Módulo 3

Sesión N° 5

ACTIVIDAD:

Auditoría y Optimización de un Pipeline de Data Wrangling

Objetivo: Procesar conjuntos de datos, abordando limpieza, estandarización, enriquecimiento, automatización y monitoreo, con documentación detallada y justificación de decisiones técnicas.

Contexto

Has recibido un pipeline (conjunto de scripts o notebooks) que integra, limpia y transforma datos de clientes provenientes de cinco fuentes distintas: archivos .csv, .xlsx y una base de datos SQL. Los procesos actuales presentan lentitud, errores frecuentes y falta de reproducibilidad.
Tu objetivo es auditar el pipeline, proponer mejoras concretas y documentar un nuevo flujo robusto y eficiente utilizando las buenas prácticas de data wrangling estudiadas.

Entregable:

- Un notebook o script con el código y comentarios.
- Un diagrama de flujo del pipeline (puede ser digital o foto de un esquema a mano).
- Un informe breve (1 – 2 páginas) justificando tus decisiones y mejoras.
- Formato: grupal.

Requerimientos:

1. Analiza los archivos y el pipeline actual
   Detecta cuellos de botella, errores recurrentes y cualquier aspecto que comprometa la calidad, velocidad o reproducibilidad del proceso.
2. Propón al menos 3 mejoras concretas
   Explica cómo implementarías mejoras en eficiencia (por ejemplo, usando chunking), en reproducibilidad (versionado, seeds), y en calidad (validaciones, logs).
3. Rediseña el pipeline
   Dibuja (puede ser a mano y escaneado) un diagrama de flujo que muestre tu nuevo proceso, destacando las etapas principales y los cambios clave.
4. Implementa el flujo de wrangling
   Desarrolla el código (en Python/pandas y/o SQL) necesario para las etapas principales de tu pipeline mejorado. Incluye comentarios explicativos.
5. Documenta y justifica cada etapa
   Explica por qué haces cada transformación, los riesgos de omitirla, y cómo mides el éxito.
6. Prueba la auditoría y el rollback
   Describe cómo podrías auditar la calidad de los datos finales y revertir cambios ante errores críticos.