Instituto Tecnológico de Costa Rica

Programa de Ciencia de los Datos - Módulo Big Data

Tarea #1

- Esteban Sáenz Villalobos (**esaenz7@gmail.com**)

- Entrega: 08 de agosto 2021, 23:00.
- Observaciones: Trabajo elaborado desde Google Colab. Ejecutar cada celda de código de forma secuencial.

Instrucciones

```
1- Para cargar el contenedor con los recursos necesarios, ejecute los
archivos build_image.sh y run_image.sh.
2- Para ejecutar el programa principal se debe aplicar el siguiente
comando:
    #spark-submit programaestudiante.py ciclista.csv ruta.csv
actividad.csv
3- Para ejecutar las pruebas del programa se debe aplicar el
siguiente comando:
    #python -m pytest -v
4- Para ejecutar las instrucciones 2 y 3 de forma automática, ejecute
el archivo run.sh.
```

- Nota 1: El código fuente en cada archivo cuenta con comentarios detallados que explican la lógica del programa.
- Nota 2: Se inluye un jupyter notebook de Google Colab con todo el código necesario, como complemento.
- Nota 3: El repositorio completo de la tarea se encuentra también en el siguiente enlace https://github.com/esaenz7/bigdataclass/tree/main/tarea1.

Detalles del trabajo

```
* Los archivos de datos están compuestos de la siguiente forma:
    1. Ciclista (50 registros)
        a. Cédula (numérico)
        b. Nombre completo (string)
        c. Provincia (San José, Alajuela, etc. Expresado como string)
    2. Ruta (15 registros)
        a. Código de ruta (identificador numérico)
        b. Nombre ruta (string)
        c. Kilómetros (numérico / decimal)
    3. Actividad (300 registros)
        a. Código de ruta
        b. Cédula
        c. Fecha (Formato YYYY-MM-DD)
```

```
* El programa consta de los siguientes archivos:
    1. procesamientodatos.py (lógica de procesamiento).
    2. programaestudiante.py (programa principal).
    3. conftest.py (contexto para las pruebas).
    4. test_programaestudiante.py (ejecución de pruebas).
* La aplicación se ejecuta por etapas (stage) ejecutadas cada una por
una función en específico.
    1. Stage1: carga de datos.
    2. Stage2: unión de los datos.
    3. Stage3: agregación de los datos.
    4. Stage4: presentación de los datos (resultados finales).
    5. Stage5: almacenamiento de los datos.
```
