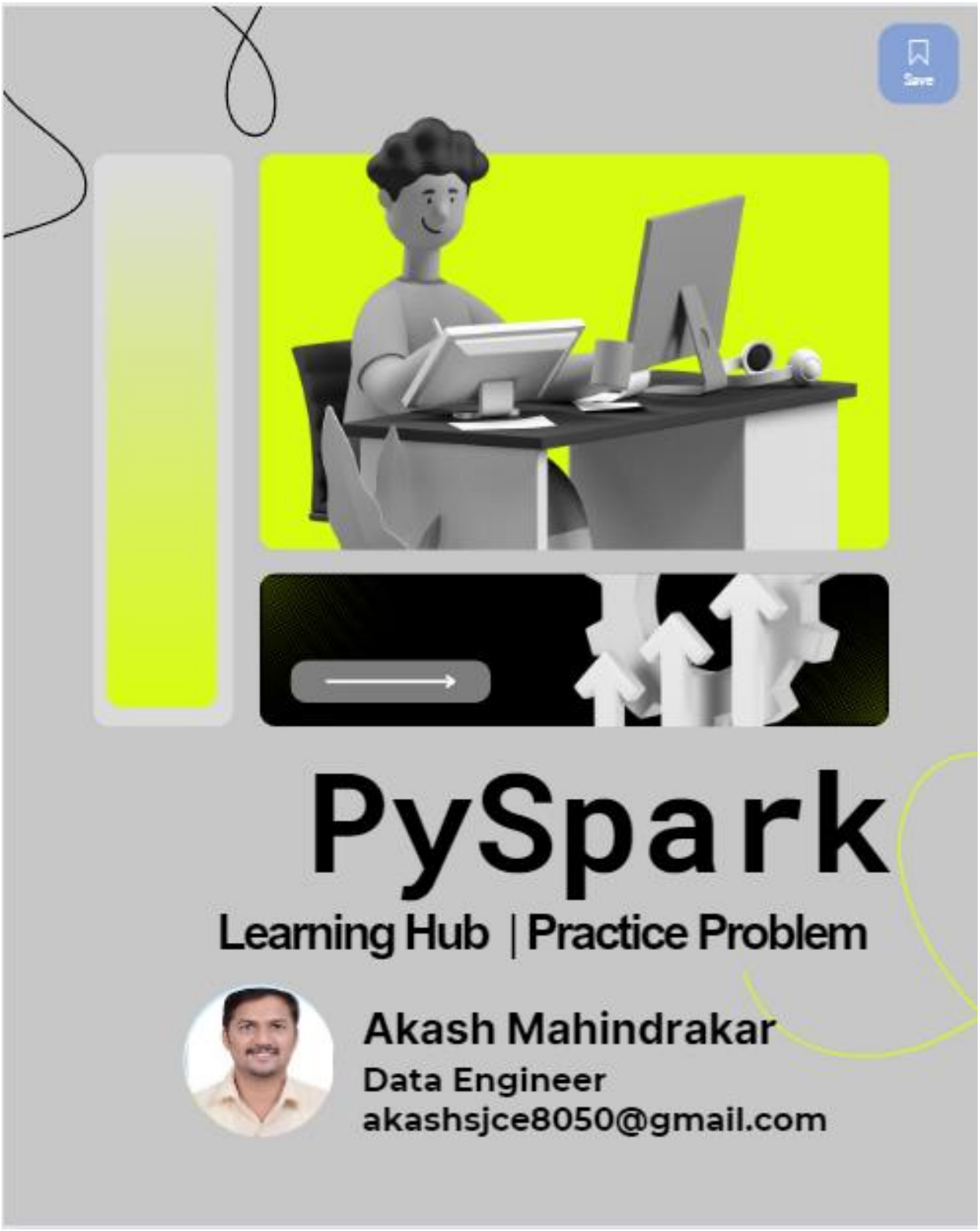
PYSPARK LEARNING HUB : DAY - 3



Step - 1 : Problem Statement

Combine Two DF

Write a Pyspark program to report the first name, last name, city, and state of each person in the Person dataframe. If the address of a personId is not present in the Address dataframe, report null instead.

Difficult Level : EASY

DataFrame:

```
# Define schema for the 'persons' table
persons_schema = StructType([
        StructField("personId", IntegerType(), True),
        StructField("lastName", StringType(), True),
        StructField("firstName", StringType(), True)
])

# Define schema for the 'addresses' table
addresses_schema = StructType([
        StructField("addressId", IntegerType(), True),
        StructField("personId", IntegerType(), True),
        StructField("city", StringType(), True),
        StructField("state", StringType(), True)
])

# Define data for the 'persons' table
persons_data = [
        (1, 'Wang', 'Allen'),
        (2, 'Alice', 'Bob')
]

# Define data for the 'addresses' table
addresses_data = [
        (1, 2, 'New York City', 'New York'),
        (2, 3, 'Leetcode', 'California')
]
```

PySpark Learning Hub

Step - 2 : Identifying The Input Data And Expected Output

INPUT

INPUT-1 persons		
PERSONID	LASTNAME	FIRSTNAME
1	Wang	Allen
2	Alice	Bob

INPUT-2 addresses			
ADDRESSID	PERSONID	CITY	STATE
1	2	New York City	New York
2	3	Leetcode	California

OUTPUT

OUTPUT			
FIRSTNAME	LASTNAME	CITY	STATE
Bob	Alice	New York City	New York
Allen	Wang		

PYSPARK Learning Hub

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
from pyspark.sql.functions import when
from pyspark.sql import functions as F
from pyspark.sql.window import Window

#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 schema for the 'persons' table
persons_schema = StructType([
        StructField("personId", IntegerType(), True),
        StructField("lastName", StringType(), True),
        StructField("firstName", StringType(), True)
])

# Define schema for the 'addresses' table
addresses_schema = StructType([
        StructField("addressId", IntegerType(), True),
        StructField("personId", IntegerType(), True),
        StructField("city", StringType(), True),
        StructField("state", StringType(), True)
])
```

PYSPARK Learning Hub

```
# Define data for the 'persons' table
persons_data = [
        (1, 'Wang', 'Allen'),
        (2, 'Alice', 'Bob')
]

# Define data for the 'addresses' table
addresses_data = [
        (1, 2, 'New York City', 'New York'),
        (2, 3, 'Leetcode', 'California')
]
```

```
# Create a PySpark DataFrame
person_df=spark.createDataFrame(persons_data,persons_schema)
address_df=spark.createDataFrame(addresses_data,addresses_schema)

person_df.show()
address_df.show()
```

```
+--------+--------+---------+
|personId|lastName|firstName|
+--------+--------+---------+
|       1|    Wang|    Allen|
|       2|   Alice|      Bob|
+--------+--------+---------+

+---------+--------+-------------+----------+
|addressId|personId|         city|     state|
+---------+--------+-------------+----------+
|        1|       2|New York City|  New York|
|        2|       3|     Leetcode|California|
+---------+--------+-------------+----------+
```

```
person_df.join(address_df,person_df.personId==address_df.personId,'left')\
        .select('firstName','lastName','city','state')\
        .show()

# Show the result DataFrame
```

```
+---------+--------+-------------+--------+
|firstName|lastName|         city|   state|
+---------+--------+-------------+--------+
|    Allen|    Wang|         null|    null|
|      Bob|   Alice|New York City|New York|
+---------+--------+-------------+--------+
```

PYSPARK Learning Hub

PYSPARK LEARNING HUB : DAY - 3

Save

Was it helpful?

follow for more!

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

Comment

SHARE YOUR THOUGHTS IN COMMENT BELOW

Share