In [ ]:
```
from pyspark.sql import SparkSession

from pyspark.sql.functions import *
from pyspark.sql.types import *

spark = SparkSession \
    .builder \
    .appName("how to read csv file") \
    .getOrCreate()
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
2023-11-26T12:56:22,835 WARN [Thread-4] org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop librar
y for your platform... using builtin-java classes where applicable
2023-11-26T12:56:24,589 WARN [Thread-4] org.apache.spark.util.Utils - Service 'SparkUI' could not bind on port 4040.
Attempting port 4041.
2023-11-26T12:56:24,590 WARN [Thread-4] org.apache.spark.util.Utils - Service 'SparkUI' could not bind on port 4041.
Attempting port 4042.
```

Cargo los datos en un dataframe de spark

In [ ]:
```
fecha_refined = spark.read.csv('/user/ort/obligatorio/fecha.csv', header=True)
```

Vista previa de las primeras filas

In [ ]:
```
fecha_refined.show()
```

```
+---+----------+----------+
| id|fecha_alta|fecha_baja|
+---+----------+----------+
|  1|  6/2/2021|      null|
|  2|  6/2/2021|      null|
|  3|  6/3/2021|      null|
|  4|  6/3/2021| 3/13/2022|
|  5|  6/4/2021| 11/5/2023|
|  6|  6/4/2021|  7/3/2022|
|  7|  6/5/2021|      null|
|  8|  6/8/2021|      null|
|  9|  6/9/2021|      null|
| 10| 6/11/2021|      null|
| 11| 6/12/2021|      null|
| 12| 6/12/2021|      null|
| 13| 6/13/2021|      null|
| 14| 6/13/2021|      null|
| 15| 6/15/2021|      null|
| 16| 6/15/2021|      null|
| 17| 6/16/2021|      null|
| 18| 6/16/2021|      null|
| 19| 6/18/2021|      null|
| 20| 6/19/2021|      null|
+---+----------+----------+
only showing top 20 rows
```

Cantidad de columnas del dataframe monthly_pays

In [ ]:
```
num_columns=len(fecha_refined.columns)
num_columns
```

Out[ ]: 3

Nombre de las columnas de monthly_pays

In [ ]:
```
fecha_refined.columns
```

Out[ ]: ['id', 'fecha_alta', 'fecha_baja']

Descripción de los datos de la tabla

In [ ]:
```
fecha_refined.describe
```

Out[ ]:
```
<bound method DataFrame.describe of DataFrame[id: string, fecha_alta: string, fecha_baja: string]>
```

Schema de la tabla

In [ ]:
```
fecha_refined.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- fecha_alta: string (nullable = true)
 |-- fecha_baja: string (nullable = true)
```

Valores Nulos o Faltantes

In [ ]:
```
total_nulos = fecha_refined.select([sum(col(c).isNull().cast("int")).alias(c) for c in fecha_refined.columns])

total_nulos.show()
```

```
+---+----------+----------+
| id|fecha_alta|fecha_baja|
+---+----------+----------+
|  0|         0|       850|
+---+----------+----------+
```

Los valores nulos corresponden a los usuarios que no se han dado de baja, por lo cual se decide mantenerlos como nulos.

In [ ]:
```
spark = SparkSession.builder.appName("EjemploConversionFecha").getOrCreate()

spark.conf.set("spark.sql.legacy.timeParserPolicy", "LEGACY")
```

```
2023-11-26T12:56:38,308 WARN [Thread-4] org.apache.spark.sql.SparkSession - Using an existing Spark session; only run time SQL configurations will take effect.
```

In [ ]:
```
fecha_refined = fecha_refined.withColumn("fecha_alta", to_date(col("fecha_alta"), "MM/dd/yyyy"))
```

In [ ]:
```
fecha_refined = fecha_refined.withColumn("fecha_baja", to_date(col("fecha_baja"), "MM/dd/yyyy"))
```

In [ ]:
```
fecha_refined.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- fecha_alta: date (nullable = true)
 |-- fecha_baja: date (nullable = true)
```

In [ ]:
```
hdfs_path = "/user/ort/obligatorio/refined/refined_fecha/"
fecha_refined.write.csv(hdfs_path, header=False, mode="overwrite")
```