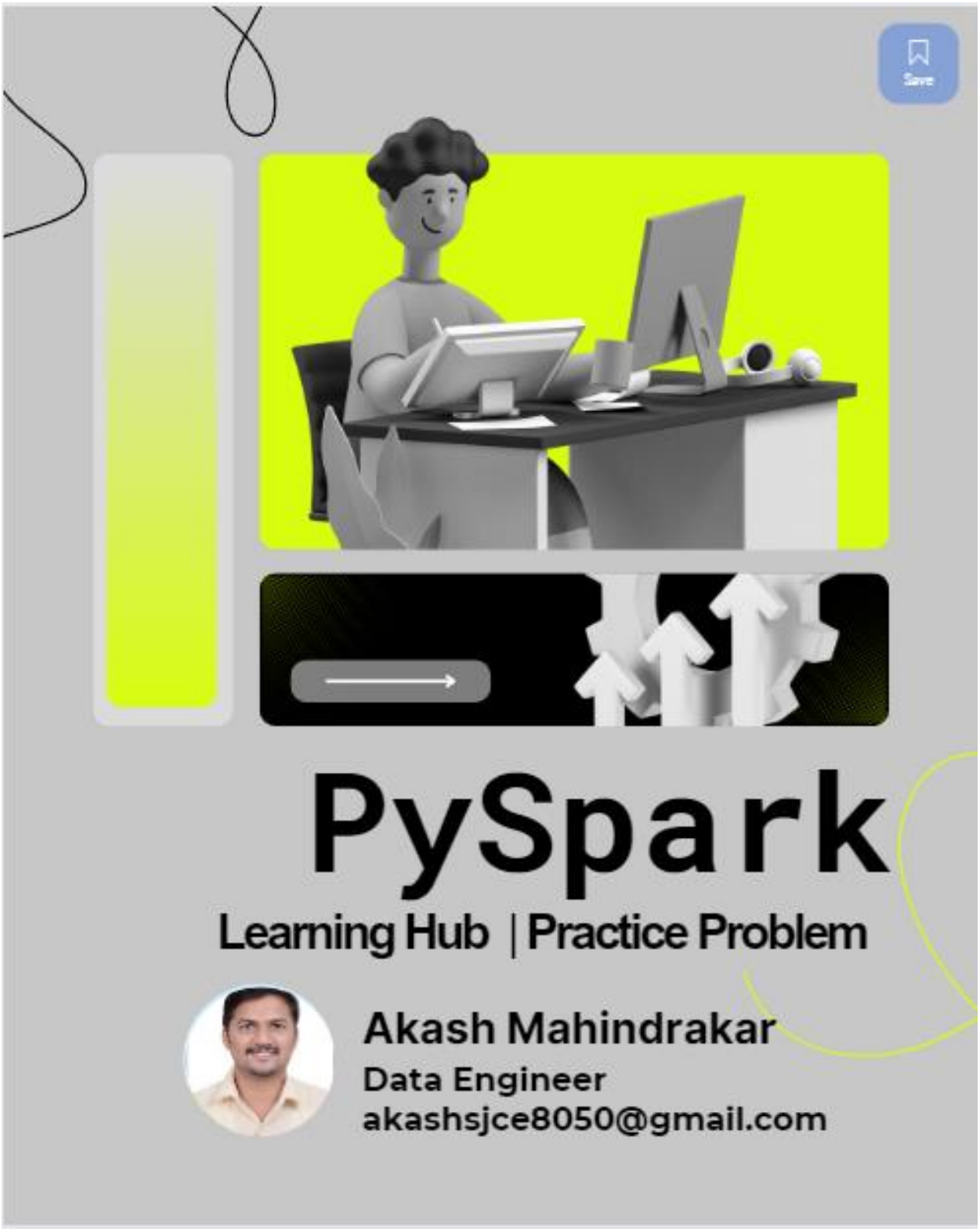
PYSPARK LEARNING HUB : DAY - 26



Step - 1 : Problem Statement

26_groupby in pyspark

Write a pyspark code perform below function

- Write the query to get the department and department wise total(sum) salary from "EmployeeDetail" table.
- Write the query to get the department and department wise total(sum) salary, display it in ascending order according to salary.
- Write the query to get the department and department wise total(sum) salary, display it in descending order according to salary.
- Write the query to get the department, total no. of departments, total(sum) salary with respect to department from "EmployeeDetail" table.

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

```
# 42. Write the query to get the department and department wise
#  total(sum) salary from "EmployeeDetail" table.

from pyspark.sql.functions import sum

emp_df.groupby(col('Department'))\
      .agg(sum('Salary').alias("sum_of_salary")).show()
```

```
+----------+-------------+
|Department|sum_of_salary|
+----------+-------------+
|        HR|    1010000.0|
|   Payroll|     500000.0|
|        IT|    1600000.0|
+----------+-------------+
```

```
# 43. Write the query to get the department and department wise total(sum) salary, display it in ascending order according to salary.

emp_df.groupby(col("Department"))\
      .agg(sum("Salary").alias("sum_of_salary"))\
      .orderBy(col('sum_of_salary').asc()).show()
```

```
+----------+-------------+
|Department|sum_of_salary|
+----------+-------------+
|   Payroll|     500000.0|
|        HR|    1010000.0|
|        IT|    1600000.0|
+----------+-------------+
```

```
# 44. Write the query to get the department and department wise total(sum)
salary,display it in descending order according to salary.

emp_df.groupby(col("Department"))\
      .agg(sum("Salary").alias("sum_of_salary"))\
      .orderBy(col('sum_of_salary').desc()).show()
```

```
+----------+--------------+
|Department|sum_of_salary|
+----------+--------------+
|        IT|     1600000.0|
|        HR|     1010000.0|
|   Payroll|      500000.0|
+----------+--------------+
```

```
# 45. Write the query to get the department, total no. of
departments,total(sum) salary with respect to department from
"EmployeeDetail" table.

from pyspark.sql.functions import count
emp_df.groupby(col("Department"))\
        .agg(count("Salary").alias("count"),\
         sum("Salary").alias("sum_od_Salary")).show()
```

```
+----------+-----+--------------+
|Department|count|sum_od_Salary|
+----------+-----+--------------+
|        HR|    2|     1010000.0|
|   Payroll|    1|      500000.0|
|        IT|    2|     1600000.0|
+----------+-----+--------------+
```

PYSPARK LEARNING HUB : DAY - 26

Save

Was it helpful?

follow for more!

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

Comment

SHARE YOUR THOUGHTS IN COMMENT BELOW

Share