Προχωρημένα Θέματα Βάσεων Δεδομένων

Μαρία Κοιλαλού | Μυρτώ Ορφανάκου
AM:03119211 | AM:

13 Ιανουαρίου 2024

1 Create DataFrame

```
    schema1 = "`DR_NO` STRING, \
    `Date Rptd` STRING, \
    `DATE OCC` STRING, \
    `TIME OCC` INTEGER, \
    `AREA` INTEGER, \
    `AREA NAME` STRING, \
    `Rpt Dist No` INTEGER, \
    `Part 1-2` INTEGER, \
    `Crm Cd` INTEGER, \
    `Crm Cd Desc` STRING, \
    `Mocodes` STRING, \
    `Vict Age` INTEGER, \
    `Vict Sex` STRING, \
    `Vict Descent` STRING, \
    `Premis Cd` INTEGER, \
    `Premis Desc` STRING, \
    `Weapon Used Cd` INTEGER, \
    `Weapon Desc` STRING, \
    `Status` STRING, \
    `Status Desc` STRING, \
    `Crm Cd 1` INTEGER, \
    `Crm Cd 2` INTEGER, \
    `Crm Cd 3` INTEGER, \
    `Crm Cd 4` INTEGER, \
    `LOCATION` STRING, \
    `Cross Street` STRING, \
    `LAT` DOUBLE, \
    `LON` DOUBLE"


data1 = spark.read.csv("/user/ubuntu/ta/advanced-db/data/crime_data_2010.csv", header=True, schema=schema1)
data2 = spark.read.csv("/user/ubuntu/ta/advanced-db/data/crime_data_2020.csv", header=True, schema=schema1)
df = data1.union(data2).distinct()

df = df.withColumn("Date Rptd", to_date(col("Date Rptd"), "MM/dd/yyyy hh:mm:ss a")) \
 .withColumn("DATE OCC", to_date(col("DATE OCC"), "MM/dd/yyyy hh:mm:ss a"))

df.count()
print(f"Total number of rows: {df.count()}")

df.printSchema()
```

```
Total number of rows: 2913595
root
 |-- DR_NO: string (nullable = true)
```

```
|-- Date Rptd: date (nullable = true)
|-- DATE OCC: date (nullable = true)
|-- TIME OCC: integer (nullable = true)
|-- AREA: integer (nullable = true)
|-- AREA NAME: string (nullable = true)
|-- Rpt Dist No: integer (nullable = true)
|-- Part 1-2: integer (nullable = true)
|-- Crm Cd: integer (nullable = true)
|-- Crm Cd Desc: string (nullable = true)
|-- Mocodes: string (nullable = true)
|-- Vict Age: integer (nullable = true)
|-- Vict Sex: string (nullable = true)
|-- Vict Descent: string (nullable = true)
|-- Premis Cd: integer (nullable = true)
|-- Premis Desc: string (nullable = true)
|-- Weapon Used Cd: integer (nullable = true)
|-- Weapon Desc: string (nullable = true)
|-- Status: string (nullable = true)
|-- Status Desc: string (nullable = true)
|-- Crm Cd 1: integer (nullable = true)
|-- Crm Cd 2: integer (nullable = true)
|-- Crm Cd 3: integer (nullable = true)
|-- Crm Cd 4: integer (nullable = true)
|-- LOCATION: string (nullable = true)
|-- Cross Street: string (nullable = true)
|-- LAT: double (nullable = true)
|-- LON: double (nullable = true)
```