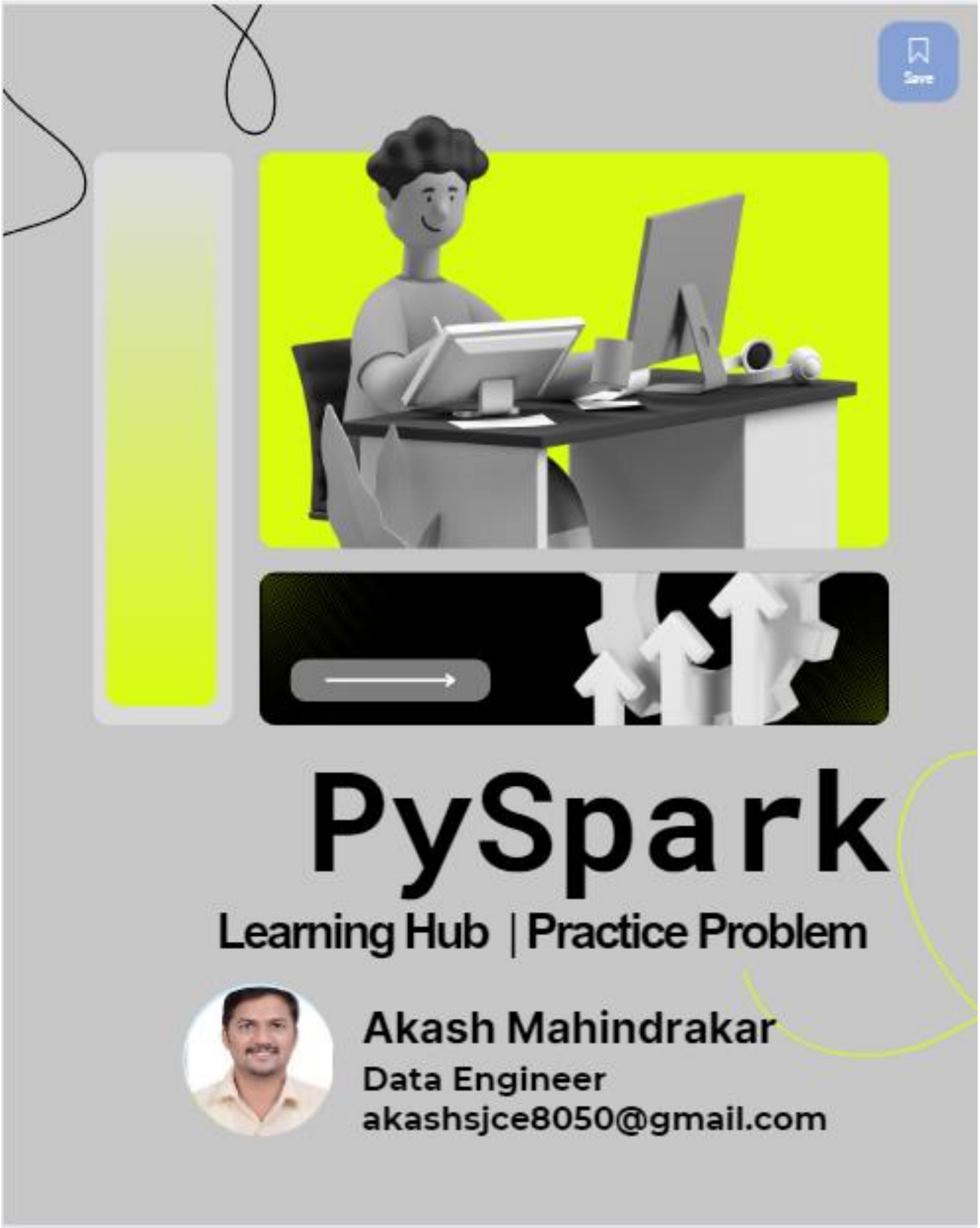
PYSPARK LEARNING HUB : DAY - 22



PySpark

Learning Hub | Practice Problem

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

WWW.LINKEDIN.COM/IN/AKASHMAHINDRAKAR

Step - 1 : Problem Statement

22_Date in pyspark

Write a pyspark code perform below function

- Get the first name, current date, joiningdate and diff between current date and joining date in months.
- Get the first name, current date, joiningdate and diff between current date and joining date in days.
- Get all employee details from EmployeeDetail table whose joining year is 2013

Difficult Level : EASY

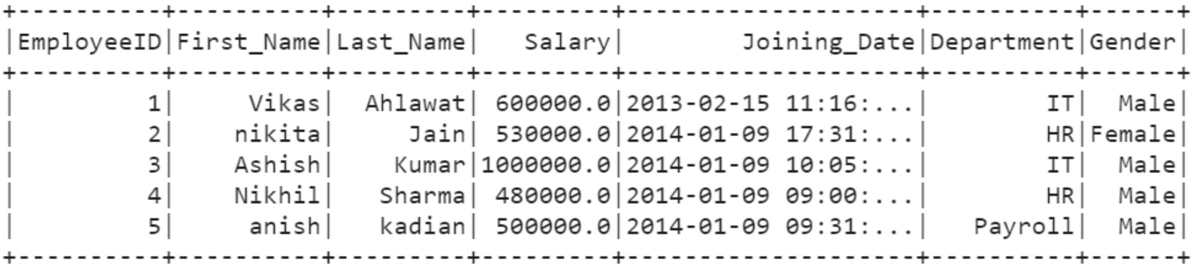
DataFrame:

```
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)
])
```

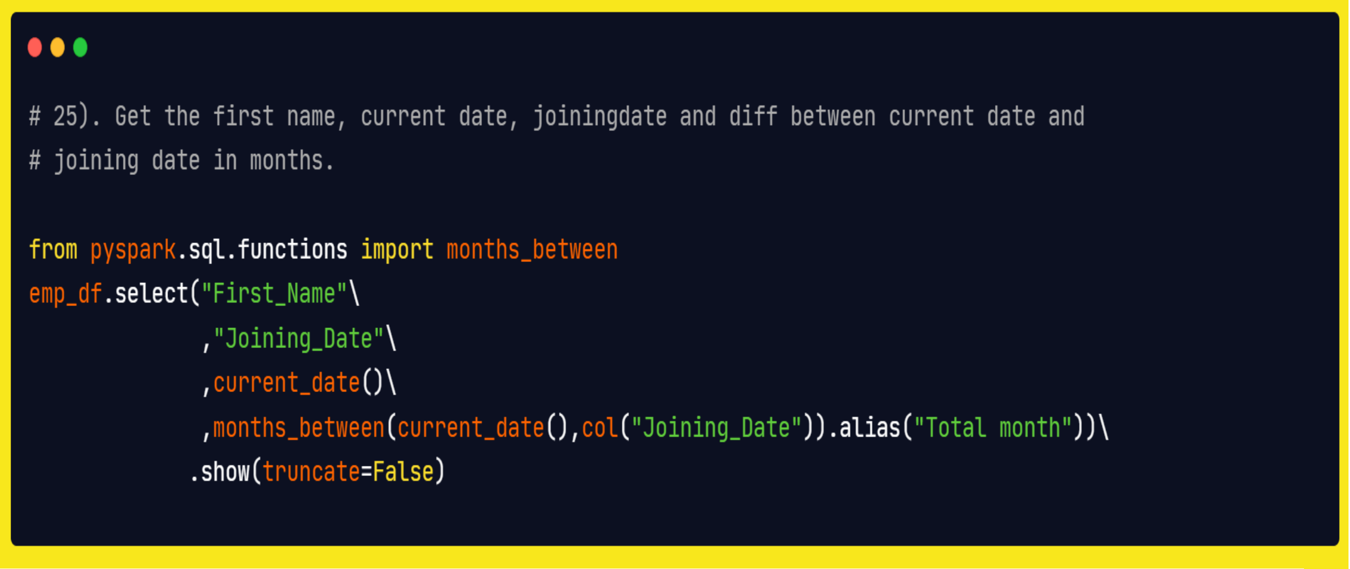
Step - 2 : Writing the pyspark code to solve the

```
# 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)
])
```

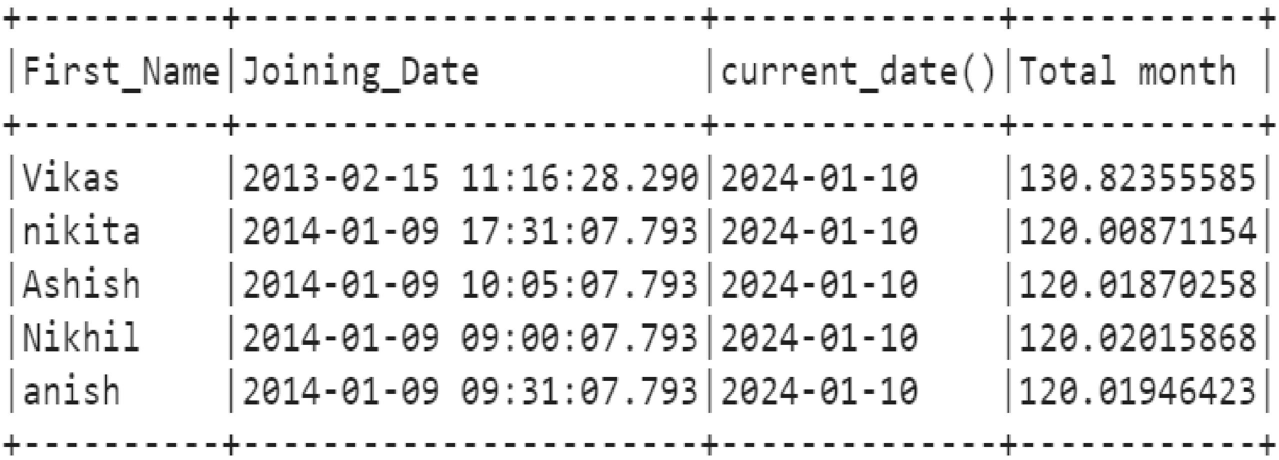
emp_df=spark.createDataFrame(data,schema)



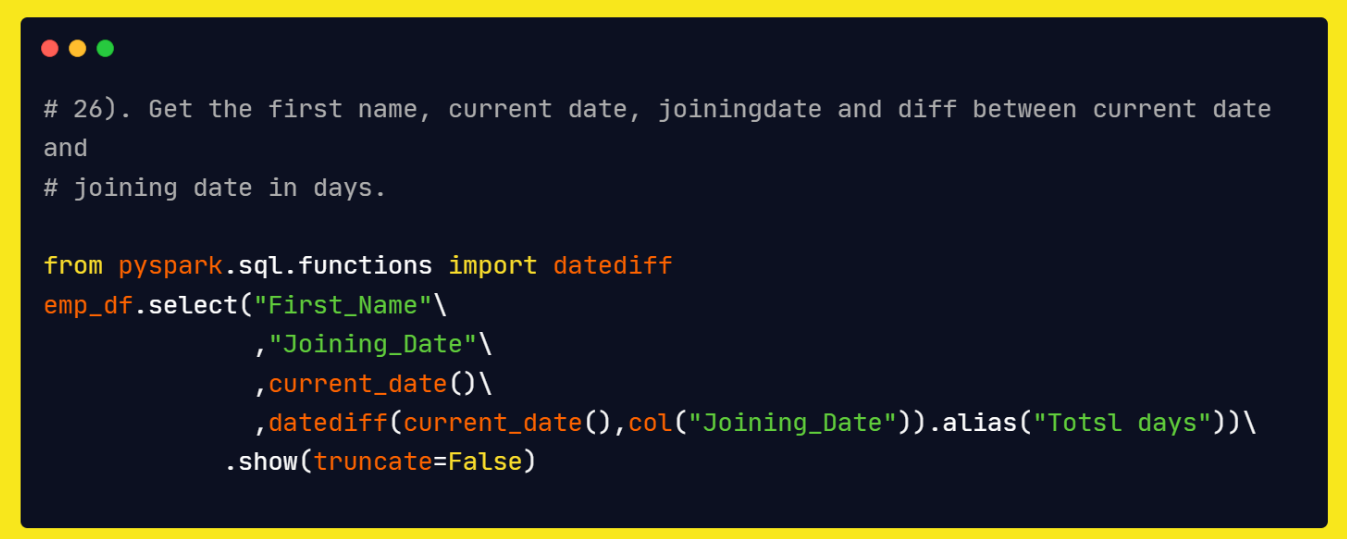
```
+----------+----------+---------+---------+--------------------+----------+------+
|EmployeeID|First_Name|Last_Name|   Salary|        Joining_Date|Department|Gender|
+----------+----------+---------+---------+--------------------+----------+------+
|         1|     Vikas|  Ahlawat| 600000.0|2013-02-15 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|
+----------+----------+---------+---------+--------------------+----------+------+
```



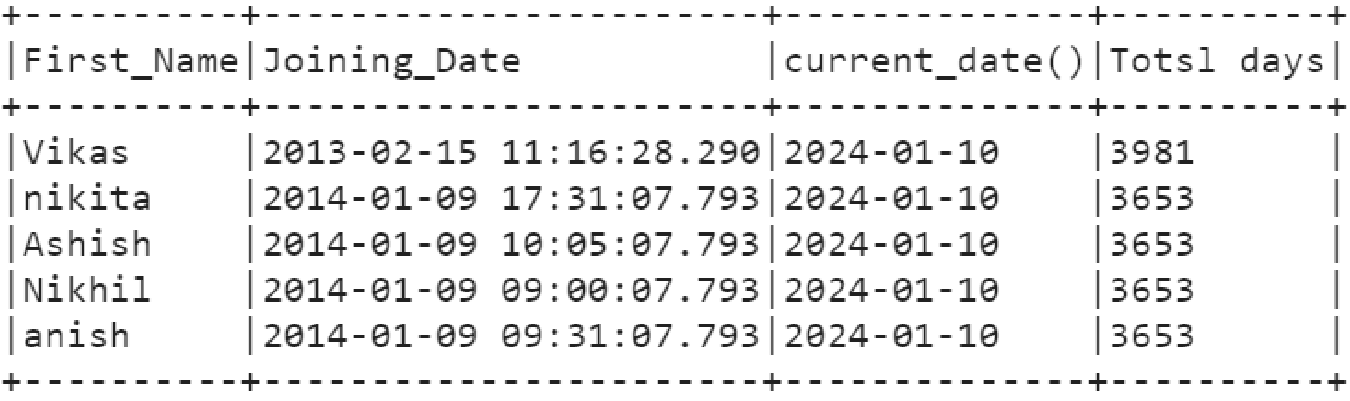
```
# 25). Get the first name, current date, joiningdate and diff between current date and
# joining date in months.

from pyspark.sql.functions import months_between
emp_df.select("First_Name"\
              ,"Joining_Date"\
              ,current_date()\
              ,months_between(current_date(),col("Joining_Date")).alias("Total month"))\
             .show(truncate=False)
```



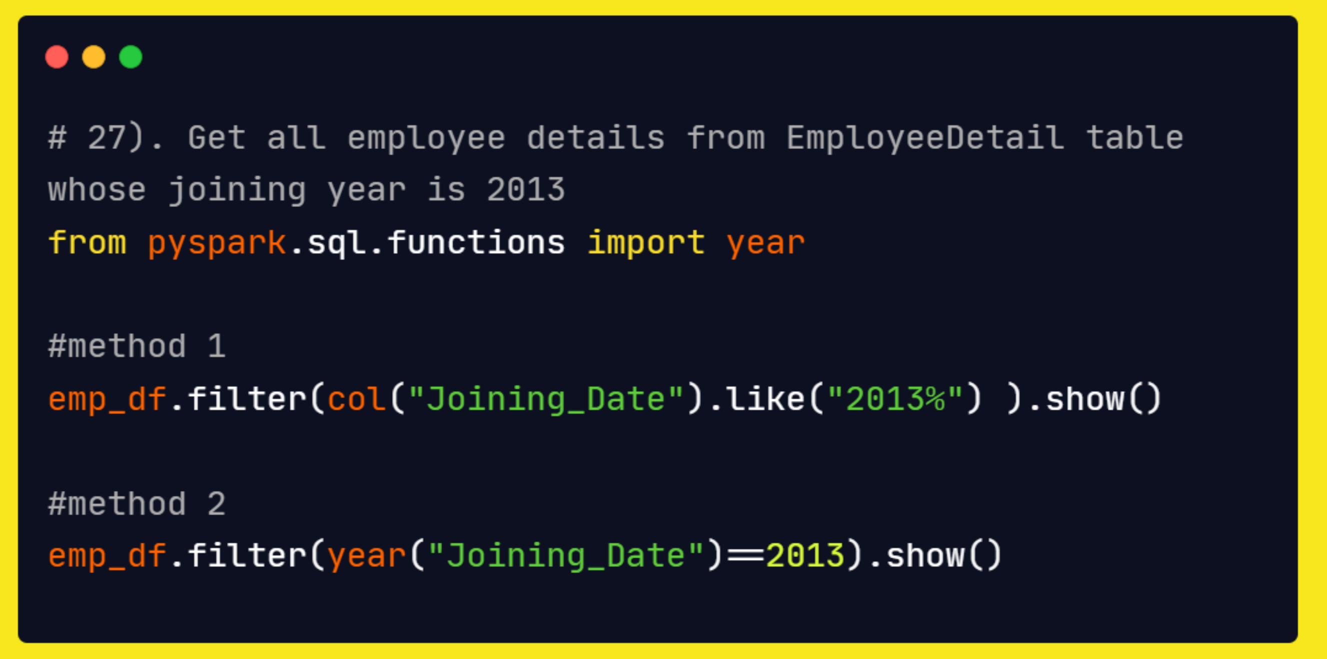
```
+----------+-----------------------+--------------+------------+
|First_Name|Joining_Date           |current_date()|Total month |
+----------+-----------------------+--------------+------------+
|Vikas     |2013-02-15 11:16:28.290|2024-01-10    |130.82355585|
|nikita    |2014-01-09 17:31:07.793|2024-01-10    |120.00871154|
|Ashish    |2014-01-09 10:05:07.793|2024-01-10    |120.01870258|
|Nikhil    |2014-01-09 09:00:07.793|2024-01-10    |120.02015868|
|anish     |2014-01-09 09:31:07.793|2024-01-10    |120.01946423|
+----------+-----------------------+--------------+------------+
```



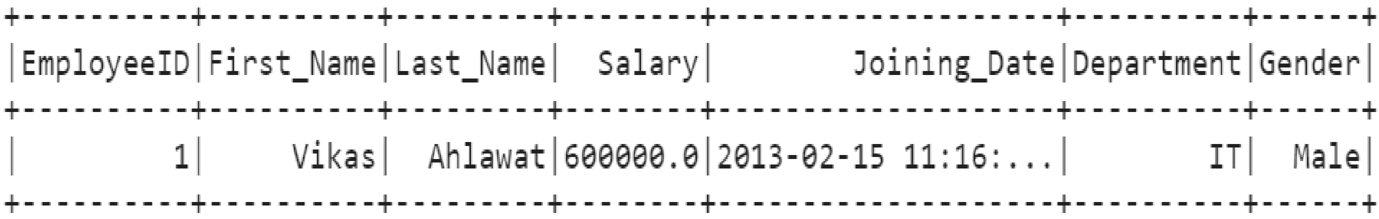
```
# 26). Get the first name, current date, joiningdate and diff between current date
and
# joining date in days.

from pyspark.sql.functions import datediff
emp_df.select("First_Name"\
              ,"Joining_Date"\
              ,current_date()\
              ,datediff(current_date(),col("Joining_Date")).alias("Totsl days"))\
            .show(truncate=False)
```



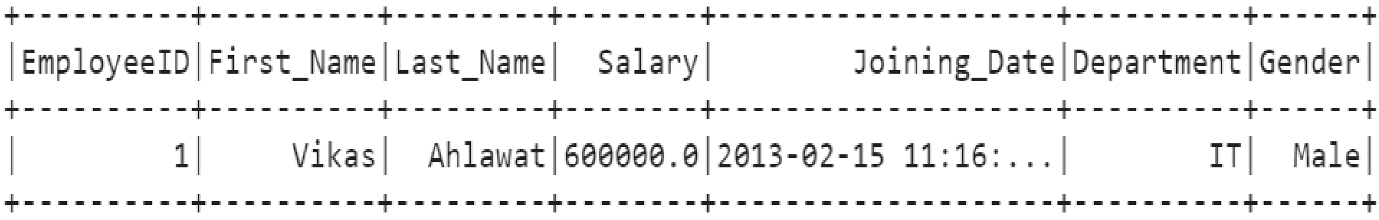
```
+----------+-----------------------+--------------+----------+
|First_Name|Joining_Date           |current_date()|Totsl days|
+----------+-----------------------+--------------+----------+
|Vikas     |2013-02-15 11:16:28.290|2024-01-10    |3981      |
|nikita    |2014-01-09 17:31:07.793|2024-01-10    |3653      |
|Ashish    |2014-01-09 10:05:07.793|2024-01-10    |3653      |
|Nikhil    |2014-01-09 09:00:07.793|2024-01-10    |3653      |
|anish     |2014-01-09 09:31:07.793|2024-01-10    |3653      |
+----------+-----------------------+--------------+----------+
```



```
# 27). Get all employee details from EmployeeDetail table
whose joining year is 2013
from pyspark.sql.functions import year

#method 1
emp_df.filter(col("Joining_Date").like("2013%") ).show()

#method 2
emp_df.filter(year("Joining_Date")==2013).show()
```



```
+----------+----------+---------+--------+--------------------+----------+------+
|EmployeeID|First_Name|Last_Name|  Salary|        Joining_Date|Department|Gender|
+----------+----------+---------+--------+--------------------+----------+------+
|         1|     Vikas|  Ahlawat|600000.0|2013-02-15 11:16:...|        IT|  Male|
+----------+----------+---------+--------+--------------------+----------+------+
```



```
+----------+----------+---------+--------+--------------------+----------+------+
|EmployeeID|First_Name|Last_Name|  Salary|        Joining_Date|Department|Gender|
+----------+----------+---------+--------+--------------------+----------+------+
|         1|     Vikas|  Ahlawat|600000.0|2013-02-15 11:16:...|        IT|  Male|
+----------+----------+---------+--------+--------------------+----------+------+
```

Was it helpful?

follow for more!

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

SHARE YOUR THOUGHTS IN COMMENT BELOW