Informe Técnico – Pipeline de Calidad y Métricas COVID-19

Autor: Edisson Ariel Guamán Parra

Proyecto Final Python – 2025

1. Arquitectura del pipeline

El pipeline fue implementado en **Dagster** siguiendo un enfoque de orquestación basado en *assets* y *asset checks*.

- Se consumen los datos abiertos de **Our World in Data (OWID)** sobre la pandemia de COVID-19.
- El flujo se compone de múltiples pasos: lectura, filtrado, limpieza, validación, cálculo de métricas y exportación de reportes.
- Se diseñaron **dos trabajos principales**:
    - lectura_chequeos: encargado de cargar los datos crudos y aplicar validaciones de entrada.
    - procesar_datos: encargado de filtrar, limpiar, calcular métricas epidemiológicas y generar un reporte consolidado.

2. Assets creados y diseño

Principales *assets*

- **datos_crudos / datos_procesar**: lectura directa del CSV OWID.
- **datos_filtrados_ecuador_spain**: selección de países de interés (Ecuador y España).
- **datos_limpios_ecuador_spain**: eliminación de nulos en new_cases y people_vaccinated.
- **datos_esenciales_ecuador_spain**: columnas clave para análisis (location, date, casos, vacunados, población).
- **metrica_incidencia_7d**: incidencia acumulada a 7 días por 100,000 habitantes.
- **metrica_factor_crec_7d**: ratio de crecimiento semanal de casos (comparación entre ventanas consecutivas).

- **reporte_excel_covid**: consolidación de resultados en un archivo Excel con tres hojas.

Justificación de diseño

- Se usó un **patrón modular**: cada transformación corresponde a un *asset* reutilizable.
- Se separaron **validaciones de entrada y salida** en *asset checks*, para asegurar calidad en todo el pipeline.
- Se eligió **Ecuador y España** como países comparativos para reducir volumen de datos y obtener insights claros.

3. Decisiones de validación

Entrada

- **Regla 1 – Columnas clave no nulas:** asegura que location, date y population estén completos, evitando inconsistencias de registros sin país o fecha.
- **Regla 2 – Muertes totales no negativas:** valida que total_deaths no tenga valores nulos ni negativos, lo que sería ilógico epidemiológicamente.

Salida

- **Regla 3 – Sin duplicados en datos esenciales:** evita duplicidad de registros por país-fecha.
- **Regla 4 – Incidencia 7d en rango válido:** se asegura que la incidencia acumulada no sea negativa ni supere valores extremos (2000 casos por 100k hab.).

4. Descubrimientos relevantes

- Se detectaron **valores nulos en muertes** (total_deaths) en algunos países y fechas. Esto refleja que ciertos gobiernos no reportaron datos completos.
- En **Ecuador y España**, las series de new_cases y people_vaccinated presentan **lagunas temporales** (días sin registro).
- Se observaron **ventanas de alta incidencia** en España durante 2021 y en Ecuador durante el primer semestre de 2020.
- El **factor de crecimiento** evidencia periodos donde los contagios crecían aceleradamente (factor > 1.5) y otros donde decrecían (factor < 1).

---

5. Consideraciones de arquitectura

- **Pandas** fue seleccionado como motor principal por su facilidad para manipular DataFrames y su integración nativa con Dagster.
- **DuckDB** no fue necesario, ya que el volumen de datos (~200 MB) es manejable en memoria.
- **Soda Core** (framework de validación externa) fue descartado en favor de los *asset checks* nativos de Dagster, que integran mejor los reportes de calidad con el pipeline.

6. Resultados

Métricas implementadas

Métrica	**Descripción**
Incidencia 7d	Casos promedio diarios por 100,000 habitantes en ventana móvil de 7 días.
Factor de crecimiento	Razón entre los casos de una semana y la semana anterior.

Interpretación:

- Valores de **incidencia altos** indican presión sobre el sistema de salud.
- Un **factor > 1** señala expansión de la pandemia, mientras que **< 1** indica reducción de contagios.

Resumen de control de calidad

nombre_regla	**estado**	**filas_afectadas**	**notas**
columnas_clave_no_nulas	OK	0	Sin problemas
total_deaths_no_negativos	ERROR	> 0	Existen nulos o negativos en muertes
check_duplicados_esenciales	OK	0	Sin duplicados detectados
check_incidencia_7d_validos	OK	0	Incidencia dentro de rangos esperados

7. Conclusión

El pipeline implementado en Dagster cumple con los objetivos de:

1. **Ingesta confiable de datos** de OWID.
2. **Limpieza y validación automática** mediante *asset checks*.
3. **Generación de métricas epidemiológicas clave** (incidencia y factor de crecimiento).
4. **Reporte consolidado en Excel** para análisis exploratorio.

La arquitectura propuesta es modular, escalable y permite incorporar más países o validaciones en el futuro.