NeuralWorks

Challenge Data Engineer

Instrucciones

- Debes entregar tu solución en un repositorio GitHub
- En el repositorio deben estar todos los archivos utilizados para la resolución de tu desafío.
- La solución debe estar implementada en un notebook .ipynb utilizando python 3, indicando claramente la pregunta que estás resolviendo. No serán revisados otros lenguajes como R o similar.
- Recuerda que no estamos en tu cabeza! Escribe los supuestos que estás asumiendo.
- Para este desafío te recomendamos que describas claramente cómo mejorar cada parte de tu ejercicio en caso de que tenga opción de mejora.
- Debes enviar el link al repositorio vía mail a loreto@neuralworks.cl contestando el correo en el que se te envió este enunciado.

Problema

NeuralWorks en su constante búsqueda por realizar proyectos entretenidos ha decidido explorar los patrones de movimientos de las personas para lograr una ciudad libre de autos. Hemos conseguido la data de diferentes fuentes para entender el origen-destino de las personas para poder entender cómo eficientar el transporte público, locaciones de puntos estratégicos, turnos entre personas con viajes similares, etc.

Como Data Engineer en este proyecto te ha tocado apoyar con la creación de turnos entre personas con viajes similares. Para esto hemos adjuntado un CSV con una muestra de la data de los viajes realizados en 3 regiones. La Data Scientist del equipo necesita tu ayuda para poder hacer un análisis profundo de los datos y construir un modelo de machine learning. Te ha pedido cumplir con los siguientes requisitos:

1. Procesos automatizados para ingerir y almacenar los datos bajo demanda
    a. Los viajes que son similares en términos de origen, destino y hora del día deben agruparse. Describa el enfoque que utilizó para agregar viajes similares.
2. Un servicio que es capaz de proporcionar la siguiente funcionalidad:
    a. Devuelve el promedio semanal de la cantidad de viajes para un área definida por un bounding box y la región
    b. Informar sobre el estado de la ingesta de datos sin utilizar una solución de polling
3. La solución debe ser escalable a 100 millones de entradas. Se recomienda simplificar los datos mediante un modelo de datos. Agregue pruebas de que la solución es escalable.
4. La solución debe estar escrita en Python usando una base de datos SQL
5. Puntos de bonificación si incluye su solución en contenedores y si dibuja cómo configuraría la aplicación en GCP

Evaluación

Se tomará en cuenta para la evaluación:

- Elegancia y creatividad de la solución
- Orden del código
- Diseño de la solución
- Documentación