



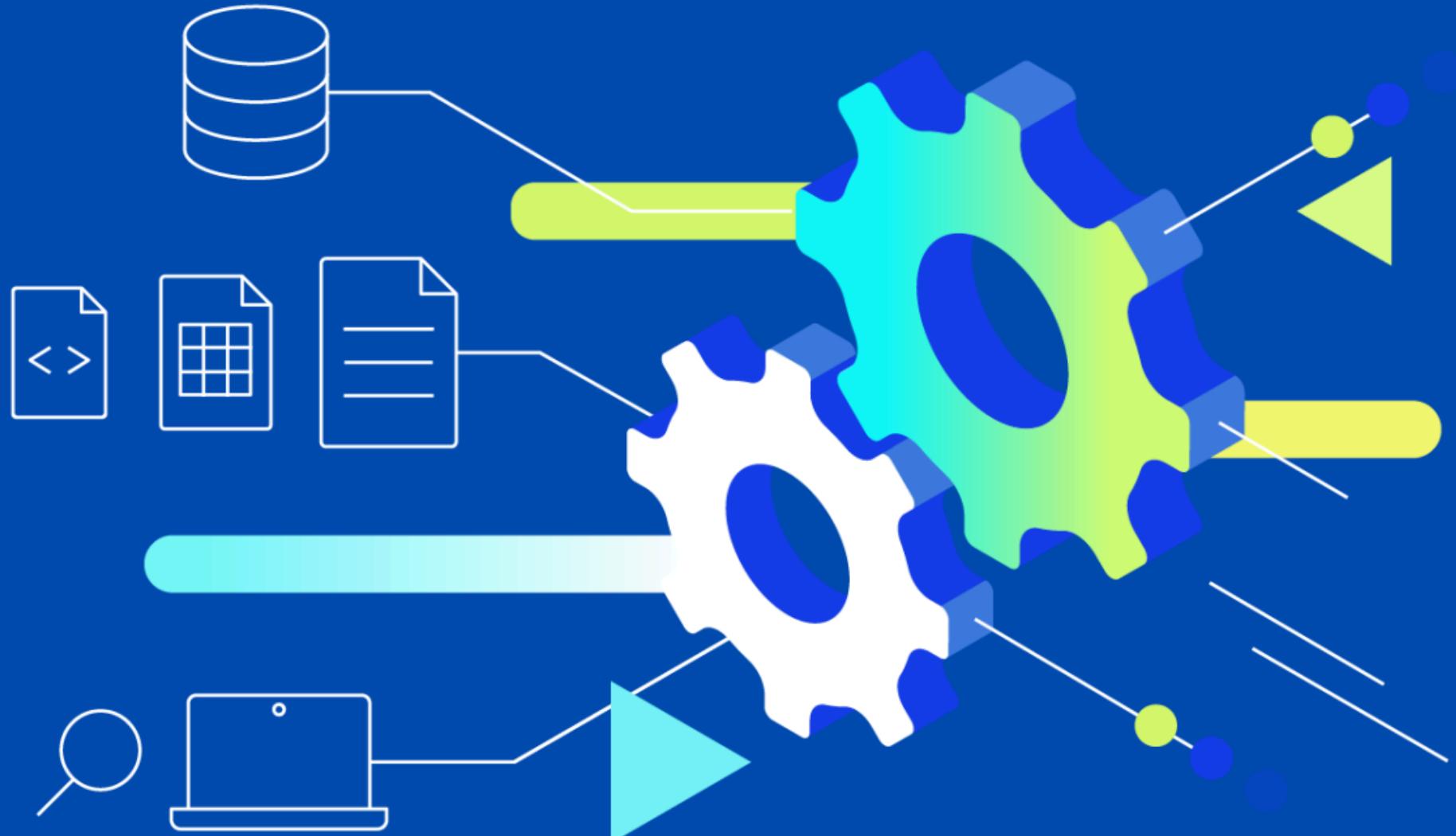
Proyecto ETL con Azure Data Factory

Integrantes:

Iván Torres
Juan Camilo Roman
Juan Pablo Toro
Miguel Ordoñez

Materia:

Infraestructura & Arquitectura de TI



Agenda

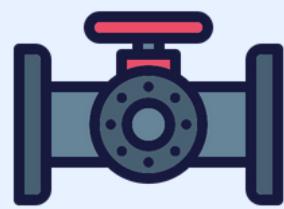


- Objetivo del proyecto
- Contexto
- Arquitectura general
- Construcción del pipeline (SQL + Mongo)
- Unificación final
- Resultados
- Conclusiones y lecciones aprendidas

Objetivo General

Construir un proceso completo ETL utilizando Azure Data Factory para integrar datos relacionales, no relacionales y archivos CSV, generando como salida salida.csv con 81 variables finales de propiedades.

Entregables



Pipeline ETL en ADF



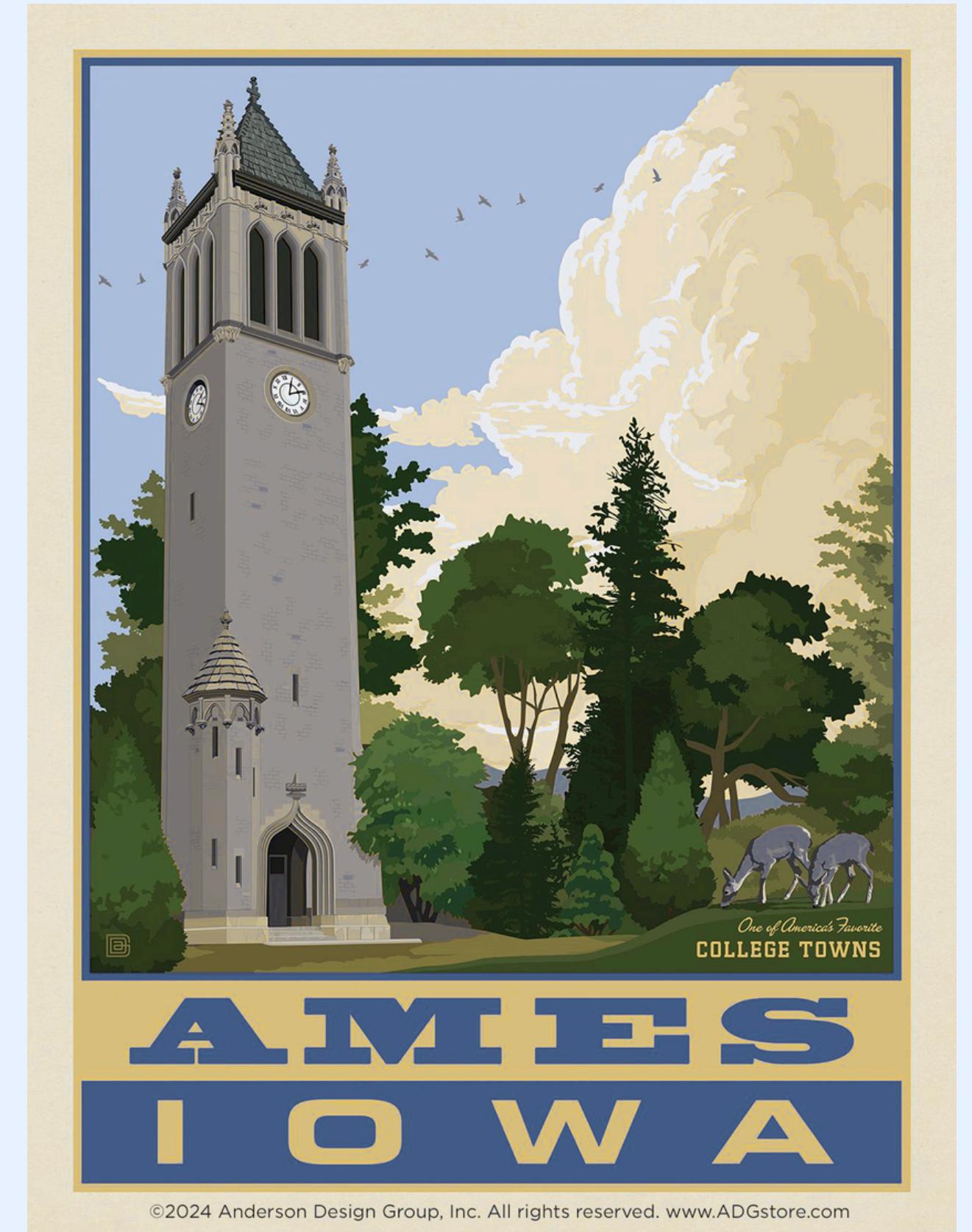
Archivo final salida.csv



Informe del proceso y lecciones
aprendidas

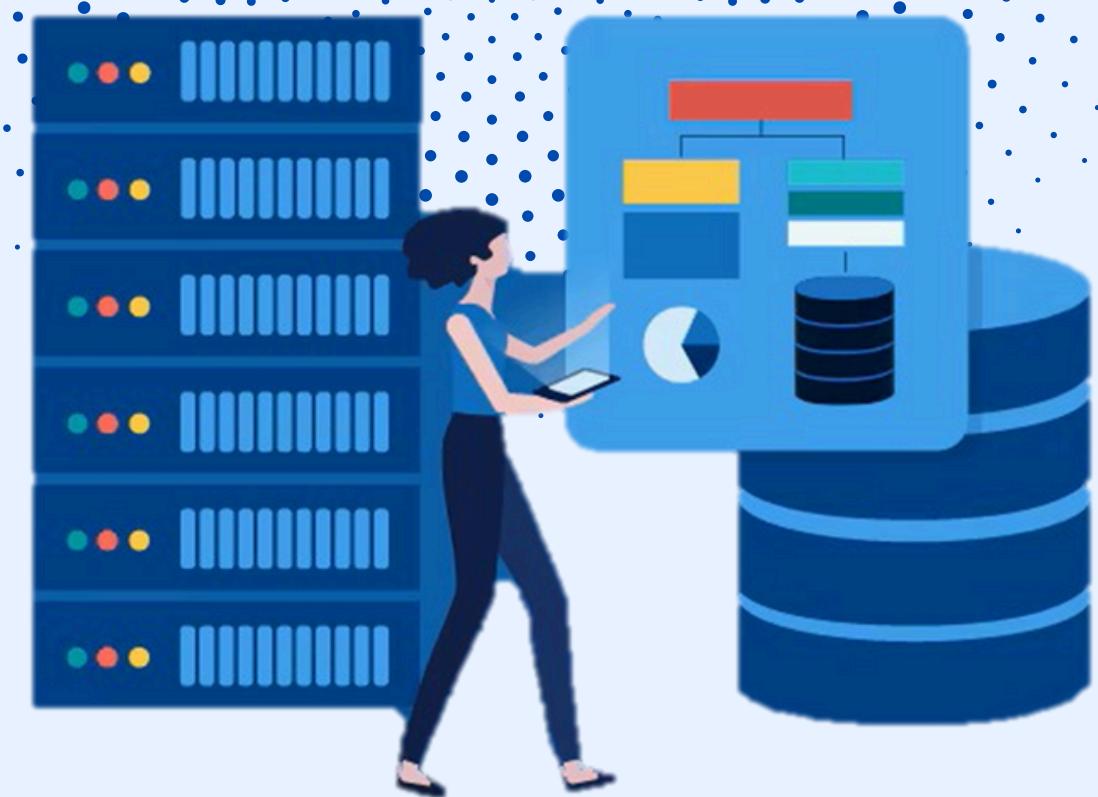
Contexto del dataset Ames

- El conjunto de datos describe propiedades residenciales en Ames, Iowa.
- 2930 propiedades con más de 80 variables estructurales.

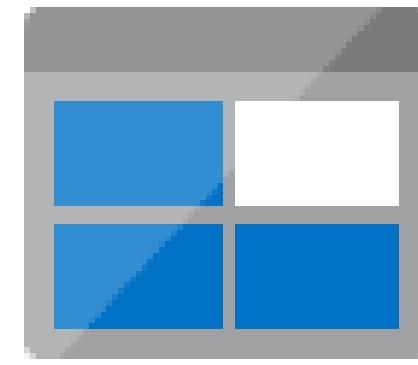


©2024 Anderson Design Group, Inc. All rights reserved. www.ADGstore.com

Arquitectura General



Linked Services



Azure Blob Storage

Linked Services



Azure Database for
PostgreSQL

Linked Services



MongoDB Atlas

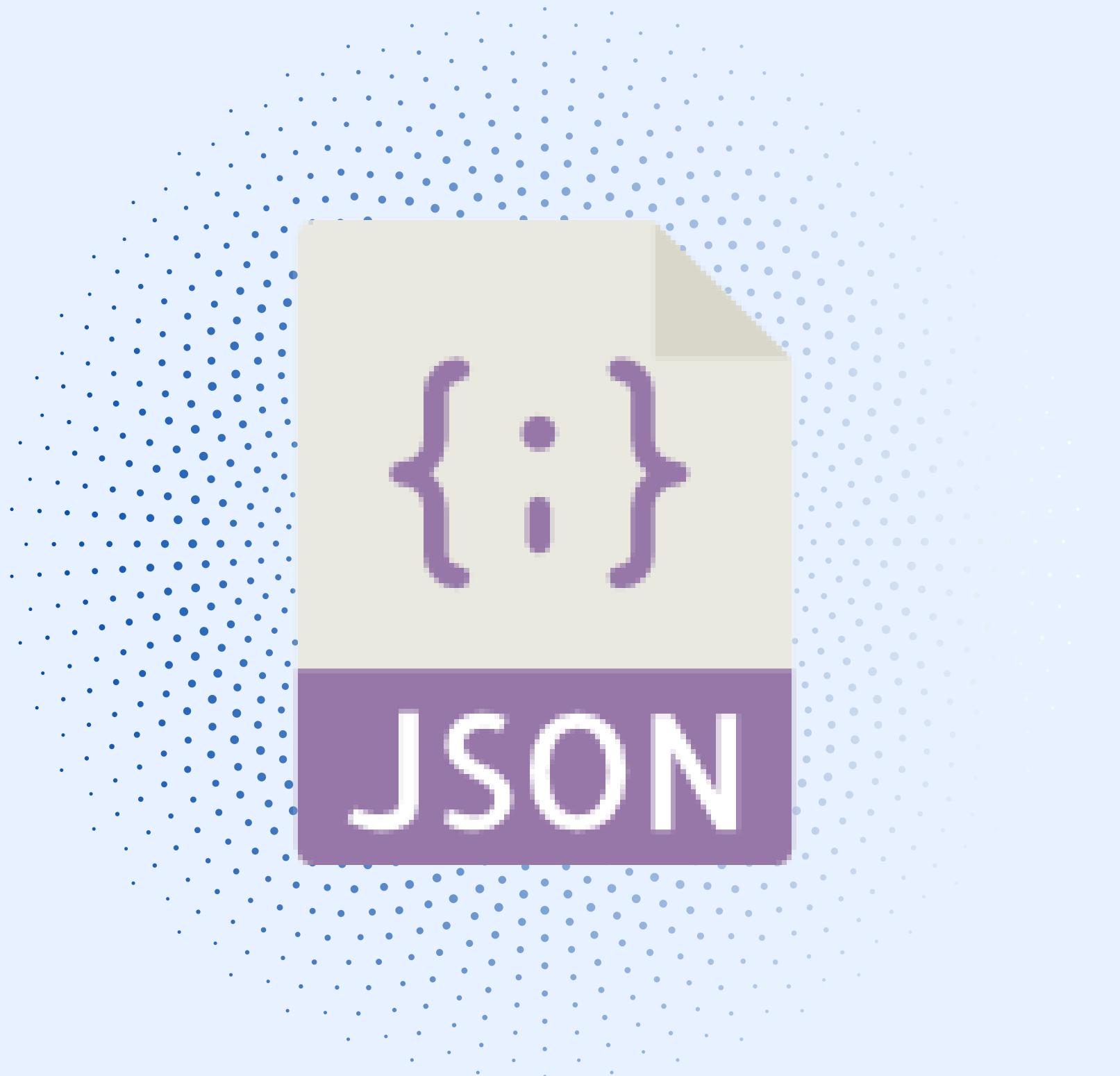
Datasets



Datasets



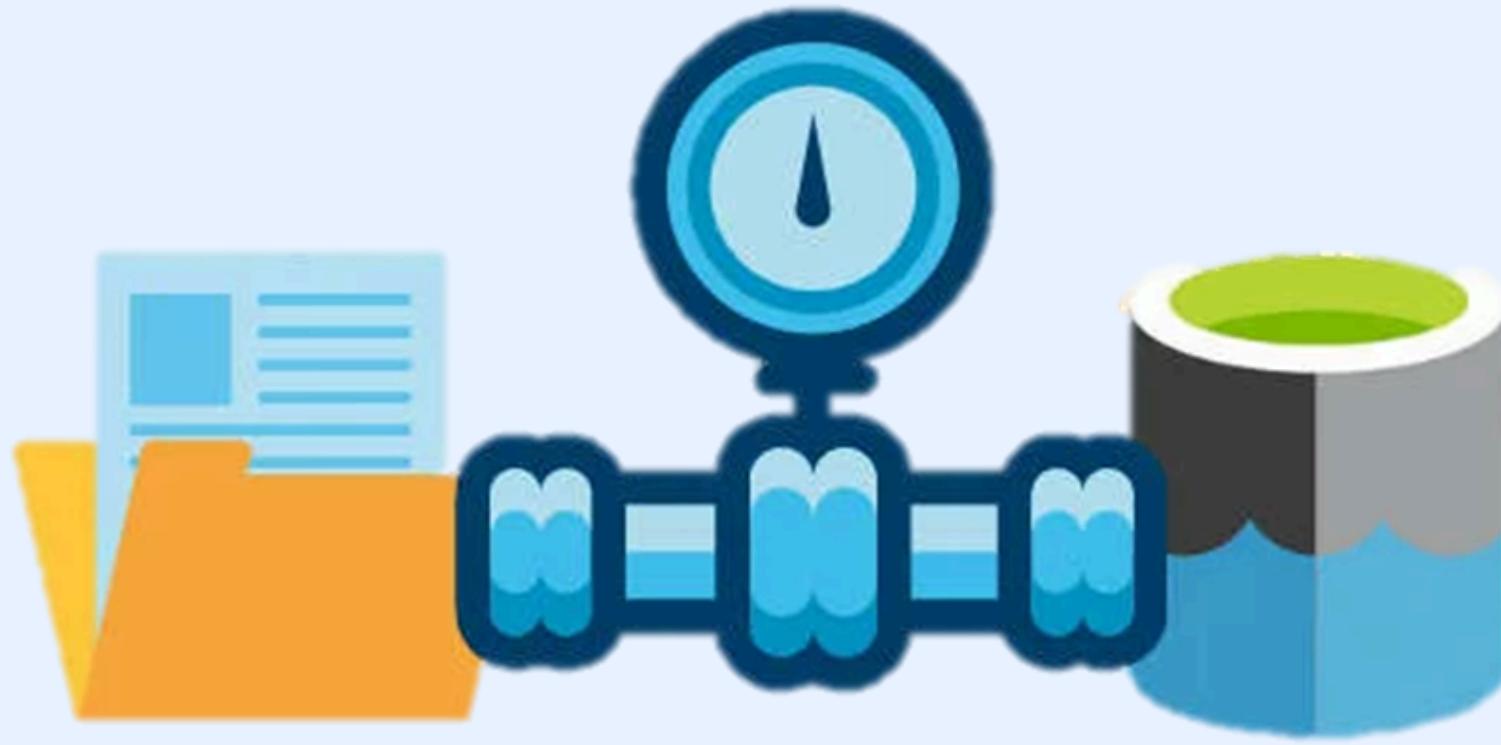
Datasets



Pipelines

Pipelines creados:

- Ingest SQL
- Ingest Mongo
- Transformaciones JSON
- Unificación final



Desarrollo del Proyecto



DataFlow SQL + CSV

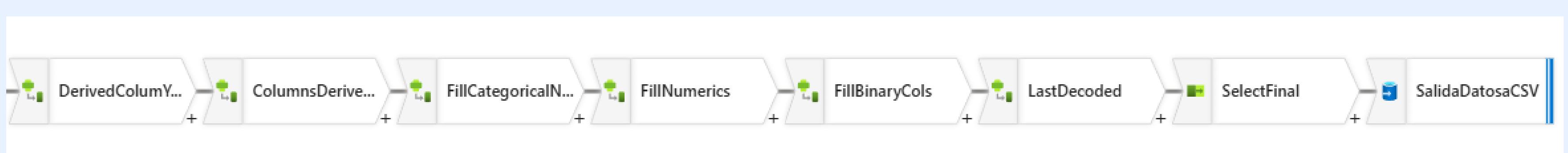
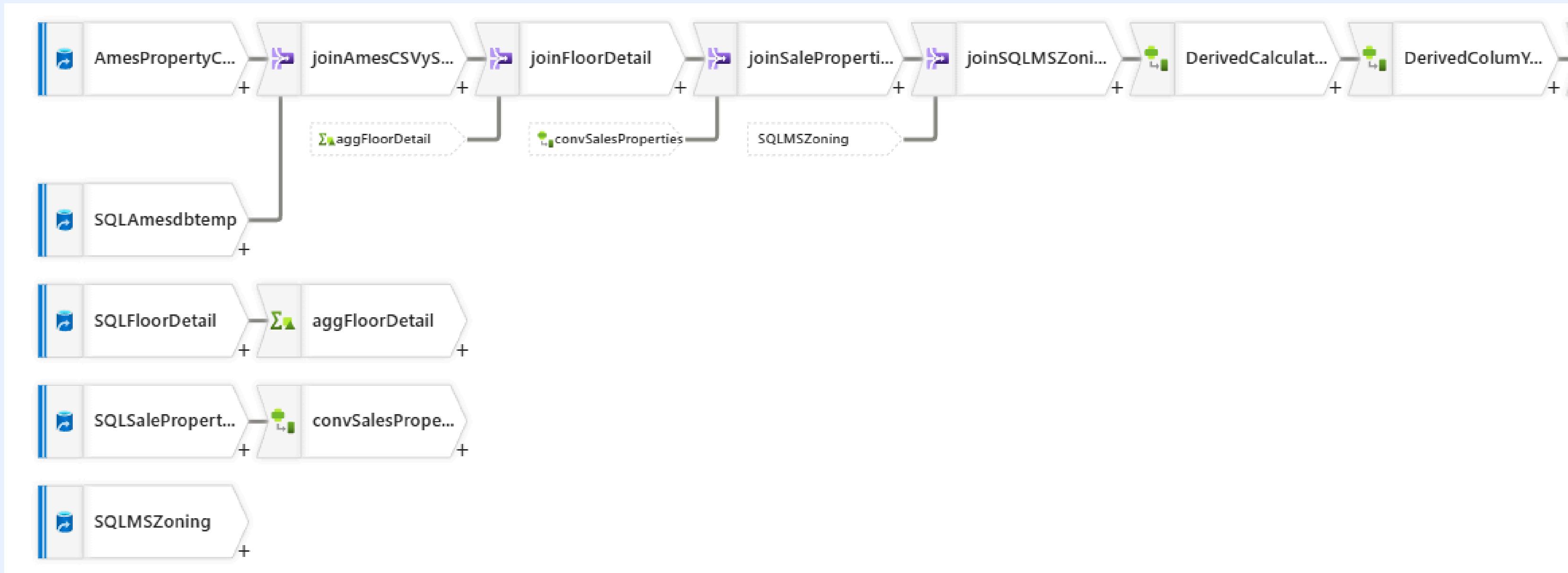
Incluyó:

- Uniones por PID
- Cálculo de GrLivArea
- Agregación de baños/dormitorios
- Limpieza NA / 0
- Fechas MoSold / YrSold
- Normalización de columnas

Ejemplos de transformaciones:

- GrLivArea
- YearRemodAdd corregido
- Suma de baños y dormitorios





DataFlow Mongo

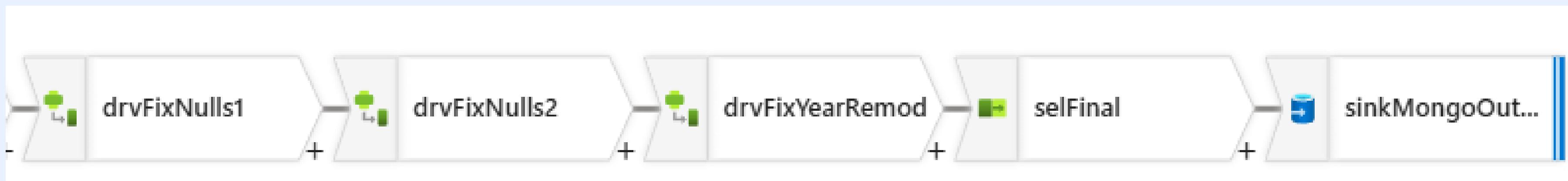
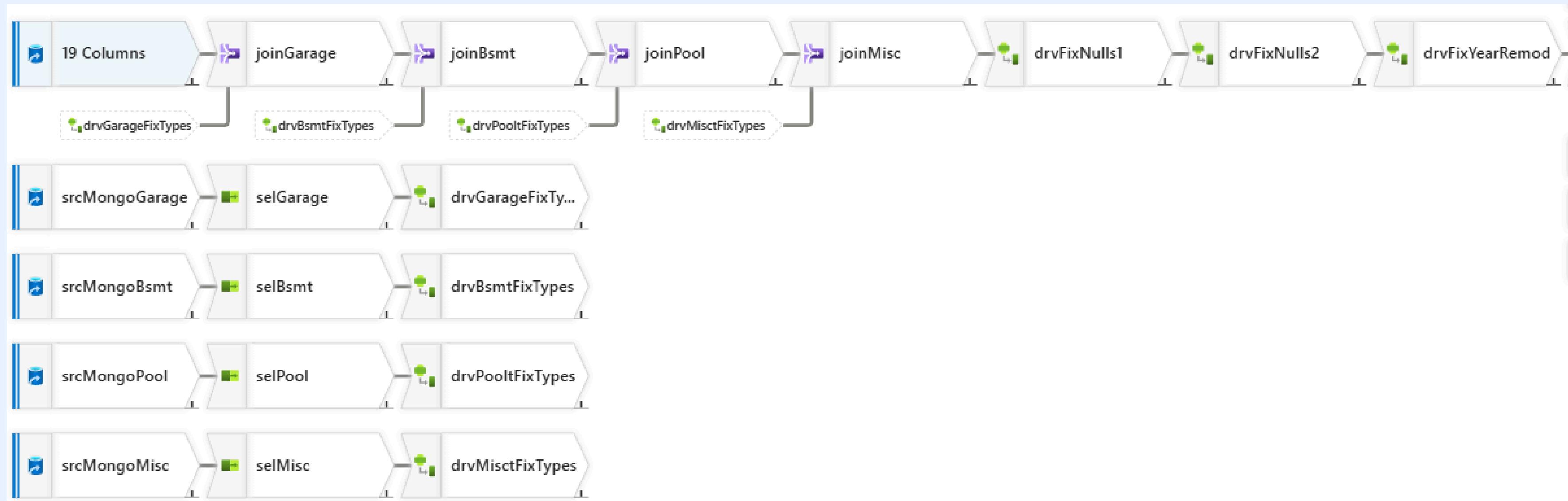
Consolidación de Garage + Bsmt + Pool + Misc:

- Renombrado
- Cast de tipos
- Nulos NA / 0
- Joins por PID

Procesamiento individual por colección:

- Garage
- Bsmt
- Pool
- Misc



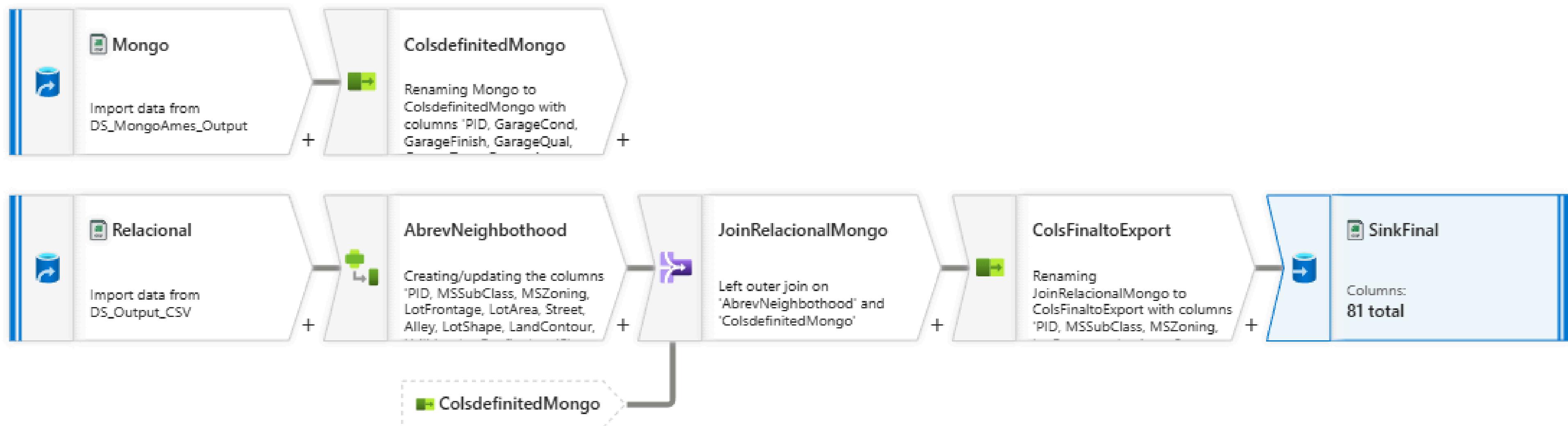


Unificación Final

Lo que ocurre aquí:

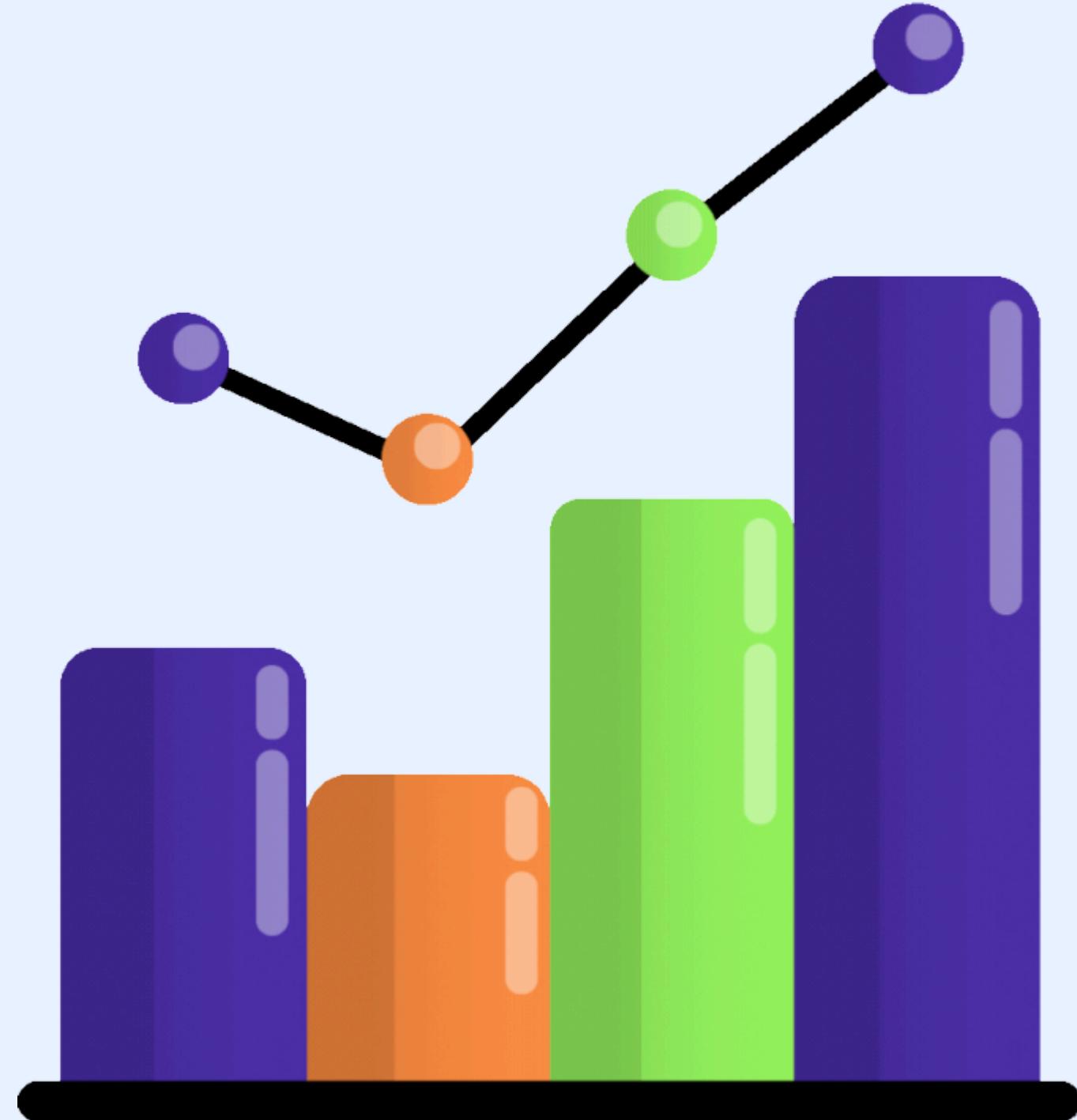
- Se unen relacional + mongo
- Se aplican abreviaciones (Neighborhood, Condition...)
- Se preparan las 81 columnas





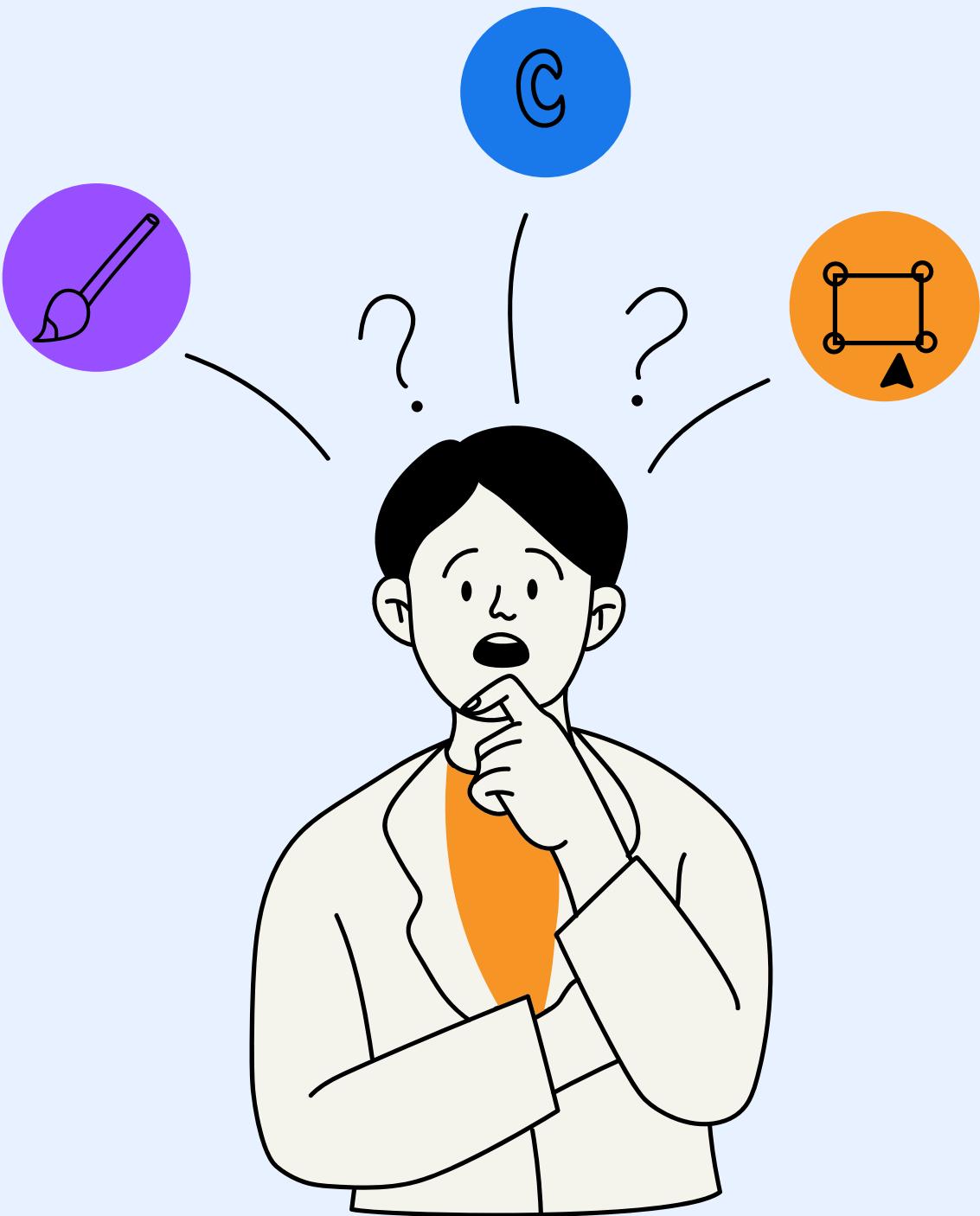
Interpretación de resultados

- 2930 propiedades consolidadas sin pérdida de registros
- 81 columnas finales limpias, tipadas y ordenadas
- Nulos corregidos (NA / 0) → dataset 100% consistente
- Variables derivadas clave (GrLivArea, baños totales, etc.)
- Integración completa de SQL + CSV + Mongo
- Dataset listo para modelos de ML y análisis avanzado



Justificación de decisiones técnicas:

- Usamos LEFT JOIN para no perder propiedades sin datos en Mongo.
- Separamos SQL y Mongo en ramas distintas para facilitar desarrollo paralelo.
- Elegimos DataFlows en lugar de CopyActivity porque requeríamos cálculos y derivaciones complejas.



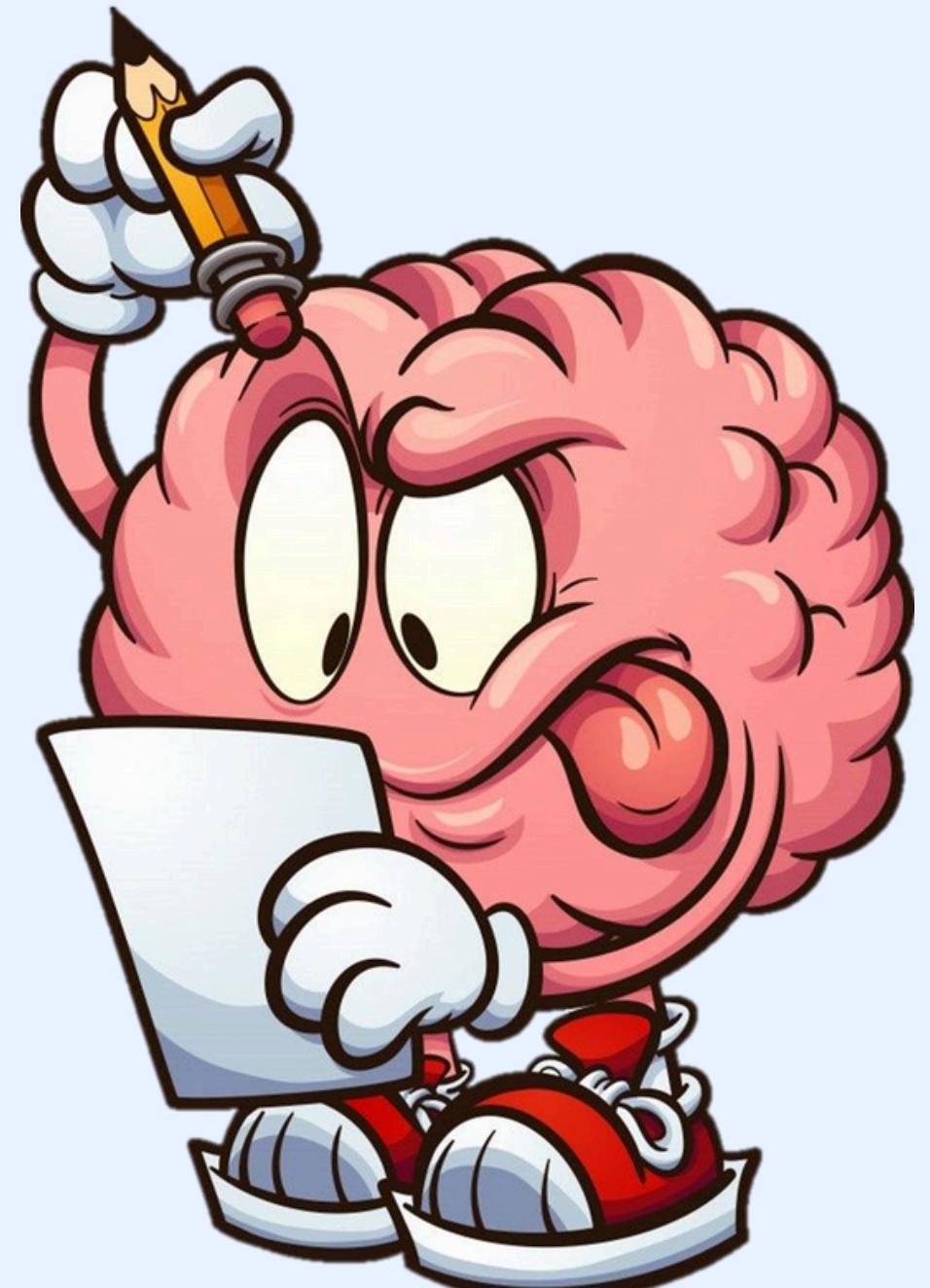
Lecciones aprendidas

Aprendizajes:

- Separar SQL y Mongo simplificó el desarrollo
- DataFlows permiten transformaciones muy potentes
- NA / 0 fue clave para evitar errores
- Joins por PID mantuvieron integridad
- Ordenar las 81 columnas fue la parte más delicada

Superación de dificultades:

- Versionamos el proyecto con ramas para evitar conflictos.
- Probamos cada dataflow por separado antes de unificar
- Validamos conteo de columnas y tipos en cada etapa



Conclusiones

ETL completo y profesional

- ✓ Multi-fuente
- ✓ Transformaciones completas
- ✓ 81 columnas finales
- ✓ Pipeline reproducible



*!Gracias por su
atención;*