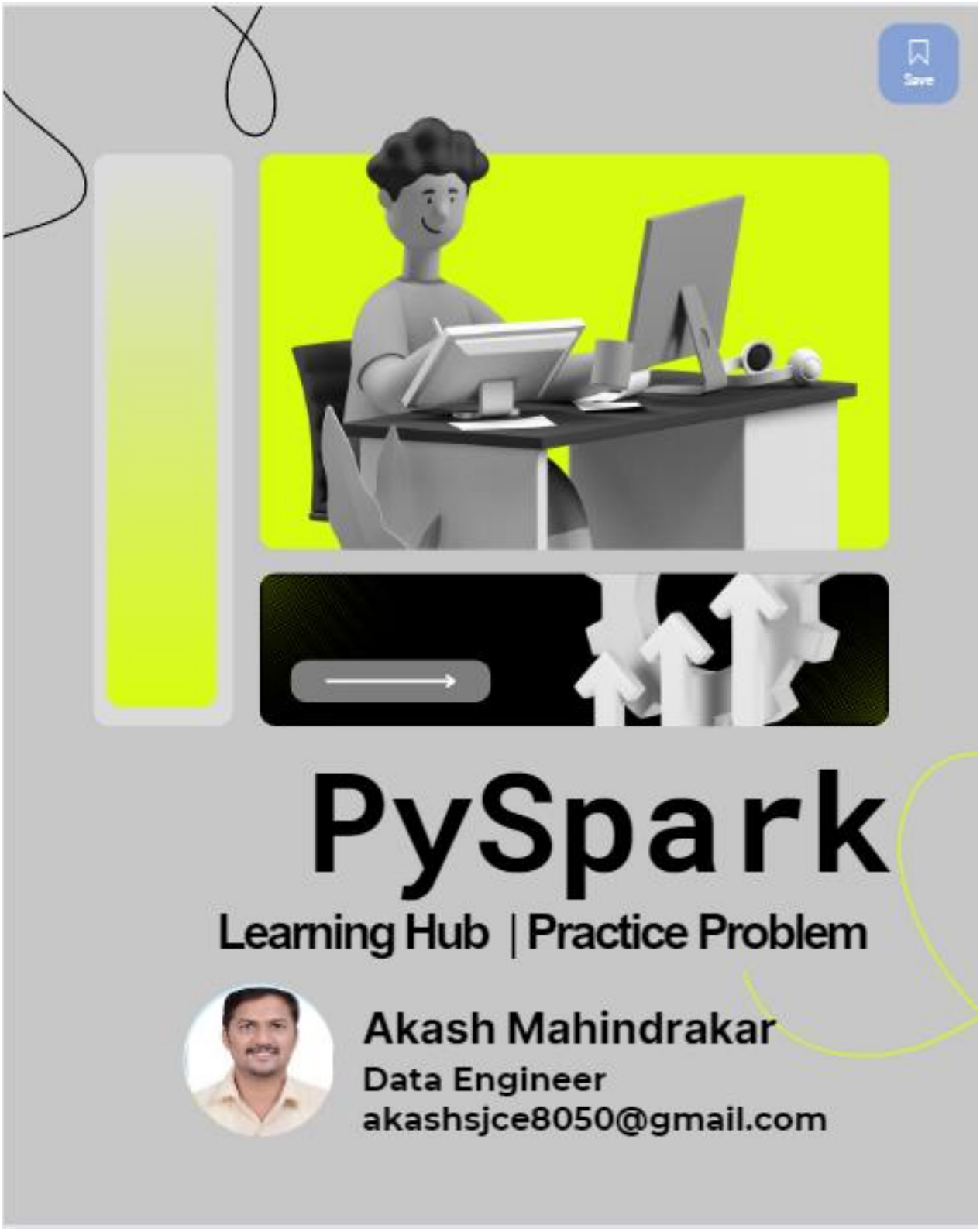
PYSPARK LEARNING HUB : DAY - 24



Step - 1 : Problem Statement

24_Trim and case in pyspark

Write a pyspark code perform below function

- Select all employee detail with First name not in "Vikas","Ashish", and "Nikhil".
- Select first name from "EmployeeDetail" df after removing white spaces from right side
- Select first name from "EmployeeDetail" table after removing white spaces from left side
- Display first name and Gender as M/F.(if male then M, if Female then F)

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|
+----------+----------+---------+---------+--------------------+----------+------+
```

```python
# 33. Select all employee detail with First name not in "Vikas","Ashish", and "Nikhil".

from pyspark.sql.functions import col,lower
emp_df.filter(~lower(col("First_Name")).isin("vikas","ashish","nikhil")).show()

# 34. Select first name from "EmployeeDetail" df after removing white spaces from
# right side

from pyspark.sql.functions import rtrim
emp_df.select(rtrim(col("First_Name"))).show()
```

```
+----------+----------+---------+-------+--------------------+----------+------+
|EmployeeID|First_Name|Last_Name|  Salary|        Joining_Date|Department|Gender|
+----------+----------+---------+-------+--------------------+----------+------+
|         2|    nikita|     Jain|530000.0|2014-01-09 17:31:...|        HR|Female|
|         5|     anish|   kadian|500000.0|2014-01-09 09:31:...|   Payroll|  Male|
+----------+----------+---------+-------+--------------------+----------+------+

+-----------------+
|rtrim(First_Name)|
+-----------------+
|            Vikas|
|           nikita|
|           Ashish|
|           Nikhil|
|            anish|
+-----------------+
```

PYSPARK LEARNING HUB : DAY - 24

```python
# 35. Select first name from "EmployeeDetail" table after removing white spaces from
# left side

from pyspark.sql.functions import ltrim
emp_df.select(ltrim(col("First_Name"))).show()

# 36. Display first name and Gender as M/F.(if male then M, if Female then F)

from pyspark.sql.functions import when
emp_df.withColumn("Gen",when(col("Gender")=="Female","F")\
                        .when(col("Gender")=="Male","M")
                ).select("First_Name","Gen").show()
```

```
+------------------+
|ltrim(First_Name)|
+------------------+
|             Vikas|
|            nikita|
|            Ashish|
|            Nikhil|
|             anish|
+------------------+
```

```
+----------+---+
|First_Name|Gen|
+----------+---+
|     Vikas|  M|
|    nikita|  F|
|    Ashish|  M|
|    Nikhil|  M|
|     anish|  M|
+----------+---+
```

PYSPARK LEARNING HUB : DAY - 24

Save

Was it helpful?

follow for more!

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

Comment

SHARE YOUR THOUGHTS IN COMMENT BELOW

Share