YGROUP: Desafío Ingeniero de Datos

Introducción

En esta parte, la dividiremos en dos fases. La primera consiste en un proyecto para hacer en casa. Su objetivo es validar tus conocimientos y capacidad con tecnologías de Data Analytics así como en el desarrollo de software y sus buenas prácticas. Para realizarlo tienes 48hs desde que se te envía este documento.

La segunda fase consiste en una entrevista oral, que se basa en el código entregado e intenta evaluar tus habilidades en el escalado de este tipo de soluciones, así como tu capacidad de resolución frente a un potencial cliente.

La idea es que tengas en cuenta el desarrollo de la primera fase con respecto a la segunda.

Enunciado

Para la primera fase lo que queremos es que usando el conjunto de datos de BIXI Montreal (https://www.kaggle.com/aubertsigouin/biximtl), crees el código necesario para obtener los resultados a los análisis pedidos. Para completar satisfactoriamente debes realizar al menos los puntos marcados como obligatorios. Se puede desarrollar en cualquier lenguaje que te sea cómodo.

Requerimientos y restricciones

Es necesario:

- Escribir un README que explique cómo ejecutar la solución.
- Proveer una forma automatizada para resolver las dependencias necesarias
- Proveer un repositorio GIT sobre el que se está realizando el trabajo. Si el repositorio nos lo envías en un fichero ZIP, entonces se tomará éste como el de entrega. Por el contrario, si nos envías un link a un repositorio público, se tomará como para la entrega el repositorio en el estado que esté a la hora de entrega, en el Branch principal que nos indiques.

No se puede:

- Usar más de 5 librerías externas (sin contar las dependencias que ellas mismas tengan)
- Tener requerimientos de infraestructura (e.j.: colas de mensajes, servidores web, bases de datos relacionales o no relacionales)

Excepciones al último punto son:

- Que todo esté hecho en containers se orquestre a través de Docker compose. Pero aún en ese caso el trabajo no se considerará si ocupa demasiada memoria, tarda mucho en ejecutarse o tiene errores al hacerlo
- Que la base de datos sea embebida (e.j.: SQLite, H2). Pero el mismo código debe proveer las facilidades para crearla y bajo ningún punto de vista debe el usuario interactuar con la base de datos, en ninguna de sus fases (creación, población/actualización, borrado)

Presentación

No hay restricción para la presentación de los datos y resultados. Puede ser tanto a través de dashboards, como exportación a ficheros o por consola. La implementación debe estar justificada y se deben mantener las restricciones anteriores.

Puntos a realizar

Obligatorios

- Histograma de tiempos de viaje para un año dado
- Listado del Top N de estaciones más utilizadas para un año dado. Dividirlo en:
    - Estaciones de salida
    - Estaciones de llegada
    - En general
- Listado del Top N de viajes más comunes para un año dado. Donde un viaje se define por su estación de salida y de llegada
- Identificación de horas punta para un año determinado sin tener en cuenta el día. Es decir, si es día de semana, fin de semana, festivo o temporada del año.

Deseables

- Pruebas unitarias sobre los distintos módulos/objetos/funciones
- Comparación de utilización del sistema entre dos años cualesquiera. La utilización del sistema se puede medir como:
    - Cantidad de viajes totales
    - Tiempo total de utilización del sistema
    - Cantidad de viajes por estaciones/bicicletas disponibles
- Capacidad instalada total (suma de la capacidad total de cada estación)
- Cambio en la capacidad instalada entre dos años puntuales

Ideales

- Ampliación de la cobertura de la red entre dos años puntuales. La misma se puede medir como el área total que generan las estaciones
- Comparación de densidad de la red para un par de años puntuales. La densidad de la red se mide como el área que abarcan todas las estaciones, dividida la cantidad de estaciones
- Velocidad promedio de los ciclistas para un año determinado
- Cantidad de bicicletas totales para un momento dado. Considerando la misma como la cantidad de bicicletas que hay en todas las estaciones activas para ese momento, más todos los viajes que se estén realizando