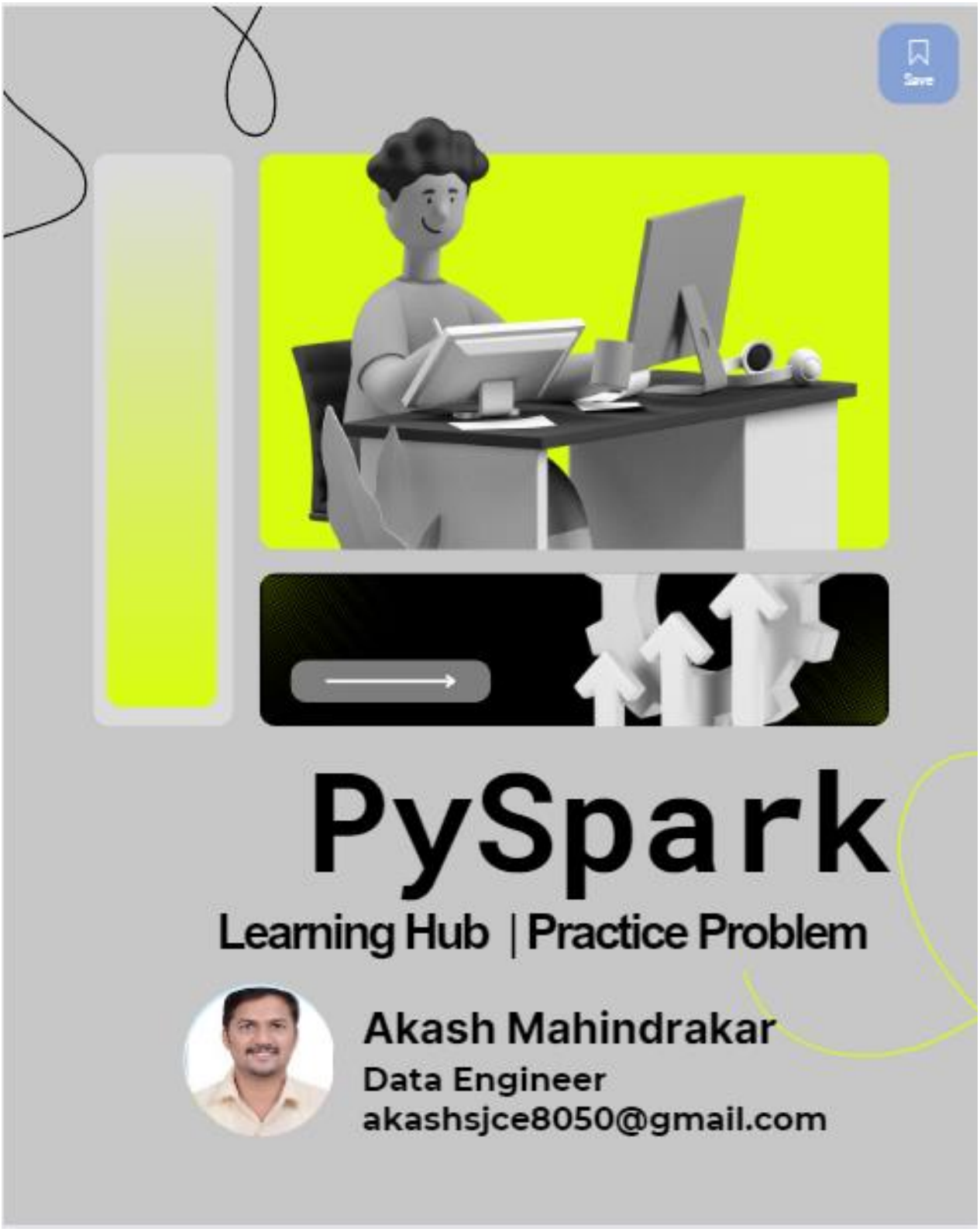
# PYSPARK LEARNING HUB : DAY - 21



## Step - 1 : Problem Statement

### 21_Date in pyspark

Write a pyspark code perform below function

- Get only Year part of "JoiningDate"
- Get only Month part of "JoiningDate".
- Get only date part of "JoiningDate".
- Get the current system date using DataFrame API
- Get the current UTC date and time using DataFrame API

**Difficult Level :** EASY

**DataFrame:**

```
data = [
        [1, "Vikas", "Ahlawat", 600000.0, "2013-02-15 11:16:28.290", "IT", "Male"],
        [2, "nikita", "Jain", 530000.0, "2014-01-09 17:31:07.793", "HR", "Female"],
        [3, "Ashish", "Kumar", 1000000.0, "2014-01-09 10:05:07.793", "IT", "Male"],
        [4, "Nikhil", "Sharma", 480000.0, "2014-01-09 09:00:07.793", "HR", "Male"],
        [5, "anish", "kadian", 500000.0, "2014-01-09 09:31:07.793", "Payroll", "Male"],
]
# Create a schema for the DataFrame
schema = StructType([
        StructField("EmployeeID", IntegerType(), True),
        StructField("First_Name", StringType(), True),
        StructField("Last_Name", StringType(), True),
        StructField("Salary", DoubleType(), True),
        StructField("Joining_Date", StringType(), True),
        StructField("Department", StringType(), True),
        StructField("Gender", StringType(), True)
])
```

## Step - 2 : Writing the pyspark code to solve

```
# Creating Spark Session
from pyspark.sql import SparkSession
from pyspark.sql.types import
StructType,StructField,IntegerType,StringType

#creating spark session
spark = SparkSession. \
builder. \
config('spark.shuffle.useOldFetchProtocol', 'true'). \
config('spark.ui.port','0'). \
config("spark.sql.warehouse.dir", "/user/itv008042/warehouse"). \
enableHiveSupport(). \
master('yarn'). \
getOrCreate()

# Create a list of rows from the image
data = [
        [1, "Vikas", "Ahlawat", 600000.0, "2013-02-15 11:16:28.290", "IT", "Male"],
        [2, "nikita", "Jain", 530000.0, "2014-01-09 17:31:07.793", "HR", "Female"],
        [3, "Ashish", "Kumar", 1000000.0, "2014-01-09 10:05:07.793", "IT", "Male"],
        [4, "Nikhil", "Sharma", 480000.0, "2014-01-09 09:00:07.793", "HR", "Male"],
        [5, "anish", "kadian", 500000.0, "2014-01-09 09:31:07.793", "Payroll", "Male"],
]

# Create a schema for the DataFrame
schema = StructType([
        StructField("EmployeeID", IntegerType(), True),
        StructField("First_Name", StringType(), True),
        StructField("Last_Name", StringType(), True),
        StructField("Salary", DoubleType(), True),
        StructField("Joining_Date", StringType(), True),
        StructField("Department", StringType(), True),
        StructField("Gender", StringType(), True)
])
```

# PYSPARK LEARNING HUB : DAY - 21

**emp_df=spark.createDataFrame(data,schema)**

```
+----------+----------+---------+---------+--------------------+----------+------+
|EmployeeID|First_Name|Last_Name|   Salary|        Joining_Date|Department|Gender|
+----------+----------+---------+---------+--------------------+----------+------+
|         1|     Vikas|  Ahlawat| 600000.0|2013-02-15 11:16:...|        IT|  Male|
|         2|    nikita|     Jain| 530000.0|2014-01-09 17:31:...|        HR|Female|
|         3|    Ashish|    Kumar|1000000.0|2014-01-09 10:05:...|        IT|  Male|
|         4|    Nikhil|   Sharma| 480000.0|2014-01-09 09:00:...|        HR|  Male|
|         5|     anish|   kadian| 500000.0|2014-01-09 09:31:...|   Payroll|  Male|
+----------+----------+---------+---------+--------------------+----------+------+
```

```python
# Get only Year part of "JoiningDate"
emp_df.select(date_format(to_timestamp(col("joining_date")),"yyyy")).show(truncate=False)

# Get only Month part of "JoiningDate".
emp_df.select(date_format(to_timestamp(col("joining_date")),"MM")).show(truncate=False)
emp_df.select(date_format(to_timestamp(col("joining_date")),"MMM")).show(truncate=False)

# Get only date part of "JoiningDate".
emp_df.select(date_format(to_timestamp(col("joining_date")),"dd")).show(truncate=False)

# Get the current system date using DataFrame API
system_date_df = spark.range(1).select(current_date().alias("system_date"))
system_date_df.show(truncate=False)

# Get the current UTC date and time using DataFrame API
utc_date_time_df = spark.range(1).select(current_timestamp().alias("utc_date_time"))
utc_date_time_df.show(truncate=False)
```

```
+-----------------------------------------------+
|date_format(to_timestamp(joining_date), yyyy)|
+-----------------------------------------------+
|2013                                           |
|2014                                           |
|2014                                           |
|2014                                           |
|2014                                           |
+-----------------------------------------------+

+-----------------------------------------------+
|date_format(to_timestamp(joining_date), MM)|
+-----------------------------------------------+
|02                                             |
|01                                             |
|01                                             |
|01                                             |
|01                                             |
+-----------------------------------------------+

+-----------------------------------------------+
|date_format(to_timestamp(joining_date), MMM)|
+-----------------------------------------------+
|Feb                                            |
|Jan                                            |
|Jan                                            |
|Jan                                            |
|Jan                                            |
+-----------------------------------------------+
```

```
+-------------------------------------------+
|date_format(to_timestamp(joining_date), dd)|
+-------------------------------------------+
|15                                         |
|09                                         |
|09                                         |
|09                                         |
|09                                         |
+-------------------------------------------+

+-----------+
|system_date|
+-----------+
|2024-01-09 |
+-----------+

+-----------------------+
|utc_date_time          |
+-----------------------+
|2024-01-09 21:14:36.535|
+-----------------------+
```

# PYSPARK LEARNING HUB : DAY - 21

Save

## Was it helpful?

follow for more!

**Akash Mahindrakar**
Data Engineer
akashsjce8050@gmail.com

Comment

SHARE YOUR THOUGHTS IN COMMENT BELOW

Share