PySpark UDF

(a.k.a User Defined Function) is the most useful feature of Spark SQL & DataFrame that is used to extend the PySpark build in capabilities.

https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/#converting-udf (https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/#converting-udf)

In [1]:
```
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
from pyspark.sql.types import *

spark = SparkSession\
    .builder\
    .appName("chapter-06-convert-datetime-utf")\
    .getOrCreate()

```

unix_timestamp()

In [2]:
```
df = spark.createDataFrame(
    [("11/25/1991",), ("01/24/1991",), ("02/03/1919",)],
    ['date_str']
)
```

In [3]:
```
df.show()
```

```
+----------+
|  date_str|
+----------+
|11/25/1991|
|01/24/1991|
|02/03/1919|
+----------+
```

In [4]:
```
df_a = df.select(
    'date_str',
    F.from_unixtime(F.unix_timestamp('date_str', 'MM/dd/yyyy')).alia
)
```

In [5]:
```
df_a.printSchema()
```

```
root
 |-- date_str: string (nullable = true)
 |-- date: string (nullable = true)
```

It is wrong that `date` datatype is still `string` , but its value is in correct `datetime` format

In [6]:
```
df_a.show()
```

```
+----------+-------------------+
|  date_str|               date|
+----------+-------------------+
|11/25/1991|1991-11-25 00:00:00|
|01/24/1991|1991-01-24 00:00:00|
|02/03/1919|1919-02-03 00:00:00|
+----------+-------------------+
```

to_date()

In [7]:
```
df_b = df.select(
    'date_str',
    F.to_date('date_str', 'MM/dd/yyyy').alias('date')
)
```

In [8]:
```
df_b.show()
```

```
+----------+----------+
|  date_str|      date|
+----------+----------+
|11/25/1991|1991-11-25|
|01/24/1991|1991-01-24|
|02/03/1919|1919-02-03|
+----------+----------+
```

In [9]:
```
df_b.printSchema()
```

```
root
 |-- date_str: string (nullable = true)
 |-- date: date (nullable = true)
```

to_timestamp()

In [10]:
```
df = spark.createDataFrame(
    [("11/25/1991 01:30:10",), ("01/24/1991 11:30:10",), ("02/03/191
    ['date_str']
)
```

In [11]:
```
df_c = df.select(
    'date_str',
    F.to_timestamp('date_str', 'MM/dd/yyyy HH:mm:SS').alias('date')
)
```

In [12]:
```
df_c.show(truncate=False)
```

```
+-------------------+---------------------+
|date_str           |date                 |
+-------------------+---------------------+
|11/25/1991 01:30:10|1991-11-25 01:30:00.1|
|01/24/1991 11:30:10|1991-01-24 11:30:00.1|
|02/03/1919 21:30:10|1919-02-03 21:30:00.1|
+-------------------+---------------------+
```

In [13]:
```
df_c.printSchema()
```

```
root
 |-- date_str: string (nullable = true)
 |-- date: timestamp (nullable = true)
```

```
spark.sparkContext._conf.setAll([("spark.sql.legacy.timeParserPolicy", "LEGACY")])

spark.sparkContext._conf.getAll()
```

UDF - to_date()

In [14]:
```
df2 = spark.createDataFrame(
    [("11/25/1991",), ("1/24/1991",), ("2/3/1919",)],
    ['date_str']
)
```

In [15]:
```
df2.show()
```

```
+----------+
|  date_str|
+----------+
|11/25/1991|
| 1/24/1991|
|  2/3/1919|
+----------+
```

In [16]:
```
from datetime import datetime
udf_to_date =  F.udf (lambda x: datetime.strptime(x, '%m/%d/%Y'), Da
```

In [17]:
```
df2_a = df2.withColumn('date', udf_to_date(F.col('date_str')))
```

In [18]:
```
1 df2_a.show()
```

```
+----------+----------+
|  date_str|      date|
+----------+----------+
|11/25/1991|1991-11-25|
| 1/24/1991|1991-01-24|
|  2/3/1919|1919-02-03|
+----------+----------+
```

In [19]:
```
1 df2_a.printSchema()
```

```
root
 |-- date_str: string (nullable = true)
 |-- date: date (nullable = true)
```

UDF - to_datetime()

In [20]:
```
1 # udf_to_datetime =  F.udf (lambda x: datetime.strptime(x, '%m/%d/%
2 udf_to_datetime =  F.udf (lambda x: datetime.strptime(x, '%m/%d/%Y %
```

In [21]:
```
1 df3 = spark.createDataFrame(
2     [("11/25/1991 1:15",), ("1/24/1991 12:30",), ("2/3/1919 18:00",)
3     ['datetime_str']
4 )
```

In [22]:
```
1 df3_a = df3.withColumn('timestamp', udf_to_datetime(F.col('datetime_
```

In [23]:
```
1 df3_a.show()
```

```
+---------------+-------------------+
|   datetime_str|          timestamp|
+---------------+-------------------+
|11/25/1991 1:15|1991-11-25 01:15:00|
|1/24/1991 12:30|1991-01-24 12:30:00|
| 2/3/1919 18:00|1919-02-03 18:00:00|
+---------------+-------------------+
```

In [24]:
```
1 df3_a.printSchema()
```

```
root
 |-- datetime_str: string (nullable = true)
 |-- timestamp: timestamp (nullable = true)
```