M1203SparkSQL

June 4, 2022

[1]:
```
jdbcDF = spark.read \
    .format("jdbc") \
    .option("url", "jdbc:mysql://bd.arcelia.net/datosabiertos") \
    .option("driver","com.mysql.cj.jdbc.Driver") \
    .option("dbtable", "(select ORIGEN, SECTOR, ENTIDAD_UM, SEXO, EDAD,␣
 ↪YEAR(FECHA_INGRESO) AS ANIO, MONTH(FECHA_INGRESO) AS MES FROM COVID19MEXICO␣
 ↪LIMIT 10000) as t") \
    .option("user", "usabierto01") \
    .option("password", "datos21%") \
    .load()

jdbcDF
```

[1]:
```
DataFrame[ORIGEN: int, SECTOR: int, ENTIDAD_UM: string, SEXO: int, EDAD: int,
ANIO: int, MES: int]
```

[2]:
```
jdbcDF.cache()
```

[2]:
```
DataFrame[ORIGEN: int, SECTOR: int, ENTIDAD_UM: string, SEXO: int, EDAD: int,
ANIO: int, MES: int]
```

[3]:
```
jdbcDF.createOrReplaceTempView("dfCovid")
df2 = spark.sql("SELECT SEXO, EDAD, EDAD * 12 AS MESES FROM dfCovid")

df2.show()
```

```
[Stage 0:>                                                          (0 + 1) / 1]

+----+----+-----+
|SEXO|EDAD|MESES|
+----+----+-----+
|   2|  41|  492|
|   1|  66|  792|
|   2|  29|  348|
|   1|  40|  480|
|   2|  34|  408|
|   1|  48|  576|
|   1|  60|  720|
|   1|  20|  240|
|   2|  47|  564|
|   1|  40|  480|
|   2|  12|  144|
|   1|  33|  396|
|   1|  32|  384|
|   2|  22|  264|
|   2|  54|  648|
|   1|  26|  312|
|   1|  32|  384|
|   2|  55|  660|
|   1|  51|  612|
|   2|  31|  372|
+----+----+-----+
only showing top 20 rows
```

[4]:
```
spark \
.sql("SELECT SEXO, EDAD, COUNT(*) nreg FROM dfCovid GROUP BY SEXO, EDAD ORDER␣
 ↪BY 1,2") \
.show()
```

```
[Stage 1:>                                                          (0 + 1) / 1]

+----+----+----+
|SEXO|EDAD|nreg|
+----+----+----+
|   1|   0|   9|
|   1|   1| 165|
|   1|   2| 134|
|   1|   3| 110|
|   1|   4|  96|
|   1|   5|  87|
|   1|   6| 101|
|   1|   7|  90|
|   1|   8|  79|
|   1|   9|  73|
|   1|  10|  87|
|   1|  11|  69|
|   1|  12|  40|
|   1|  13|  54|
|   1|  14|  40|
|   1|  15|  43|
|   1|  16|  27|
|   1|  17|  30|
|   1|  18|  36|
|   1|  19|  39|
+----+----+----+
only showing top 20 rows
```

[5]:
```
spark.catalog.listTables()
```

```
ivysettings.xml file not found in HIVE_HOME or
HIVE_CONF_DIR,/etc/hive/conf.dist/ivysettings.xml will be used
```

[5]:
```
[Table(name='covid_avro', database='default', description=None,
tableType='EXTERNAL', isTemporary=False),
 Table(name='covid_avro_s4s1', database='default', description=None,
tableType='EXTERNAL', isTemporary=False),
 Table(name='covid_parquet', database='default', description=None,
tableType='EXTERNAL', isTemporary=False),
 Table(name='covid_particion', database='default', description=None,
tableType='EXTERNAL', isTemporary=False),
 Table(name='sirilo', database='default', description=None,
tableType='EXTERNAL', isTemporary=False),
 Table(name='sirilo_avro', database='default', description=None,
tableType='EXTERNAL', isTemporary=False),
 Table(name='trade_hive', database='default', description=None,
tableType='MANAGED', isTemporary=False),
 Table(name='dfcovid', database=None, description=None, tableType='TEMPORARY',
isTemporary=True)]
```

[6]:
```
spark.sql("show tables").show()
```

```
+--------+---------------+-----------+
|database|      tableName|isTemporary|
+--------+---------------+-----------+
| default|     covid_avro|      false|
| default|covid_avro_s4s1|      false|
| default|  covid_parquet|      false|
| default|covid_particion|      false|
| default|         sirilo|      false|
| default|    sirilo_avro|      false|
| default|     trade_hive|      false|
|        |        dfcovid|       true|
+--------+---------------+-----------+
```

[7]:
```
spark.sql('describe dfcovid').show()
```

```
+----------+---------+-------+
|  col_name|data_type|comment|
+----------+---------+-------+
|    ORIGEN|      int|   null|
|    SECTOR|      int|   null|
|ENTIDAD_UM|   string|   null|
|      SEXO|      int|   null|
|      EDAD|      int|   null|
|      ANIO|      int|   null|
|       MES|      int|   null|
+----------+---------+-------+
```

[8]:

```
import pandas as pd

dfc = pd.read_csv("https://raw.githubusercontent.com/omarmendoza564/datos/main/
↪datos/201128CatalogosEntidades.csv" \
                  ,header = 0,dtype = {'CLAVE_ENTIDAD':␣
↪str,'ENTIDAD_FEDERATIVA': str, 'ABREVIATURA': str } \
                  ,keep_default_na=False)
```

[9]:

```
#dfc
dfcat = spark.createDataFrame(dfc)

dfcat.show()
```

```
[Stage 6:>                                                          (0 + 1) / 1]

+-------------+--------------------+-----------+
|CLAVE_ENTIDAD|  ENTIDAD_FEDERATIVA|ABREVIATURA|
+-------------+--------------------+-----------+
|           01|      AGUASCALIENTES|         AS|
|           02|     BAJA CALIFORNIA|         BC|
|           03| BAJA CALIFORNIA SUR|         BS|
|           04|            CAMPECHE|         CC|
|           05|COAHUILA DE ZARAGOZA|         CL|
|           06|              COLIMA|         CM|
|           07|             CHIAPAS|         CS|
|           08|           CHIHUAHUA|         CH|
|           09|    CIUDAD DE MÉXICO|         DF|
|           10|             DURANGO|         DG|
|           11|          GUANAJUATO|         GT|
|           12|            GUERRERO|         GR|
|           13|             HIDALGO|         HG|
|           14|             JALISCO|         JC|
|           15|              MÉXICO|         MC|
|           16| MICHOACÁN DE OCAMPO|         MN|
|           17|             MORELOS|         MS|
|           18|             NAYARIT|         NT|
|           19|          NUEVO LEÓN|         NL|
|           20|              OAXACA|         OC|
+--------------+--------------------+-----------+
only showing top 20 rows
```

[10]:

```
#Join
from pyspark.sql.functions import desc
jdbcDF.join(dfcat, jdbcDF.ENTIDAD_UM == dfcat.CLAVE_ENTIDAD).select("*").
↪sort(desc("ENTIDAD_FEDERATIVA")).show()
```

```
+------+------+----------+----+----+----+---+-------------+------------------+--
---------+
|ORIGEN|SECTOR|ENTIDAD_UM|SEXO|EDAD|ANIO|MES|CLAVE_ENTIDAD|ENTIDAD_FEDERATIVA|AB
REVIATURA|
+------+------+----------+----+----+----+---+-------------+------------------+--
---------+
|     1|     4|        32|   2|  17|2020|  1|           32|         ZACATECAS|
ZS|
|     1|    12|        32|   2|  50|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     4|        32|   1|  53|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     4|        32|   2|  26|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     4|        32|   2|  51|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     4|        32|   2|  58|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     4|        32|   1|  51|2020|  1|           32|         ZACATECAS|
ZS|
|     1|    12|        32|   2|   2|2020|  1|           32|         ZACATECAS|
ZS|
|     1|    12|        32|   1|   4|2020|  1|           32|         ZACATECAS|
ZS|
|     1|    12|        32|   1|   2|2020|  1|           32|         ZACATECAS|
ZS|
|     1|    12|        32|   1|  24|2020|  1|           32|         ZACATECAS|
ZS|
|     1|    12|        32|   2|   1|2020|  1|           32|         ZACATECAS|
ZS|
|     1|    12|        32|   2|  33|2020|  1|           32|         ZACATECAS|
ZS|
|     1|    12|        32|   1|   9|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     6|        32|   2|  23|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     6|        32|   1|  24|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     4|        32|   2|  13|2020|  1|           32|         ZACATECAS|
ZS|
|     1|    12|        32|   1|  11|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     6|        32|   2|  15|2020|  1|           32|         ZACATECAS|
ZS|
|     1|     5|        32|   1|  11|2020|  1|           32|         ZACATECAS|
ZS|
+------+------+----------+----+----+----+---+-------------+------------------+--
---------+
only showing top 20 rows
```

[11]:
```
jdbcDF.join(dfcat, jdbcDF.ENTIDAD_UM == dfcat.CLAVE_ENTIDAD) \
    .select("*").sort(desc("ENTIDAD_FEDERATIVA")).count()
```

[11]: 10000

[ ]:
```

```

[ ]:
```

```

[ ]:
```

```