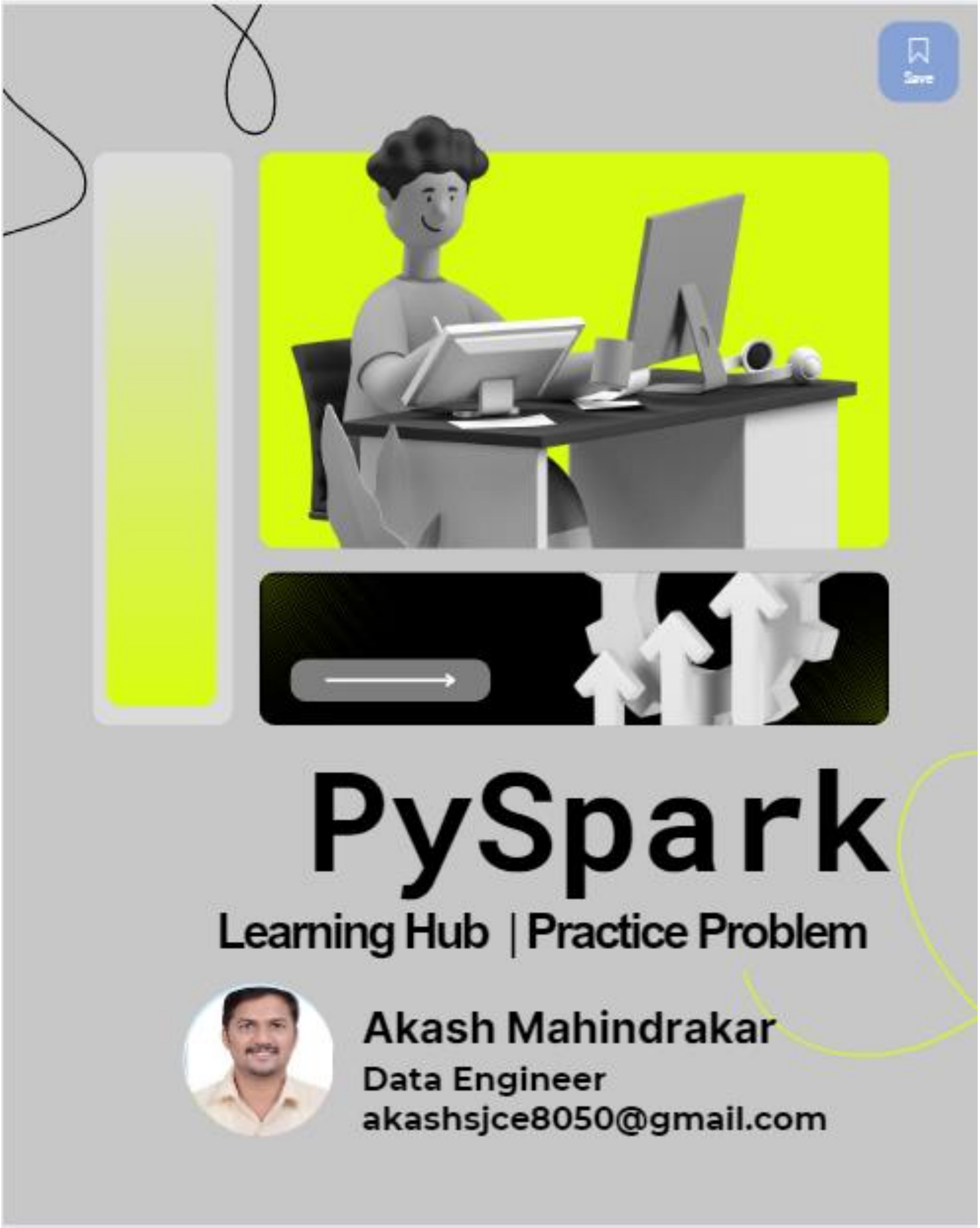
PYSPARK LEARNING HUB : DAY - 11



PySpark

Learning Hub | Practice Problem

Akash Mahindrakar

Data Engineer

akashsjce8050@gmail.com

WWW.LINKEDIN.COM/IN/AKASHMAHINDRAKAR

Step - 1 : Problem Statement

11_Find Customer Referee

Find the names of the customer that are not referred by the customer with id = 2.

Return the result table in any order

Difficult Level : EASY

DataFrame:

```
# Define the schema for the Customer table
schema = StructType([
        StructField("id", IntegerType(), True),
        StructField("name", StringType(), True),
        StructField("referee_id", IntegerType(), True)
])

# Create an RDD with the data
data = [
        (1, 'Will', None),
        (2, 'Jane', None),
        (3, 'Alex', 2),
        (4, 'Bill', None),
        (5, 'Zack', 1),
        (6, 'Mark', 2)
]
```

Step - 2 : Identifying The Input Data And Expected Output

INPUT

INPUT		
ID	NAME	REFEREE_ID
1	Will	
2	Jane	
3	Alex	2
4	Bill	
5	Zack	1
6	Mark	2

OUTPUT

OUTPUT
NAME
Will
Jane
Bill
Zack

Step - 3 : 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()

# Define the schema for the Customer table
schema = StructType([
        StructField("id", IntegerType(), True),
        StructField("name", StringType(), True),
        StructField("referee_id", IntegerType(), True)
])

# Create an RDD with the data
data = [
        (1, 'Will', None),
        (2, 'Jane', None),
        (3, 'Alex', 2),
        (4, 'Bill', None),
        (5, 'Zack', 1),
        (6, 'Mark', 2)
]
```

```
# Create a PySpark DataFrame
customer_df = spark.createDataFrame(data ,schema )
```

```
# Filter customers not referred by customer with id = 2
result_df = customer_df.filter((col("referee_id").isNull()) |
(col("referee_id") != 2))

# Select only the 'name' column
result_df = result_df.select("name")
```

```
+------+
| name|
+------+
| Will|
| Jane|
| Bill|
| Zack|
+------+
```

Save

Was it helpful?

follow for more!

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

Comment

SHARE YOUR THOUGHTS IN COMMENT BELOW

Share