Actividad 2

November 20, 2022

[28]:

```
from pyspark.sql import SparkSession
from pyspark.sql.types import *
import pyspark.sql.functions as F
spark = SparkSession.builder.appName("actividad_2").getOrCreate()
```

[37]:

```
def toUri(path):
    return path
```

[35]:

```
schema_df = StructType([
    StructField("PROVINCIA", LongType(), False),
    StructField("MUNICIPIO", LongType(), False),
    StructField("ESTACION", LongType(), False),
    StructField("MAGNITUD", LongType(), False),
    StructField("PUNTO_MUESTREO", StringType(), False),
    StructField("ANO", LongType(), False),
    StructField("MES", LongType(), False),
    StructField("DIA", LongType(), False),
    StructField("H01", LongType(), False),
    StructField("V01", StringType(), False),
    StructField("H02", LongType(), False),
    StructField("V02", StringType(), False),
    StructField("H03", LongType(), False),
    StructField("V03", StringType(), False),
    StructField("H04", LongType(), False),
    StructField("V04", StringType(), False),
    StructField("H05", LongType(), False),
    StructField("V05", StringType(), False),
    StructField("H06", LongType(), False),
    StructField("V06", StringType(), False),
    StructField("H07", LongType(), False),
    StructField("V07", StringType(), False),
    StructField("H08", LongType(), False),
    StructField("V08", StringType(), False),
    StructField("H09", LongType(), False),
    StructField("V09", StringType(), False),
    StructField("H10", LongType(), False),
    StructField("V10", StringType(), False),
    StructField("H11", LongType(), False),
    StructField("V11", StringType(), False),
    StructField("H12", LongType(), False),
    StructField("V12", StringType(), False),
    StructField("H13", LongType(), False),
    StructField("V13", StringType(), False),
    StructField("H14", LongType(), False),
    StructField("V14", StringType(), False),
    StructField("H15", LongType(), False),
    StructField("V15", StringType(), False),
    StructField("H16", LongType(), False),
    StructField("V16", StringType(), False),
    StructField("H17", LongType(), False),
    StructField("V17", StringType(), False),
    StructField("H18", LongType(), False),
    StructField("V18", StringType(), False),
    StructField("H19", LongType(), False),
    StructField("V19", StringType(), False),
    StructField("H20", LongType(), False),
    StructField("V20", StringType(), False),
    StructField("H21", LongType(), False),
    StructField("V21", StringType(), False),
    StructField("H22", LongType(), False),
    StructField("V22", StringType(), False),
    StructField("H23", LongType(), False),
    StructField("V23", StringType(), False),
    StructField("H24", LongType(), False),
    StructField("V24", StringType(), False),

])
```

[44]:

```
df1 = spark.read.csv (toUri('../data/Anio202012/ene_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df2 = spark.read.csv (toUri('../data/Anio202012/feb_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df3 = spark.read.csv (toUri('../data/Anio202012/mar_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df4 = spark.read.csv (toUri('../data/Anio202012/abr_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df5 = spark.read.csv (toUri('../data/Anio202012/may_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df6 = spark.read.csv (toUri('../data/Anio202012/jun_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df7 = spark.read.csv (toUri('../data/Anio202012/jul_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
```

```
df8 = spark.read.csv (toUri('../data/Anio202012/ago_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df9 = spark.read.csv (toUri('../data/Anio202012/ene_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df10 = spark.read.csv (toUri('../data/Anio202012/oct_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df11 = spark.read.csv (toUri('../data/Anio202012/nov_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)
df12= spark.read.csv (toUri('../data/Anio202012/dic_mo20.
 ↪csv'),header=True,sep=';',schema=schema_df)

df12.printSchema()
```

```
root
 |-- PROVINCIA: long (nullable = true)
 |-- MUNICIPIO: long (nullable = true)
 |-- ESTACION: long (nullable = true)
 |-- MAGNITUD: long (nullable = true)
 |-- PUNTO_MUESTREO: string (nullable = true)
 |-- ANO: long (nullable = true)
 |-- MES: long (nullable = true)
 |-- DIA: long (nullable = true)
 |-- H01: long (nullable = true)
 |-- V01: string (nullable = true)
 |-- H02: long (nullable = true)
 |-- V02: string (nullable = true)
 |-- H03: long (nullable = true)
 |-- V03: string (nullable = true)
 |-- H04: long (nullable = true)
 |-- V04: string (nullable = true)
 |-- H05: long (nullable = true)
 |-- V05: string (nullable = true)
 |-- H06: long (nullable = true)
 |-- V06: string (nullable = true)
 |-- H07: long (nullable = true)
 |-- V07: string (nullable = true)
 |-- H08: long (nullable = true)
 |-- V08: string (nullable = true)
 |-- H09: long (nullable = true)
 |-- V09: string (nullable = true)
 |-- H10: long (nullable = true)
 |-- V10: string (nullable = true)
 |-- H11: long (nullable = true)
 |-- V11: string (nullable = true)
 |-- H12: long (nullable = true)
 |-- V12: string (nullable = true)
 |-- H13: long (nullable = true)
 |-- V13: string (nullable = true)
 |-- H14: long (nullable = true)
 |-- V14: string (nullable = true)
 |-- H15: long (nullable = true)
 |-- V15: string (nullable = true)
 |-- H16: long (nullable = true)
 |-- V16: string (nullable = true)
 |-- H17: long (nullable = true)
 |-- V17: string (nullable = true)
 |-- H18: long (nullable = true)
 |-- V18: string (nullable = true)
 |-- H19: long (nullable = true)
 |-- V19: string (nullable = true)
 |-- H20: long (nullable = true)
 |-- V20: string (nullable = true)
 |-- H21: long (nullable = true)
 |-- V21: string (nullable = true)
 |-- H22: long (nullable = true)
 |-- V22: string (nullable = true)
 |-- H23: long (nullable = true)
 |-- V23: string (nullable = true)
 |-- H24: long (nullable = true)
 |-- V24: string (nullable = true)
```

[45]:

```
df = df1.union(df2)
df = df.union(df3)
df = df.union(df4)
df = df.union(df5)
df = df.union(df6)
df = df.union(df7)
df = df.union(df8)
df = df.union(df9)
df = df.union(df10)
df = df.union(df11)
df = df.union(df12)
```

[47]:

```
# PUNTO 3 Carga los datos de calidad de aire de todos los meses del 2020 a un␣
 ↪dataframe de Spark.
print (df.count())
```

```
55656
```

[48]:

```
df.printSchema()
```

```
root
 |-- PROVINCIA: long (nullable = true)
 |-- MUNICIPIO: long (nullable = true)
 |-- ESTACION: long (nullable = true)
 |-- MAGNITUD: long (nullable = true)
 |-- PUNTO_MUESTREO: string (nullable = true)
 |-- ANO: long (nullable = true)
 |-- MES: long (nullable = true)
 |-- DIA: long (nullable = true)
 |-- H01: long (nullable = true)
 |-- V01: string (nullable = true)
 |-- H02: long (nullable = true)
 |-- V02: string (nullable = true)
 |-- H03: long (nullable = true)
 |-- V03: string (nullable = true)
 |-- H04: long (nullable = true)
 |-- V04: string (nullable = true)
 |-- H05: long (nullable = true)
 |-- V05: string (nullable = true)
 |-- H06: long (nullable = true)
 |-- V06: string (nullable = true)
 |-- H07: long (nullable = true)
 |-- V07: string (nullable = true)
 |-- H08: long (nullable = true)
 |-- V08: string (nullable = true)
 |-- H09: long (nullable = true)
 |-- V09: string (nullable = true)
 |-- H10: long (nullable = true)
 |-- V10: string (nullable = true)
 |-- H11: long (nullable = true)
 |-- V11: string (nullable = true)
 |-- H12: long (nullable = true)
 |-- V12: string (nullable = true)
 |-- H13: long (nullable = true)
 |-- V13: string (nullable = true)
 |-- H14: long (nullable = true)
 |-- V14: string (nullable = true)
 |-- H15: long (nullable = true)
 |-- V15: string (nullable = true)
 |-- H16: long (nullable = true)
 |-- V16: string (nullable = true)
 |-- H17: long (nullable = true)
 |-- V17: string (nullable = true)
 |-- H18: long (nullable = true)
 |-- V18: string (nullable = true)
 |-- H19: long (nullable = true)
 |-- V19: string (nullable = true)
 |-- H20: long (nullable = true)
 |-- V20: string (nullable = true)
 |-- H21: long (nullable = true)
 |-- V21: string (nullable = true)
```

```
 |-- H22: long (nullable = true)
 |-- V22: string (nullable = true)
 |-- H23: long (nullable = true)
 |-- V23: string (nullable = true)
 |-- H24: long (nullable = true)
 |-- V24: string (nullable = true)
```

[63]:
```
df.show(2)
```

```
+---------+---------+--------+--------+--------------+----+---+---+---+---+---+-
--+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+---+---+---+---+
|PROVINCIA|MUNICIPIO|ESTACION|MAGNITUD|PUNTO_MUESTREO| ANO|MES|DIA|H01|V01|H02|V
02|H03|V03|H04|V04|H05|V05|H06|V06|H07|V07|H08|V08|H09|V09|H10|V10|H11|V11|H12|V
12|H13|V13|H14|V14|H15|V15|H16|V16|H17|V17|H18|V18|H19|V19|H20|V20|H21|V21|H22|V
22|H23|V23|H24|V24|
+---------+---------+--------+--------+--------------+----+---+---+---+---+---+-
--+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+---+---+---+---+
|       28|       79|       4|       1| 28079004_1_38|2020|  1|  1|  7|  V|  8|
V|  9|  V|  8|  V|  6|  V|  6|  V|  5|  V|  5|  V|  4|  V|  5|  V|  6|  V|  8|
V| 13|  V| 14|  V| 13|  V| 12|  V| 11|  V| 10|  V| 10|  V| 12|  V| 14|  V| 12|
V| 11|  V|  9|  V|
|       28|       79|       4|       1| 28079004_1_38|2020|  1|  2|  8|  V|  8|
V|  7|  V|  6|  V|  5|  V|  5|  V|  5|  V|  9|  V| 10|  V|  9|  V|  8|  V| 12|
V| 16|  V| 16|  V| 14|  V| 12|  V| 11|  V| 10|  V| 11|  V| 14|  V| 14|  V| 15|
V| 12|  V| 10|  V|
+---------+---------+--------+--------+--------------+----+---+---+---+---+---+-
--+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+---+---+---+---+
only showing top 2 rows
```

[53]:
```
df.createOrReplaceTempView('sqlAire')
```

[61]:
```
#PUNTO 4 Indica el número de estaciones distintas que hay en los ficheros.

puntos_muestreo=spark.sql('select DISTINCT punto_muestreo from sqlAire ')
print("Estaciones diferentes :",puntos_muestreo.count())
```

```
Estaciones diferentes : 153
```

[89]:

```python
#PUNTO 5 Indica el número de los distintas MAGNITUDES que se miden.

magnitudes=spark.sql('select DISTINCT magnitud from sqlAire order by magnitud')
print("Magnitudes diferentes :",magnitudes.count())
```

```
Magnitudes diferentes : 14
```

[90]:

```python
magnitudes.show()
```

```
+---------+
|magnitud|
+---------+
|        1|
|        6|
|        7|
|        8|
|        9|
|       10|
|       12|
|       14|
|       20|
|       30|
|       35|
|       42|
|       43|
|       44|
+---------+
```

[87]:

```python
spark.sql("""
select Magnitud, count(magnitud) as CountMagnitud
from sqlAire
group by Magnitud
order by Magnitud
""").show()
```

```
+---------+-------------+
|Magnitud|CountMagnitud|
+---------+-------------+
|        1|         3590|
|        6|         3588|
|        7|         8726|
|        8|         8726|
|        9|         2541|
|       10|         4736|
|       12|         8726|
|       14|         5132|
|       20|         2197|
|       30|         2197|
|       35|         2197|
|       42|         1100|
|       43|         1100|
|       44|         1100|
+---------+-------------+
```

[96]:

```
##PUNTO 6 Indica el número de filas que hay para el día 18-01-2020.

spark.sql("""
                select count(*) as Numero_de_registros
                from sqlAire
                where DIA = 18 and MES = 01 and ANO=2020

                """).show()
```

```
+-------------------+
|Numero_de_registros|
+-------------------+
|                306|
+-------------------+
```

[103]:

```
#PUNTO 7 Averigua la media de dióxido de azufre a las 12h de cada día. A modo␣
 ↪de ejemplo el resultado debería mostrar
spark.sql("""
                select ANO, mes, dia, avg(h12) as media_12h
                from sqlAire
                where magnitud = 1 and V12="V"
                group by ano,mes,dia
                order by mes,dia,ano
                """).show()
```

```
+----+---+---+---------+
| ANO|mes|dia|media_12h|
+----+---+---+---------+
|2020|  1|  1|      9.5|
|2020|  1|  2|     11.2|
|2020|  1|  3|      8.6|
|2020|  1|  4|      5.4|
|2020|  1|  5|      6.0|
|2020|  1|  6|      7.0|
|2020|  1|  7|     11.1|
|2020|  1|  8|     12.4|
|2020|  1|  9|     12.8|
|2020|  1| 10|      6.5|
|2020|  1| 11|      5.6|
|2020|  1| 12|      8.0|
|2020|  1| 13|      9.0|
|2020|  1| 14|      8.2|
|2020|  1| 15|      8.1|
|2020|  1| 16|      8.2|
|2020|  1| 17|      5.9|
|2020|  1| 18|      5.5|
|2020|  1| 19|      4.5|
|2020|  1| 20|      4.6|
+----+---+---+---------+
only showing top 20 rows
```

[ ]:

```

```