Data Engineer Challenge

Ene 2024

1. Descripción del problema:

Se desea construir un sistema de análisis y visualización de los datos de uso del sistema de transporte público de bicicletas en la ciudad de Buenos Aires. Para esto, se utilizará un conjunto de datos disponible en data.buenosaires.gob.ar.

2. Objetivo:

Diseñar y alimentar una base de datos relacional con datos acerca de los viajes realizados a través del sistema de transporte público de bicicletas de la ciudad de Buenos Aires a lo largo del 2023, para luego, mediante una herramienta de visualización, se pueda graficar el volumen de viajes realizados entre las distintas estaciones que conforman la red.

3. Requerimientos técnicos:

- Descargar los conjuntos de datos de los viajes realizados en bicicletas públicas durante 2023 en Buenos Aires y de los usuarios de sistema desde data.buenosaires.gob.ar/dataset/bicicletas-publicas, y de las estaciones públicas de bicicletas desde data.buenosaires.gob.ar/dataset/estaciones-bicicletas-publicas
- Crear una base de datos relacional y diseñar las tablas necesarias para almacenar los datos de los conjuntos de datos.
- Crear un proceso de ETL, en cualquier lenguaje, que realice la limpieza y preprocesamiento de los datos y los inserte en la base de datos relacional.
- Utilizar una herramienta de visualización para que se conecte a la base de datos relacional y muestre el volumen de viajes realizados entre las distintas estaciones a lo largo del año.
- Crear consultas SQL que permitan obtener información relevante para el análisis de tendencias de uso del sistema público de bicicletas en la ciudad de Buenos Aires, teniendo en cuenta:
    - Comunas de las que más viajes parten
    - Comunas que más viajes reciben
    - Relación entre la edad de los usuarios y la cantidad de viajes que realizaron.
    - Relación entre la cantidad de viajes realizados y la fecha de alta de los usuarios.
    - Relación entre la duración de los viajes y el modelo de bicicleta.
    - Usuarios más activos durante 2023.
- Subir el proyecto a GitHub, GitLab o cualquier otro repositorio público o privado, con el README correspondiente acerca de cómo realizar las pruebas.

4. Puntos extras:

- Utilizar técnicas de optimización de consultas SQL para mejorar el rendimiento del sistema.

- Utilizar técnicas de visualización avanzadas para crear visualizaciones interactivas y atractivas que permitan a los usuarios explorar y analizar los datos de manera más eficiente.
- Utilizar herramientas de análisis de big data para analizar conjuntos de datos más grandes o para realizar análisis más complejos.