Evaluación Ingeniero de Datos

Lea la prueba completa y desarrolle todo lo que pueda. Documente todas sus consideraciones en la resolución de los problemas planteados o alternativas de solución.

1. **Carga y cruce de información de personas:**

Desarrollar un flujograma, donde se recibe un archivo Excel que contiene: nombre completo de persona (de la forma "Juan Antonio Garcia Hernandez"), número de DUI y otros campos demográficos.

Los nombres deben ser clasificados como: primer nombre, segundo nombre, tercer nombre, primer apellido, segundo apellido, apellido de casada.

Se requiere un reporte Excel donde estén clasificados los registros del archivo original indicando si el registro fue encontrado en alguna base interna, ya sea como cliente de banco, como persona en lista de control, ambos casos, o indicar si no encontró en ninguna (registro nuevo). Realice operaciones de limpieza de cadenas para maximizar la efectividad del cruce masivo.

Cartera de clientes del banco: información general de los clientes.

Lista de control: información de personas con las que el banco no debería establecer relación.

Consideraciones:

\* Plantear que herramientas se utilizarían.

\* El nombre completo puede tener caracteres especiales, espacios en blanco, números, mayúsculas, minúsculas.

\* Usar expresiones regulares de ser necesario.

\* Indicar si se detecta alguna inconsistencia. Es decir, comente si existen escenarios ambiguos en el resultado del proceso que deban ser verificados manualmente.

**2. Realizar el siguiente ETL:**

Realizar un ETL que cargue en una tabla el archivo "fuente.xlsx"

Se requiere un reporte Excel donde estén clasificados las personas basadas en la cartera donde encontró (tabla clientes, tabla lista\_control) o indicar si no encontró en ninguna cartera.

Consideraciones:

\* Utilizar las tablas de la base sqlite proporcionada (ev\_ing\_dat.db)

\* Usar python, spark, sql, Power BI, según considere necesario.

\* Documentar la solución, los pasos del proceso.

\* Usa bitácora para el proceso (log y tabla de bitácora)

\* Versionar la solución usando git.

\* Calendarizar el programa para que se ejecute todos los 1° de cada mes a las 10 a.m.

**Reporte esperado:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Nombre completo** | **Primer nombre** | **Segundo nombre** | **Tercer nombre** | **Primer apellido** | **Segundo apellido** | **Apellido de casada** | **Clasificación** |
| Juan Antonio Garcia Hernandez | Juan | Antonio |  | Garcia | Hernandez |  | Lista de control |
| María Lopez de Hernandez | María |  |  | López |  | De Hernandez | Cartera de Clientes |
| Sofía Gomez | Sofía |  |  | Gomez |  |  | No se encontró en las BD |

1. **Carga de listas de control ONU y OFAC.**

Se tienen dos URLs y se requiere un proceso que cargue esta información en diferentes tablas relacionales según considere.

ONU: <https://scsanctions.un.org/resources/xml/en/consolidated.xml>

OFAC: <https://www.treasury.gov/ofac/downloads/sdn.xml>

* Analizar la estructura de los archivos xml y considerar que estos son actualizados de forma periódica.
* Crear las tablas y relaciones que considere necesario.
* Utilizar Python y sqlite.
* Compartir evidencias del funcionamiento, scripts usados (Python, sql, etc)
* Versionar el proceso usando git