



INSTITUTO TECNOLÓGICO DEL AZUAY

PROYECTO PYTHON

Nombres: Luis David Rey Jordán

Curso: N6A

29/08/2025

Introducción

El presente informe documenta el desarrollo e implementación de un proyecto orientado al análisis exploratorio de datos (EDA) sobre el impacto de la pandemia de COVID-19, utilizando Dagster como orquestador de datos y diversas herramientas del ecosistema de Python. El objetivo principal es construir un flujo de trabajo automatizado que permita la extracción, transformación, almacenamiento y análisis de información proveniente de fuentes confiables, generando resultados en formatos reutilizables como archivos CSV y Excel.

A través de este sistema se busca facilitar la comprensión de las métricas relacionadas con la evolución de la pandemia, integrando procesos de perfilamiento de datos, generación de reportes estructurados y validaciones de calidad. La arquitectura desarrollada prioriza la escalabilidad, la transparencia en cada etapa del pipeline y la posibilidad de extender el análisis a otros conjuntos de datos en el futuro.

OBJETIVOS

Objetivo General

Desarrollar un flujo de trabajo automatizado para la extracción, procesamiento, análisis y generación de reportes de datos relacionados con la pandemia de COVID-19, utilizando Dagster como orquestador de datos y librerías de análisis en Python.

Objetivos Específicos

1. Automatizar la ingesta de datos provenientes de la base de datos pública de *Our World in Data (OWID)* en formato CSV.
2. Procesar y limpiar los datos para garantizar consistencia y calidad en el análisis posterior.
3. Implementar un perfilado de datos que permita identificar estadísticas descriptivas, distribuciones y posibles valores atípicos.
4. Definir métricas clave (casos confirmados, fallecimientos, tasas de vacunación, entre otras) que faciliten la interpretación del impacto de la pandemia.
5. Generar reportes en diferentes formatos (CSV y Excel) para la consulta y el uso en entornos externos al pipeline.
6. Configurar un sistema de orquestación en Dagster que estructure las tareas en activos (assets) y permita su ejecución secuencial y controlada.
7. Garantizar la reproducibilidad del flujo de datos, documentando las dependencias y configuraciones necesarias para su despliegue.

ASSETS

1. leer_datos

- Objetivo: Extraer los datos de COVID-19 desde un archivo CSV remoto de Our World in Data (OWID).
- Tareas principales:
- Descarga el dataset usando requests.
- Carga el contenido en un DataFrame de pandas.
- Estandariza la columna de país (country → location).
- Salida: Un DataFrame en bruto sin limpieza adicional.

2. datos_limpios

- Objetivo: Asegurar la consistencia básica de los datos.
- Tareas principales:
- Convierte la columna de fechas a formato datetime.
- Elimina registros con fechas inválidas o futuras.
- Filtra poblaciones con valores negativos o nulos.
- Rellena nulos en new_cases y people_vaccinated con 0.
- Elimina duplicados por país y fecha.
- Salida: Un DataFrame depurado y validado.

3. chequeos_entrada

- Objetivo: Ejecutar verificaciones de integridad sobre los datos limpios.
- Chequeos realizados:
- No existen fechas posteriores al día actual.
- Las columnas clave (location, date, population) no tienen valores nulos.
- No existen duplicados por país y fecha.
- La población es estrictamente positiva.
- No hay valores negativos en new_cases.
- Salida: Resultados de validación (AssetCheckResult) y el DataFrame validado.

4. datos_procesados

- Objetivo: Filtrar y preparar los datos para el análisis comparativo de países.
- Tareas principales:
- Selecciona únicamente los países de interés (Ecuador y Perú).
- Conserva solo las columnas relevantes: location, date, new_cases, people_vaccinated, population.
- Salida: DataFrame reducido y enfocado en el análisis.

5. tabla_perfilado

- Objetivo: Generar un resumen estadístico básico de los datos procesados.
- Tareas principales:
- Calcula valores mínimos y máximos de new_cases, people_vaccinated y date.
- Determina el porcentaje de datos faltantes en cada variable.
- Exporta la tabla de perfilado en formato CSV (tabla_perfilado.csv).
- Salida: Un DataFrame de perfilado con información descriptiva.

6. metrica_incidencia_7d

- Objetivo: Calcular la incidencia acumulada a 7 días por cada 100.000 habitantes.
- Tareas principales:
- Calcula la incidencia diaria: $(\text{new_cases} / \text{population}) * 100000$.
- Aplica un promedio móvil de 7 días para suavizar la serie.
- Renombra columnas para su presentación (pais, fecha).
- Exporta los resultados a CSV (metrica_incidencia_7d.csv).
- Salida: DataFrame con la incidencia acumulada a 7 días.

7. metrica_factor_crec_7d

- Objetivo: Evaluar el crecimiento semanal de casos de COVID-19.
- Tareas principales:
- Calcula casos acumulados en ventanas de 7 días (casos_semana).
- Desplaza los valores una semana atrás para comparaciones.
- Calcula el factor de crecimiento: $\text{casos_semana} / \text{casos_semana_prev}$.
- Exporta los resultados a CSV (metrica_factor_crec_7d.csv).
- Salida: DataFrame con los valores del factor de crecimiento semanal.









































8. chequeos_salida

- Objetivo: Validar la consistencia de las métricas calculadas.
- Chequeos realizados:
- Verificar que la incidencia acumulada a 7 días (incidencia_7d) esté en el rango 0 a 2000.
- Salida: Resultados de validación y el DataFrame de métricas.

9. reporte_excel_covid

- Objetivo: Consolidar y exportar los resultados finales en un único archivo Excel.
- Tareas principales:
- Exporta en diferentes hojas los datos procesados, la incidencia acumulada y el factor de crecimiento.
- El archivo generado se guarda como reporte_covid.xlsx.
- Salida: Un archivo Excel con todos los resultados organizados.

Todos funcionan correctamente al probarse en dagster

Assets				View lineage	Reload definition
Search and filter assets				Materialize selected (9)	0:04
<input checked="" type="checkbox"/>	Asset name	Code location / Asset group	Status		
<input checked="" type="checkbox"/>	 chequeos_entrada Ejecuta chequeos de integridad sobre el DataFrame limpio	 eda_covid.py  default	 Materialized 30 ago, 5:38 p.m.		
<input checked="" type="checkbox"/>	 chequeos_salida Chequeos de salida sobre métricas calculadas	 eda_covid.py  default	 Materialized 30 ago, 5:38 p.m.		
<input checked="" type="checkbox"/>	 datos_limpios Limpieza de datos: fechas válidas, nulos manejados y consistencia básica.	 eda_covid.py  default	 Materialized 30 ago, 5:38 p.m.		
<input checked="" type="checkbox"/>	 datos_procesados Filtrar, limpiar y seleccionar columnas esenciales para Ecuador y Perú.	 eda_covid.py  default	 Materialized 30 ago, 5:38 p.m.		
<input checked="" type="checkbox"/>	 leer_datos Leer CSV remoto de OVID y retornar DataFrame sin limpiar.	 eda_covid.py  default	 Materialized 30 ago, 5:38 p.m.		
<input checked="" type="checkbox"/>	 metrica_factor_crec_7d Factor de crecimiento semanal (7 días)	 eda_covid.py  default	 Materialized 30 ago, 5:38 p.m.		
<input checked="" type="checkbox"/>	 metrica_incendencia_7d Incidencia acumulada a 7 días por 100 mil habitantes	 eda_covid.py  default	 Materialized 30 ago, 5:38 p.m.		
<input checked="" type="checkbox"/>	 reporte_excel_covid Exportar resultados finales (procesados y métricas) a Excel	 eda_covid.py  default	 Materialized 30 ago, 5:38 p.m.		

Conclusión

El presente análisis implementa un flujo de trabajo completo para la gestión y procesamiento de datos COVID-19, desde la extracción de información remota hasta la generación de métricas clave y reportes consolidados. Cada asset del pipeline contribuye a asegurar la calidad y consistencia de los datos, realizando limpieza, filtrado y chequeos de integridad que garantizan resultados confiables.

Se calcularon métricas relevantes como la incidencia acumulada a 7 días por 100.000 habitantes y el factor de crecimiento semanal, facilitando el seguimiento de la evolución de la pandemia en Ecuador y Perú. Además, la exportación de los resultados a un archivo Excel permite un acceso sencillo para análisis adicionales, visualización o toma de decisiones.