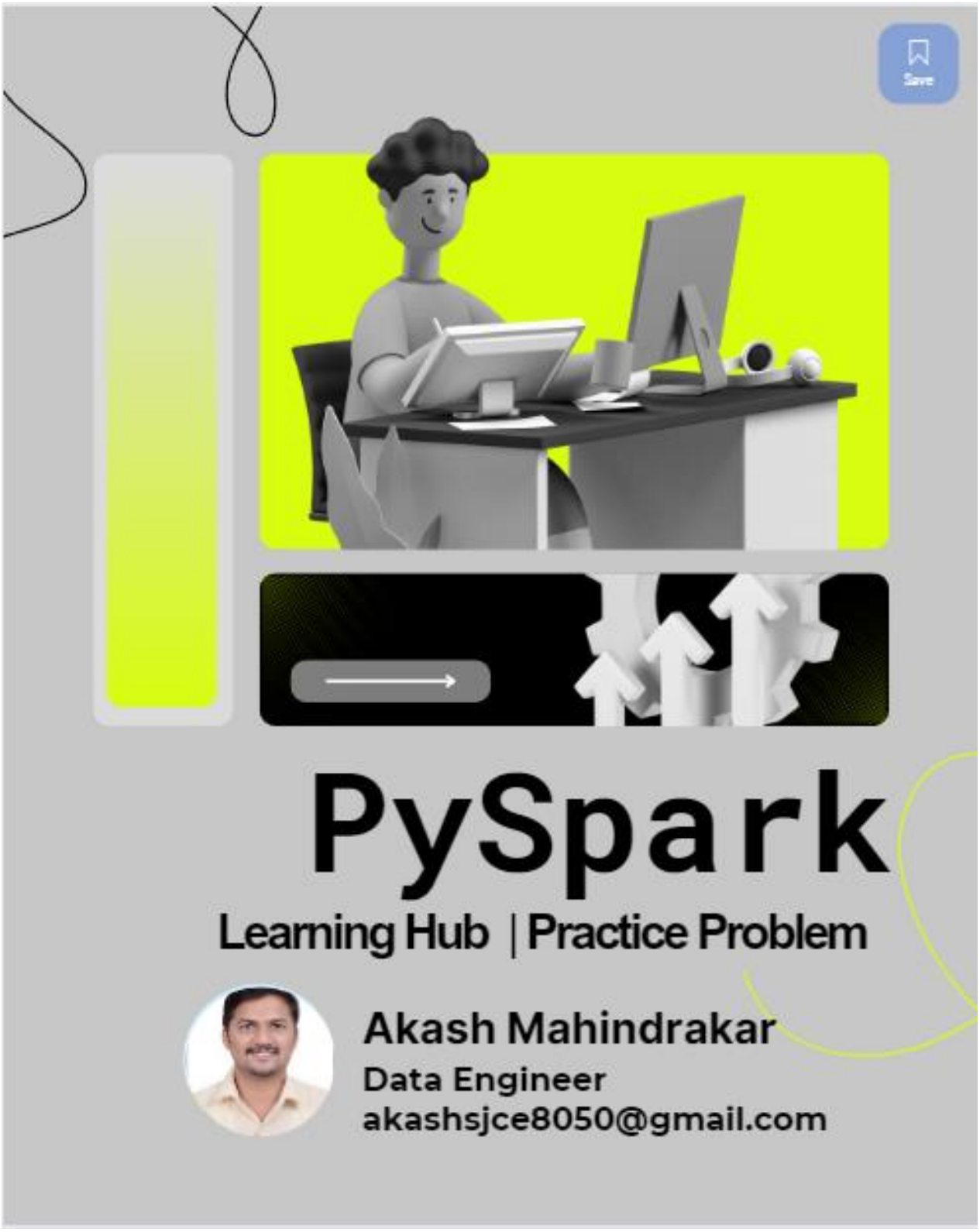
PYSPARK LEARNING HUB : DAY - 16



Step - 1 : Problem Statement

16_Select in pyspark

Write a pyspark code perform below function

- Write a pyspark code to get all employee detail.
- Write a query to get only "FirstName" column from emp_df
- Write a Pyspark code to get FirstName in upper case as "First Name".
- Write a pyspark code to get FirstName in lower case

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)
])

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

#1. Write a pyspark code to get all employee detail

```
emp_df.show()
```

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

 # 2. Write a query to get only "FirstName" column from emp_df

```
# Method 1
emp_df.select("First_Name").show()

# Method 2
emp_df.select(col("First_Name")).show()

# Method 3
emp_df.createOrReplaceTempView("emp_table")
spark.sql("select First_Name from emp_table").show()
```

```
+----------+
|First_Name|
+----------+
|     Vikas|
|    nikita|
|    Ashish|
|    Nikhil|
|     anish|
+----------+

+----------+
|First_Name|
+----------+
|     Vikas|
|    nikita|
|    Ashish|
|    Nikhil|
|     anish|
+----------+

+----------+
|First_Name|
+----------+
|     Vikas|
|    nikita|
|    Ashish|
|    Nikhil|
|     anish|
+----------+
```

arning Hub

 # 3. Write a Pyspark code to get FirstName in upper case as "First Name".

```
emp_df.select(upper("First_Name")).show()
```

```
+-----------------+
|upper(First_Name)|
+-----------------+
|            VIKAS|
|           NIKITA|
|           ASHISH|
|           NIKHIL|
|            ANISH|
+-----------------+
```

#4. Write a pyspark code to get FirstName in lower case

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

```
+-----------------+
|lower(First_Name)|
+-----------------+
|            vikas|
|           nikita|
|           ashish|
|           nikhil|
|            anish|
+-----------------+
```

PySpark Learning Hub

PYSPARK LEARNING HUB : DAY - 16

Save

Was it helpful?

follow for more!

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

Comment

SHARE YOUR THOUGHTS IN COMMENT BELOW

Share