Prueba Técnica – Ingeniero de Datos

Objetivo:

Evaluar la capacidad del candidato para:

- Identificar problemas de calidad de datos en un conjunto de datos.
- Aplicar técnicas de limpieza, validación y auditoría de datos.
- Documentar hallazgos y proponer estrategias de mejora.
- Utilizar herramientas para depuración.
- Aplicar principios de gobierno de datos.

Tiempo estimado para resolución:

- 4 a 6 horas

Entregables esperados:

1. Informe técnico con:
    - Análisis exploratorio y resumen de hallazgos de calidad.
    - Validaciones realizadas y problemas detectados.
    - Estrategia de limpieza y supuestos adoptados.
    - Métricas de calidad antes y después de la limpieza.
2. Script en Python o notebook que:
    - Ingesta el archivo dataset_hospital.json.
    - Realiza las limpiezas necesarias.
    - Aplica validaciones cruzadas entre tablas.
    - Exporta las versiones limpias de las tablas.
3. (Opcional pero valorado) Implementación de pruebas automáticas para validar integridad de datos.

Actividades a realizar:

Parte 1 – Análisis de calidad de datos (Exploración)

1. Identifica y describe los principales problemas de calidad en las tablas pacientes y citas_medicas.

Parte 2 – Limpieza y validación

2. Aplica un proceso de limpieza que resuelva los problemas hallados en el punto anterior. Justifica tus decisiones.
3. Aplica validaciones cruzadas entre campos.

Parte 3 – Indicadores de calidad y documentación

4. Crea un resumen con indicadores de calidad de datos antes y después de la limpieza.
5. Documenta claramente:
    - Supuestos adoptados durante la limpieza.
    - Reglas de validación implementadas.
    - Recomendaciones de mejora para asegurar la calidad futura de los datos.

Bonus (opcional, se valora positivamente)

- Implementar pruebas automáticas con pytest, great_expectations o cualquier otro framework de validación de datos.
- Simular una migración de los datos limpios a una estructura destino (por ejemplo, un Data Warehouse).

Requisitos técnicos

- Python.
- Puede usar Jupyter Notebook o script .py
- Enviar en un archivo comprimido .zip que contenga:
    - Informe en PDF
    - Script o notebook
    - Datasets limpios exportados.