mm4cwadxd

March 3, 2025

1 1. Introducción a DataFrames:

1.1 Conceptos básicos:

Los DataFrames son una abstracción de datos estructurados, organizados en filas y columnas, con un esquema definido. Esta estructura facilita la manipulación y el análisis de datos utilizando las APIs de Spark SQL.

1.2 Creación de DataFrames:

- **Desde RDDs**: Los DataFrames pueden crearse a partir de RDDs (colecciones distribuidas de datos). La creación de un DataFrame desde un RDD permite trabajar con datos no estructurados transformándolos en un formato tabular.
- **Desde archivos**: Spark SQL permite la creación de DataFrames desde varios formatos de archivos, como CSV, JSON y Parquet. Puedes cargar estos archivos directamente en un DataFrame utilizando la API de Spark SQL. También se pueden usar otros formatos de archivo.
- **Desde tablas Hive**: Puedes crear DataFrames a partir de tablas existentes en Hive, aprovechando el metastore de Hive.
- **Otras fuentes**: Spark puede leer datos de diversas fuentes incluyendo bases de datos relacionales mediante JDBC, NoSQL, ORC, y otros sistemas de almacenamiento.

1.2.1 1. Cargar datos desde un RDD:

Para convertir un RDD en un DataFrame, se utiliza la función `toDF()` o `createDataFrame()`.

- **toDF()**: Infiere el esquema del DataFrame a partir del RDD, normalmente usado con una tupla o lista en Python. https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.toDF.h (similar, la original no está documentada en https://spark.apache.org/docs/latest/api/python/reference/api/
- **createDataFrame()**: Permite especificar explícitamente el esquema (`StructType`) del DataFrame. https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.Spark

[0]:
```
# Inicializamos sesión

from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("RDDtoDF").getOrCreate()
```

[0]:
```
import sys
print("Python version: ", sys.version)
```

```
Python version:  3.9.21 (main, Dec  4 2024, 08:53:34)
[GCC 9.4.0]
```

[0]:
```
from pyspark import SparkContext
sc = SparkContext.getOrCreate()
print("Spark version: ", sc.version)
```

```
Spark version:  3.3.2
```

[0]:
```
## Ejemplo con toDF():
rdd = spark.sparkContext.parallelize([("Alice", 34), ("Bob", 23)])
df = rdd.toDF(["name", "age"])
df.show()

## Ejemplo sobre rdd cargado de un fichero (error porque no consigue inferir el␣
 ↪esquema)
# lines = sc.textFile("dbfs:/FileStore/u.data")
df_peliculas = spark.read.option("header", "false").csv("dbfs:/FileStore/u.
 ↪data")
# df_peliculas = lines.toDF()
df_peliculas.show()
```

```
+-----+---+
| name|age|
+-----+---+
|Alice| 34|
|  Bob| 23|
+-----+---+

+--------------------+
|                 _c0|
+--------------------+
|196\t242\t3\t8812…|
|186\t302\t3\t8917…|
|22\t377\t1\t87888…|
|244\t51\t2\t88060…|
|166\t346\t1\t8863…|
|298\t474\t4\t8841…|
|115\t265\t2\t8811…|
|253\t465\t5\t8916…|
|305\t451\t3\t8863…|
| 6\t86\t3\t883603013|
|62\t257\t2\t87937…|
|286\t1014\t5\t879…|
|200\t222\t5\t8760…|
|210\t40\t3\t89103…|
|224\t29\t3\t88810…|
|303\t785\t3\t8794…|
```

```
|122\t387\t5\t8792…|
|194\t274\t2\t8795…|
|291\t1042\t4\t874…|
|234\t1184\t2\t892…|
+--------------------+
only showing top 20 rows
```

[0]:
```
## Ejemplo con createDataFrame() especificando el esquema:
from pyspark.sql import Row
from pyspark.sql.types import StructType, StructField, StringType, IntegerType

esquema = StructType([
            StructField("usuario",IntegerType(),False),
            StructField("pelicula",IntegerType(),False),
            StructField("rating", IntegerType(),False),
            StructField("timestamp",IntegerType(),False)
])

lineas_rdd = sc.textFile("dbfs:/FileStore/u.data")

# Convertimos un rdd de filas en un rdd de listas de 4 strings
lineas_rdd_cadenas = lineas_rdd.map(lambda x: x.split())

# Como el esquema espera enteros, convertimos el rdd a listas de 4 enteros
lineas_rdd_enteros = lineas_rdd_cadenas.map(lambda x: [int(x[0]), int(x[1]),␣
 ↪int(x[2]), int(x[3])])

df_peliculas = spark.createDataFrame(lineas_rdd_enteros,esquema)
df_peliculas.show()
```

```
+-------+--------+------+---------+
|usuario|pelicula|rating|timestamp|
+-------+--------+------+---------+
|    196|     242|     3|881250949|
|    186|     302|     3|891717742|
|     22|     377|     1|878887116|
|    244|      51|     2|880606923|
|    166|     346|     1|886397596|
|    298|     474|     4|884182806|
|    115|     265|     2|881171488|
|    253|     465|     5|891628467|
|    305|     451|     3|886324817|
|      6|      86|     3|883603013|
|     62|     257|     2|879372434|
|    286|    1014|     5|879781125|
|    200|     222|     5|876042340|
|    210|      40|     3|891035994|
```

```
|    224|      29|     3|888104457|
|    303|     785|     3|879485318|
|    122|     387|     5|879270459|
|    194|     274|     2|879539794|
|    291|    1042|     4|874834944|
|    234|    1184|     2|892079237|
+-------+--------+------+---------+
only showing top 20 rows
```

[0]:
```
# Mismo caso pero definiendo la función para el map
def leerFila(fila):
    lista = fila.split()
    return [int(lista[0]),int(lista[1]),int(lista[2]),int(lista[3])]

lineas_rdd_filas = sc.textFile("dbfs:/FileStore/u.data")
lineas_rdd_enteros2 = lineas_rdd_filas.map(leerFila)

df_peliculas2 = spark.createDataFrame(lineas_rdd_enteros2,esquema)
df_peliculas2.show()
```

```
+-------+--------+------+---------+
|usuario|pelicula|rating|timestamp|
+-------+--------+------+---------+
|    196|     242|     3|881250949|
|    186|     302|     3|891717742|
|     22|     377|     1|878887116|
|    244|      51|     2|880606923|
|    166|     346|     1|886397596|
|    298|     474|     4|884182806|
|    115|     265|     2|881171488|
|    253|     465|     5|891628467|
|    305|     451|     3|886324817|
|      6|      86|     3|883603013|
|     62|     257|     2|879372434|
|    286|    1014|     5|879781125|
|    200|     222|     5|876042340|
|    210|      40|     3|891035994|
|    224|      29|     3|888104457|
|    303|     785|     3|879485318|
|    122|     387|     5|879270459|
|    194|     274|     2|879539794|
|    291|    1042|     4|874834944|
|    234|    1184|     2|892079237|
+-------+--------+------+---------+
only showing top 20 rows
```

1.2.2 2. Cargar datos desde ficheros CSV:

Sintaxis: Se utiliza `spark.read.csv()`. Se pueden especificar opciones como `header` para indicar si el archivo tiene encabezado e `inferSchema` para que Spark infiera los tipos de datos.

- `header` indica si la primera línea del archivo CSV contiene los nombres de las columnas.
- `inferSchema` permite a Spark determinar automáticamente los tipos de datos de cada columna.

https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrameReader.csv.ht

Consideraciones

- **Rutas de archivos**: Asegúrate de proporcionar las rutas correctas a tus archivos CSV y Parquet.
- **Esquema**: Si no se utiliza `inferSchema` al leer archivos CSV, el esquema del DataFrame debe especificarse explícitamente.
- **DataFrames**: Los DataFrames proporcionan una forma de procesar y analizar datos estructurados. A diferencia de los RDDs, los DataFrames están basados en un esquema, es decir, conocen los nombres y tipos de las columnas de un conjunto de datos.

[0]:
```
# Desde un fichero CSV
df=spark.read.option("header", "true").option("inferSchema", "true").csv("dbfs:/
  ↪FileStore/olive.csv")
df.show()
```

```
+-------+----+---------------+------------------+------------+-------+------
--------------+------------------+-------+--------------------+----------+---
--------------+------------+
|Country|Year|Beginning Stocks|Domestic Consumption|Ending Stocks|Exports|Feed
Waste Dom. Cons.|Food Use Dom. Cons.|Imports|Industrial Dom.
Cons.|Production|Total Distribution|Total Supply|
+-------+----+---------------+------------------+------------+-------+------
--------------+------------------+-------+--------------------+----------+---
--------------+------------+
|Algeria|1964|              0|                15|           0|      3|
0|                15|      0|                 0|        18|
18|          18|
|Algeria|1965|              0|                12|           0|      5|
0|                12|      0|                 0|        17|
17|          17|
|Algeria|1966|              0|                16|           0|      0|
0|                16|      0|                 0|        16|
16|          16|
|Algeria|1967|              0|                15|           0|      7|
0|                15|      0|                 0|        22|
22|          22|
|Algeria|1968|              0|                11|           0|      7|
0|                11|      0|                 0|        18|
18|          18|
```

```
|Algeria|1969|              0|                  19|            0|      3|
0|                 19|      0|                   0|        22|
22|          22|
|Algeria|1970|              0|                  12|            0|      1|
0|                 12|      0|                   0|        13|
13|          13|
|Algeria|1971|              0|                  20|            2|      1|
0|                 20|      0|                   0|        23|
23|          23|
|Algeria|1972|              2|                  13|            2|      2|
0|                 13|      0|                   0|        15|
17|          17|
|Algeria|1973|              2|                  11|            3|      4|
0|                 11|      0|                   0|        16|
18|          18|
|Algeria|1974|              3|                  10|            1|      0|
0|                 10|      0|                   0|         8|
11|          11|
|Algeria|1975|              1|                  16|            3|      0|
0|                 16|      0|                   0|        18|
19|          19|
|Algeria|1976|              3|                  15|            3|      0|
0|                 15|      0|                   0|        15|
18|          18|
|Algeria|1977|              3|                   8|            0|      0|
0|                  8|      0|                   0|         5|
8|           8|
|Algeria|1978|              0|                  13|            1|      0|
0|                 13|      0|                   0|        14|
14|          14|
|Algeria|1979|              1|                  11|            0|      0|
0|                 11|      0|                   0|        10|
11|          11|
|Algeria|1980|              0|                  18|            0|      0|
0|                 18|      0|                   0|        18|
18|          18|
|Algeria|1981|              0|                  15|            0|      0|
0|                 15|      0|                   0|        15|
15|          15|
|Algeria|1982|              0|                  16|            0|      0|
0|                 16|      0|                   0|        16|
16|          16|
|Algeria|1983|              0|                  13|            0|      0|
0|                 13|      0|                   0|        13|
13|          13|
+-------+----+---------------+--------------------+-------------+-------+------
---------------+------------------+-------+--------------------+----------+---
---------------+------------+
```

only showing top 20 rows

1.2.3 3. Cargar datos desde ficheros Parquet:

Se utiliza `spark.read.parquet()`.
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrameReader.parqu

Consideraciones

- **Esquema**: El esquema se almacena en el mismo archivo.

[0]:
```
# Desde un único fichero parquet
df=spark.read.parquet("dbfs:/FileStore/palm.parquet")
df.show(1000)
```

```
+------------------+----+-------------+---------------+--------------------+--
----------+-------+--------------------+------------------+-------+----------
----------+----------+-----------------+------------+-----+
|           Country|Year|Area Harvested|Beginning Stocks|Domestic
Consumption|Ending Stocks|Exports|Feed Waste Dom. Cons.|Food Use Dom.
Cons.|Imports|Industrial Dom. Cons.|Production|Total Distribution|Total
Supply|Yield|
+------------------+----+-------------+---------------+--------------------+--
----------+-------+--------------------+------------------+-------+----------
----------+----------+-----------------+------------+-----+
|       Afghanistan|1964|          0.0|            0.0|                 0.0|
0.0|    0.0|                 0.0|               0.0|    0.0|
0.0|       0.0|                0.0|         0.0|  0.0|
|       Afghanistan|1965|          0.0|            0.0|                 0.0|
0.0|    0.0|                 0.0|               0.0|    0.0|
0.0|       0.0|                0.0|         0.0|  0.0|
|       Afghanistan|1966|          0.0|            0.0|                 0.0|
0.0|    0.0|                 0.0|               0.0|    0.0|
0.0|       0.0|                0.0|         0.0|  0.0|
|       Afghanistan|1967|          0.0|            0.0|                 0.0|
0.0|    0.0|                 0.0|               0.0|    0.0|
0.0|       0.0|                0.0|         0.0|  0.0|
|       Afghanistan|1968|          0.0|            0.0|                 0.0|
0.0|    0.0|                 0.0|               0.0|    0.0|
0.0|       0.0|                0.0|         0.0|  0.0|
|       Afghanistan|1969|          0.0|            0.0|                 0.0|
0.0|    0.0|                 0.0|               0.0|    0.0|
0.0|       0.0|                0.0|         0.0|  0.0|
|       Afghanistan|1970|          0.0|            0.0|                 0.0|
0.0|    0.0|                 0.0|               0.0|    0.0|
0.0|       0.0|                0.0|         0.0|  0.0|
|       Afghanistan|1971|          0.0|            0.0|                 0.0|
0.0|    0.0|                 0.0|               0.0|    0.0|
```

```
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1972|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1973|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1974|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1975|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1976|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1977|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1978|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1979|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1980|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1981|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1982|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1983|          0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|       Afghanistan|1984|          0.0|            0.0|               10.0|
0.0|    0.0|                 0.0|             10.0|   10.0|
0.0|       0.0|             10.0|        10.0|  0.0|
|       Afghanistan|1985|          0.0|            0.0|               10.0|
0.0|    0.0|                 0.0|             10.0|   10.0|
0.0|       0.0|             10.0|        10.0|  0.0|
|       Afghanistan|1986|          0.0|            0.0|               10.0|
0.0|    0.0|                 0.0|             10.0|   10.0|
0.0|       0.0|             10.0|        10.0|  0.0|
|       Afghanistan|1987|          0.0|            0.0|               15.0|
0.0|    0.0|                 0.0|             15.0|   15.0|
```

```
0.0|       0.0|             15.0|        15.0|  0.0|
|       Afghanistan|1988|          0.0|            0.0|               20.0|
0.0|    0.0|                 0.0|              20.0|   20.0|
0.0|       0.0|             20.0|        20.0|  0.0|
|       Afghanistan|1989|          0.0|            0.0|               25.0|
0.0|    0.0|                 0.0|              25.0|   25.0|
0.0|       0.0|             25.0|        25.0|  0.0|
|       Afghanistan|1990|          0.0|            0.0|               30.0|
0.0|    0.0|                 0.0|              30.0|   30.0|
0.0|       0.0|             30.0|        30.0|  0.0|
|       Afghanistan|1991|          0.0|            0.0|               35.0|
0.0|    0.0|                 0.0|              35.0|   35.0|
0.0|       0.0|             35.0|        35.0|  0.0|
|       Afghanistan|1992|          0.0|            0.0|               40.0|
0.0|    0.0|                 0.0|              40.0|   40.0|
0.0|       0.0|             40.0|        40.0|  0.0|
|       Afghanistan|1993|          0.0|            0.0|               45.0|
0.0|    0.0|                 0.0|              45.0|   45.0|
0.0|       0.0|             45.0|        45.0|  0.0|
|       Afghanistan|1994|          0.0|            0.0|               50.0|
0.0|    0.0|                 0.0|              50.0|   50.0|
0.0|       0.0|             50.0|        50.0|  0.0|
|       Afghanistan|1995|          0.0|            0.0|               50.0|
0.0|    0.0|                 0.0|              50.0|   50.0|
0.0|       0.0|             50.0|        50.0|  0.0|
|       Afghanistan|1996|          0.0|            0.0|               52.0|
0.0|    0.0|                 0.0|              52.0|   52.0|
0.0|       0.0|             52.0|        52.0|  0.0|
|       Afghanistan|1997|          0.0|            0.0|               60.0|
0.0|    0.0|                 0.0|              60.0|   60.0|
0.0|       0.0|             60.0|        60.0|  0.0|
|       Afghanistan|1998|          0.0|            0.0|               60.0|
0.0|    0.0|                 0.0|              60.0|   60.0|
0.0|       0.0|             60.0|        60.0|  0.0|
|       Afghanistan|1999|          0.0|            0.0|               65.0|
0.0|    0.0|                 0.0|              65.0|   65.0|
0.0|       0.0|             65.0|        65.0|  0.0|
|       Afghanistan|2000|          0.0|            0.0|               55.0|
5.0|    0.0|                 0.0|              55.0|   60.0|
0.0|       0.0|             60.0|        60.0|  0.0|
|       Afghanistan|2001|          0.0|            5.0|               45.0|
10.0|    0.0|                  0.0|             45.0|   50.0|
0.0|       0.0|             55.0|        55.0|  0.0|
|       Afghanistan|2002|          0.0|           10.0|               40.0|
0.0|    0.0|                 0.0|              40.0|   30.0|
0.0|       0.0|             40.0|        40.0|  0.0|
|       Afghanistan|2003|          0.0|            0.0|              125.0|
20.0|    0.0|                  0.0|             125.0|  145.0|
```

```
0.0|       0.0|            145.0|       145.0|  0.0|
|       Afghanistan|2004|          0.0|            20.0|              150.0|
55.0|    0.0|                 0.0|             150.0|  185.0|
0.0|       0.0|            205.0|       205.0|  0.0|
|       Afghanistan|2005|          0.0|            55.0|              151.0|
55.0|    0.0|                 0.0|             151.0|  151.0|
0.0|       0.0|            206.0|       206.0|  0.0|
|       Afghanistan|2006|          0.0|            55.0|              135.0|
65.0|    0.0|                 0.0|             135.0|  145.0|
0.0|       0.0|            200.0|       200.0|  0.0|
|       Afghanistan|2007|          0.0|            65.0|              125.0|
56.0|    0.0|                 0.0|             125.0|  116.0|
0.0|       0.0|            181.0|       181.0|  0.0|
|       Afghanistan|2008|          0.0|            56.0|              117.0|
23.0|    0.0|                 0.0|             117.0|   84.0|
0.0|       0.0|            140.0|       140.0|  0.0|
|       Afghanistan|2009|          0.0|            23.0|              118.0|
5.0|    0.0|                  0.0|            118.0|  100.0|
0.0|       0.0|            123.0|       123.0|  0.0|
|       Afghanistan|2010|          0.0|             5.0|              105.0|
5.0|    0.0|                  0.0|            105.0|  105.0|
0.0|       0.0|            110.0|       110.0|  0.0|
|       Afghanistan|2011|          0.0|             5.0|              110.0|
0.0|    0.0|                  0.0|            110.0|  105.0|
0.0|       0.0|            110.0|       110.0|  0.0|
|       Afghanistan|2012|          0.0|             0.0|              115.0|
0.0|    0.0|                  0.0|            115.0|  115.0|
0.0|       0.0|            115.0|       115.0|  0.0|
|       Afghanistan|2013|          0.0|             0.0|              115.0|
0.0|    0.0|                  0.0|            115.0|  115.0|
0.0|       0.0|            115.0|       115.0|  0.0|
|       Afghanistan|2014|          0.0|             0.0|              115.0|
32.0|    0.0|                 0.0|             115.0|  147.0|
0.0|       0.0|            147.0|       147.0|  0.0|
|       Afghanistan|2015|          0.0|            32.0|              120.0|
18.0|    0.0|                 0.0|             120.0|  106.0|
0.0|       0.0|            138.0|       138.0|  0.0|
|       Afghanistan|2016|          0.0|            18.0|              130.0|
23.0|    0.0|                 0.0|             130.0|  135.0|
0.0|       0.0|            153.0|       153.0|  0.0|
|       Afghanistan|2017|          0.0|            23.0|              140.0|
29.0|    0.0|                 0.0|             140.0|  146.0|
0.0|       0.0|            169.0|       169.0|  0.0|
|       Afghanistan|2018|          0.0|            29.0|              145.0|
36.0|    0.0|                 0.0|             145.0|  152.0|
0.0|       0.0|            181.0|       181.0|  0.0|
|       Afghanistan|2019|          0.0|            36.0|              150.0|
34.0|    0.0|                 0.0|             150.0|  148.0|
```

```
0.0|       0.0|            184.0|       184.0|  0.0|
|       Afghanistan|2020|          0.0|            34.0|               155.0|
36.0|    0.0|                 0.0|             155.0|  157.0|
0.0|       0.0|            191.0|       191.0|  0.0|
|       Afghanistan|2021|          0.0|            36.0|               205.0|
38.0|    0.0|                 0.0|             205.0|  207.0|
0.0|       0.0|            243.0|       243.0|  0.0|
|       Afghanistan|2022|          0.0|            38.0|               215.0|
43.0|    0.0|                 0.0|             215.0|  220.0|
0.0|       0.0|            258.0|       258.0|  0.0|
|       Afghanistan|2023|          0.0|            43.0|               220.0|
43.0|    0.0|                 0.0|             220.0|  220.0|
0.0|       0.0|            263.0|       263.0|  0.0|
|           Algeria|1964|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1965|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1966|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1967|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1968|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1969|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1970|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1971|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1972|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1973|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1974|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1975|          0.0|             0.0|                 0.0|
0.0|    0.0|                   0.0|               0.0|    0.0|
```

```
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1976|           0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1977|           0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1978|           0.0|            0.0|                4.0|
0.0|    0.0|                 0.0|              4.0|    4.0|
0.0|       0.0|              4.0|         4.0|  0.0|
|           Algeria|1979|           0.0|            0.0|                5.0|
0.0|    0.0|                 0.0|              5.0|    5.0|
0.0|       0.0|              5.0|         5.0|  0.0|
|           Algeria|1980|           0.0|            0.0|                3.0|
0.0|    0.0|                 0.0|              3.0|    3.0|
0.0|       0.0|              3.0|         3.0|  0.0|
|           Algeria|1981|           0.0|            0.0|                6.0|
0.0|    0.0|                 0.0|              6.0|    6.0|
0.0|       0.0|              6.0|         6.0|  0.0|
|           Algeria|1982|           0.0|            0.0|                3.0|
0.0|    0.0|                 0.0|              3.0|    3.0|
0.0|       0.0|              3.0|         3.0|  0.0|
|           Algeria|1983|           0.0|            0.0|                3.0|
0.0|    0.0|                 0.0|              3.0|    3.0|
0.0|       0.0|              3.0|         3.0|  0.0|
|           Algeria|1984|           0.0|            0.0|                4.0|
0.0|    0.0|                 0.0|              4.0|    4.0|
0.0|       0.0|              4.0|         4.0|  0.0|
|           Algeria|1985|           0.0|            0.0|                1.0|
0.0|    0.0|                 0.0|              1.0|    1.0|
0.0|       0.0|              1.0|         1.0|  0.0|
|           Algeria|1986|           0.0|            0.0|                0.0|
0.0|    0.0|                 0.0|              0.0|    0.0|
0.0|       0.0|              0.0|         0.0|  0.0|
|           Algeria|1987|           0.0|            0.0|               10.0|
0.0|    0.0|                 0.0|             10.0|   10.0|
0.0|       0.0|             10.0|        10.0|  0.0|
|           Algeria|1988|           0.0|            0.0|               10.0|
0.0|    0.0|                 0.0|             10.0|   10.0|
0.0|       0.0|             10.0|        10.0|  0.0|
|           Algeria|1989|           0.0|            0.0|               10.0|
0.0|    0.0|                 0.0|             10.0|   10.0|
0.0|       0.0|             10.0|        10.0|  0.0|
|           Algeria|1990|           0.0|            0.0|               10.0|
0.0|    0.0|                 0.0|             10.0|   10.0|
0.0|       0.0|             10.0|        10.0|  0.0|
|           Algeria|1991|           0.0|            0.0|               10.0|
0.0|    0.0|                 0.0|             10.0|   10.0|
```

```
0.0|       0.0|             10.0|        10.0|  0.0|
|           Algeria|1992|          0.0|             0.0|                15.0|
0.0|    0.0|                  0.0|               15.0|   15.0|
0.0|       0.0|             15.0|        15.0|  0.0|
|           Algeria|1993|          0.0|             0.0|                20.0|
0.0|    0.0|                  0.0|               20.0|   20.0|
0.0|       0.0|             20.0|        20.0|  0.0|
|           Algeria|1994|          0.0|             0.0|                13.0|
0.0|    0.0|                  0.0|               13.0|   13.0|
0.0|       0.0|             13.0|        13.0|  0.0|
|           Algeria|1995|          0.0|             0.0|                10.0|
0.0|    0.0|                  0.0|               10.0|   10.0|
0.0|       0.0|             10.0|        10.0|  0.0|
|           Algeria|1996|          0.0|             0.0|                20.0|
0.0|    0.0|                  0.0|               20.0|   20.0|
0.0|       0.0|             20.0|        20.0|  0.0|
|           Algeria|1997|          0.0|             0.0|                22.0|
0.0|    0.0|                  0.0|               22.0|   22.0|
0.0|       0.0|             22.0|        22.0|  0.0|
|           Algeria|1998|          0.0|             0.0|                18.0|
0.0|    0.0|                  0.0|               18.0|   18.0|
0.0|       0.0|             18.0|        18.0|  0.0|
|           Algeria|1999|          0.0|             0.0|                19.0|
0.0|    0.0|                  0.0|               19.0|   19.0|
0.0|       0.0|             19.0|        19.0|  0.0|
|           Algeria|2000|          0.0|             0.0|                61.0|
0.0|    0.0|                  0.0|               61.0|   61.0|
0.0|       0.0|             61.0|        61.0|  0.0|
|           Algeria|2001|          0.0|             0.0|                87.0|
0.0|    0.0|                  0.0|               87.0|   87.0|
0.0|       0.0|             87.0|        87.0|  0.0|
|           Algeria|2002|          0.0|             0.0|               120.0|
0.0|    0.0|                  0.0|              120.0|  120.0|
0.0|       0.0|            120.0|       120.0|  0.0|
|           Algeria|2003|          0.0|             0.0|               100.0|
0.0|    0.0|                  0.0|              100.0|  100.0|
0.0|       0.0|            100.0|       100.0|  0.0|
|           Algeria|2004|          0.0|             0.0|               129.0|
0.0|    0.0|                  0.0|              129.0|  129.0|
0.0|       0.0|            129.0|       129.0|  0.0|
|           Algeria|2005|          0.0|             0.0|                75.0|
0.0|    0.0|                  0.0|               75.0|   75.0|
0.0|       0.0|             75.0|        75.0|  0.0|
|           Algeria|2006|          0.0|             0.0|                98.0|
0.0|    0.0|                  0.0|               98.0|   98.0|
0.0|       0.0|             98.0|        98.0|  0.0|
|           Algeria|2007|          0.0|             0.0|               137.0|
0.0|    0.0|                  0.0|              137.0|  137.0|
```

```
0.0|       0.0|             137.0|       137.0|  0.0|
|           Algeria|2008|           0.0|             0.0|                56.0|
0.0|    0.0|                  0.0|               56.0|   56.0|
0.0|       0.0|             56.0|        56.0|  0.0|
|           Algeria|2009|           0.0|             0.0|                51.0|
0.0|    0.0|                  0.0|               51.0|   51.0|
0.0|       0.0|             51.0|        51.0|  0.0|
|           Algeria|2010|           0.0|             0.0|                60.0|
7.0|    0.0|                  0.0|               60.0|   67.0|
0.0|       0.0|             67.0|        67.0|  0.0|
|           Algeria|2011|           0.0|             7.0|                49.0|
16.0|    0.0|                  0.0|               49.0|   58.0|
0.0|       0.0|             65.0|        65.0|  0.0|
|           Algeria|2012|           0.0|            16.0|                80.0|
60.0|    0.0|                  0.0|               80.0|  124.0|
0.0|       0.0|            140.0|       140.0|  0.0|
|           Algeria|2013|           0.0|            60.0|                85.0|
31.0|    0.0|                  0.0|

*** WARNING: max output size exceeded, skipping output. ***

    0.0|              240.0|   71.0|                62.0|     236.0|
324.0|       324.0| 1.05|
|  Congo (Kinshasa)|2013|         268.0|            18.0|               325.0|
47.0|    5.0|                  0.0|              265.0|   77.0|
60.0|     282.0|             377.0|       377.0| 1.05|
|  Congo (Kinshasa)|2014|         273.0|            47.0|               345.0|
70.0|    6.0|                  0.0|              285.0|   86.0|
60.0|     288.0|             421.0|       421.0| 1.06|
|  Congo (Kinshasa)|2015|         277.0|            70.0|               375.0|
67.0|    7.0|                  0.0|              315.0|   88.0|
60.0|     291.0|             449.0|       449.0| 1.05|
|  Congo (Kinshasa)|2016|         275.0|            67.0|               395.0|
62.0|    8.0|                  0.0|              335.0|  110.0|
60.0|     288.0|             465.0|       465.0| 1.05|
|  Congo (Kinshasa)|2017|         278.0|            62.0|               395.0|
52.0|    9.0|                  0.0|              345.0|  103.0|
50.0|     291.0|             456.0|       456.0| 1.05|
|  Congo (Kinshasa)|2018|         279.0|            52.0|               395.0|
55.0|    9.0|                  0.0|              355.0|  114.0|
40.0|     293.0|             459.0|       459.0| 1.05|
|  Congo (Kinshasa)|2019|         280.0|            55.0|               395.0|
60.0|   10.0|                  0.0|              365.0|  116.0|
30.0|     294.0|             465.0|       465.0| 1.05|
|  Congo (Kinshasa)|2020|         285.0|            60.0|               405.0|
46.0|   10.0|                  0.0|              375.0|  101.0|
30.0|     300.0|             461.0|       461.0| 1.05|
|  Congo (Kinshasa)|2021|         285.0|            46.0|               405.0|
```

```
16.0|   10.0|                0.0|             380.0|   85.0|
25.0|     300.0|            431.0|       431.0| 1.05|
|  Congo (Kinshasa)|2022|        285.0|           16.0|              420.0|
16.0|   10.0|                0.0|             390.0|  130.0|
30.0|     300.0|            446.0|       446.0| 1.05|
|  Congo (Kinshasa)|2023|        285.0|           16.0|              425.0|
16.0|   10.0|                0.0|             395.0|  135.0|
30.0|     300.0|            451.0|       451.0| 1.05|
|        Costa Rica|1964|          0.0|            0.0|               10.0|
0.0|    0.0|                 0.0|             10.0|    1.0|
0.0|       9.0|             10.0|        10.0|  0.0|
|        Costa Rica|1965|          0.0|            0.0|               12.0|
0.0|    0.0|                 0.0|             12.0|    3.0|
0.0|       9.0|             12.0|        12.0|  0.0|
|        Costa Rica|1966|          0.0|            0.0|               12.0|
0.0|    0.0|                 0.0|             12.0|    2.0|
0.0|      10.0|             12.0|        12.0|  0.0|
|        Costa Rica|1967|          0.0|            0.0|               11.0|
0.0|    0.0|                 0.0|             11.0|    1.0|
0.0|      10.0|             11.0|        11.0|  0.0|
|        Costa Rica|1968|          0.0|            0.0|               13.0|
0.0|    0.0|                 0.0|             13.0|    2.0|
0.0|      11.0|             13.0|        13.0|  0.0|
|        Costa Rica|1969|          0.0|            0.0|               11.0|
0.0|    0.0|                 0.0|             11.0|    0.0|
0.0|      11.0|             11.0|        11.0|  0.0|
|        Costa Rica|1970|          0.0|            0.0|               15.0|
0.0|    0.0|                 0.0|             15.0|    1.0|
0.0|      14.0|             15.0|        15.0|  0.0|
|        Costa Rica|1971|          0.0|            0.0|               14.0|
0.0|    0.0|                 0.0|             14.0|    0.0|
0.0|      14.0|             14.0|        14.0|  0.0|
|        Costa Rica|1972|          0.0|            0.0|               22.0|
0.0|    0.0|                 0.0|             22.0|    0.0|
0.0|      22.0|             22.0|        22.0|  0.0|
|        Costa Rica|1973|          0.0|            0.0|               22.0|
0.0|    0.0|                 0.0|             22.0|    0.0|
0.0|      22.0|             22.0|        22.0|  0.0|
|        Costa Rica|1974|          0.0|            0.0|               22.0|
0.0|    0.0|                 0.0|             22.0|    0.0|
0.0|      22.0|             22.0|        22.0|  0.0|
|        Costa Rica|1975|          0.0|            0.0|               24.0|
0.0|    0.0|                 0.0|             24.0|    0.0|
0.0|      24.0|             24.0|        24.0|  0.0|
|        Costa Rica|1976|          0.0|            0.0|               26.0|
0.0|    0.0|                 0.0|             26.0|    0.0|
0.0|      26.0|             26.0|        26.0|  0.0|
|        Costa Rica|1977|          0.0|            0.0|               28.0|
```

```
0.0|    0.0|                 0.0|              28.0|    0.0|
0.0|      28.0|             28.0|        28.0|  0.0|
|        Costa Rica|1978|          0.0|            0.0|               31.0|
0.0|    0.0|                 0.0|              31.0|    0.0|
0.0|      31.0|             31.0|        31.0|  0.0|
|        Costa Rica|1979|          0.0|            0.0|               32.0|
0.0|    0.0|                 0.0|              32.0|    0.0|
0.0|      32.0|             32.0|        32.0|  0.0|
|        Costa Rica|1980|          0.0|            0.0|               31.0|
0.0|    0.0|                 0.0|              31.0|    0.0|
0.0|      31.0|             31.0|        31.0|  0.0|
|        Costa Rica|1981|          0.0|            0.0|               35.0|
0.0|    0.0|                 0.0|              35.0|    0.0|
0.0|      35.0|             35.0|        35.0|  0.0|
|        Costa Rica|1982|          0.0|            0.0|               40.0|
0.0|    0.0|                 0.0|              40.0|    0.0|
0.0|      40.0|             40.0|        40.0|  0.0|
|        Costa Rica|1983|          0.0|            0.0|               40.0|
0.0|    0.0|                 0.0|              40.0|    0.0|
0.0|      40.0|             40.0|        40.0|  0.0|
|        Costa Rica|1984|          0.0|            0.0|               40.0|
0.0|    0.0|                 0.0|              40.0|    0.0|
0.0|      40.0|             40.0|        40.0|  0.0|
|        Costa Rica|1985|          0.0|            0.0|               45.0|
0.0|    0.0|                 0.0|              45.0|    0.0|
0.0|      45.0|             45.0|        45.0|  0.0|
|        Costa Rica|1986|          0.0|            0.0|               54.0|
0.0|    0.0|                 0.0|              54.0|    0.0|
0.0|      54.0|             54.0|        54.0|  0.0|
|        Costa Rica|1987|          0.0|            0.0|               46.0|
0.0|   10.0|                 0.0|              46.0|    0.0|
0.0|      56.0|             56.0|        56.0|  0.0|
|        Costa Rica|1988|          0.0|            0.0|               44.0|
0.0|   14.0|                 0.0|              44.0|    0.0|
0.0|      58.0|             58.0|        58.0|  0.0|
|        Costa Rica|1989|          0.0|            0.0|               59.0|
0.0|   14.0|                 0.0|              59.0|    0.0|
0.0|      73.0|             73.0|        73.0|  0.0|
|        Costa Rica|1990|          0.0|            0.0|               52.0|
0.0|   12.0|                 0.0|              52.0|    0.0|
0.0|      64.0|             64.0|        64.0|  0.0|
|        Costa Rica|1991|          0.0|            0.0|               44.0|
0.0|   14.0|                 0.0|              44.0|    0.0|
0.0|      58.0|             58.0|        58.0|  0.0|
|        Costa Rica|1992|          0.0|            0.0|               47.0|
0.0|   14.0|                 0.0|              47.0|    0.0|
0.0|      61.0|             61.0|        61.0|  0.0|
|        Costa Rica|1993|          0.0|            0.0|               70.0|
```

```
0.0|   14.0|                 0.0|              70.0|    0.0|
0.0|      84.0|             84.0|        84.0|  0.0|
|        Costa Rica|1994|          0.0|            0.0|               46.0|
0.0|   42.0|                 0.0|              46.0|    0.0|
0.0|      88.0|             88.0|        88.0|  0.0|
|        Costa Rica|1995|          0.0|            0.0|               37.0|
0.0|   56.0|                 0.0|              37.0|    0.0|
0.0|      93.0|             93.0|        93.0|  0.0|
|        Costa Rica|1996|          0.0|            0.0|               26.0|
0.0|   71.0|                 0.0|              26.0|    0.0|
0.0|      97.0|             97.0|        97.0|  0.0|
|        Costa Rica|1997|          0.0|            0.0|               36.0|
0.0|   65.0|                 0.0|              36.0|    0.0|
0.0|     101.0|            101.0|       101.0|  0.0|
|        Costa Rica|1998|          0.0|            0.0|               38.0|
0.0|   71.0|                 0.0|              38.0|    0.0|
0.0|     109.0|            109.0|       109.0|  0.0|
|        Costa Rica|1999|         40.0|            0.0|               40.0|
22.0|   75.0|                   0.0|             40.0|    0.0|
0.0|     137.0|            137.0|       137.0| 3.43|
|        Costa Rica|2000|         40.0|           22.0|               45.0|
53.0|   79.0|                   0.0|             45.0|    5.0|
0.0|     150.0|            177.0|       177.0| 3.75|
|        Costa Rica|2001|         42.0|           53.0|               45.0|
64.0|   83.0|                   0.0|             45.0|   11.0|
0.0|     128.0|            192.0|       192.0| 3.05|
|        Costa Rica|2002|         43.0|           64.0|               35.0|
55.0|  107.0|                   0.0|             35.0|    2.0|
0.0|     131.0|            197.0|       197.0| 3.05|
|        Costa Rica|2003|         47.0|           55.0|               30.0|
7.0|  172.0|                   0.0|             30.0|    3.0|
0.0|     151.0|            209.0|       209.0| 3.21|
|        Costa Rica|2004|         50.0|            7.0|               45.0|
8.0|  133.0|                   0.0|             45.0|    4.0|
0.0|     175.0|            186.0|       186.0|  3.5|
|        Costa Rica|2005|         53.0|            8.0|               65.0|
44.0|  103.0|                   0.0|             55.0|    8.0|
10.0|     196.0|            212.0|       212.0|  3.7|
|        Costa Rica|2006|         54.0|           44.0|               75.0|
26.0|  138.0|                   0.0|             55.0|    5.0|
20.0|     190.0|            239.0|       239.0| 3.52|
|        Costa Rica|2007|         53.0|           26.0|               80.0|
30.0|  121.0|                   0.0|             55.0|    6.0|
25.0|     199.0|            231.0|       231.0| 3.76|
|        Costa Rica|2008|         55.0|           30.0|               86.0|
30.0|  126.0|                   0.0|             60.0|    6.0|
26.0|     206.0|            242.0|       242.0| 3.75|
|        Costa Rica|2009|         57.0|           30.0|               92.0|
```

```
46.0|  132.0|                0.0|             62.0|   13.0|
30.0|     227.0|            270.0|       270.0| 3.98|
|       Costa Rica|2010|         60.0|           46.0|               96.0|
49.0|  168.0|                0.0|             65.0|   25.0|
31.0|     242.0|            313.0|       313.0| 4.03|
|       Costa Rica|2011|         64.0|           49.0|               97.0|
49.0|  182.0|                0.0|             65.0|   23.0|
32.0|     256.0|            328.0|       328.0|  4.0|
|       Costa Rica|2012|         75.0|           49.0|              102.0|
108.0|  170.0|                0.0|             70.0|   31.0|
32.0|     300.0|            380.0|       380.0|  4.0|
|       Costa Rica|2013|         78.0|           108.0|              102.0|
79.0|  154.0|                0.0|             70.0|   24.0|
32.0|     203.0|            335.0|       335.0|  2.6|
|       Costa Rica|2014|         69.0|           79.0|               95.0|
21.0|  159.0|                0.0|             65.0|    8.0|
30.0|     188.0|            275.0|       275.0| 2.73|
|       Costa Rica|2015|         72.0|           21.0|               85.0|
41.0|  147.0|                0.0|             60.0|    1.0|
25.0|     251.0|            273.0|       273.0| 3.49|
|       Costa Rica|2016|         77.0|           41.0|               70.0|
45.0|  174.0|                0.0|             60.0|    1.0|
10.0|     247.0|            289.0|       289.0| 3.21|
|       Costa Rica|2017|         77.0|           45.0|               60.0|
24.0|  211.0|                0.0|             50.0|    5.0|
10.0|     245.0|            295.0|       295.0| 3.18|
|       Costa Rica|2018|         72.0|           24.0|               50.0|
19.0|  211.0|                0.0|             40.0|   13.0|
10.0|     243.0|            280.0|       280.0| 3.38|
|       Costa Rica|2019|         76.0|           19.0|               45.0|
13.0|  232.0|                0.0|             35.0|   11.0|
10.0|     260.0|            290.0|       290.0| 3.42|
|       Costa Rica|2020|         76.0|           13.0|               49.0|
21.0|  217.0|                0.0|             40.0|    8.0|
9.0|     266.0|            287.0|       287.0|  3.5|
|       Costa Rica|2021|         77.0|           21.0|               50.0|
52.0|  187.0|                0.0|             40.0|    3.0|
10.0|     265.0|            289.0|       289.0| 3.44|
|       Costa Rica|2022|         77.0|           52.0|               52.0|
40.0|  240.0|                0.0|             42.0|   10.0|
10.0|     270.0|            332.0|       332.0| 3.51|
|       Costa Rica|2023|         78.0|           40.0|               50.0|
40.0|  235.0|                0.0|             40.0|   10.0|
10.0|     275.0|            325.0|       325.0| 3.53|
|     Cote d'Ivoire|1964|          0.0|            0.0|               16.0|
0.0|    0.0|                 0.0|             16.0|    1.0|
0.0|      15.0|             16.0|        16.0|  0.0|
|     Cote d'Ivoire|1965|          0.0|            0.0|               19.0|
```

```
0.0|    0.0|                 0.0|              19.0|    4.0|
0.0|      15.0|             19.0|        19.0|  0.0|
|     Cote d'Ivoire|1966|          0.0|            0.0|               16.0|
0.0|    0.0|                 0.0|              16.0|    1.0|
0.0|      15.0|             16.0|        16.0|  0.0|
|     Cote d'Ivoire|1967|          0.0|            0.0|               26.0|
0.0|    0.0|                 0.0|              26.0|    2.0|
0.0|      24.0|             26.0|        26.0|  0.0|
|     Cote d'Ivoire|1968|          0.0|            0.0|               47.0|
0.0|    1.0|                 0.0|              47.0|    1.0|
0.0|      47.0|             48.0|        48.0|  0.0|
|     Cote d'Ivoire|1969|          0.0|            0.0|               46.0|
0.0|   13.0|                 0.0|              46.0|    0.0|
0.0|      59.0|             59.0|        59.0|  0.0|
|     Cote d'Ivoire|1970|          0.0|            0.0|               64.0|
0.0|   28.0|                 0.0|              64.0|    0.0|
0.0|      92.0|             92.0|        92.0|  0.0|
|     Cote d'Ivoire|1971|          0.0|            0.0|               44.0|
0.0|   48.0|                 0.0|              44.0|    0.0|
0.0|      92.0|             92.0|        92.0|  0.0|
|     Cote d'Ivoire|1972|          0.0|            0.0|               85.0|
0.0|   55.0|                 0.0|              85.0|    0.0|
0.0|     140.0|            140.0|       140.0|  0.0|
|     Cote d'Ivoire|1973|          0.0|            0.0|               48.0|
0.0|  102.0|                 0.0|              48.0|    0.0|
0.0|     150.0|            150.0|       150.0|  0.0|
|     Cote d'Ivoire|1974|          0.0|            0.0|               57.0|
0.0|   94.0|                 0.0|              57.0|    0.0|
0.0|     151.0|            151.0|       151.0|  0.0|
|     Cote d'Ivoire|1975|          0.0|            0.0|               34.0|
0.0|   92.0|                 0.0|              34.0|    0.0|
0.0|     126.0|            126.0|       126.0|  0.0|
|     Cote d'Ivoire|1976|          0.0|            0.0|               56.0|
0.0|   79.0|                 0.0|              56.0|    0.0|
0.0|     135.0|            135.0|       135.0|  0.0|
|     Cote d'Ivoire|1977|          0.0|            0.0|               63.0|
0.0|   75.0|                 0.0|              63.0|    0.0|
0.0|     138.0|            138.0|       138.0|  0.0|
|     Cote d'Ivoire|1978|          0.0|            0.0|               75.0|
0.0|   49.0|                 0.0|              75.0|    0.0|
0.0|     124.0|            124.0|       124.0|  0.0|
|     Cote d'Ivoire|1979|          0.0|            0.0|              102.0|
0.0|   80.0|                 0.0|             102.0|    0.0|
0.0|     182.0|            182.0|       182.0|  0.0|
|     Cote d'Ivoire|1980|          0.0|            0.0|               84.0|
0.0|   63.0|                 0.0|              84.0|    0.0|
0.0|     147.0|            147.0|       147.0|  0.0|
|     Cote d'Ivoire|1981|          0.0|            0.0|               92.0|
```

```
0.0|   68.0|                 0.0|              92.0|    0.0|
0.0|     160.0|            160.0|       160.0|  0.0|
|     Cote d'Ivoire|1982|          0.0|             0.0|              104.0|
0.0|   58.0|                 0.0|             104.0|    0.0|
0.0|     162.0|            162.0|       162.0|  0.0|
|     Cote d'Ivoire|1983|          0.0|             0.0|              111.0|
0.0|   56.0|                 0.0|             111.0|    0.0|
0.0|     167.0|            167.0|       167.0|  0.0|
|     Cote d'Ivoire|1984|          0.0|             0.0|              108.0|
0.0|   56.0|                 0.0|              98.0|    0.0|
10.0|     164.0|            164.0|       164.0|  0.0|
|     Cote d'Ivoire|1985|          0.0|             0.0|              123.0|
0.0|   94.0|                 0.0|             112.0|    0.0|
11.0|     217.0|            217.0|       217.0|  0.0|
|     Cote d'Ivoire|1986|          0.0|             0.0|              107.0|
0.0|  120.0|                 0.0|              64.0|    0.0|
43.0|     227.0|            227.0|       227.0|  0.0|
|     Cote d'Ivoire|1987|         91.0|             0.0|              133.0|
0.0|   65.0|                 0.0|              71.0|    0.0|
62.0|     198.0|            198.0|       198.0| 2.18|
|     Cote d'Ivoire|1988|         93.0|             0.0|              151.0|
0.0|   52.0|                 0.0|              79.0|    0.0|
72.0|     203.0|            203.0|       203.0| 2.18|
|     Cote d'Ivoire|1989|        106.0|             0.0|              131.0|
0.0|  144.0|                 0.0|              85.0|    0.0|
46.0|     275.0|            275.0|       275.0| 2.59|
|     Cote d'Ivoire|1990|        133.0|             0.0|              150.0|
0.0|  128.0|                 0.0|              95.0|    0.0|
55.0|     278.0|            278.0|       278.0| 2.09|
|     Cote d'Ivoire|1991|        143.0|             0.0|              148.0|
0.0|  133.0|                 0.0|              98.0|    0.0|
50.0|     281.0|            281.0|       281.0| 1.97|
|     Cote d'Ivoire|1992|        148.0|             0.0|              135.0|
18.0|  140.0|                  0.0|              95.0|    0.0|
40.0|     293.0|            293.0|       293.0| 1.98|
|     Cote d'Ivoire|1993|        154.0|            18.0|              125.0|
5.0|  184.0|                 0.0|              88.0|    0.0|
37.0|     296.0|            314.0|       314.0| 1.92|
|     Cote d'Ivoire|1994|        157.0|             5.0|              160.0|
16.0|  120.0|                  0.0|             110.0|    5.0|
50.0|     286.0|            296.0|       296.0| 1.82|
|     Cote d'Ivoire|1995|        159.0|            16.0|              200.0|
14.0|  106.0|                  0.0|             125.0|    0.0|
75.0|     304.0|            320.0|       320.0| 1.91|
|     Cote d'Ivoire|1996|        163.0|            14.0|              200.0|
9.0|   90.0|                 0.0|             120.0|    0.0|
80.0|     285.0|            299.0|       299.0| 1.75|
|     Cote d'Ivoire|1997|        166.0|             9.0|              214.0|
```

```
1.0|   99.0|                 0.0|              139.0|   10.0|
75.0|     295.0|             314.0|       314.0| 1.78|
|     Cote d'Ivoire|1998|         166.0|             1.0|               242.0|
1.0|   73.0|                 0.0|              167.0|   10.0|
75.0|     305.0|             316.0|       316.0| 1.84|
|     Cote d'Ivoire|1999|         166.0|             1.0|               281.0|
1.0|   17.0|                 0.0|              200.0|   18.0|
81.0|     280.0|             299.0|       299.0| 1.69|
|     Cote d'Ivoire|2000|         159.0|             1.0|               205.0|
55.0|    1.0|                  0.0|              155.0|   12.0|
50.0|     248.0|             261.0|       261.0| 1.56|
|     Cote d'Ivoire|2001|         164.0|              55.0|               205.0|
126.0|    7.0|                 0.0|              155.0|   23.0|
50.0|     260.0|             338.0|       338.0| 1.59|
|     Cote d'Ivoire|2002|         176.0|              126.0|               202.0|
185.0|    1.0|                 0.0|              155.0|   28.0|
47.0|     234.0|             388.0|       388.0| 1.33|
|     Cote d'Ivoire|2003|         172.0|              185.0|               195.0|
192.0|  115.0|                 0.0|              150.0|    9.0|
45.0|     308.0|             502.0|       502.0| 1.79|
+------------------+----+--------------+---------------+--------------------+--
-----------+-------+--------------------+------------------+-------+----------
-----------+----------+------------------+------------+-----+
only showing top 1000 rows
```

2 2. API de DataFrames:

Para realizar transformaciones en un DataFrame en Spark con Python, se utilizan diversas funciones que permiten modificar, seleccionar, o agregar datos. Esta es la sintaxis y ejemplos de uso de algunas de las transformaciones más comunes:

2.1 Transformaciones:

- **select**: Permite seleccionar columnas específicas de un DataFrame.
- **filter**: Permite filtrar filas basadas en una condición.
- **withColumn**: Permite añadir nuevas columnas o modificar las existentes.
- **Otras transformaciones**: La API incluye otras transformaciones para manipular los datos como groupBy, sort y join. También permite crear funciones definidas por el usuario para manipulación personalizada de datos.

2.2 Acciones:

- **show**: Muestra las primeras filas de un DataFrame.
- **count**: Cuenta el número de filas en un DataFrame.
- **collect**: Retorna todos los elementos de un DataFrame al driver (cuidado con el uso en grandes datasets).

- **Otras acciones**: Incluyen take, takeSample y describe para obtener información y estadísticas sobre los DataFrames.

2.3 Consideraciones:

- **Inmutabilidad**: Los DataFrames son inmutables; cada transformación crea un nuevo DataFrame.
- **show()**: La función `show()` se utiliza para mostrar una muestra de los datos resultantes tras una transformación.
- **Importaciones**: Algunas funciones requieren importaciones adicionales desde `pyspark.sql.functions`, como `col`, `lit`, `expr`, `avg`, `count`, etc.
- **Expresiones SQL**: Puedes usar expresiones SQL con `expr()` y `selectExpr()` para transformaciones más complejas.
- **Columnas**: Las columnas se pueden referenciar usando su nombre como string, usando la notación de corchetes sobre el DataFrame o con la función `col()`.

2.3.1 1. select():

Se utiliza para seleccionar un subconjunto de columnas de un DataFrame. También se puede usar `selectExpr()` para seleccionar columnas con expresiones SQL. https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.select.html

[0]:
```
df=spark.read.option("header", "true").option("inferSchema", "true").csv("dbfs:/
  ↪FileStore/olive.csv")

df.select("Country", "Year", "Production").show()
df.select('*').show()
df.select(df.Country, (df.Production * 1000).alias('Production (Tm)')).show()
```

```
+-------+----+----------+
|Country|Year|Production|
+-------+----+----------+
|Algeria|1964|        18|
|Algeria|1965|        17|
|Algeria|1966|        16|
|Algeria|1967|        22|
|Algeria|1968|        18|
|Algeria|1969|        22|
|Algeria|1970|        13|
|Algeria|1971|        23|
|Algeria|1972|        15|
|Algeria|1973|        16|
|Algeria|1974|         8|
|Algeria|1975|        18|
|Algeria|1976|        15|
|Algeria|1977|         5|
|Algeria|1978|        14|
|Algeria|1979|        10|
|Algeria|1980|        18|
```

```
|Algeria|1981|        15|
|Algeria|1982|        16|
|Algeria|1983|        13|
+-------+----+----------+
only showing top 20 rows

+-------+----+---------------+-------------------+-------------+-------+------
--------------+------------------+-------+--------------------+----------+---
--------------+------------+
|Country|Year|Beginning Stocks|Domestic Consumption|Ending Stocks|Exports|Feed
Waste Dom. Cons.|Food Use Dom. Cons.|Imports|Industrial Dom.
Cons.|Production|Total Distribution|Total Supply|
+-------+----+---------------+-------------------+-------------+-------+------
--------------+------------------+-------+--------------------+----------+---
--------------+------------+
|Algeria|1964|              0|                 15|            0|      3|
0|                15|      0|                  0|        18|
18|          18|
|Algeria|1965|              0|                 12|            0|      5|
0|                12|      0|                  0|        17|
17|          17|
|Algeria|1966|              0|                 16|            0|      0|
0|                16|      0|                  0|        16|
16|          16|
|Algeria|1967|              0|                 15|            0|      7|
0|                15|      0|                  0|        22|
22|          22|
|Algeria|1968|              0|                 11|            0|      7|
0|                11|      0|                  0|        18|
18|          18|
|Algeria|1969|              0|                 19|            0|      3|
0|                19|      0|                  0|        22|
22|          22|
|Algeria|1970|              0|                 12|            0|      1|
0|                12|      0|                  0|        13|
13|          13|
|Algeria|1971|              0|                 20|            2|      1|
0|                20|      0|                  0|        23|
23|          23|
|Algeria|1972|              2|                 13|            2|      2|
0|                13|      0|                  0|        15|
17|          17|
|Algeria|1973|              2|                 11|            3|      4|
0|                11|      0|                  0|        16|
18|          18|
|Algeria|1974|              3|                 10|            1|      0|
0|                10|      0|                  0|         8|
11|          11|
```

```
|Algeria|1975|              1|                16|            3|      0|
0|                16|      0|                 0|        18|
19|          19|
|Algeria|1976|              3|                15|            3|      0|
0|                15|      0|                 0|        15|
18|          18|
|Algeria|1977|              3|                 8|            0|      0|
0|                 8|      0|                 0|         5|
8|           8|
|Algeria|1978|              0|                13|            1|      0|
0|                13|      0|                 0|        14|
14|          14|
|Algeria|1979|              1|                11|            0|      0|
0|                11|      0|                 0|        10|
11|          11|
|Algeria|1980|              0|                18|            0|      0|
0|                18|      0|                 0|        18|
18|          18|
|Algeria|1981|              0|                15|            0|      0|
0|                15|      0|                 0|        15|
15|          15|
|Algeria|1982|              0|                16|            0|      0|
0|                16|      0|                 0|        16|
16|          16|
|Algeria|1983|              0|                13|            0|      0|
0|                13|      0|                 0|        13|
13|          13|
+-------+----+---------------+-------------------+-------------+-------+------
---------------+------------------+-------+--------------------+----------+---
---------------+------------+
only showing top 20 rows

+-------+---------------+
|Country|Production (Tm)|
+-------+---------------+
|Algeria|          18000|
|Algeria|          17000|
|Algeria|          16000|
|Algeria|          22000|
|Algeria|          18000|
|Algeria|          22000|
|Algeria|          13000|
|Algeria|          23000|
|Algeria|          15000|
|Algeria|          16000|
|Algeria|           8000|
|Algeria|          18000|
|Algeria|          15000|
```

```
|Algeria|           5000|
|Algeria|          14000|
|Algeria|          10000|
|Algeria|          18000|
|Algeria|          15000|
|Algeria|          16000|
|Algeria|          13000|
+-------+---------------+
only showing top 20 rows
```

2.3.2 2. filter() o where():

Se utiliza para filtrar filas basadas en una condición. `filter()` y `where()` son sinónimos. https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.filter.html

[0]:
```
df.filter((df["Country"] == "Spain") & (df["Year"] < 1984)).select("Country",␣
  ↪"Year", "Production").show()
```

```
+-------+----+----------+
|Country|Year|Production|
+-------+----+----------+
|  Spain|1964|       110|
|  Spain|1965|       324|
|  Spain|1966|       437|
|  Spain|1967|       259|
|  Spain|1968|       480|
|  Spain|1969|       358|
|  Spain|1970|       475|
|  Spain|1971|       341|
|  Spain|1972|       440|
|  Spain|1973|       447|
|  Spain|1974|       333|
|  Spain|1975|       455|
|  Spain|1976|       390|
|  Spain|1977|       361|
|  Spain|1978|       500|
|  Spain|1979|       433|
|  Spain|1980|       479|
|  Spain|1981|       300|
|  Spain|1982|       666|
|  Spain|1983|       258|
+-------+----+----------+
```

2.3.3 3. withColumn():

Se utiliza para añadir una nueva columna o reemplazar una existente. La función `lit()` crea una columna con un valor literal y `expr()` permite usar expresiones SQL.

https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.withColumn.

[0]:
```
from pyspark.sql.functions import lit, expr
from datetime import datetime

df=spark.read.option("header", "true").option("inferSchema", "true").csv("dbfs:/
 ↪FileStore/olive.csv")
df_Spain = df.filter((df["Country"] == "Spain") | (df["Country"] == "European␣
 ↪Union")).select("Country", "Year", "Production")

df_Spain = df_Spain.withColumn("Production (Tm)", df_Spain["Production"]*1000)
df_Spain = df_Spain.withColumn("Region", expr("CASE WHEN Year <= 1990 THEN␣
 ↪'Spain' ELSE 'EU' END"))

current_year = datetime.now().year
df_Spain = df_Spain.withColumn("Diff_Years", lit(current_year) -␣
 ↪df_Spain["Year"])

df_Spain.show(df_Spain.count())
```

```
+--------------+----+----------+---------------+------+----------+
|       Country|Year|Production|Production (Tm)|Region|Diff_Years|
+--------------+----+----------+---------------+------+----------+
|European Union|1964|         0|              0| Spain|        61|
|European Union|1965|         0|              0| Spain|        60|
|European Union|1966|         0|              0| Spain|        59|
|European Union|1967|         0|              0| Spain|        58|
|European Union|1968|         0|              0| Spain|        57|
|European Union|1969|         0|              0| Spain|        56|
|European Union|1970|         0|              0| Spain|        55|
|European Union|1971|         0|              0| Spain|        54|
|European Union|1972|         0|              0| Spain|        53|
|European Union|1973|         0|              0| Spain|        52|
|European Union|1974|         0|              0| Spain|        51|
|European Union|1975|         0|              0| Spain|        50|
|European Union|1976|         0|              0| Spain|        49|
|European Union|1977|         0|              0| Spain|        48|
|European Union|1978|         0|              0| Spain|        47|
|European Union|1979|         0|              0| Spain|        46|
|European Union|1980|         0|              0| Spain|        45|
|European Union|1981|         0|              0| Spain|        44|
|European Union|1982|         0|              0| Spain|        43|
|European Union|1983|         0|              0| Spain|        42|
|European Union|1984|         0|              0| Spain|        41|
|European Union|1985|         0|              0| Spain|        40|
|European Union|1986|         0|              0| Spain|        39|
|European Union|1987|         0|              0| Spain|        38|
|European Union|1988|         0|              0| Spain|        37|
```

European Union	1989	0	0	Spain	36
European Union	1990	0	0	Spain	35
European Union	1991	0	0	EU	34
European Union	1992	0	0	EU	33
European Union	1993	0	0	EU	32
European Union	1994	0	0	EU	31
European Union	1995	0	0	EU	30
European Union	1996	0	0	EU	29
European Union	1997	0	0	EU	28
European Union	1998	0	0	EU	27
European Union	1999	1867	1867000	EU	26
European Union	2000	1871	1871000	EU	25
European Union	2001	2402	2402000	EU	24
European Union	2002	1944	1944000	EU	23
European Union	2003	2415	2415000	EU	22
European Union	2004	2350	2350000	EU	21
European Union	2005	2025	2025000	EU	20
European Union	2006	2132	2132000	EU	19
European Union	2007	2235	2235000	EU	18
European Union	2008	2110	2110000	EU	17
European Union	2009	2450	2450000	EU	16
European Union	2010	2500	2500000	EU	15
European Union	2011	2700	2700000	EU	14
European Union	2012	1625	1625000	EU	13
European Union	2013	2483	2483000	EU	12
European Union	2014	1435	1435000	EU	11
European Union	2015	2324	2324000	EU	10
European Union	2016	1752	1752000	EU	9
European Union	2017	2188	2188000	EU	8
European Union	2018	2264	2264000	EU	7
European Union	2019	1925	1925000	EU	6
European Union	2020	2051	2051000	EU	5
European Union	2021	2272	2272000	EU	4
European Union	2022	1392	1392000	EU	3
European Union	2023	1415	1415000	EU	2
Spain	1964	110	110000	Spain	61
Spain	1965	324	324000	Spain	60
Spain	1966	437	437000	Spain	59
Spain	1967	259	259000	Spain	58
Spain	1968	480	480000	Spain	57
Spain	1969	358	358000	Spain	56
Spain	1970	475	475000	Spain	55
Spain	1971	341	341000	Spain	54
Spain	1972	440	440000	Spain	53
Spain	1973	447	447000	Spain	52
Spain	1974	333	333000	Spain	51
Spain	1975	455	455000	Spain	50
Spain	1976	390	390000	Spain	49

	Spain	1977	361	361000	Spain	48
	Spain	1978	500	500000	Spain	47
	Spain	1979	433	433000	Spain	46
	Spain	1980	479	479000	Spain	45
	Spain	1981	300	300000	Spain	44
	Spain	1982	666	666000	Spain	43
	Spain	1983	258	258000	Spain	42
	Spain	1984	689	689000	Spain	41
	Spain	1985	397	397000	Spain	40
	Spain	1986	493	493000	Spain	39
	Spain	1987	691	691000	Spain	38
	Spain	1988	376	376000	Spain	37
	Spain	1989	551	551000	Spain	36
	Spain	1990	639	639000	Spain	35
	Spain	1991	0	0	EU	34
	Spain	1992	0	0	EU	33
	Spain	1993	0	0	EU	32
	Spain	1994	0	0	EU	31
	Spain	1995	0	0	EU	30
	Spain	1996	0	0	EU	29
	Spain	1997	0	0	EU	28
	Spain	1998	0	0	EU	27
	Spain	1999	0	0	EU	26
	Spain	2000	0	0	EU	25
	Spain	2001	0	0	EU	24
	Spain	2002	0	0	EU	23
	Spain	2003	0	0	EU	22
	Spain	2004	0	0	EU	21
	Spain	2005	0	0	EU	20
	Spain	2006	0	0	EU	19
	Spain	2007	0	0	EU	18
	Spain	2008	0	0	EU	17
	Spain	2009	0	0	EU	16
	Spain	2010	0	0	EU	15
	Spain	2011	0	0	EU	14
	Spain	2012	0	0	EU	13
	Spain	2013	0	0	EU	12
	Spain	2014	0	0	EU	11
	Spain	2015	0	0	EU	10
	Spain	2016	0	0	EU	9
	Spain	2017	0	0	EU	8
	Spain	2018	0	0	EU	7
	Spain	2019	0	0	EU	6
	Spain	2020	0	0	EU	5
	Spain	2021	0	0	EU	4
	Spain	2022	0	0	EU	3
	Spain	2023	0	0	EU	2

2.3.4 4. withColumnRenamed():

Se utiliza para renombrar una columna existente.
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.withColumnl

[0]:
```
df_Spain.withColumnRenamed("Country", "País").withColumnRenamed("Year", "Año").
↪show()
```

```
+--------------+----+----------+---------------+------+----------+
|          País| Año|Production|Production (Tm)|Region|Diff_Years|
+--------------+----+----------+---------------+------+----------+
|European Union|1964|         0|              0| Spain|        61|
|European Union|1965|         0|              0| Spain|        60|
|European Union|1966|         0|              0| Spain|        59|
|European Union|1967|         0|              0| Spain|        58|
|European Union|1968|         0|              0| Spain|        57|
|European Union|1969|         0|              0| Spain|        56|
|European Union|1970|         0|              0| Spain|        55|
|European Union|1971|         0|              0| Spain|        54|
|European Union|1972|         0|              0| Spain|        53|
|European Union|1973|         0|              0| Spain|        52|
|European Union|1974|         0|              0| Spain|        51|
|European Union|1975|         0|              0| Spain|        50|
|European Union|1976|         0|              0| Spain|        49|
|European Union|1977|         0|              0| Spain|        48|
|European Union|1978|         0|              0| Spain|        47|
|European Union|1979|         0|              0| Spain|        46|
|European Union|1980|         0|              0| Spain|        45|
|European Union|1981|         0|              0| Spain|        44|
|European Union|1982|         0|              0| Spain|        43|
|European Union|1983|         0|              0| Spain|        42|
+--------------+----+----------+---------------+------+----------+
only showing top 20 rows
```

2.3.5 5. groupBy():

Se utiliza para agrupar filas con valores iguales en una columna y realizar operaciones de agregación. Se combina con funciones de agregación como `count()`, `sum()`, `avg()`, `min()`, `max()`.
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.groupBy.htm

[0]:
```
# Podemos realizarlo de dos formas: utilizando funciones de F o con␣
↪diccionarios.
# La primera es más clara y permite realizar varias agregaciones sobre la misma␣
↪columna.
```

```
df=spark.read.parquet("dbfs:/FileStore/palm.parquet")

from pyspark.sql import functions as F
df.select(df.Country, df.Year, df.Production) \
    .groupBy("Country") \
    .agg(
        F.sum("Production").alias("TotalProd"),
        F.max("Production").alias("MaxProd")
    ) \
    .orderBy("TotalProd",ascending=False) \
.show(4)

df.select(df.Country, df.Year, df.Production) \
    .groupBy("Country") \
    .agg({"Production": "sum"}) \
    .withColumnRenamed("sum(Production)", "TotalProd") \
    .orderBy("TotalProd", ascending=False) \
.show(4)
```

```
+---------+---------+-------+
|  Country|TotalProd|MaxProd|
+---------+---------+-------+
|Indonesia| 721653.0|47000.0|
| Malaysia| 552873.0|20800.0|
| Thailand|  49854.0| 3450.0|
|  Nigeria|  43015.0| 1500.0|
+---------+---------+-------+
only showing top 4 rows

+---------+---------+
|  Country|TotalProd|
+---------+---------+
|Indonesia| 721653.0|
| Malaysia| 552873.0|
| Thailand|  49854.0|
|  Nigeria|  43015.0|
+---------+---------+
only showing top 4 rows
```

2.3.6 6. sort() o orderBy():

Se utiliza para ordenar las filas del DataFrame. `sort()` y `orderBy()` son equivalentes y pueden usar el orden ascendente (`asc`) o descendente (`desc`).
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.sort.html
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.orderBy.html

[0]:
```
from pyspark.sql import functions as F
df=spark.read.parquet("dbfs:/FileStore/palm.parquet")
df = df.select(df.Country, df.Year, df.Production).groupBy("Country","Year").
 ↪agg(F.sum("Production").alias("TotalProd"))
df.sort("Country").show(5)
df.sort(df["Country"].desc()).show(5)
df.orderBy("TotalProd").show(5)
df.orderBy(df["TotalProd"].desc()).show(5)
```

```
+-----------+----+---------+
|    Country|Year|TotalProd|
+-----------+----+---------+
|Afghanistan|2022|      0.0|
|Afghanistan|2021|      0.0|
|Afghanistan|1999|      0.0|
|Afghanistan|2009|      0.0|
|Afghanistan|2012|      0.0|
+-----------+----+---------+
only showing top 5 rows

+--------+----+---------+
| Country|Year|TotalProd|
+--------+----+---------+
|Zimbabwe|1965|      0.0|
|Zimbabwe|1964|      0.0|
|Zimbabwe|1991|      0.0|
|Zimbabwe|1978|      0.0|
|Zimbabwe|2013|      0.0|
+--------+----+---------+
only showing top 5 rows

+-----------------+----+---------+
|          Country|Year|TotalProd|
+-----------------+----+---------+
|      El Salvador|1975|      0.0|
|            Italy|2023|      0.0|
|Former Yugoslavia|1971|      0.0|
|            Haiti|2012|      0.0|
|          Ireland|1976|      0.0|
+-----------------+----+---------+
only showing top 5 rows

+---------+----+---------+
|  Country|Year|TotalProd|
+---------+----+---------+
|Indonesia|2023|  47000.0|
|Indonesia|2022|  46500.0|
```

```
|Indonesia|2020|  43500.0|
|Indonesia|2019|  42500.0|
|Indonesia|2021|  42000.0|
+---------+----+---------+
only showing top 5 rows
```

2.3.7 7. drop():

Se utiliza para eliminar una o varias columnas del DataFrame.
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.drop.html

[0]:
```
df_Spain.show(5)
df_Spain_reducido = df_Spain.drop("Year", "Diff_Years")
df_Spain_reducido.show(5)
```

```
+--------------+----+----------+---------------+------+----------+
|       Country|Year|Production|Production (Tm)|Region|Diff_Years|
+--------------+----+----------+---------------+------+----------+
|European Union|1964|         0|              0| Spain|        61|
|European Union|1965|         0|              0| Spain|        60|
|European Union|1966|         0|              0| Spain|        59|
|European Union|1967|         0|              0| Spain|        58|
|European Union|1968|         0|              0| Spain|        57|
+--------------+----+----------+---------------+------+----------+
only showing top 5 rows

+--------------+----------+---------------+------+
|       Country|Production|Production (Tm)|Region|
+--------------+----------+---------------+------+
|European Union|         0|              0| Spain|
|European Union|         0|              0| Spain|
|European Union|         0|              0| Spain|
|European Union|         0|              0| Spain|
|European Union|         0|              0| Spain|
+--------------+----------+---------------+------+
only showing top 5 rows
```

2.3.8 8. distinct():

Se utiliza para eliminar las filas duplicadas del DataFrame.
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.distinct.html

[0]:
```
df_Spain_reducido.distinct().show()
```

```
+--------------+----------+---------------+------+
|       Country|Production|Production (Tm)|Region|
+--------------+----------+---------------+------+
```

```
|European Union|         0|              0|    EU|
|European Union|      2188|        2188000|    EU|
|European Union|      2500|        2500000|    EU|
|European Union|      2110|        2110000|    EU|
|European Union|      2324|        2324000|    EU|
|European Union|      1944|        1944000|    EU|
|European Union|      2025|        2025000|    EU|
|European Union|      2350|        2350000|    EU|
|European Union|      2483|        2483000|    EU|
|European Union|         0|              0| Spain|
|European Union|      1435|        1435000|    EU|
|European Union|      2132|        2132000|    EU|
|European Union|      2415|        2415000|    EU|
|European Union|      2700|        2700000|    EU|
|European Union|      2235|        2235000|    EU|
|European Union|      2402|        2402000|    EU|
|European Union|      1752|        1752000|    EU|
|European Union|      1871|        1871000|    EU|
|European Union|      1867|        1867000|    EU|
|European Union|      1625|        1625000|    EU|
+--------------+----------+---------------+------+
only showing top 20 rows
```

2.3.9 9. join():

Se utiliza para combinar dos DataFrames basados en una o varias columnas en común. Se puede especificar el tipo de join: 'inner', 'outer', 'left_outer', 'right_outer', o 'leftsemi'.
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.join.html

[0]:
```python
from pyspark.sql.types import StructType, StructField, IntegerType, StringType
from pyspark.sql import functions as F

# Definir el esquema del DataFrame de ratings
esquemaRatings = StructType([
    StructField("UserID", IntegerType(), True),
    StructField("MovieID", IntegerType(), True),
    StructField("Rating", IntegerType(), True),
    StructField("Timestamp", IntegerType(), True)
])

# Cargar el archivo de ratings u-data en un DataFrame, con separador el␣
 ↪tabulador \t
dfRatings = spark.read.csv("dbfs:/FileStore/u.data", sep="\t",␣
 ↪schema=esquemaRatings, header=False)

# Definir esquema para el DataFrame de películas
esquemaPeliculas = StructType([
```

```
    StructField("MovieID", IntegerType(), True),
    StructField("Title", StringType(), True),
    StructField("ReleaseDate", StringType(), True),
    StructField("EmptyColumn", StringType(), True),
    StructField("IMDB_URL", StringType(), True),
    StructField("Unknown", IntegerType(), True),
    StructField("Action", IntegerType(), True),
    StructField("Adventure", IntegerType(), True),
    StructField("Animation", IntegerType(), True),
    StructField("Children", IntegerType(), True),
    StructField("Comedy", IntegerType(), True),
    StructField("Crime", IntegerType(), True),
    StructField("Documentary", IntegerType(), True),
    StructField("Drama", IntegerType(), True),
    StructField("Fantasy", IntegerType(), True),
    StructField("FilmNoir", IntegerType(), True),
    StructField("Horror", IntegerType(), True),
    StructField("Musical", IntegerType(), True),
    StructField("Mystery", IntegerType(), True),
    StructField("Romance", IntegerType(), True),
    StructField("SciFi", IntegerType(), True),
    StructField("Thriller", IntegerType(), True),
    StructField("War", IntegerType(), True),
    StructField("Western", IntegerType(), True)
])

# Cargar el archivo de películas en un DataFrame, con separador |
dfPeliculas = spark.read.csv("dbfs:/FileStore/u.item", sep="|",␣
 ↪schema=esquemaPeliculas, header=False)

# Mostrar las 10 películas con más votos
dfRatingsNombres = dfRatings.join(dfPeliculas,on="MovieID",how="inner")
dfRatingsAgrupados = dfRatingsNombres.groupBy("Title").agg(F.count("Title").
 ↪alias("Ratings")).orderBy("Ratings",ascending=False)
dfRatingsAgrupados.show(10)
```

```
+--------------------+-------+
|               Title|Ratings|
+--------------------+-------+
|    Star Wars (1977)|    583|
|      Contact (1997)|    509|
|        Fargo (1996)|    508|
|Return of the Jed…|    507|
|    Liar Liar (1997)|    485|
|English Patient, …|    481|
|       Scream (1996)|    478|
|    Toy Story (1995)|    452|
```

```
|Air Force One (1997)|    431|
|Independence Day …|    429|
+--------------------+-------+
only showing top 10 rows
```

2.3.10 10. union() o unionAll():

Se utiliza para combinar dos DataFrames con las mismas columnas. `union()` elimina duplicados, `unionAll()` no.
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.union.html y https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.unionAll.ht

[0]:
```
# Crear dos dataframes con los mismos campos (películas más veces puntuadas y␣
 ↪películas mejor puntuadas)
data_best = [
    ("Inception", 200, 8.5),
    ("Avatar", 150, 7.8),
    ("Titanic", 180, 7.9)
]
dfRatingsBest = spark.createDataFrame(data_best, ["Title", "Ratings",␣
 ↪"MeanRating"])

data_top = [
    ("The Dark Knight", 220, 9.0),
    ("Inception", 200, 8.5),
    ("Interstellar", 130, 8.6)
]
dfRatingsTop = spark.createDataFrame(data_top, ["Title", "Ratings",␣
 ↪"MeanRating"])

# Hacer la unión y mostrarlos.
dfUnion = dfRatingsBest.union(dfRatingsTop)
dfUnion.show()

# Películas más veces puntuadas
# dfRatingsBest (Title, Ratings, MeanRating)

# Películas mejor puntuadas, con más de 100 votos
# dfRatingsTop (Title, Ratings, MeanRating)

# dfRatingsBest.union(dfRatingsTop).show()
```

```
+--------------+-------+----------+
|         Title|Ratings|MeanRating|
+--------------+-------+----------+
|     Inception|    200|       8.5|
|        Avatar|    150|       7.8|
```

```
|        Titanic|    180|       7.9|
|The Dark Knight|    220|       9.0|
|      Inception|    200|       8.5|
|   Interstellar|    130|       8.6|
+---------------+-------+----------+
```

2.3.11 11. map():

Se utiliza para aplicar una función a cada fila del DataFrame, convirtiéndolo a RDD.
La función map() se aplica a RDDs (Resilient Distributed Datasets), no directamente a DataFrames.
Para usar map en un DataFrame, primero debes convertirlo a un RDD usando df.rdd.
https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.RDD.map.html

[0]:
```python
# Obtener una lista de nombres de películas y un diccionario con el número de␣
 ↪votos en cada puntuación:

# Nos quedamos con las columnas MovieID y Rating
dfRatings = dfRatings.select("MovieID", "Rating")

# Convertir el DataFrame de ratings a un RDD de filas
rddFilas = dfRatings.rdd

# Convertir el RDD de filas a un RDD de tuplas
rddTuplas = rddFilas.map(lambda fila: (fila[0], (fila[1],1)))

# Función para crear un diccionario con el número de votos para cada puntuación
def crearRatingDict(tuplas):
    RatingDict = {}
    for rating, cont in tuplas:
        if rating in RatingDict:
            RatingDict[rating] += cont
        else:
            RatingDict[rating] = cont
    return RatingDict

# Agrupar por MovieID y agregar las puntuaciones
rddRatingsAgrupados = rddTuplas.groupByKey().mapValues(crearRatingDict)

# Volver a convertir a dataframe y hacer join con películas para obtener el␣
 ↪nombre
dfRatingDict = rddRatingsAgrupados.toDF(["MovieID", "RatingDict"])
dfJoined = dfPeliculas.join(dfRatingDict, on="MovieID", how="inner")

# Mostrar 10 películas con su nombre y puntuaciones
dfJoined.select("Title", "RatingDict").show(10, truncate=False)
```