Programación Avanzada

Taller Grupal 1

El propósito de esta actividad es integrar los datos desde el año 2022 hasta el 2024, para realizar la actividad debe emplear alguna de las estrategias que se revisó anteriormente (csvkit y pandas).

Le pido que considere la siguiente forma de trabajo, primero trabajar para crear un DataFrame de Spark para el año 2022, luego crear otro para el año 2023 y finalmente uno para el 2024.

Por cada DataFrame creado imprima el schema ¿nota algo extraño en alguno de ellos? Si es así, busque información sobre el método drop de spark y úselo.

Ahora debe unir cada DataFrame utilizando la función *unionByName*.

Una vez que tiene el DataFrame con todos los datos debe:

- Realizar 3 consultas básicas
- Almacenar el DataFrame en un archivo con el formato parquet.
- Crear un nuevo DataFrame leyendo el archivo parquet y aplicando el siguiente esquema.

#	Field name	Type
1	arma	String
2	autoidentificacion_etnica	String
3	codigo_canton	Integer
4	codigo_circuito	String
5	codigo_distrito	String
6	codigo_iccs	String
7	codigo_parroquia	Integer
8	codigo_provincia	String
9	codigo_subcircuito	String
10	condicion	String
11	edad	String
12	estado_civil	String
13	estatus_migratorio	String
14	fecha_detencion_aprehension	Date
15	genero	String
16	hora_detencion_aprehension	Timestamp
17	lugar	String
18	movilizacion	String

#	Field name	Type
19	nacionalidad	String
20	nivel_de_instruccion	String
21	nombre_canton	String
22	nombre_circuito	String
23	nombre_distrito	String
24	nombre_parroquia	String
25	nombre_provincia	String
26	nombre_subcircuito	String
27	nombre_subzona	String
28	nombre_zona	String
29	numero_detenciones	Integer
30	presunta_flagrancia	String
31	presunta_infraccion	String
32	presunta_modalidad	String
33	presunta_subinfraccion	String
34	sexo	String
35	tipo	String
36	tipo_arma	String
37	tipo_lugar	String

- Almacenar este último DataFrame en un nuevo archivo parquet.