



Postgrado
en Informática
UMSA



GEORGINA QUIROZ MENDOZA
CARLOS ALMARAZ ESCOBAR
DILAN CONDORI ALEJO
ALVARO JURADO ALFARO
JAIME MONTECINOS MARQUEZ

ADQUISICIÓN Y COMPRENSIÓN DE DATOS - M7





Postgrado
en Informática
UMSA

Introducción

- **Objetivo:** Construir un pipeline de datos para ingestión, limpieza, integración y análisis de información climática.
- **Alcance:** Integración de fuentes externas (API + históricos), control de calidad y almacenamiento estructurado.





Postgrado
en Informática
UMSA



Fuentes de Datos

- API OpenWeather → Datos en tiempo real de 9 ciudades de Bolivia.
- API Meteostat
- CSV Históricos:
[Environment_Temperature_change_E_All_Data_NOFLAG + FAOSTAT_data_en_11-1-2024](#).
- Justificación: diversidad de fuentes asegura riqueza y comparabilidad.





Postgrado
en Informática
UMSA



Ingesta de Datos

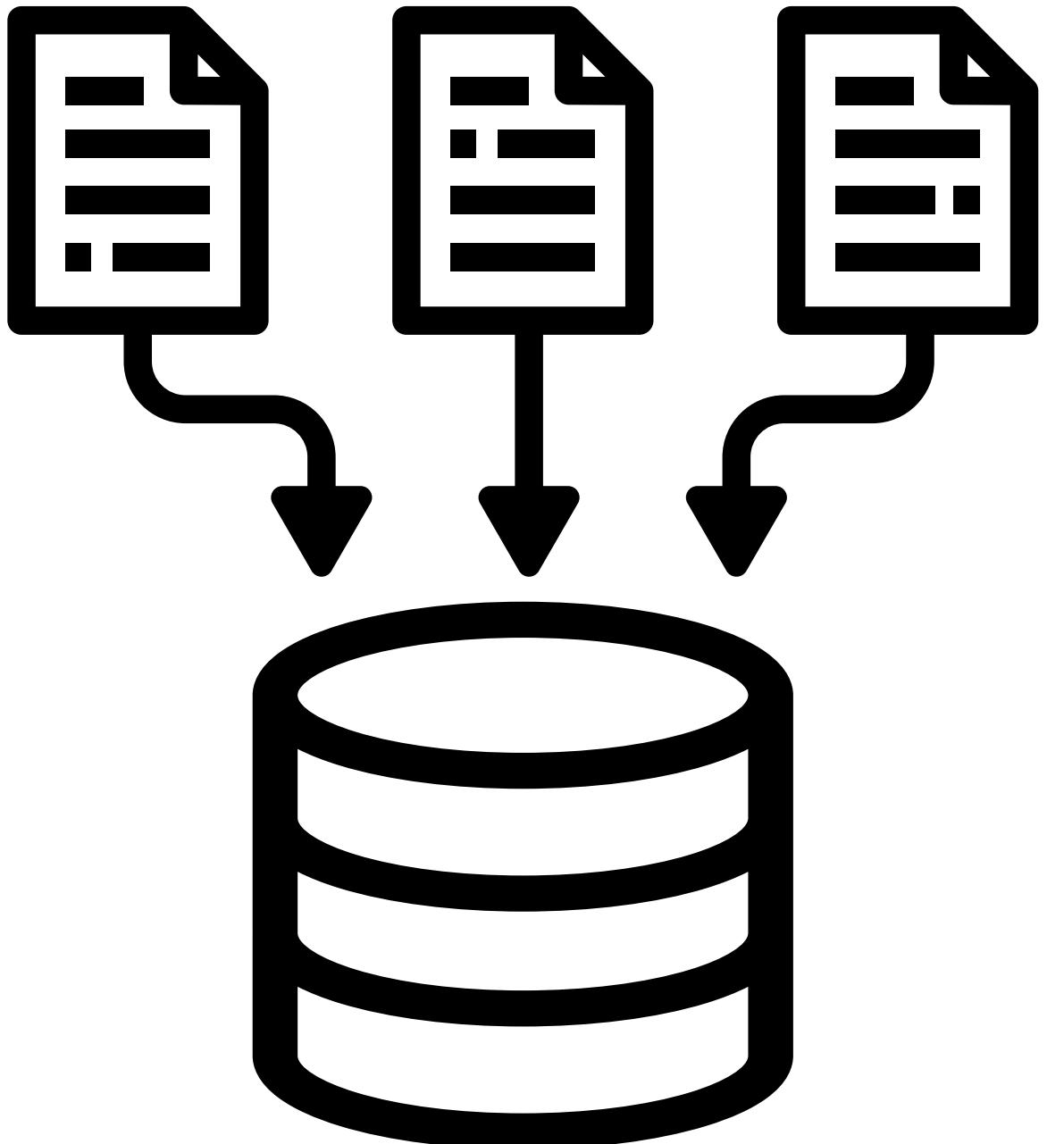
API OpenWeather y Mateostat:

- Consumo con `requests`.
- Respuesta en `JSON` convertida a `DataFrame`.

CSV Históricos:

- Cargados con `pandas.read_csv()`.
- Normalización de fechas y unidades.

Control de errores: manejo de excepciones en conexión a API y validación de archivos CSV.





Postgrado
en Informática
UMSA



Calidad de Datos

Limpieza:

- Normalización de columnas (nombres, tipos de dato).
- Eliminación de duplicados y nulos.

Transformacion

- Conversión JSON → DataFrame.
- Integración con CSVs históricos.

Validación:

- Valores de temperatura, humedad y viento dentro de rangos razonables.

Diccionario de datos:

- Definición de variables: temp_c, humidity, wind_speed, anomalia_temp.

Profiling:

- Reporte HTML con estadísticas y gráficos automáticos.





Postgrado
en Informática
UMSA

Procesamiento y Clasificación

Clasificación y enriquecimiento

- Temperatura → frío, templado, cálido, caluroso.
- Viento → calma, brisa, ventoso, fuerte.

Derivación de variables

Cálculo de anomalías
climáticas respecto a la
media histórica.





Postgrado
en Informática
UMSA



Almacenamiento de Datos

- CSV limpio →
`/data/processed/dataset_clima_limpio.csv.`
- Excel procesado → fácil intercambio con analistas.
- SQLite DB (`weather.db`) → tabla `fact_weather`.
- Justificación: repositorios estructurados garantizan trazabilidad y reutilización.



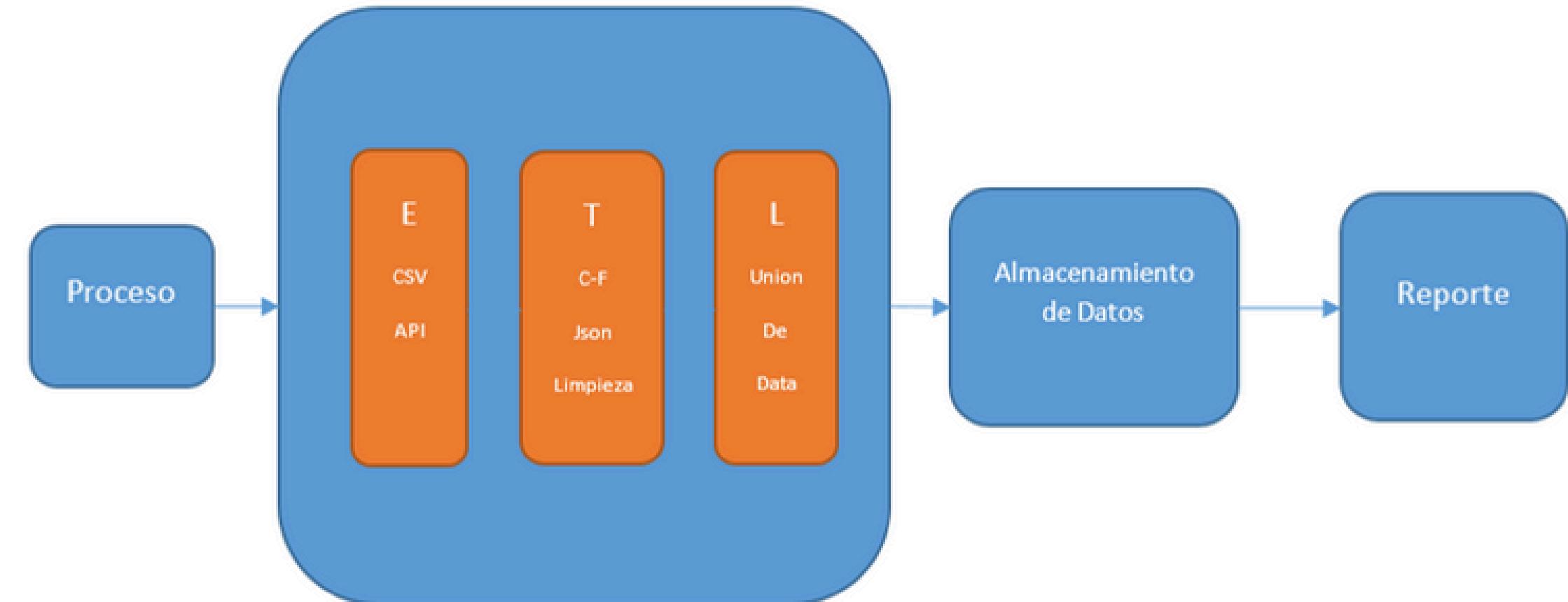


Arquitectura de Solución

Batch Processing (24 horas)

Justificación:

- **Datos climáticos cambian cada pocas horas.**
- **Formato de Datos Diferentes.**
- **Facil de orquestar con Airflow/cron.**



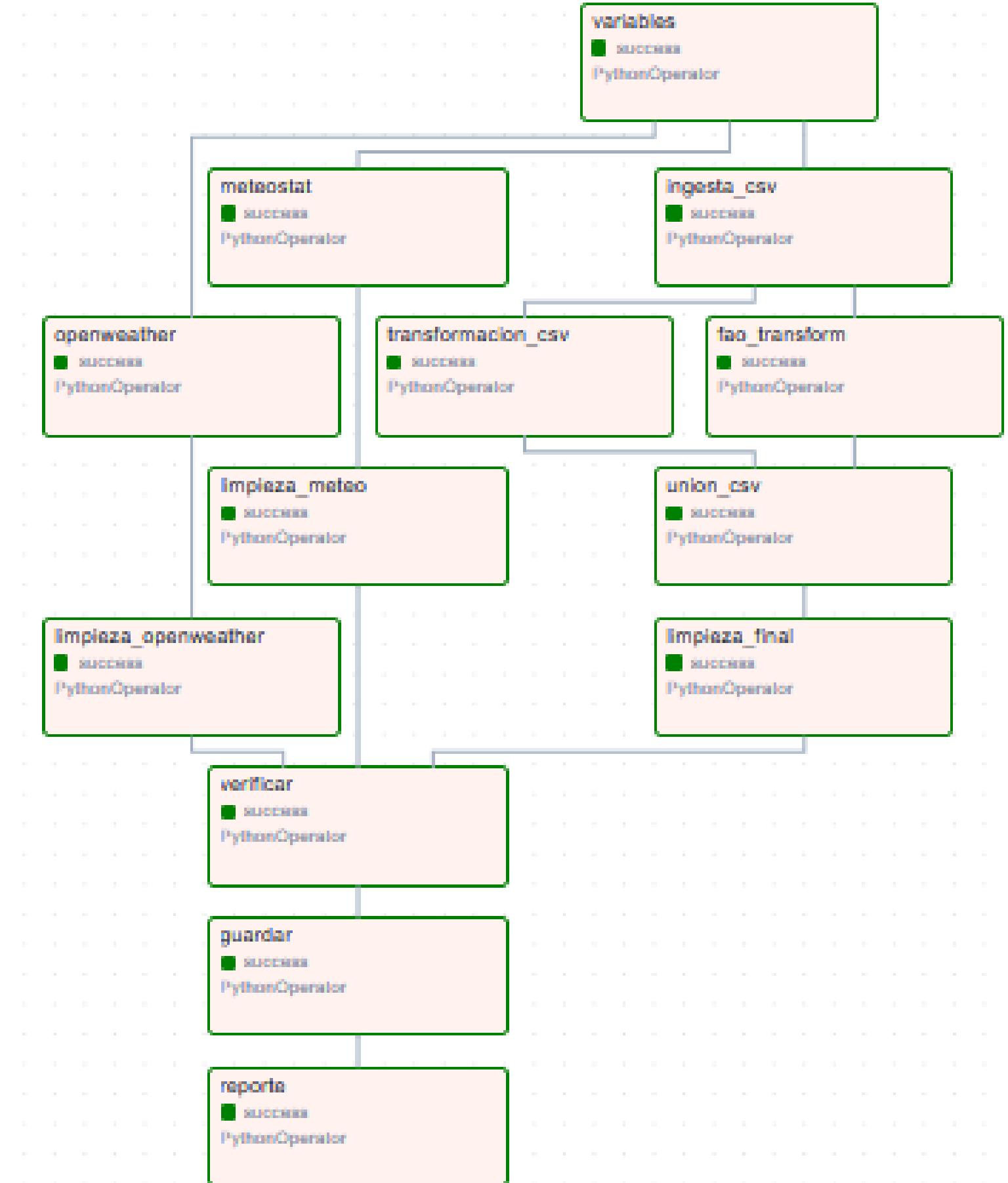


Pipeline Orquestado

Arquitectura orquestada con:

1. Extraccion (API + CSV).
2. Trasformacion (Trasnformacion, Limpieza).
 - a. T-Enviroment
 - b. T-FAO
 - c. T-Open
 - d. T-Meteostat
3. Carga.
4. Almacenamiento de Datos .db, .csv

Reporte html





Postgrado
en Informática
UMSA

Data Quality Report y Documentación

Reporte HTML generado automáticamente:

- **Conteo de filas/columnas.**
- **Distribución de variables.**
- **Porcentaje de nulos y duplicados.**
- **Histogramas y series temporales.**

Documentación técnica:

- **Pasos de ingesta, limpieza y almacenamiento.**
- **Diccionario de datos.**

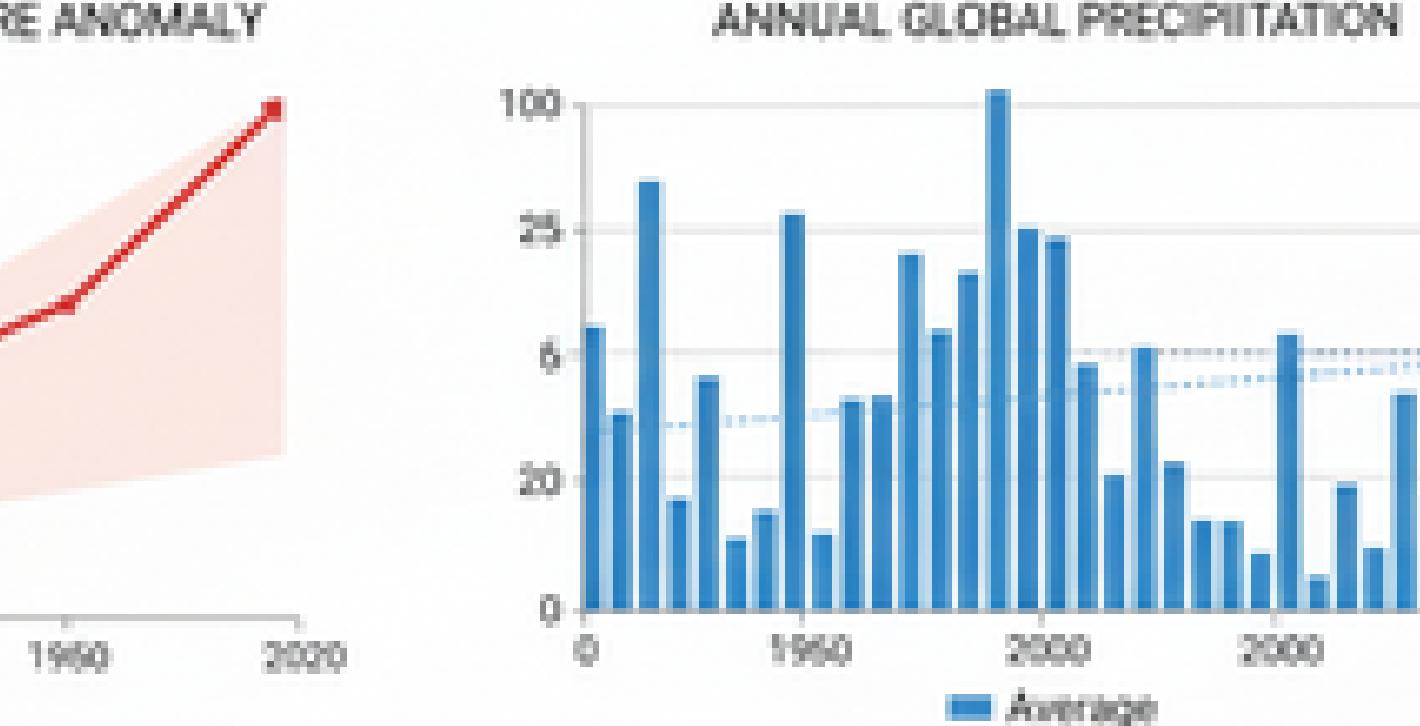
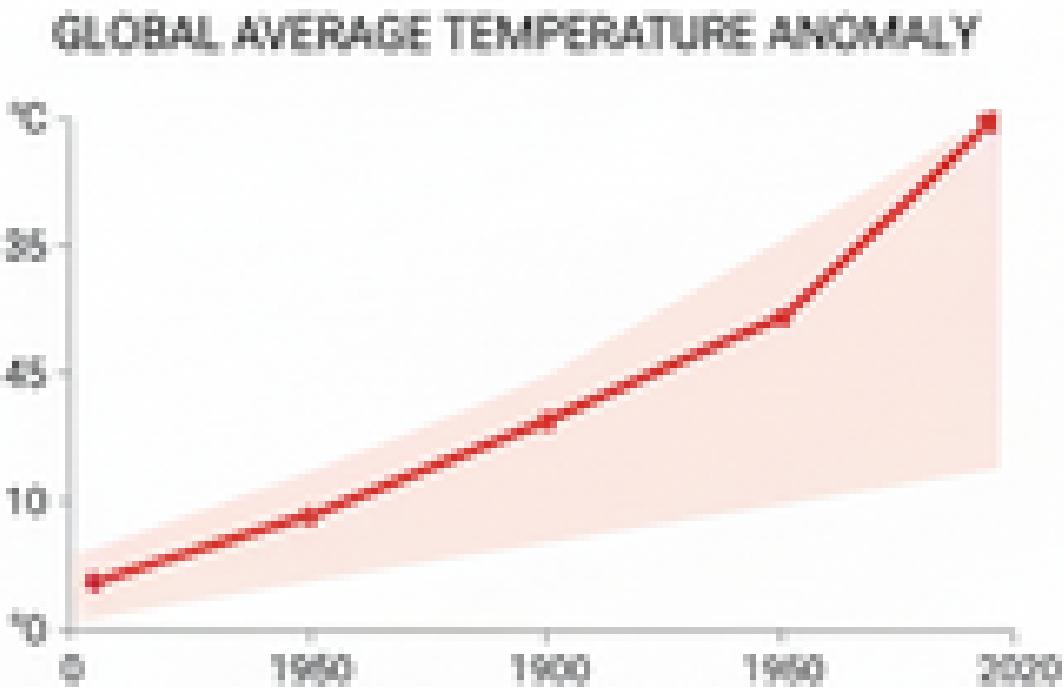




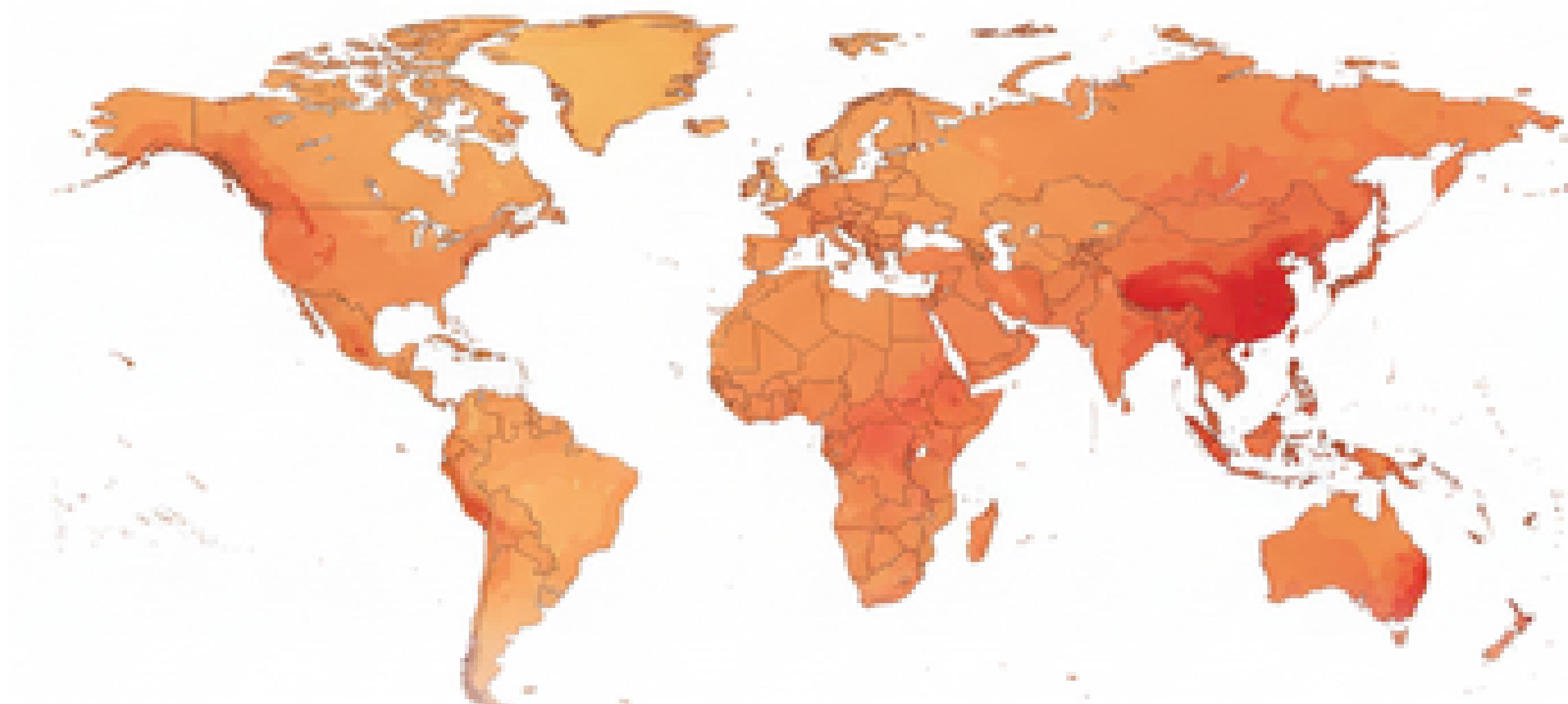
Postgrado
en Informática
UMSA

Resultados

- Pipeline funcional desde múltiples fuentes.
- Dataset limpio, validado y enriquecido.
- Reporte de calidad publicado.
- Arquitectura reproducible y escalable.



REGIONAL TEMPERATURE CHANGE
(2000 vs 1900)





Limitaciones

Costos de la API OpenWeather:

- El plan gratuito tiene un límite de requests/día y acceso restringido a endpoints avanzados.
- Para escalabilidad y más ciudades, se requiere migrar a planes de pago.

Dependencia de Meteostat:

- Posibles desconexiones o interrupciones en estaciones meteorológicas.
- Riesgo de brechas en series temporales por datos faltantes.

Features limitados en datos históricos:

- Variables restringidas (ej. no siempre hay radiación solar, presión detallada o viento a distintos niveles).
- Resolución temporal limitada (ej. datos diarios en lugar de horarios).



Conclusiones y Futuras Mejoras

- Se logró implementar un ETL robusto y documentado.
- Se cumplieron los 7 lineamientos: ingesta, calidad, procesamiento, orquestación, almacenamiento, arquitectura y reporte.
- Orquestación en producción con Airflow.

Mejoras:

- Integración con PostgreSQL / Data Warehouse.
- Dashboards interactivos (Power BI, Grafana).
- Extensión a streaming con Kafka.



**Gracias por su
atención**