Mod13_Sesion3_Streaming_Pyspark

September 23, 2023

[1]:

```
from pyspark.sql.types import StructType, StructField, StringType, DateType
from pyspark.sql.functions import to_date
```

[2]:

```
custom_schema = StructType([
    StructField("Palabra", StringType(), True),
    StructField("Timestamp", StringType(), True),
    StructField("Timestamp1", StringType(), True)
])
```

[3]:

```
df = spark.read.csv("/tmp/dcd/streamdata/*.csv", custom_schema)

df.printSchema()

df.show()
```

```
root
 |-- Palabra: string (nullable = true)
 |-- Timestamp: string (nullable = true)
 |-- Timestamp1: string (nullable = true)
```

```
+--------+--------------------+--------------------+
| Palabra|           Timestamp|          Timestamp1|
+--------+--------------------+--------------------+
|ALMACENA|2023-09-23T16:32:…|2023-09-23T16:32:…|
|STREMING|2023-09-23T16:32:…|2023-09-23T16:32:…|
| INTENTO|2023-09-23T16:32:…|2023-09-23T16:32:…|
| Ramirez|2023-09-23T16:26:…|2023-09-23T16:26:…|
| RAMIREZ|2023-09-23T16:32:…|2023-09-23T16:32:…|
| SEGUNDO|2023-09-23T16:32:…|2023-09-23T16:32:…|
|  Sergio|2023-09-23T16:26:…|2023-09-23T16:26:…|
|  SERGIO|2023-09-23T16:32:…|2023-09-23T16:32:…|
|  IBARRA|2023-09-23T16:32:…|2023-09-23T16:32:…|
|  Ibarra|2023-09-23T16:26:…|2023-09-23T16:26:…|
|   SALIR|2023-09-23T16:33:…|2023-09-23T16:33:…|
|   Salir|2023-09-23T16:27:…|2023-09-23T16:27:…|
|   batch|2023-09-23T16:26:…|2023-09-23T16:26:…|
|   Batch|2023-09-23T16:26:…|2023-09-23T16:26:…|
|   DATOS|2023-09-23T16:32:…|2023-09-23T16:32:…|
|    Otro|2023-09-23T16:26:…|2023-09-23T16:26:…|
|    HOLA|2023-09-23T16:32:…|2023-09-23T16:32:…|
|    HOLA|2023-09-23T16:26:…|2023-09-23T16:26:…|
|     LOS|2023-09-23T16:32:…|2023-09-23T16:32:…|
|     VER|2023-09-23T16:32:…|2023-09-23T16:32:…|
+--------+--------------------+--------------------+
only showing top 20 rows
```

[4]:

```
from pyspark.sql.functions import to_timestamp
df = df.withColumn("Timestamp", to_timestamp("Timestamp", "yyyy-MM-dd'T'HH:mm:
 ↪ss.SSS'Z'"))

df.printSchema()

df.show()
```

```
root
 |-- Palabra: string (nullable = true)
 |-- Timestamp: timestamp (nullable = true)
 |-- Timestamp1: string (nullable = true)



+--------+--------------------+--------------------+
| Palabra|           Timestamp|          Timestamp1|
+--------+--------------------+--------------------+
|ALMACENA|2023-09-23 16:32:…|2023-09-23T16:32:…|
|STREMING|2023-09-23 16:32:…|2023-09-23T16:32:…|
| INTENTO|2023-09-23 16:32:…|2023-09-23T16:32:…|
| Ramirez|2023-09-23 16:26:…|2023-09-23T16:26:…|
| RAMIREZ|2023-09-23 16:32:…|2023-09-23T16:32:…|
| SEGUNDO|2023-09-23 16:32:…|2023-09-23T16:32:…|
|  Sergio|2023-09-23 16:26:…|2023-09-23T16:26:…|
|  SERGIO|2023-09-23 16:32:…|2023-09-23T16:32:…|
|  IBARRA|2023-09-23 16:32:…|2023-09-23T16:32:…|
|  Ibarra|2023-09-23 16:26:…|2023-09-23T16:26:…|
|   SALIR|2023-09-23 16:33:…|2023-09-23T16:33:…|
|   Salir|2023-09-23 16:27:…|2023-09-23T16:27:…|
|   batch|2023-09-23 16:26:…|2023-09-23T16:26:…|
|   Batch|2023-09-23 16:26:…|2023-09-23T16:26:…|
|   DATOS|2023-09-23 16:32:…|2023-09-23T16:32:…|
|    Otro|2023-09-23 16:26:…|2023-09-23T16:26:…|
|    HOLA|2023-09-23 16:32:…|2023-09-23T16:32:…|
|    HOLA|2023-09-23 16:26:…|2023-09-23T16:26:…|
|     LOS|2023-09-23 16:32:…|2023-09-23T16:32:…|
|     VER|2023-09-23 16:32:…|2023-09-23T16:32:…|
+--------+--------------------+--------------------+
only showing top 20 rows
```

[5]:

```python
from pyspark.sql.functions import year, month, dayofmonth, hour, minute, second

df = df.withColumn("Año", year(df["timestamp"]))

df = df.withColumn("Mes", month(df["timestamp"]))

df = df.withColumn("Día", dayofmonth(df["timestamp"]))

df = df.withColumn("Hora", hour(df["timestamp"]))

df = df.withColumn("Minuto", minute(df["timestamp"]))

df = df.withColumn("Segundo", second(df["timestamp"]))

# Mostrar el resultado

df.show(truncate=False)
```

```
+--------+-----------------------+------------------------+----+---+---+----+---
---+-------+
|Palabra |Timestamp              |Timestamp1              |Año
|Mes|Día|Hora|Minuto|Segundo|
+--------+-----------------------+------------------------+----+---+---+----+---
---+-------+
|ALMACENA|2023-09-23 16:32:57.008|2023-09-23T16:32:57.008Z|2023|9  |23 |16  |32
|57     |
|STREMING|2023-09-23 16:32:46.075|2023-09-23T16:32:46.075Z|2023|9  |23 |16  |32
|46     |
|INTENTO |2023-09-23 16:32:46.075|2023-09-23T16:32:46.075Z|2023|9  |23 |16  |32
|46     |
|Ramirez |2023-09-23 16:26:26.904|2023-09-23T16:26:26.904Z|2023|9  |23 |16  |26
|26     |
|RAMIREZ |2023-09-23 16:32:46.075|2023-09-23T16:32:46.075Z|2023|9  |23 |16  |32
|46     |
|SEGUNDO |2023-09-23 16:32:46.075|2023-09-23T16:32:46.075Z|2023|9  |23 |16  |32
|46     |
|Sergio  |2023-09-23 16:26:26.904|2023-09-23T16:26:26.904Z|2023|9  |23 |16  |26
|26     |
|SERGIO  |2023-09-23 16:32:46.075|2023-09-23T16:32:46.075Z|2023|9  |23 |16  |32
|46     |
|IBARRA  |2023-09-23 16:32:46.075|2023-09-23T16:32:46.075Z|2023|9  |23 |16  |32
|46     |
|Ibarra  |2023-09-23 16:26:26.904|2023-09-23T16:26:26.904Z|2023|9  |23 |16  |26
|26     |
|SALIR   |2023-09-23 16:33:07.915|2023-09-23T16:33:07.915Z|2023|9  |23 |16  |33
|7      |
|Salir   |2023-09-23 16:27:04.299|2023-09-23T16:27:04.299Z|2023|9  |23 |16  |27
|4      |
|batch   |2023-09-23 16:26:55.001|2023-09-23T16:26:55.001Z|2023|9  |23 |16  |26
|55     |
|Batch   |2023-09-23 16:26:55.001|2023-09-23T16:26:55.001Z|2023|9  |23 |16  |26
|55     |
|DATOS   |2023-09-23 16:32:57.008|2023-09-23T16:32:57.008Z|2023|9  |23 |16  |32
|57     |
|Otro    |2023-09-23 16:26:55.001|2023-09-23T16:26:55.001Z|2023|9  |23 |16  |26
|55     |
|HOLA    |2023-09-23 16:32:35.001|2023-09-23T16:32:35.001Z|2023|9  |23 |16  |32
|35     |
|HOLA    |2023-09-23 16:26:15.001|2023-09-23T16:26:15.001Z|2023|9  |23 |16  |26
|15     |
|LOS     |2023-09-23 16:32:57.008|2023-09-23T16:32:57.008Z|2023|9  |23 |16  |32
|57     |
|VER     |2023-09-23 16:32:57.008|2023-09-23T16:32:57.008Z|2023|9  |23 |16  |32
|57     |
+--------+-----------------------+------------------------+----+---+---+----+---
---+-------+
only showing top 20 rows
```

```
[6]: from pyspark.sql.functions import count
     df.groupBy("Palabra").agg(count("*").alias("Count")).show()
```

```
[Stage 7:==============================>                              (1 + 1) / 2]

+--------+-----+
| Palabra|Count|
+--------+-----+
|    Otro|    1|
| INTENTO|    1|
|ALMACENA|    1|
|   DATOS|    1|
|  Sergio|    1|
|  Ibarra|    1|
|   batch|    1|
|   Salir|    1|
|  IBARRA|    1|
|STREMING|    1|
| RAMIREZ|    1|
|   Batch|    1|
|   SALIR|    1|
| Ramirez|    1|
|  SERGIO|    1|
| SEGUNDO|    1|
|    null|    7|
|     VER|    1|
|       Y|    1|
|      DE|    1|
+--------+-----+
only showing top 20 rows
```

[ ]:

[ ]:

[ ]:

[ ]: