CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Data Engineering

Módulo Nº2 - Unidad Nº 1

Procesamiento de datos

Objetivos

- Implementar técnicas para procesamiento para limpiar y estandarizar los datos.
- Aplicar técnicas de procesamiento para enriquecer los datos y obtener información relevante.

Consigna

En la consigna anterior, has guardado datos crudos en formato Parquet.
Ahora, deberás leer esos archivos y aplicar tareas de procesamiento o transformación de datos con Pandas. Esas tareas de procesamiento pueden ser:

- Eliminación de duplicados
- Eliminación o reemplazo de nulos
- Conversión de tipos de datos de columnas
- Renombrar columnas
- Formatear columnas de tipo fecha
- Crear nuevas columnas a partir de alguna lógica (Por ejemplo, una columna booleana que indique si una temperatura está por arriba de un límite)
- Cruzar dataframes usando JOINS
- Aplicar agregaciones por medio de GROUP BY y funciones como MAX, MIN, AVG, etc.

Deberás realizar al menos 4 tareas de transformación.

El resultado del procesamiento debe ser guardado en una, o varias, tablas de una base de datos OLAP, Cabe aclarar que tenés que realizar la creación de las tablas desde Python con la librería SQLAlchemy.

Tenés que usar una base de datos Postgres de Aiven.

Podes realizar la consigna sobre el mismo script o notebook, o bien sobre uno nuevo. Tené en cuenta que en la entrega, tenés que presentar todos los scripts.

Formato de presentación:

- Jupyter notebook (archivo .ipynb) o archivo Python (.py).
- Renombrar los archivos a entregar con nombre y apellido, seguido del nombre que consideren necesario. Por ejemplo: GuidoFranco_TP3.py.
- El nombre de las tablas a crear en la base de datos debe tener como prefijo tu nombre y apellido, por ejemplo: guidofranco_estaciones, guidofranco_valores, etc.
- El programa será ejecutado por el tutor desde una plataforma como Google Colab, o bien de forma local.

Criterios de evaluación

1. Calidad del código y presentación
    a. El código debe estar bien estructurado y seguir buenas prácticas de programación en Python.
    b. El funcionamiento del código debe estar documentado de forma clara y concisa.
    c. El trabajo debe ser entregado de manera ordenada y legible.
2. Implementación de técnicas de limpieza y enriquecimiento de datos.
3. Almacenamiento de los resultados en base de datos OLAP.

Centro de e-Learning SCEU UTN - BA. Medrano 951 2do piso
(1179) // Tel. +54 11 7078- 8073 / Fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Anexo

¿Como crear una tabla en una base de datos con Python?

```python
from sqlalchemy import create_engine

# Las credenciales no deben estar en el código fuente!!!
user = "avnadmin"
pwd = "AVNS_Kg_VRFPdd_nM5YwXo4n"
host = "pg-3730f3e-datadev.aivencloud.com"
port = 15191
dbname = "defaultdb"

# Crear string de conexión a partir de los datos (user, pwd, host, port, dbname, etc)
conn_string = f"postgresql://{user}:{pwd}@{host}:{port}/{dbname}?sslmode=require"
engine = create_engine(conn_string)

with engine.begin() as conn:
    conn.execute("CREATE TABLE IF NOT EXISTS ....;")
```

¿Cómo cargar datos a una tabla con Pandas?

Se recomienda usar el método to_sql de la librería Pandas para cargar los datos en la base de datos OLAP. Tenés que prestar atención a los parámetros **“if_exists”** y **“method”**.

- **“if_exists”** debería tener el valor **“append”**, no se recomienda usar el valor “replace” ya que pandas elimina la tabla y la vuelve a crear con otro esquema.
- El parámetro **“method”** debería el valor **“multi”** para que incluya varios registros en la misma sentencia INSERT y así acelere el proceso.

Se sugiere revisar el video sobre Almacenamiento de Data Warehouse.