Este script en PySpark realiza un análisis exploratorio de datos (EDA) sobre un conjunto de datos de cáncer infantil en la ciudad de Pereira en Colombia. Utiliza Spark DataFrames para la manipulación y análisis de datos.

**Funciones:**

**Carga de datos:** Importa un archivo CSV desde HDFS.

**Exploración de datos:** Muestra la estructura y estadísticas básicas del conjunto de datos.

**Limpieza de datos:** Elimina filas con valores nulos.

**Transformaciones:** Agrupa los datos por nacionalidad y calcula la edad promedio.

**Consultas SQL:**

Filtra niños de nacionalidad venezolana.

Filtra niños de estrato 1.

Filtra niños de sexo masculino.

Filtra niños con tipo de salud subsidiada.

**Instrucciones para ejecutar:**

Tener instalado: Apache Spark, Hadoop con HDFS, Python con PySpark.

1. Antes de ejecutar el código, el archivo **dj7a-2nxi.csv** debe estar en **HDFS**.

2. Guarda el script como **bigdatat3**.**py** y ejecútalo.

**Analiza los resultados:**

1. Se imprimirá el esquema de los datos.

2. Se mostrarán las primeras filas del dataset.

3. Se realizarán transformaciones y consultas SQL sobre los datos.

4. Se realiza un análisis de los datos obtenidos.