Proyecto Final – Ciencia de Datos en Python

Proyecto: Ingeniería de Datos con Python
Tema: Python, Pandas, SQL, ETL, AWS
Fecha y Hora de Entrega: 10/04/2022 25:55
Formato de Entrega: Archivos de Construcción y Video.
Grupo: Grupos de 2 o 3 personas
Calificación: Presentación por medio de Vídeo

ATENCIÓN: Todos los proyectos consumen tiempo, así que trate de empezar lo más pronto que pueda. Recuerde que el proyecto es en GRUPOS, es decir, no puede trabajar junto a otros compañeros de otros grupos.

DESCROPCIÓN: Para este proyecto usted deberá desarrollar un pipeline de ingeniería utilizando Python, SQL y AWS como herramientas de desarrollo, su proyecto debe contar con los siguientes componentes:

- Scope: Scope del Proyecto y descripción de fuentes de información,
- Exploración: Exploración de la data para definir el modelo de datos,
- Modelo de datos: Deberá definer el modelo de datos que usará para su Proyecto ya sea un DW o un DL,
- Procesamiento: deberá definer todos su Código como un conjunto de scripts en Python, los cuales extraigan, transofrmen y cargen la data.
- Analitica: Deberá plantear al menos 5 preguntas de análisis que puedan ser resueltas con la estructura que definio.

- Reporte: Un documento de Markdown el cual incluya todos los elementos solicitados anteriormente.

DETALLES TECNICOS: A continuación se describen los detalles técnicos mínimos que su projecto debe cumplir:

- Como fuentes de infomrmación deberá usilizar al menos una base de datos montada en RDS y dos archivos externos almacenados en S3.

- El procesamiento puede realizarlo en un maquina local o en una instancia de EC2 corriendo Python.

- La salida deberá ser sobre Redshift si su salida es en DW o sobre S3 si su salida es un DL.

- El Notebook le servirá para constuir todos los procedimientos, sin embargo su proyecto debe correr con scripts de Python.

- Notar que no puede usar SQL para hacer la construcción de ningína estructura salvo para leer de tablas almacenadas en las bases de datos es decir SELECT * FROM tabla.

Puede utilizar las fuentes de información y diseño que deseen para resolver el problema planteado, a continuación se le comparten algunas fuentes de datos:

- Google Datasets: https://datasetsearch.research.google.com
- Kaggle Datasets: https://www.kaggle.com/datasets
- Github Datasets: https://github.com/awesomedata/awesome-public-datasets
- Data.Gov: https://catalog.data.gov/dataset
- 21 Free Datasources: https://www.dataquest.io/blog/free-datasets-for-projects/
- KdNuggets: https://www.kdnuggets.com/datasets/index.html
- ICS Datasets: https://archive.ics.uci.edu/ml/datasets.php
- Reddit Dataset: https://www.reddit.com/r/datasets/

ENTREGA: Como entrega deberá publicar todos los archivos utilizados por medio de un link de Git, incluyendo el reporte de desarrollo, los notebooks utiilzados, los y los scripts finales. Adicinalmente deberá hacer un video de 5 a 7 minitos máximo donde explique todos los pasos que realizó para desarrollar su proyecto, es decir describir todos los elementos de su proyecto.