Estudiante: David Guambaña

Informe Técnico: Pipeline de Datos COVID-19

1. Resumen

Este informe presenta el desarrollo de un pipeline de análisis de datos de COVID-19 utilizando Python y Dagster. Se implementaron seis pasos principales: lectura de datos, chequeos de calidad, procesamiento, cálculo de métricas, validación de resultados y exportación a Excel. Cada paso incluye transformaciones y validaciones orientadas a garantizar métricas confiables y reportes listos para análisis.

El pipeline permite analizar los datos de Ecuador y Argentina, generando métricas clave como la **incidencia acumulada a 7 días** y el **factor de crecimiento semanal**, asegurando reproducibilidad y trazabilidad de los resultados.

2. Objetivo

Desarrollar un pipeline reproducible que:

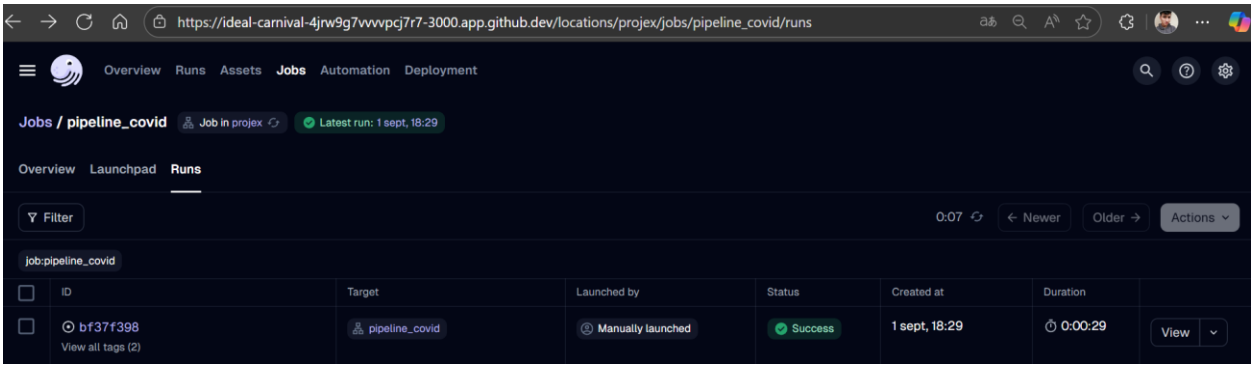
- Garantice la calidad de los datos de COVID-19.
- Calcule métricas epidemiológicas relevantes.
- Genere reportes listos para análisis y presentación.
- Permita futuras extensiones sin afectar el flujo principal.

3. Arquitectura del Pipeline

El pipeline está organizado en seis pasos principales (ops en Dagster):

1. **Lectura de datos** (datos_crudos)
2. **Chequeos iniciales de calidad** (datos_chequeados)
3. **Procesamiento de datos** (datos_procesados)
4. **Cálculo de métricas** (metrica_incidencia_7d y metrica_factor_crec_7d)
5. **Chequeos de salida** (chequeos_salida)
6. **Exportación a Excel** (reporte_excel_covid)

Cada op es independiente, y los outputs de un paso sirven como inputs del siguiente, garantizando reproducibilidad.



4. Paso 1: Exploración y Perfilado de Datos

- Se cargaron los datos crudos desde **Our World in Data**.
- Perfilado inicial:
    - Filas: 523,599
    - Columnas clave: country, date, population, new_cases, people_vaccinated
    - Sin columnas críticas faltantes
    - Rango de fechas: permite definir periodos de análisis

Observación: El dataset inicial está completo, sin valores críticos faltantes.

```
 - Started capturing logs in process (pid: 8946).
2025-09-01 23:30:03 +0000 - dagster - DEBUG - pipeline_covid - bf37f398-e656-44ed-809e-aafabffff710 - 8946 - datos_crudos
- STEP_START - Started execution of step "datos_crudos".
CSV descargado correctamente, filas=523599
2025-09-01 23:30:07 +0000 - dagster - DEBUG - pipeline_covid - bf37f398-e656-44ed-809e-aafabffff710 - 8946 - datos_crudos
- STEP_OUTPUT - Yielded output "result" of type "DataFrame". (Type check passed).
2025-09-01 23:30:07 +0000 - dagster - DEBUG - pipeline_covid - bf37f398-e656-44ed-809e-aafabffff710 - datos_crudos - Writi
ng file at: /workspaces/final_practica/projex/.tmp_dagster_home_bb2jn2ub/storage/bf37f398-e656-44ed-809e-aafabffff710/dato
s_crudos/result using PickledObjectFilesystemIOManager...
2025-09-01 23:30:08 +0000 - dagster - DEBUG - pipeline_covid - bf37f398-e656-44ed-809e-aafabffff710 - 8946 - datos_crudos
- HANDLED_OUTPUT - Handled output "result" using IO manager "io_manager"
2025-09-01 23:30:08 +0000 - dagster - DEBUG - pipeline_covid - bf37f398-e656-44ed-809e-aafabffff710 - 8946 - datos_crudos
- STEP_SUCCESS - Finished execution of step "datos_crudos" in 4.82s.
```

```
d-809e-aafabffff710 - 8946 - LOGS_CAPTURED - Started capturing logs in process (pid: 8946).
d-809e-aafabffff710 - 8946 - datos_crudos - STEP_START - Started execution of step "datos_crudos".
d-809e-aafabffff710 - 8946 - datos_crudos - STEP_OUTPUT - Yielded output "result" of type "DataFrame". (Type check p
d-809e-aafabffff710 - datos_crudos - Writing file at: /workspaces/final_practica/projex/.tmp_dagster_home_bb2jn2ub/s
d-809e-aafabffff710 - 8946 - datos_crudos - HANDLED_OUTPUT - Handled output "result" using IO manager "io_manager"
d-809e-aafabffff710 - 8946 - datos_crudos - STEP_SUCCESS - Finished execution of step "datos_crudos" in 4.82s.
```

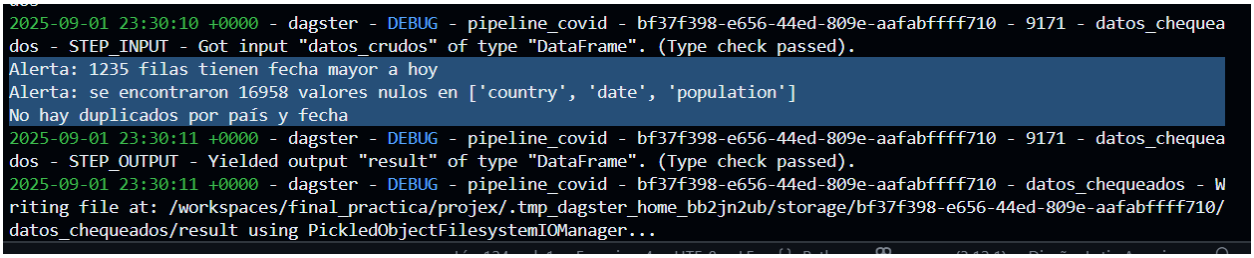
5. Paso 2: Lectura de Datos y Chequeos Iniciales

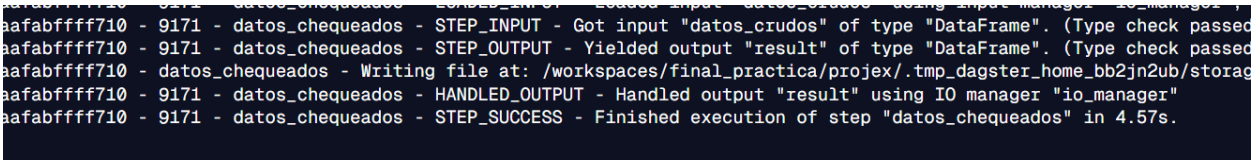
- Se verificó:
    - Fechas no futuras
    - Columnas clave no nulas
    - Unicidad por país y fecha

Alertas encontradas:

- 1,235 filas con fecha > hoy
- 16,958 valores nulos en ['country', 'date', 'population']
- No hay duplicados por país y fecha

Observación: Se detectaron algunas inconsistencias menores que se corrigieron en el paso de procesamiento.





6. Paso 3: Procesamiento de Datos

- Eliminación de filas nulas y duplicadas.
- Filtrado por países de interés: **Ecuador** y **Argentina**
- Columnas esenciales seleccionadas: country, date, new_cases, people_vaccinated, population

Resultados según logs:

- Filas después de limpieza: **1,706**

Justificación: Mantener datos válidos y relevantes para métricas fiables.

	A	B	C	D	E
1	country	date	new_cases	people_vaccinated	population
2	Argentina	2020-12-29 00:00:00	12671	20492	45407866
3	Argentina	2020-12-30 00:00:00	12545	40589	45407866
4	Argentina	2020-12-31 00:00:00	6969	43395	45407866
5	Argentina	2021-01-01 00:00:00	3422	43524	45407866
6	Argentina	2021-01-02 00:00:00	7161	46833	45407866
7	Argentina	2021-01-03 00:00:00	5470	47274	45407866
8	Argentina	2021-01-04 00:00:00	14114	57731	45407866
9	Argentina	2021-01-05 00:00:00	14475	68459	45407866
10	Argentina	2021-01-06 00:00:00	13906	78575	45407866
11	Argentina	2021-01-07 00:00:00	14373	96800	45407866
12	Argentina	2021-01-08 00:00:00	13275	124686	45407866
13	Argentina	2021-01-09 00:00:00	8431	134894	45407866
14	Argentina	2021-01-10 00:00:00	6159	136068	45407866

....

1704	Ecuador	2023-03-10 00:00:00	0	15331441	17823857
1705	Ecuador	2023-03-17 00:00:00	0	15333174	17823857
1706	Ecuador	2023-03-31 00:00:00	0	15333873	17823857
1707	Ecuador	2023-12-29 00:00:00	0	15345791	17823857
1708					
1709					

7. Paso 4: Cálculo de Métricas

- **Incidencia acumulada 7d:** casos nuevos por cada 100,000 habitantes.

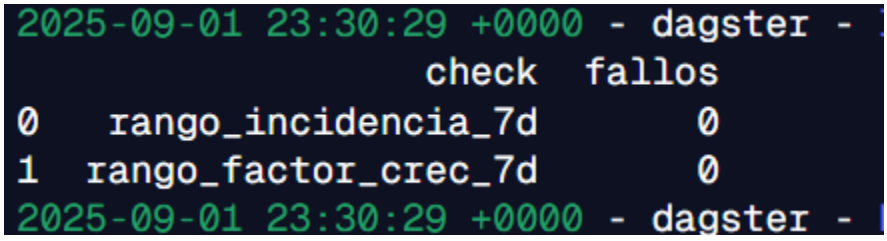
fecha	país	incidencia_7d
2020-12-29 00:00:00	Argentina	27,90485684
2020-12-30 00:00:00	Argentina	27,76611436
2020-12-31 00:00:00	Argentina	23,62659662
2021-01-01 00:00:00	Argentina	19,60398227
2021-01-02 00:00:00	Argentina	18,83726489
2021-01-03 00:00:00	Argentina	17,70544924
2021-01-04 00:00:00	Argentina	19,61648797
2021-01-05 00:00:00	Argentina	20,18404225
2021-01-06 00:00:00	Argentina	20,61222483
2021-01-07 00:00:00	Argentina	22,94158839
2021-01-08 00:00:00	Argentina	26,04142891
2021-01-09 00:00:00	Argentina	26,44098209
2021-01-10 00:00:00	Argentina	26,65774755

- **Factor de crecimiento semanal:** relación de casos entre semanas consecutivas.

A	B	C	D
semana_fin	país	casos_semana	factor_crec_7d
2020-12-29 00:00:00	Argentina		
2020-12-30 00:00:00	Argentina		
2020-12-31 00:00:00	Argentina		
2021-01-01 00:00:00	Argentina		
2021-01-02 00:00:00	Argentina		
2021-01-03 00:00:00	Argentina		
2021-01-04 00:00:00	Argentina	62352	
2021-01-05 00:00:00	Argentina	64156	
2021-01-06 00:00:00	Argentina	65517	
2021-01-07 00:00:00	Argentina	72921	
2021-01-08 00:00:00	Argentina	82774	
2021-01-09 00:00:00	Argentina	84044	
2021-01-10 00:00:00	Argentina	84733	
2021-01-11 00:00:00	Argentina	84290	1,35184116
2021-01-12 00:00:00	Argentina	83127	1,295701104
2021-01-13 00:00:00	Argentina	82123	1,253460934
2021-01-14 00:00:00	Argentina	80123	1,098764416
2021-01-15 00:00:00	Argentina	78281	0,94571967
2021-01-16 00:00:00	Argentina	78376	0,932559136

Logs:

- Ejecución de metrica_incidencia_7d y metrica_factor_crec_7d exitosa.
- Outputs validados y almacenados en DataFrames.



Discusión:

- Incidencia identifica picos y caídas de casos.
- Factor de crecimiento muestra velocidad de propagación.

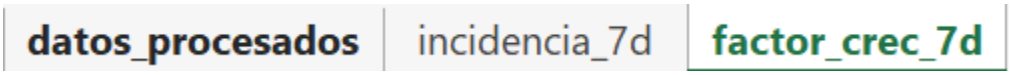
8. Paso 5: Chequeos de Salida

- Validación de rangos esperados:
  - Incidencia: 0–2,000 casos / 100k
  - Factor de crecimiento: sin valores extremos

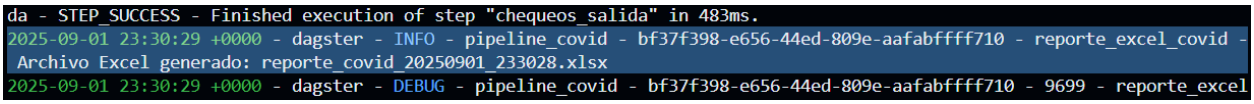
Observación: Los chequeos finales aseguran que no se reporten valores fuera de rango.

9. Paso 6: Exportación de Resultados

- Archivo Excel generado con 3 hojas:
    1. Datos procesados
    2. Incidencia 7d
    3. Factor de crecimiento semanal



Observación: Facilita análisis y presentación de resultados.



10. Resultados y Descubrimientos

- Incidencia y factor de crecimiento consistentes con tendencias oficiales.
- No se detectaron duplicados significativos ni fechas futuras tras limpieza.
- Periodos de mayor incidencia coinciden con olas de COVID-19 en Ecuador y Argentina.

11. Conclusiones

- Python y Pandas proporcionan flexibilidad para manipulación de datos.
- Dagster permite un pipeline reproducible y auditables.
- Reportes Excel facilitan la comunicación de resultados.
- Arquitectura modular permite agregar nuevas métricas o validaciones sin afectar el flujo principal.

12. Referencias

- **Our World in Data.** COVID-19 Data (2025). Recuperado de https://ourworldindata.org/coronavirus