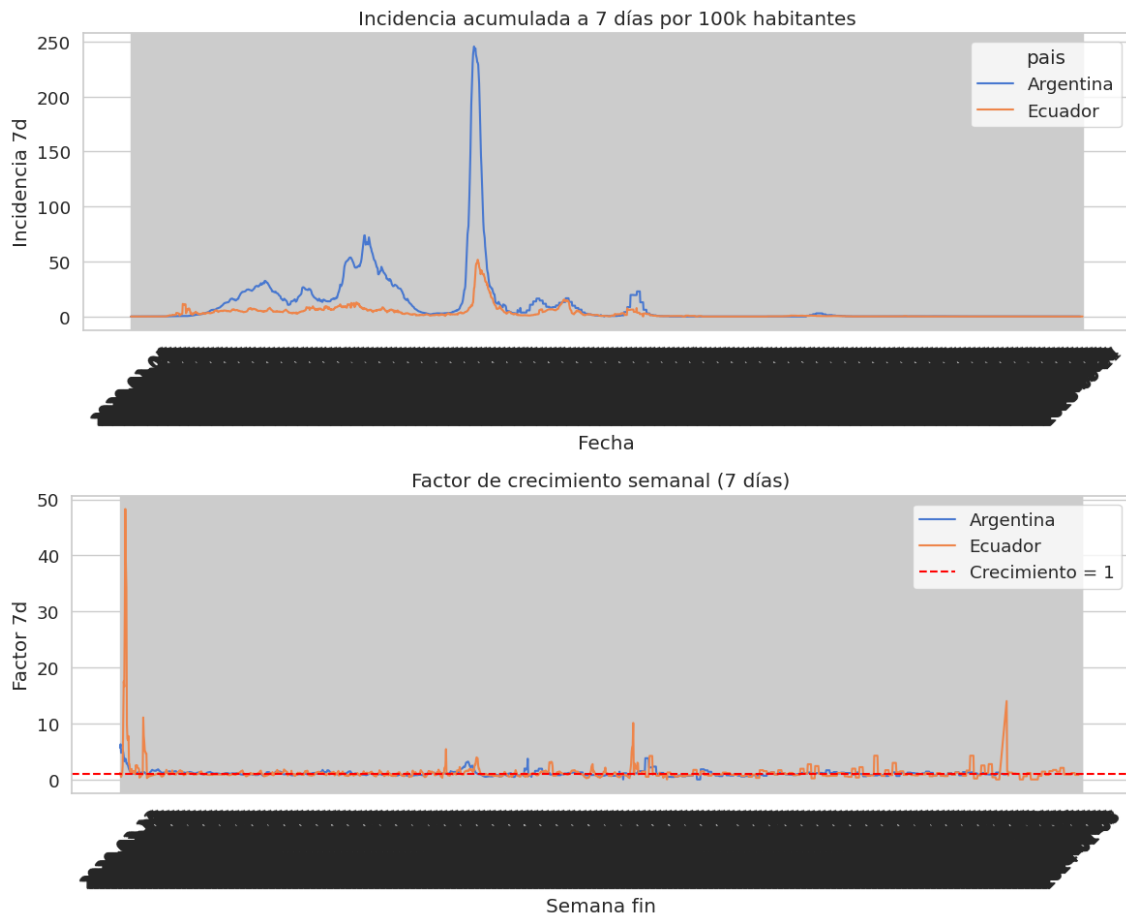
- Generación de reporte_covid.xlsx con hojas:
 - datos_procesados
 - incidencia_7d

- factor_crec_7d
- Exportación adicional de CSVs ligeros por métrica.

Ubicación: data/outputs/

7. Análisis y Visualización de Datos COVID-19

Se observa que los primeros días de cada serie contienen valores nulos o infinitos debido a que el cálculo de la media móvil de 7 días (incidencia) y del factor de crecimiento requiere datos de semanas previas. Para el análisis y visualización, estas filas fueron ignoradas.



Incidencia 7d por 100k habitantes:

- Argentina tuvo mayor incidencia promedio (10.86) y picos más altos (hasta 245) que Ecuador (promedio 2.99, máximo 51.6).
- La mayoría de los días muestran valores bajos, pero existen días con picos importantes.
- La gráfica muestra tendencia estable en Ecuador y más fluctuaciones en Argentina.

Factor de crecimiento semanal (7d):

- Valores cercanos a 1 predominan, indicando estabilidad.

- Argentina y Ecuador presentan semanas de crecimiento (>1) y decrecimiento (<1).
- No se encontraron valores extremos en ninguna métrica.

Conclusión:

Argentina presenta mayor variabilidad e incidencia que Ecuador, mientras que el factor de crecimiento semanal muestra estabilidad en la mayoría de los periodos.