Practica 3. Cargar y manipular datos

Nombre: _____ **Fecha**: _____

Apellido Paterno Apellido Materno Nombre(s)

Calificación: _____

Objetivo: El participante pondrá en practica las herramientas basicas de spark RDD, DataFrame, Ingesta

Instrucciones: *Desarrollar los siguientes ejercicios*

Cargar datos de vuelos

En este ejercicio, cargará algunos datos de vuelos de aerolíneas desde un archivo CSV.

url

https://raw.githubusercontent.com/omarmendoza564/datos/main/datos/flights-larger.csv

Notas sobre el formato CSV:

Los campos están separados por una coma y los datos faltantes se indican con la cadena 'NA'.

Diccionario de datos:

- month,dayofmonth,dayofweek,carrier,flight,origin,mile,depart,duration,delay
- month - mes (entero entre 1 y 12)
- dayofmonth - día del mes (entero entre 1 y 31)
- dayofweek - día de la semana (entero; 1 = lunes y 7 = domingo)
- carrier - aerolinea (código IATA)
- origin - aeropuerto de origen (código IATA)
- mile - distancia (millas)
- depart - hora de salida (hora decimal)
- duration - duración esperada (minutos)
- delay - retraso (minutos)

2. Guardar el archivo en HDFS
3. Cargar el archivo en un notebook de jupyter
4. Listar del dataframe
    1) Numero de registros
    2) Estructura
    3) Nombre de las Columnas
    4) Tipos de datos
    5) Ver los primeros 20 registros
    6) Descricipcion Estadistica
    7) Descripcion estadistica de una sola columna (delay)
    8) Realizar un agrupamiento
    9) Mostrar la filas ordenas por un campo
    10) Generar una consulta SQL desde el DataFrame
    11) Generar un agrupamiento que muestre funciones de agregacion, minimo tres (sum, max, min, avg)
    12) Guardar el resultado en una tabla de hive con un directorio hdfs específico
    13) Listar las tablas de la base de datos
    14) Mostrar el esquema de la nueva tabla

Entregar un archivo PDF